IMPROVED CONVERGENCE AND GENERALIZATION ANALYSIS FOR ONLINE KNOWLEDGE DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Online knowledge distillation enables distributed models to learn collaboratively without requiring a pretrained teacher. However, the evolving and mutually dependent distillation targets complicate both convergence and generalization analysis. We study online distillation over decentralized networks with heterogeneous model architectures and personalized local distributions. We establish a finite-round convergence bound by controlling objective drift through the neighbors' local gradients and stochastic noise. This analysis does not rely on uniform bounds on gradients or objective drift used in existing analyses and yields tighter guarantees in wide-range parameter regimes. By combining our convergence and generalization analyses, we bound the personalized test risks of the local models produced after a finite number of training rounds. The resulting bound explicitly characterizes the roles of distribution discrepancies, dataset sizes, the trade-off between original supervised and distillation losses, and ensemble weights to construct teachers, including how these effects propagate through the network. This explicit parameter dependence enables tunable-parameter optimization based on a surrogate of the generalization bound. Using ensemble weigh optimization as example, our experiments demonstrate that this approach can substantially improve average personalized test accuracy across a range of heterogeneous settings.

1 INTRODUCTION

Knowledge distillation (KD) transfers knowledge from a teacher model to a student by training the student to match the teacher's predictions on shared inputs (Hinton et al., 2015). Generally, soft predictions can convey information beyond ground-truth labels, such as relationships among classes, providing additional supervision for student learning. Conventional offline distillation relies on one or more pretrained teachers that remain fixed during student training. It therefore separates teacher and student training into distinct stages, with knowledge transferred primarily in one direction. In contrast, online distillation enables multiple models to learn from one another throughout training without requiring a separately pretrained teacher (Anil et al., 2018; Zhang et al., 2018; Jeong et al., 2018; Abourayya et al., 2025a; Fang et al., 2026). Each node acts as both a student and a part of the teacher. Such collaboration through shared outputs can accommodate heterogeneous model architectures and reduce communication when the exchanged outputs are smaller than the model parameter or gradient used in conventional federated and decentralized learning methods.

However, online distillation introduces mutual dependence among peers as their models evolve during training. Analyses developed for fixed teachers do not directly address these interactions. Furthermore, existing online-distillation analyses leave room for tighter finite-round convergence bounds and more explicit generalization guarantees for the trained models. To address these gaps, this paper develops improved convergence and generalization analyses for online distillation.

Overview of online distillation. We consider a network of nodes $\mathcal{N} := \{1, \dots, N\}$ that communicate without a central server. Node i has a local training dataset $\hat{\mathcal{D}}_i$ of i.i.d. samples from its data distribution $\mathcal{P}^{(i)}$. The nodes also have access to a shared, *unlabeled reference* dataset $\hat{\mathcal{D}}^{\mathrm{re}}$. Let $\mathcal{X}$ be the input space, $\mathcal{Y} := \{1, \dots, M\}$ the label space, and $\Delta_{\mathcal{Y}}$ the probability simplex over $\mathcal{Y}$. Node i uses a model $h_{i,\theta} : \mathcal{X} \to \Delta_{\mathcal{Y}}$ with parameter $\theta \in \Theta_i \subset \mathbb{R}^{d_i}$. Note that model architectures and data distributions may differ across nodes.

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.

At the beginning of round t, node i has parameter θ_i^t. It forms its ensemble teacher labels by averaging its neighbors' predictions using fixed ensemble weights $w_{ij} \geq 0$ with $\sum_{j=1}^N w_{ij} = 1$. For $j \neq i$, we set $w_{ij} = 0$ when there is no communication link from node j to i. A self-weight $w_{ii} > 0$ is also permitted. Thus, the weight matrix $W := (w_{ij})$ can also describe the communication graph of the entire network.

We consider online distillation where teachers are formed by averaging prediction probabilities (Anil et al., 2018; Cho et al., 2023).[1] Node i optimizes the composite empirical loss

$$f_{i,\boldsymbol{\theta}^t}(\theta) := \frac{\rho_i}{|\hat{\mathcal{D}}_i|} \sum_{(x,y)\in\hat{\mathcal{D}}_i} \ell\big(h_{i,\theta}(x), e_y\big) + \frac{1-\rho_i}{|\hat{\mathcal{D}}^{\mathrm{re}}|} \sum_{x\in\hat{\mathcal{D}}^{\mathrm{re}}} d\left(h_{i,\theta}(x), \sum_{j=1}^N w_{ij} h_{j,\theta_j^t}(x)\right). \tag{1}$$

Here, $\boldsymbol{\theta}^t := (\theta_j^t)_{j=1}^N$, $\ell : \Delta_{\mathcal{Y}} \times \Delta_{\mathcal{Y}} \to \mathbb{R}_+$ is the original supervised loss function, $d : \Delta_{\mathcal{Y}} \times \Delta_{\mathcal{Y}} \to \mathbb{R}_+$ is the distillation loss function, and e_y is the one-hot vector for class y. The trade-off factor $\rho_i \in [0, 1]$ balances original local supervision and distillation.

Within each round, node i performs $K_i \geq 1$ local stochastic-gradient steps with stepsize $\eta_i > 0$: $\theta_i^{t,k+1} = \theta_i^{t,k} - \eta_i \widetilde{\nabla} f_{i,\boldsymbol{\theta}^t}(\theta_i^{t,k})$, where $k = 0, \ldots, K_i - 1$ and $\widetilde{\nabla} f_{i,\boldsymbol{\theta}^t}(\theta_i^{t,k})$ is a minibatch estimator of $\nabla f_{i,\boldsymbol{\theta}^t}(\theta_i^{t,k})$. The local step counts and stepsizes may differ across nodes but remain fixed across rounds. All ensemble labels $\sum_{j=1}^N w_{ij} h_{j,\theta_j^t}(x)$ are evaluated with the round-start parameters θ_j^t but remain fixed throughout the local updates. The ensemble labels are refreshed at the next communication round. In the following we briefly summarize existing work on convergence and generalization analyses. More details are provided in Appendix A.

Algorithms 1 and 2 in Appendix B specify the online-distillation procedure analyzed in this work. Although formulations and implementation details vary across prior work, this procedure follows the common principle of learning from both local supervision and peers' evolving outputs (Anil et al., 2018). In our framework, each local update takes a stochastic-gradient step on the composite loss. Methods that instead optimize the supervised and distillation losses in separate stages fall outside the scope of our analysis (Liu et al., 2024).

Note that the composite loss in (1) poses two related analytical challenges. First, each node's updates change both its own model and the supervision available to its neighbors, whose subsequent updates can influence that node in turn. This feedback produces coupled and time-varying local objectives, requiring a convergence analysis that accounts for their evolution during training. Second, minimizing the composite objective need not minimize a node's original supervised loss. This raises the question of what generalization guarantees can be established for the models produced by the training procedure.

Existing convergence analysis. Kernel-based analyses establish convergence and generalization guarantees under a common kernel and prescribed ensemble weights for constructing distillation targets (Park et al., 2023; 2024). For more general models, asymptotic stationarity has been established without finite-round convergence rates (Bistritz et al., 2020; Cho et al., 2023). Some complementary work studies two-stage distillation (Taya et al., 2022; Liu et al., 2024; Ma et al., 2026) and asynchronous distillation (Fang et al., 2026). These analyses concern different update protocols and different stationarity criterion to the framework we consider, so their guarantees do not directly apply here. The finite-round analyses most closely related to ours are those of Tan et al. (2022) and Abourayya et al. (2025a), which rely on uniform bounds on expected stochastic-gradient norms and on round-to-round drift in local objectives, respectively. These uniform bounds simplify the treatment of inter-node interactions, but such constants may be conservative. We instead control objective drift through neighbors' local gradients and stochastic noise, retaining more information about the neighbors' movements.

Existing generalization analysis. Some existing generalization bounds concern ensembles of local risk minimizers rather than models produced by the distillation procedure (Lin et al., 2020; Cho

[1] Other formulations average logits before applying softmax (Lin et al., 2020) or study distillation for regression (Park et al., 2024). Our approach to controlling distillation objective drift through neighbors' model updates in the convergence analysis can be adapted to these settings under suitable adjustments to corresponding assumptions and constants in the bounds.

et al., 2022; 2023; Wang et al., 2023; Jang et al., 2025). Other analyses consider generic models without explicitly relating their generalization performance to the training results (Shao et al., 2024). Moreover, the generalization bound of Shao et al. (2024) upper-bounds a node's local risk using teacher-dependent quantities, such as the teacher's risk, without explicitly characterizing how these quantities depend on the algorithm parameters. In particular, changing the ensemble weights also changes the models learned under those weights and hence their risks. These teacher-dependent quantities therefore cannot generally be treated as fixed when selecting the ensemble weights. This motivates combining generalization and convergence analyses to derive an explicit upper bound on each node's personal generalization risk at the output of the training algorithm.

Our contributions.

- We derive an improved finite-round convergence bound by controlling objective drift through the neighbors' local gradients, and stochastic noise. Our analysis accommodates node-specific local step counts, tradeoff factors and stepsizes without relying on the uniform bound on expected stochastic-gradient norms used by Tan et al. (2022) or the uniform objective-drift bound used by Abourayya et al. (2025a). We identify parameter regimes in which our analysis yields tighter bounds.
- We propose a new generalization analysis approach, combining it with our convergence analysis to bound the expected personalized test risks of models produced by online distillation after a finite number of rounds. Our bound characterizes the roles of distribution discrepancies across nodes, dataset sizes, the trade-off factor, and the ensemble weight matrix, while accounting for how their effects propagate through the network. This characterization provides a basis for selecting tunable parameters.
- We use a surrogate of the generalization bound to guide the selection of ensemble weights. Numerical experiments show that the resulting choices can substantially improve the average test accuracy across wide settings.

2 Convergence Analysis

In this section, we establish a new convergence bound for Algorithm 1 with heterogeneous local step counts, stepsizes and tradeoff factors. It is based on controlling the objective drift through neighbors' local gradients and stochastic noise. We will compare it with alternative analyses based on assumptions on bounded gradient norm and bounded objective drift from Tan et al. (2022) and Abourayya et al. (2025a).

2.1 Convergence Bound

Unless otherwise specified, all expectations are over training randomness, conditioned on the datasets $\{\hat{\mathcal{D}}_l\}_{l=1}^N$ and $\hat{\mathcal{D}}^{\mathrm{re}}$, and all equations involving random quantities are understood to hold almost surely.

We first state the assumptions required for our analysis.

Assumption 1. *$\exists L > 0$ s.t. $\forall i, t$, and $\forall \theta, \theta' \in \mathbb{R}^{d_i}$, $\|\nabla f_{i,\boldsymbol{\theta}^t}(\theta) - \nabla f_{i,\boldsymbol{\theta}^t}(\theta')\| \leq L\|\theta - \theta'\|$.*

Assumption 2. *$\exists \sigma_{sgd}, C_{sgd} \geq 0$, s.t. $\forall i, t$, and $\forall k \in \{1, \ldots, K_i\}$, $\mathbb{E}[\xi_{i,t,k} \mid \mathcal{F}_{i,t,k}] = 0$ and $\mathbb{E}[\|\xi_{i,t,k}\|^2 \mid \mathcal{F}_{i,t,k}] \leq \sigma_{sgd}^2 + C_{sgd}\|\nabla f_{i,\boldsymbol{\theta}^t}(\theta_i^{t,k})\|^2$, where $\xi_{i,t,k} := \widetilde{\nabla} f_{i,\boldsymbol{\theta}^t}(\theta_i^{t,k}) - \nabla f_{i,\boldsymbol{\theta}^t}(\theta_i^{t,k})$, and $\mathcal{F}_{i,t,k}$ is the sigma-field generated by the training history up to the beginning of round t, together with the randomness used by node i in local steps before step k of round t.*

Assumption 3. *$\forall i, t, \exists f_{inf} \geq 0$, s.t. $\forall \theta \in \mathbb{R}^{d_i}$, $f_{i,\boldsymbol{\theta}^t}(\theta) \geq f_{inf}$.*

Assumption 4. *$\exists L_d > 0$, s.t., for any admissible u, u', v, $|d(v, u) - d(v, u')| \leq L_d\|u - u'\|$.*

Assumption 5. *$\exists L_h > 0$, s.t. $\forall i$, $\forall \theta^{(1)}, \theta^{(2)} \in \mathbb{R}^{d_i}$ and $\forall x \in \mathcal{X}$, $\|h_{i,\theta^{(1)}}(x) - h_{i,\theta^{(2)}}(x)\| \leq L_h\|\theta^{(1)} - \theta^{(2)}\|$.*

Assumption 6. *$\forall i, 0 < \eta_i < \frac{2}{L(1+C_{\mathrm{sgd}})}$.*

Let $\boldsymbol{K} = (K_1, \ldots, K_N)$ and $K_{\max} := \max_i K_i$. We measure the stationarity of Algorithm 1 at the beginning of communication rounds by

$$S_0 := \frac{1}{TN} \sum_{i=1}^{N} \sum_{t=0}^{T-1} \mathbb{E} \|\nabla f_{i,\boldsymbol{\theta}^t}(\theta_i^t)\|^2. \tag{2}$$

Let $C_{\eta_i} := \eta_i - \frac{L(1+C_{\mathrm{sgd}})\eta_i^2}{2}$ and $A_{\boldsymbol{K}} := \frac{1}{TN} \sum_{i=1}^{N} \frac{\mathbb{E} f_i^0(\theta_i^{0,0}) - f_{\inf}}{K_i C_{\eta_i}} + \frac{1}{N} \sum_{i=1}^{N} \frac{L\eta_i^2 \sigma_{\mathrm{sgd}}^2}{2C_{\eta_i}}$. The term $A_{\boldsymbol{K}}$ captures the standard stochastic gradient descent (SGD) contribution from the initial objective gaps and stochastic-gradient noise. For compactness of writing a bound on S_0, we also define, for $R \geq 0$,

$$\Phi_{\boldsymbol{K}}(R) := \min \left\{ \left(\max_i K_i \right) R, \left[\left(\max_i \alpha_i \right) \sqrt{R} + \sqrt{C_{\mathrm{sgd}} \left(\max_i \beta_i \right)^2 R + \frac{\sigma_{\mathrm{sgd}}^2}{2} \left(\frac{1}{N} \sum_{i=1}^{N} \beta_i^2 \right)} \right]^2 \right\},$$

where $\alpha_i := 1 + L\eta_i \sqrt{\frac{(K_i - 1)(2K_i - 1)}{6}}$ and $\beta_i := L\eta_i \frac{K_i - 1}{\sqrt{K_i}}$ reflect the effect of local updates within each round. Note that both branches of $\Phi_{\boldsymbol{K}}$ are strictly increasing in R, and so is $\Phi_{\boldsymbol{K}}$.

Theorem 1. *Under Assumptions 1-6, $S_0 \leq \Phi_{\boldsymbol{K}}(Q)$, where $Q := A_{\boldsymbol{K}} + B r_{\boldsymbol{K}} + \tilde{B} \sqrt{\sigma_{\mathrm{sgd}}^2 + C_{\mathrm{sgd}} r_{\boldsymbol{K}}^2}$, $r_{\boldsymbol{K}} := \frac{B + \tilde{B}\sqrt{C_{\mathrm{sgd}}}}{2} + \frac{1}{2} \sqrt{\left(B + \tilde{B}\sqrt{C_{\mathrm{sgd}}} \right)^2 + 4 \left(A_{\boldsymbol{K}} + \tilde{B} \sigma_{\mathrm{sgd}} \right)}$, $B := L_d L_h \sqrt{\frac{1}{N} \sum_{j=1}^{N} \Gamma_j^2}$, and $\tilde{B} := L_d L_h \sqrt{\frac{1}{N} \sum_{j=1}^{N} \frac{\Gamma_j^2}{K_j}}$, $\Gamma_j := K_j \eta_j \sum_{i=1}^{N} \frac{(1-\rho_i) w_{ij}}{K_i C_{\eta_i}}$.*

The proof is given in Appendix C.2.1. The terms in Q beyond $A_{\boldsymbol{K}}$ account for drift in the distillation objective caused by updates to the nodes' models. The coefficient Γ_j quantifies how node j contributes to this drift: it combines node j's influence on each recipient i, measured by $(1 - \rho_i) w_{ij}$, with the nodes' local step counts and stepsizes. Holding all other quantities fixed, the drift penalty is nondecreasing in the influence coefficients $(1 - \rho_i) w_{ij}$.

For comparison, we next state a FedProto-style bound in Theorem 2, which requires an additional assumption on the upper bound of expected norm of stochastic gradients (Tan et al., 2022).

Assumption 7. *$\exists U \geq 0$ s.t. $\forall i, t, k$, $\mathbb{E}\left[\|\widetilde{\nabla} f_{i,\boldsymbol{\theta}^t}(\theta_i^{t,k})\| \right] \leq U$.*

Theorem 2. *Under Assumptions 1-6 and Assumption 7, $S_0 \leq \Phi_{\boldsymbol{K}}(\tilde{Q})$, where $\tilde{Q} := A_{\boldsymbol{K}} + \frac{L_d L_h U}{N} \sum_{j=1}^{N} \Gamma_j$.*

Note that Assumptions 1-6 do not imply Assumption 7, which is shown by a counterexample in Appendix C.5. The proof of Theorem 2 is based on an extension of the techniques in Tan et al. (2022) to our setting. The details are given in Appendix C.2.2.

Alternatively, there exists also a FEDMOSAIC-style bound as stated in Theorem 3, which relies on another assumption that replaces Assumptions 4 and 5 with direct control of drift in the objective (Abourayya et al., 2025a).

Assumption 8. *$\exists \tilde{U} \geq 0$ s.t. $\forall i, t, \theta$, $f_{i,\boldsymbol{\theta}^{t+1}}(\theta) - f_{i,\boldsymbol{\theta}^t}(\theta) \leq \tilde{U}$.*

Theorem 3. *Under Assumptions 1-3, 6, and 8, $S_0 \leq \Phi_{\boldsymbol{K}} \left(A_{\boldsymbol{K}} + \frac{\tilde{U}}{N} \sum_{i=1}^{N} \frac{1}{K_i C_{\eta_i}} \right)$.*

The proof of Theorem 3 is based on an extension of the techniques in Abourayya et al. (2025a) to our setting. The details are given in Appendix C.2.3.

We remark that the original convergence bounds in Tan et al. (2022) and Abourayya et al. (2025a) were derived for individual nodes. To facilitate comparison with Theorem 1, Theorems 2 and 3 instead bound averaged stationarity over nodes. Theorems 2 and 3 also extend the corresponding analysis techniques to heterogeneous local step counts K_i and stepsizes η_i, which are not explicitly treated in the original bounds.

Interpretation. In all three bounds, the terms added to $A_{\boldsymbol{K}}$ capture the effect of objective drift induced by the neighbors' evolving models. The three bounds differ in how they control this drift. Theorem 1 relates it to the neighbors' local gradients and stochastic noise, allowing the drift estimate to reflect reductions in the neighbors' gradient magnitudes during training. The other two arguments use uniform bounds on neighbor movement or objective drift, expressed through fixed U or $\tilde{U}$, respectively, which do not capture such reductions. When these constants are conservative, Theorem 1 can yield a tighter bound on S_0. The comparisons below detail when this improvement occurs.

2.2 Comparing Theorem 1 against Theorem 2

First, the proposed bound $\Phi_{\boldsymbol{K}}(Q)$ in Theorem 1 does not require Assumption 7. It is applicable without requiring a uniform bound on expected stochastic-gradient norms.

Next, we show that $\Phi_{\boldsymbol{K}}(Q)$ can be tighter than the bound $\Phi_{\boldsymbol{K}}(\tilde{Q})$ in Theorem 2 in a wide range of scenarios. Note when $B = 0$, then $\tilde{B} = 0$ and $Q = \tilde{Q} = A_{\boldsymbol{K}}$, so the two bounds coincide. Therefore, we focus on non-trivial cases where $B > 0$, which can be satisfied when $\exists i$ s.t. $\rho_i < 1$.

A sufficient condition for $\Phi_{\boldsymbol{K}}(Q) \leq \Phi_{\boldsymbol{K}}(\tilde{Q})$ is given below. Its proof is given in Appendix C.3.1, which also contains additional analysis on the tightness comparison between Q and $\tilde{Q}$.

Theorem 4. *Suppose* $B > 0$ *and* $\chi_\Gamma U > (\tilde{B}/B)\sigma_{\mathrm{sgd}}$. *Define* $r_u := \frac{(\chi_\Gamma U)^2 - \left((\tilde{B}/B)\sigma_{\mathrm{sgd}}\right)^2}{\chi_\Gamma U + (\tilde{B}/B)\sqrt{\sigma_{\mathrm{sgd}}^2 + C_{\mathrm{sgd}}((\chi_\Gamma U)^2 - \left((\tilde{B}/B)\sigma_{\mathrm{sgd}}\right)^2)}}$ *and* $E_{\max} := \frac{\chi_\Gamma U r_u^2}{(\tilde{B}/B)\sigma_{\mathrm{sgd}} + \left(1 + (\tilde{B}/B)\sqrt{C_{\mathrm{sgd}}}\right) r_u}$. *If* $\tilde{Q} \leq E_{\max}$, *then* $Q \leq \tilde{Q}$, *and consequently* $\Phi_{\boldsymbol{K}}(Q) \leq \Phi_{\boldsymbol{K}}(\tilde{Q})$.

For fixed local step counts, stepsizes, and problem constants, increasing T reduces the initial objective-gap contribution to $\tilde{Q}$ while leaving $E_{\max}$ unchanged. Theorem 4 can therefore certify an improvement achieved by Theorem 1 once this initial objective-gap contribution is sufficiently small, provided the remaining noise and drift contributions lie below $E_{\max}$. When all local step counts increase proportionally, with remaining parameters and constants held fixed, $E_{\max}$ increases, while $\tilde{Q}$ decreases, as shown in Appendix C.4. The sufficient condition in Theorem 4 therefore becomes easier to satisfy. Furthermore, teacher influence also affects the comparison. Consider a proportional reduction $\Gamma_j \mapsto c\Gamma_j$ for every node, where $0 < c < 1$, with the remaining quantities held fixed. This scales both B and $\tilde{B}$ by c, while leaving $\tilde{B}/B$ and χ_Γ unchanged. Consequently, $E_{\max}$ remains unchanged and $\tilde{Q}$ decreases, also making the sufficient condition easier to satisfy.

The advantage of Theorem 1 stems from its refined control of the neighbors' shifts, as detailed in Appendix C.3.2.

2.3 Comparing Theorem 1 against Theorem 3

Theorem 1 controls objective drift through the neighbors' local gradients and stochastic noise, whereas Theorem 3 uses a uniform drift constant. Retaining information about the updates that change the neighbors' predictions can yield a tighter bound when the uniform drift estimate is conservative. Appendix C.6 compares the used assumptions and gives the exact condition under which this improvement occurs.

3 Generalization Analysis and Ensemble Weight Matrix Optimization

In this section, we derive a generalization bound for the models produced by Algorithm 1. We then illustrate its application to ensemble-weight optimization.

3.1 Generalization Bound

We consider generalization performance of the output models $\hat{\boldsymbol{\theta}} := (\hat{\theta}_j)_{j=1}^N$, selected as $\hat{\theta}_j := \theta_j^{t^\star}$ where $t^\star \in \arg\min_{0 \leq t \leq T-1} \frac{1}{N}\sum_{i=1}^N \mathbb{E}\left\|\nabla f_{i,\boldsymbol{\theta}^t}(\theta_i^t)\right\|^2$. We consider a common trade-off factor

$0 < \rho \leq 1$ and the case where the training and test data of node i follow $\mathcal{P}^{(i)}$. We denote the reference data distribution by $\mathcal{P}^{\mathrm{re}}$, and the its input marginal distribution by $\mathcal{P}_X^{\mathrm{re}}$. We allow the local input marginal distributions $\mathcal{P}_X^{(i)}$ to differ across nodes and from $\mathcal{P}_X^{\mathrm{re}}$, but assume that all nodes and the reference set share the same conditional label distribution $\mathcal{P}_{Y|X}$. We consider classification tasks and evaluate generalization using the expected total variation distance

$$R_i(\theta) := \mathbb{E}_{X\sim\mathcal{P}_X^{(i)}}\left[\mathrm{TV}\left(\mathcal{P}_{Y|X}, \mathcal{Q}_{i,\theta,X}\right)\right], \tag{3}$$

where $\mathrm{TV}\left(\mathcal{P}_{Y|X}, \mathcal{Q}_{i,\theta,X}\right) := \sup_{A\in\mathcal{B}_\mathcal{Y}} \left|\mathcal{P}_{Y|X}(A) - \mathcal{Q}_{i,\theta,X}(A)\right|$ and $\mathcal{B}_\mathcal{Y}$ is the Borel sigma-field on $\mathcal{Y}$. The hypothesis $h_{i,\theta}(X)$ induces the probability measure $\mathcal{Q}_{i,\theta,X}$ through $\mathcal{Q}_{i,\theta,X}(A) = \sum_{l\in A}[h_{i,\theta}(X)]_l, \forall A \in \mathcal{B}_\mathcal{Y}$, where $[h_{i,\theta}(X)]_l$ denotes the l-th component of the prediction vector. Denote the input-distribution discrepancies by $A_{ij} := \mathrm{TV}\left(\mathcal{P}_X^{(i)}, \mathcal{P}_X^{(j)}\right)$ and $b_i := \mathrm{TV}\left(\mathcal{P}_X^{(i)}, \mathcal{P}_X^{\mathrm{re}}\right)$.

To translate the stationarity guarantee in Theorem 1 into a generalization bound, we impose the following Polyak-Łojasiewicz (PL) condition.

Assumption 9. *$\forall i$, $\exists \mu_i > 0$ s.t., for every model tuple $\hat{\boldsymbol{\theta}}$, $\frac{1}{2\mu_i}\left\|\nabla f_{i,\hat{\boldsymbol{\theta}}}(\hat{\theta}_i)\right\|^2 \geq f_{i,\hat{\boldsymbol{\theta}}}(\hat{\theta}_i) - \inf_{\theta\in\Theta_i} f_{i,\hat{\boldsymbol{\theta}}}(\theta)$.*

We also assume an upper bound on each node's expected minimum empirical composite loss over its parameter space Θ_i.

Assumption 10. *For every node i, there exists $\epsilon_i \geq 0$ s.t., $\mathbb{E}\left[\inf_{\theta\in\Theta_i} f_{i,\hat{\boldsymbol{\theta}}}(\theta)\right] \leq \epsilon_i$.*

The Rademacher complexity associated with node i is $\Re_i := \mathbb{E}_{\hat{\mathcal{D}}_i,\boldsymbol{\sigma}_i}\left[\sup_{\theta_i\in\Theta_i} \frac{1}{|\hat{\mathcal{D}}_i|}\sum_{l=1}^{|\hat{\mathcal{D}}_i|} \sigma_l [h_{i,\theta_i}(x_l)]_{y_l}\right]$, where $\boldsymbol{\sigma}_i = (\sigma_l)_{l=1}^{|\hat{\mathcal{D}}_i|}$ consists of independent Rademacher random variables.

Define the loss function-dependent residual r, which equals $(1-\rho)M/4$ when ℓ and d are KL divergences, or equals $(1-\rho)/8$ when ℓ is KL divergence and d is squared Euclidean norm.

For $0 < \delta < 1$, collect the sampling, and distribution-shift penalties in $\boldsymbol{B} := (B_i)_{i=1}^N$, where

$$\begin{aligned} B_i := &\left[\frac{\log(N/\delta)}{2}\left(\frac{\rho^2\big(1+(1-\rho)w_{ii}\big)^2}{|\hat{\mathcal{D}}_i|} + \rho^2(1-\rho)^2\sum_{j\neq i}\frac{w_{ij}^2}{|\hat{\mathcal{D}}_j|} + \frac{(1-\rho)^2(2-\rho)^2}{|\hat{\mathcal{D}}^{\mathrm{re}}|}\right)\right]^{1/2} \\ &+ (1-\rho)\sum_j w_{ij}\big[2(2-\rho)A_{ij} + b_j\big] + (1-\rho)^2\sum_{j,k} w_{ij}w_{jk}\big[2A_{ik} + b_k\big]. \end{aligned} \tag{4}$$

Define $\boldsymbol{R} := \left(\mathbb{E}R_i(\hat{\theta}_i)\right)_{i=1}^N$, $\boldsymbol{\epsilon} := (\epsilon_i)_{i=1}^N$, $\boldsymbol{\Re} := (\Re_i)_{i=1}^N$, and $\boldsymbol{\lambda} := \left(\max_j \frac{\left[(I-(1-\rho)W)^{-1}\right]_{ij}}{\mu_j}\right)$.

Theorem 5 (Network-dependent generalization)**.** *Suppose the conditions of Theorem 1 and Assumptions 9 and 10 hold. Then, for every $0 < \delta < 1$, with probability at least $1-\delta$ over the datasets $\{\hat{\mathcal{D}}_l\}_{l=1}^N$ and $\hat{\mathcal{D}}^{\mathrm{re}}$,*

$$\boldsymbol{R} \preceq (I - qW)^{-1}(\boldsymbol{\epsilon} + 2\rho\boldsymbol{\Re}) + (I - q^2W^2)^{-1}\boldsymbol{B} + \frac{q+r}{\rho}\mathbf{1} + \frac{N\Phi_{\boldsymbol{K}}(Q)}{2}\boldsymbol{\lambda}, \tag{5}$$

Here $\preceq$ denotes componentwise inequality, I is the $N \times N$ identity matrix, and $\mathbf{1}$ is the all-ones vector. Both inverse matrices exist and are elementwise nonnegative.

The proof is given in Appendix C.7.1. We first use the triangle inequality for total variation to bound each node's local risk in terms of its neighbors' risks, the expected distillation losses, and distribution discrepancies. We then apply the McDiarmid's inequality and Rademacher complexity to bound the expected composite losses by their empirical counterparts evaluated at the output models, together with statistical penalties. This yields a coupled system of inequalities in which each node's risk is bounded by an ensemble of other nodes' risks and additional terms, eventually leading to the bound in (5).

This bound characterizes how each node's generalization guarantee depends on its interactions with the network. Through distillation, a node can utilize information learned from the other nodes' private data, while its guarantee also depends on the optimization residuals, hypothesis set complexities, sampling penalties, and distribution mismatches. In particular, we make the following several observations.

Distribution-compatible and student-aware teachers. In (4), the one-hop term $2(1-\rho)(2-\rho)\sum_j w_{ij}A_{ij}$ favors teachers whose input distributions are closer to node i's, with other contributions held fixed. Along the information-flow path $k \to j \to i$, the two-hop term $2(1-\rho)^2\sum_{j,k} w_{ij}w_{jk}A_{ik}$ measures the mismatch between the original source k and the recipient i. Thus, when choosing its own teachers, node j can also account for the distributions of its downstream students. For $0 < \rho < 1$, the ratio of the two-hop to one-hop coefficients is $(1-\rho)/(2-\rho) < 1/2$, which discounts the indirect impact.

Reference mismatch and statistical reliability. In (4), the terms $(1-\rho)\sum_j w_{ij}b_j$ and $(1-\rho)^2\sum_{j,k} w_{ij}w_{jk}b_k$ penalize reliance on teachers whose local input distributions differ from the reference distribution, with an additional factor $1-\rho$ for indirect teachers. A large reference mismatch can hinder the transfer of a teacher's knowledge through distillation, but it does not by itself imply that the teacher cannot provide useful information. For statistical reliability, the factors $w_{ij}^2/|\hat{\mathcal{D}}_j|$ discourage concentrating weight on teachers with small local datasets, whereas the Rademacher terms account for hypothesis-set complexity. The coefficient $(1+(1-\rho)w_{ii})^2$ combines the contribution from local supervision with the self-teacher contribution, which is larger than the non-self coefficients $((1-\rho)w_{ij})^2$. Finally, increasing the reference-set size reduces the related sampling penalty, but it does not change the distribution discrepancies b_j.

Propagation through the network. The inverse matrices describe how the node-dependent residuals and penalties discussed above contribute to each node's generalization guarantee through information propagation in the network. We note the Neumann expansions $\big(I-(1-\rho)W\big)^{-1} = \sum_{\tau=0}^{\infty}(1-\rho)^\tau W^\tau$ and $\big(I-(1-\rho)^2W^2\big)^{-1} = \sum_{\tau=0}^{\infty}(1-\rho)^{2\tau}W^{2\tau}$. Each $(W^\tau)_{ij}$ sums products of weights along dependency walks of length τ from i to j, corresponding to information flow from j to i. Longer walks are discounted by powers of $1-\rho$. The even powers in the second expansion arise from the two-step risk recursion, but the resulting term still captures odd-hop effects because $\boldsymbol{B}$ contains one-hop terms. Overall, W has two coupled roles: it influences the quality of each node's local ensemble teacher and controls how information propagates through the network.

The role of ρ. Row stochasticity gives $\Big\|\big(I-(1-\rho)W\big)^{-1}\Big\|_\infty = \frac{1}{\rho}$ and $\Big\|\big(I-(1-\rho)^2W^2\big)^{-1}\Big\|_\infty = \frac{1}{\rho(2-\rho)}$, where $\|\cdot\|_\infty$ denotes the induced matrix infinity norm. As ρ approaches zero, these norms grow and the additive term $(1-\rho)/\rho$ diverges, so the bound can become loose, which is consistent with weaker reliance on labeled supervision. As ρ approaches one, local supervision receives greater weight and the impact of information propagation vanish. At $\rho = 1$, the bound reduces to the plain form $\mathbb{E}R_i(\hat{\theta}_i) \le \frac{s_i}{2\mu_i} + \epsilon_i + 2\Re_i + \sqrt{\frac{\log(N/\delta)}{2|\hat{\mathcal{D}}_i|}}$, which contains only the local optimization residual, minimum-loss term, hypothesis-set complexity, and sampling penalty, without network contribution. Thus, ρ controls the trade-off between local supervision and knowledge transfer.

3.2 Application of Generalization Analysis: Bound-Guided Ensemble weights

The preceding analysis shows that the ensemble weights contained in W influence both the local ensemble teachers and the propagation of information through the network. We may use the average generalization risk as objective to guide the selection of ensemble weights. However, the average of the right-hand side of (5) is not directly usable as a numerical objective since it involves quantities such as the Lipschitz constants and PL constants that are generally unavailable. We therefore derive a surrogate that removes the dependence of selected terms on W while retaining the weight-dependent contributions. It is given by

$$\frac{1}{N}\sum_i \mathbb{E}R_i(\hat{\theta}_i) \le \frac{(1-\rho)+r}{\rho} + \frac{N\tilde{\Phi}}{2\rho}\max_j \frac{1}{\mu_j} + \frac{1}{\rho}\sum_j \epsilon_j + \sum_j 2\Re_j + \frac{1}{\gamma}\sum_i \pi_i B_i, \quad (6)$$

where $\tilde{\Phi}$ is the resulting quantity by replacing $d_j, \forall j$, in $\Phi_{\boldsymbol{K}}(Q)$ by $\sum_i \frac{(1-\rho_i)}{K_i C_{\eta_i}}$. The average personal risk bound and the derivation details for (6) are given in Appendix C.8.

In (6), the only remaining W-dependent term is $\sum_{i=1}^N \pi_i B_i$. This gives a surrogate problem

$$P_1: \quad \min_W \sum_{i=1}^N \pi_i B_i$$
$$\text{s.t.} \quad W \ge 0, \quad W\mathbf{1} = \mathbf{1}, \text{ and } w_{ij} = 0 \text{ for disallowed edges.} \quad (7)$$

However, the input-distribution discrepancies A_{ij} and b_i are not directly observed. To avoid exchanging raw inputs for discrepancy estimation, we use label-marginal discrepancies as tractable proxies. In particular, we replace A_{ij} and b_i in B_i with $\mathrm{TV}\left(\mathcal{P}_Y^{(i)}, \mathcal{P}_Y^{(j)}\right)$ and $\mathrm{TV}\left(\mathcal{P}_Y^{(i)}, \mathcal{P}_Y^{\mathrm{re}}\right)$, respectively, while leaving all other terms unchanged. We denote the resulting quantity by $\tilde{B}_i$.

In implementation, we estimate the label marginals $\mathcal{P}_Y^{(i)}$ and $\mathcal{P}_Y^{\mathrm{re}}$ using empirical class proportions in the local training set, denoted by $\widehat{p}_i$, and in the reference sets, denoted by $\widehat{p}_{\mathrm{re}}$, respectively. To compute these two quantities, nodes need to share their empirical class proportions rather than individual points. We also assume that the reference set's empirical class proportions are available. For example, when nodes contribute samples to the reference set, they can report the corresponding class counts. The individual reference labels need not be disclosed or used during distillation. Let $\widehat{B}_i$ denote the resulting empirical version of $\tilde{B}_i$. We can optimize the empirical approximated counterpart of P_1 by minimizing $\sum_{i=1}^N \pi_i \widehat{B}_i$ over W while keeping the constraints unchanged. We denote the resulting problem by $\hat{P}_1$, which can be solved by standard projected gradient descent. Appendix C.9 analyzes the distribution discrepancy approximation error, the finite-sample estimation error and their effect on the original quantity $\sum_{i=1}^N \pi_i B_i$. It also shows that the generalization guarantee remains valid after replacing A_{ij} and b_i with $\mathrm{TV}(\widehat{p}_i, \widehat{p}_j)$ and $\mathrm{TV}(\widehat{p}_i, \widehat{p}_{\mathrm{re}})$, respectively, provided that the corresponding error terms are included and the confidence level is adjusted accordingly.

4 Numerical Experiments

We investigate whether selecting ensemble weights using the proposed generalization bound surrogate improves average personalized test accuracy across nodes.

Setup and evaluation. We evaluate the performance of online distillation on MNIST, FashionMNIST, and CIFAR-10. We consider 20 nodes: 15 of them have 500 private training examples each, and five have 2,500. To vary label heterogeneity, we consider $d \in \{2, 4, 6, 8, 10\}$ dominant classes per node. For $d < 10$, the nodes are divided into five groups of four, with a common dominant class set within each group and different sets across groups. Each dominant class is assigned six times the sampling probability of each non-dominant class. The case $d = 10$ uses uniform class probabilities.

Nodes share a 10,000-example reference set and exchange predictions over a complete graph. All nodes use the same architecture: a 44,426-parameter LeNet for MNIST and FashionMNIST, and a 1,116,970-parameter CNN for CIFAR-10. We sweep ρ between 0.01 and 0.95. Training lasts 15,000 rounds on MNIST and CIFAR-10 and 25,000 rounds on FashionMNIST, with one local update per round.

Each node is evaluated on the official test set. Its personalized test accuracy is the weighted average of its class-wise accuracies, using the empirical class proportions in its private training set as weights. The reported average test accuracy averages these personalized accuracies equally across the 20 nodes. Appendix D provides more implementation details and full experiment results. Throughout this section and Appendix D, the reported mean test accuracies are with 95% confidence intervals computed across independent runs.

Table 1: Best observed averaged test accuracy (%) over ρ. The best baseline is selected separately for each setting and each method selects its own ρ.

Dataset	Comparison	$d = 2$	$d = 4$	$d = 6$	$d = 8$	$d = 10$
FashionMNIST	GBOW (ours)	$\mathbf{86.13 \pm 0.08}$	$\mathbf{86.37 \pm 0.09}$	$\mathbf{85.39 \pm 0.08}$	$\mathbf{83.47 \pm 0.10}$	$\mathbf{82.44 \pm 0.10}$
	Best baseline	$85.10 \pm 0.08^{\mathrm{C}}$	$85.08 \pm 0.11^{\mathrm{C}}$	$83.75 \pm 0.06^{\mathrm{C}}$	$81.99 \pm 0.11^{\mathrm{C}}$	$82.16 \pm 0.09^{\mathrm{T}}$
MNIST	GBOW (ours)	$\mathbf{98.28 \pm 0.03}$	$\mathbf{98.25 \pm 0.02}$	$\mathbf{98.19 \pm 0.03}$	98.00 ± 0.02	$\mathbf{98.05 \pm 0.06}$
	Best baseline	$98.19 \pm 0.04^{\mathrm{C}}$	$98.20 \pm 0.02^{\mathrm{C}}$	$98.08 \pm 0.03^{\mathrm{T}}$	$\mathbf{98.05 \pm 0.02}^{\mathrm{T}}$	$98.02 \pm 0.04^{\mathrm{T}}$
CIFAR-10	GBOW (ours)	$\mathbf{63.59 \pm 0.17}$	$\mathbf{61.35 \pm 0.16}$	$\mathbf{59.87 \pm 0.17}$	$\mathbf{57.20 \pm 0.22}$	$\mathbf{56.53 \pm 0.33}$
	Best baseline	$60.34 \pm 0.23^{\mathrm{C}}$	$58.89 \pm 0.26^{\mathrm{C}}$	$57.90 \pm 0.67^{\mathrm{T}}$	$56.94 \pm 0.29^{\mathrm{T}}$	$55.61 \pm 0.32^{\mathrm{T}}$

Best-baseline labels: $^{\mathrm{C}}$ COMET; $^{\mathrm{T}}$ TSIZE.

Comparators. We call the proposed method, which selects ensemble weights by solving $\hat{P}_1$, **G**eneralization **BO**und-guided ensemble **W**eight design (**GBOW**). We compare GBOW with seven teacher-design rules: **UNIFO**, equal weights (Park et al., 2023); **ENTRO**, greater weights on lower-entropy predictions (Deng et al., 2023); **VARIA**, weights proportional to class-probability variance (Cho et al., 2022); **TSIZE**, weights proportional to training-set size (Park et al., 2024); **COMET**, a teacher formed from the nearest prediction-cluster centroid (Cho et al., 2023); **SELEC**, density-ratio-based selective sharing (Shao et al., 2024); and **CORRE**, greater weight on less correlated predictions (Shi et al., 2023). For a controlled comparison, we incorporate these ensemble-weight design methods into a common training procedure, so that our comparison assesses the effects of different weighting rules, rather than each cited system in its original form.

Performance across data partitions and datasets. As shown in Table 1, GBOW has the highest observed best-over-ρ mean in 14 of 15 settings. For example, its advantage over the strongest baseline is up to 1.64 percentage points on FashionMNIST and up to 3.25 on CIFAR-10. Further details are given in Appendix D.3, where Figures 2-4 show that GBOW achieves the highest accuracy and remains competitive across the evaluated values of ρ. Unlike all baseline methods, GBOW explicitly incorporates ρ into its weight optimization, allowing the ensemble weights to account for the trade-off between the original loss and KD loss. This leads to strong competitive performance across different settings.

Further experimental results. We present further experimental results, including the effect of model architectural heterogeneity in Appendix D.4 and communication-graph sparsity in Appendix D.5. They confirm the observed performance advantage of GBOW over a wide range of scenarios.

5 Conclusion

We have developed improved convergence and generalization analyses for online knowledge distillation over decentralized networks with heterogeneous models and data distributions. Our convergence analysis controls objective drift by accounting for neighbors' model updates, without relying on uniform bounds on expected stochastic-gradient norms or objective drift. This approach yields tighter finite-round guarantees than those obtained using related analysis techniques in expansive parameter regimes. Combining this analysis with our generalization results provides a bound on the personalized test risks of the finite-round output models, expressed solely in terms of problem constants and algorithm parameters without any learned-model-dependent term. The bound clarifies how data heterogeneity, dataset sizes, trade-off factor, and ensemble weights affect the system generalization performance through the information propagation over the network. A surrogate of this bound enables optimization over tunable parameters such as the ensemble weights. In numerical experiments, we show the optimized ensemble weights based on the generalization bound can empirically improve average personalized test accuracy.

AI USE STATEMENT

We used generative AI tools to assist with improving the manuscript's readability, code implementation and refinement, and checking correctness of analysis. We did not use these tools to generate research idea, develop theoretical framework, or provide key mathematical arguments. We also did not use AI tools for any synthetic data generation, translation, dataset cleaning or reformatting, or qualitative or thematic data analysis. The authors manually reviewed and tested all AI-assisted code and checked and revised all other AI-assisted outputs. We take responsibility for the final content of this work, including any text, claims, or artifacts produced with AI assistance.

REPRODUCIBILITY STATEMENT

We provide the materials needed to assess both the theoretical and empirical results. Appendixs 2.1 and 3.1 state the assumptions required for each theorem, and Appendix C gives detailed proofs of the key mathematical claims. Appendix D documents the implementation of our proposed method and the baseline methods, including the choices needed to reproduce the reported comparisons. Code is provided in the supplementary file.

REFERENCES

Amr Abourayya, Jens Kleesiek, Bharat Rao, and Michael Kamp. Whom to trust? adaptive collaboration in personalized federated learning. *arXiv preprint arXiv:2507.00259*, 2025a.

Amr Abourayya, Jens Kleesiek, Kanishka Rao, Erman Ayday, Bharat Rao, Geoffrey I Webb, and Michael Kamp. Little is enough: Boosting privacy by sharing only hard labels in federated semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025b.

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

Rohan Anil, Gabriel Pereyra, Alexandre Passos, Robert Ormandi, George E Dahl, and Geoffrey E Hinton. Large scale distributed neural network training through online distillation. *arXiv preprint arXiv:1804.03235*, 2018.

Ilai Bistritz, Ariana Mann, and Nicholas Bambos. Distributed distillation for on-device learning. In *Proceedings of the International Conference on Neural Information Processing Systems*, 2020.

Yae Jee Cho, Andre Manoel, Gauri Joshi, Robert Sim, and Dimitrios Dimitriadis. Heterogeneous ensemble knowledge transfer for training large models in federated learning. *arXiv preprint arXiv:2204.12703*, 2022.

Yae Jee Cho, Jianyu Wang, Tarun Chirvolu, and Gauri Joshi. Communication-efficient and model-heterogeneous personalized federated learning via clustered knowledge transfer. *IEEE Journal of Selected Topics in Signal Processing*, 17(1):234–247, 2023.

Sayantan Chowdhury, Ben Liang, Ali Tizghadam, and Ilijc Albanese. Communication-efficient loss minimization over heterogeneous data with federated hierarchical ensemble aggregation via distillation. In *OPT 2024: Optimization for Machine Learning*, 2024.

Yongheng Deng, Ju Ren, Cheng Tang, Feng Lyu, Yang Liu, and Yaoxue Zhang. A hierarchical knowledge transfer framework for heterogeneous federated learning. In *IEEE Conference on Computer Communications*, 2023.

Lucas Qingyang Fang, Tiyao Liu, Jinhao Jing, Zeji Li, Kaijie Chen, Harikrishna Kuttivelil, and Katia Obraczka. Convergence theory of knowledge distillation in asynchronous P2P gossip learning network. *arXiv preprint arXiv:2609.01952*, 2026.

Dylan J Foster, Spencer Greenberg, Satyen Kale, Haipeng Luo, Mehryar Mohri, and Karthik Sridharan. Hypothesis set stability and generalization. In *Proceedings of the International Conference on Neural Information Processing Systems*, 2019.

Shayan Mohajer Hamidi, Xizhen Deng, Renhao Tan, Linfeng Ye, and Ahmed Hussein Salamah. How to train the teacher model for effective knowledge distillation. In *European Conference on Computer Vision*, 2024.

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

Daniel Hsu, Ziwei Ji, Matus Telgarsky, and Lan Wang. Generalization bounds via distillation. *arXiv preprint arXiv:2104.05641*, 2021.

Won-Jun Jang, Hyeon-Seo Park, and Si-Hyeon Lee. Provably near-optimal federated ensemble distillation with negligible overhead. In *International Conference on Machine Learning*, 2025.

Eunjeong Jeong, Seungeun Oh, Hyesung Kim, Jihong Park, Mehdi Bennis, and Seong-Lyun Kim. Communication-efficient on-device machine learning: Federated distillation and augmentation under non-IID private data. *arXiv preprint arXiv:1811.11479*, 2018.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, 2020.

Nikola Konstantinov and Christoph Lampert. Robust learning from untrusted sources. In *International Conference on Machine Learning*, 2019.

Bingying Li and Haiyun He. On the generalization of knowledge distillation: An information-theoretic view. *arXiv preprint arXiv:2605.13143*, 2026.

Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi. Ensemble distillation for robust model fusion in federated learning. In *Proceedings of the International Conference on Neural Information Processing Systems*, 2020.

Lumin Liu, Jun Zhang, Shenghui Song, and Khaled B Letaief. Communication-efficient federated distillation: Theoretical analysis and performance enhancement. *IEEE Internet of Things Journal*, 11(23):37959–37973, 2024.

Xiao Ma, Hong Shen, Hui Tian, Wei Ke, and Wenqi Lyu. Robust decentralized federated distillation via multi-modality knowledge collaboration. *arXiv preprint arXiv:2609.07230*, 2026.

Aditya K Menon, Ankit Singh Rawat, Sashank Reddi, Seungyeon Kim, and Sanjiv Kumar. A statistical perspective on distillation. In *International Conference on Machine Learning*, 2021.

Itai Morad, Nir Shlezinger, and Yonina Eldar. SGD-based knowledge distillation with Bayesian teachers: Theory and guidelines. In *International Conference on Learning Representations*, 2026.

Sejun Park, Kihun Hong, and Ganguk Hwang. Towards understanding ensemble distillation in federated learning. In *International Conference on Machine Learning*, 2023.

Sejun Park, Kihun Hong, and Ganguk Hwang. A kernel perspective on distillation-based collaborative learning. In *Proceedings of the International Conference on Neural Information Processing Systems*, 2024.

Minh Pham, Minsu Cho, Ameya Joshi, and Chinmay Hegde. Revisiting self-distillation. *arXiv preprint arXiv:2206.08491*, 2022.

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

Mher Safaryan, Alexandra Peste, and Dan Alistarh. Knowledge distillation performs partial variance reduction. In *Proceedings of the International Conference on Neural Information Processing Systems*, 2023.

Jiawei Shao, Fangzhao Wu, and Jun Zhang. Selective knowledge sharing for privacy-preserving federated distillation without a good teacher. *Nature Communications*, 15(1):349, 2024.

Luyao Shi, Ning Jiang, Jialiang Tang, and Xinlei Huang. Correlation guided multi-teacher knowledge distillation. In *International Conference on Neural Information Processing*, 2023.

Yue Tan, Guodong Long, Lu Liu, Tianyi Zhou, Qinghua Lu, Jing Jiang, and Chengqi Zhang. FedProto: Federated prototype learning across heterogeneous clients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

Akihito Taya, Takayuki Nishio, Masahiro Morikura, and Koji Yamamoto. Decentralized and model-free federated learning: Consensus-based distillation in function space. *IEEE Transactions on Signal and Information Processing over Networks*, 8:799–814, 2022.

Haozhao Wang, Yichen Li, Wenchao Xu, Ruixuan Li, Yufeng Zhan, and Zhigang Zeng. DAFKD: Domain-aware federated knowledge distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Dezhong Yao, Wanning Pan, Yutong Dai, Yao Wan, Xiaofeng Ding, Chen Yu, Hai Jin, Zheng Xu, and Lichao Sun. FedGKD: Toward heterogeneous federated learning via global knowledge distillation. *IEEE Transactions on Computers*, 73(1):3–17, 2023.

Taesun Yeom, Taehyeok Ha, and Jaeho Lee. Generalization analysis of linear knowledge distillation. In *High-Dimensional Learning Dynamics*, 2026.

Xiaohui Zhang, Daniel Povey, and Sanjeev Khudanpur. A diversity-penalizing ensemble training method for deep learning. In *Proceedings of Interspeech*, 2015.

Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

Qingyue Zhao and Banghua Zhu. Towards the fundamental limits of knowledge transfer over finite domains. In *International Conference on Learning Representations*, 2024.

A Related works

A.1 Knowledge Distillation and Online Distillation

Knowledge distillation (KD) is a training paradigm in which a student model is trained to match the outputs of a teacher model on the same inputs, typically while also learning from the corresponding ground-truth labels (Hinton et al., 2015). In conventional offline KD, the teacher is pretrained and remains fixed during student training, which may be either a single model or an ensemble of models.

In contrast, online distillation jointly trains multiple peer models without requiring a fixed pretrained teacher. Each node learns from ground-truth supervision and the evolving predictions of other peer nodes, enabling collaboration without a separate teacher-training stage or centralized access to all training data. Early related approaches include ensemble training (Zhang et al., 2015), codistillation (Anil et al., 2018), deep mutual learning (Zhang et al., 2018), federated distillation (Jeong et al., 2018), and distributed distillation (Bistritz et al., 2020). Subsequent work develops variants with different teacher constructions, update rules, and communication protocols while retaining the principle of learning from peers during training (Taya et al., 2022; Cho et al., 2023; Park et al., 2023; Shi et al., 2023; Liu et al., 2024; Abourayya et al., 2025a;b; Ma et al., 2026; Fang et al., 2026). Empirical studies have shown that online distillation can match or outperform conventional offline distillation in some settings. Compared with federated or decentralized learning methods that exchange model parameters or gradients, collaboration through shared outputs can accommodate heterogeneous model architectures. It can also reduce communication when the exchanged outputs are substantially smaller than the corresponding model updates (Bistritz et al., 2020).

Another related line of work combines federated learning and KD (Lin et al., 2020; Cho et al., 2022; Deng et al., 2023; Yao et al., 2023; Wang et al., 2023; Chowdhury et al., 2024; Liu et al., 2024; Jang et al., 2025). These methods use parameter aggregation, knowledge distillation, or both to update one or more server-maintained models and subsequently feed the resulting models or distilled knowledge back to nodes. Online distillation differs in that nodes learn directly from one another's shared predictions without training a centralized global model, and their local models will not be reinitialized by a server-aggregated model.

A.2 Analysis for Offline Distillation

One line of work examines the information conveyed by soft labels beyond that available in hard labels. Authors in Hinton et al. (2015) argue that soft labels encode similarities among classes, while Allen-Zhu & Li (2023) shows that independently trained teachers can capture complementary feature views that are transferable to a student. From a statistical perspective, works by Menon et al. (2021); Hamidi et al. (2024); Yeom et al. (2026) suggest that distillation may improve student generalization performance compared with training on hard labels when the teacher's predictions closely approximate the Bayes conditional class probabilities.

Another line of work examines how accurately students can reproduce their teachers. For linear and deep linear models, Phuong & Lampert (2019) analyzes convergence behavior of the prediction disagreement between the student and teacher. Under suitable assumptions, the student recovers the teacher's prediction when sufficiently many distillation samples are available. The work by Zhao & Zhu (2024) characterizes the statistical limits of approximating a teacher's output distribution under different levels of access to its output information.

From an optimization perspective, Pham et al. (2022) provides empirical evidence that KD can lead students to flatter minima than their teachers trained with hard labels. The analyses in Safaryan et al. (2023) and Morad et al. (2026) show that suitable teacher supervision can reduce variance-related terms in the student's stochastic gradient descent (SGD) convergence bound.

Other studies relate the generalization properties of teachers and students. The work by Foster et al. (2019) derives student risk bounds involving the stability of the teacher's learning algorithm. In the reverse direction, Hsu et al. (2021) bounds the teacher's test risk using the student's empirical loss and hypothesis-set complexity, together with distillation discrepancy. The work by Li & He (2026) bounds the student's expected generalization gap in terms of the teacher's expected generalization gap and a divergence between their learning processes.

These analyses treat the teacher as fixed during student training and therefore do not directly address online distillation, where the teachers themselves evolve. Moreover, the nodes' learning dynamics are coupled since each node influences the predictions used to train its peers, whose subsequent updates in turn influence that node. Analyzing online distillation thus requires accounting for both evolving distillation targets and mutual dependence among the nodes.

A.3 Convergence Analysis for Online Distillation

One line of work exploits the linear structure of kernel methods. The work by Park et al. (2023) studies distributed kernel ridge regression under statistically homogeneous local data distributions. They prove that the local models converge to limiting regressors as the number of communication rounds tends to infinity, but they do not provide a finite-round error bound that explicitly depends on the number of communication rounds. Park et al. (2024) extend this analysis to statistically heterogeneous and unbalanced local datasets. For their DCL-KR algorithm, they bound the error of the models obtained after a finite number of communication rounds selected by an early-stopping rule. Nevertheless, their analysis assumes that the distillation loss in each round is minimized exactly through infinitely many gradient-descent steps.

Beyond kernel-based analyses, the work by Abourayya et al. (2025b) studies online distillation for classification, allowing heterogeneous model architectures. Each node shares its hard-label predictions on the public data points, and a server uses majority voting to generate consensus pseudo-labels for the data points. Assuming that each node's training error monotonically decreases and all nodes share the same training distribution, they show that these consensus labels eventually stabilize with high probability. However, this result does not itself establish convergence of the local models.

In the work by Bistritz et al. (2020), each node maintains auxiliary soft-decision variables to track the global averaged soft decisions. The authors prove that the gradient of each node's composite loss converges to zero almost surely as the number of communication rounds tends to infinity. The work by Cho et al. (2023) subsequently considers a server-based setting in which a central server can collect the logits of all nodes, and return the aggregated information. This server-side aggregation eliminates the decentralized consensus-tracking variables used by Bistritz et al. (2020). Using similar analytical techniques, Cho et al. (2023) also show that the gradient of each node's composite

objective with respect to its local model parameters converges to zero almost surely. However, the convergence guarantees in both works are asymptotic and do not provide a finite-round bound.

The convergence analyses most closely related to ours are those of Tan et al. (2022) and Abourayya et al. (2025a). Both derive finite-round bounds on the time-averaged expected squared gradient norm of each node's composite objective. The former assumes uniformly bounded expected stochastic-gradient norms, whereas the latter assumes uniformly bounded round-to-round drift of each node's time-varying composite objective. They rely on these assumptions to control the effects of inter-node interactions by uniform constants, which may lead to conservative convergence bounds. Our analysis does not assume uniformly bounded expected stochastic-gradient norms or rely on a uniform objective-drift bound. It characterizes inter-node interactions through the neighbors' local gradients and stochastic noise, which can lead to tighter bounds in broad parameter regimes.

Finally, there are complementary works that study different update protocols and convergence criteria. Some analyses consider two-stage methods in which nodes optimize the supervised and distillation losses sequentially (Taya et al., 2022; Liu et al., 2024; Ma et al., 2026). These updates differ from the stochastic-gradient steps on the composite loss considered here, and their guarantees do not directly establish composite-objective stationarity under our update rule. More recently, Fang et al. (2026) analyze asynchronous peer-to-peer distillation, where a randomly activated pair of nodes learns from each other's predictions using a composite loss. They study the stationarity about the original losses at the mean prediction together with the prediction disagreement. In contrast, our setting uses synchronous rounds, with each node distilling from a weighted ensemble of its neighbors. Our guarantees concern stationarity of the nodes' time-varying composite objectives. These differences in update protocols and convergence criteria prevent direct comparison with their bounds.

A.4 Generalization Analysis for Online Distillation

Park et al. (2023) and Park et al. (2024) provide generalization guarantees by showing that the resulting local models approach the population-risk minimizer for the test distribution as the amount of local and public data increases. Both analyses assume a common target distribution and ensemble weighs are prespecified. Uniform weights are adopted in Park et al. (2023), whereas Park et al. (2024) uses weights proportional to the nodes' local training set sizes. More fundamentally, their theoretical guarantees are restricted to kernel regression with a common kernel across nodes, which do not directly extend to general or heterogeneous model architectures.

The work by Cho et al. (2023) considers a general setting in which nodes may have different target distributions and model architectures and allow a general loss function. However, their generalization analysis assumes that each local model is an empirical-risk minimizer on its local training set. This assumption does not characterize the final output of online distillation, since mutual distillation can move each local model away from its original local empirical-risk minimizer. It is worth noting that some KD-based federated learning studies provide generalization bounds as well. However, these bounds also characterize the generalization performance of the ensemble of local risk minimizers, rather than the resulting models produced by the their algorithms (Lin et al., 2020; Cho et al., 2022; Wang et al., 2023; Jang et al., 2025).

Authors in Shao et al. (2024) derive a generalization bound for generic models but do not relate them to the local models produced by their training procedure either. Moreover, the bound retains model-dependent quantities, without characterizing how the algorithm determine them. Consequently, the bound does not explicitly reflect the generalization performance of the algorithm's output models, which also limits its direct use for tunable parameter optimization.

B Online Distillation Algorithm

The online distillation algorithm analysed in this paper is summarized in Algorithm 1 and Algorithm 2.

Algorithm 1 Online distillation

Require: Initial parameters $\{\theta_i^0\}_{i=1}^N$, weights W, trade-off factors $\{\rho_i\}_{i=1}^N$, local step counts $\{K_i\}_{i=1}^N$, stepsizes $\{\eta_i\}_{i=1}^N$, and rounds T

for $t = 0, \dots, T-1$ **do**
  for each node i in parallel **do**
    Independently sample $S_{\text{ref}}^{t,i,k} \subseteq \hat{\mathcal{D}}^{\text{re}}$ for $k = 0, \dots, K_i - 1$.
    Request predictions on $\bigcup_{k=0}^{K_i-1} S_{\text{ref}}^{t,i,k}$ from each node $j \neq i$ with $w_{ij} > 0$.
  end for
  Each node j sends the requested predictions $h_{j,\theta_j^t}(x)$ to the requesting nodes.
  Each node i caches the received predictions and evaluates its own round-start predictions if $w_{ii} > 0$.
  for each node i in parallel, after prediction exchange **do**
    $\theta_i^{t+1} \leftarrow \text{LocalSGD}(i, t)$ using Algorithm 2.
  end for
end for
return $\{\theta_i^T\}_{i=1}^N$

Algorithm 2 Local stochastic gradient descent: LocalSGD(i, t)

Require: Node i, round t, parameter θ_i^t, local step count K_i, stepsize η_i, trade-off factor ρ_i, reference minibatches, and cached ensemble labels

$\theta_i^{t,0} \leftarrow \theta_i^t$
for $k = 0, \dots, K_i - 1$ **do**
  Independently sample a private minibatch $S_i^{t,k} \subseteq \hat{\mathcal{D}}_i$.
  Gradient descent:

$$\theta_i^{t,k+1} \leftarrow \theta_i^{t,k} - \eta_i \frac{\rho_i}{|S_i^{t,k}|} \sum_{(x,y)\in S_i^{t,k}} \nabla_\theta \ell\big(h_{i,\theta}(x), e_y\big)\Big|_{\theta=\theta_i^{t,k}} - \eta_i \frac{1-\rho_i}{|S_{\text{ref}}^{t,i,k}|} \sum_{x\in S_{\text{ref}}^{t,i,k}} \nabla_\theta d\left(h_{i,\theta}(x), \sum_{j=1}^N w_{ij} h_{j,\theta_j^t}(x)\right)\Big|_{\theta=\theta_i^{t,k}}$$

end for
return θ_i^{t,K_i}

C Mathematical Details

C.1 Lemmas and Proofs for Convergence Analysis

In this subsection, we present the lemmas and proofs needed for the subsequent convergence analysis.

For brevity of mathematical expressions, we define the following symbols.

$$a_{i,t,k} := \mathbb{E}\|\nabla f_{i,\boldsymbol{\theta}^t}(\theta_i^{t,k})\|^2,$$

$$S := \frac{1}{N}\sum_{i=1}^N \frac{1}{TK_i} \sum_{t=0}^{T-1}\sum_{k=0}^{K_i-1} \mathbb{E}\|\nabla f_{i,\boldsymbol{\theta}^t}(\theta_i^{t,k})\|^2,$$

$$S_0 := \frac{1}{TN}\sum_{i=1}^N \sum_{t=0}^{T-1} \mathbb{E}\|\nabla f_{i,\boldsymbol{\theta}^t}(\theta_i^{t,0})\|^2,$$

$$C_{\eta_i} := \eta_i - \frac{L\,(1 + C_{\text{sgd}})\,\eta_i^2}{2}, \text{ (note that } C_{\eta_i} > 0 \text{ under Assumption 6)}$$

$$
\begin{aligned}
A_{\boldsymbol{K}} &:= \frac{1}{TN}\sum_{i=1}^{N}\frac{\mathbb{E}f_i^0(\theta_i^{0,0}) - f_{\inf}}{K_i C_{\eta_i}} + \frac{1}{N}\sum_{i=1}^{N}\frac{L\eta_i^2\sigma_{\mathrm{sgd}}^2}{2C_{\eta_i}},\\
d_j &:= \sum_{i=1}^{N}\frac{(1-\rho_i)\,w_{ij}}{K_i C_{\eta_i}},\\
\Gamma_j &:= K_j\eta_j d_j,\\
B &:= L_d L_h\sqrt{\frac{1}{N}\sum_{j=1}^{N}\Gamma_j^2},\\
\tilde{B} &:= L_d L_h\sqrt{\frac{1}{N}\sum_{j=1}^{N}\frac{\Gamma_j^2}{K_j}},\\
a_i^t(x) &:= \sum_j w_{ij} h_{j,\theta_j^{t,0}}(x),\\
G_i^t(\theta) &:= \frac{1}{|\widehat{\mathcal{D}}^{\mathrm{re}}|}\sum_{x\in\widehat{\mathcal{D}}^{\mathrm{re}}} d\big(h_{i,\theta}(x), a_i^t(x)\big)\,.
\end{aligned}
$$

Lemma 1. *Under Assumptions 1-2,*

$$
\mathbb{E}[\|\widetilde{\nabla} f_{i,\boldsymbol{\theta}^t}(\theta_i^{t,k})\|^2 \mid \mathcal{F}_{i,t,k}] \le \sigma_{sgd}^2 + (1+C_{\mathrm{sgd}})\,\|\nabla f_{i,\boldsymbol{\theta}^t}(\theta_i^{t,k})\|^2, \tag{8}
$$

and

$$
\mathbb{E}[f_{i,\boldsymbol{\theta}^t}(\theta_i^{t,k+1}) \mid \mathcal{F}_{i,t,k}] \le f_{i,\boldsymbol{\theta}^t}(\theta_i^{t,k}) - C_{\eta_i}\|\nabla f_{i,\boldsymbol{\theta}^t}(\theta_i^{t,k})\|^2 + \frac{L\eta_i^2\sigma_{sgd}^2}{2}. \tag{9}
$$

Proof. This is a standard descent lemma. Because $\widetilde{\nabla} f_{i,\boldsymbol{\theta}^t}(\theta_i^{t,k}) = \nabla f_{i,\boldsymbol{\theta}^t}(\theta_i^{t,k}) + \xi_{i,t,k}$ and $\mathbb{E}[\xi_{i,t,k} \mid \mathcal{F}_{i,t,k}] = 0$,

$$
\begin{aligned}
\mathbb{E}[\|\widetilde{\nabla} f_{i,\boldsymbol{\theta}^t}(\theta_i^{t,k})\|^2 \mid \mathcal{F}_{i,t,k}] &= \|\nabla f_{i,\boldsymbol{\theta}^t}(\theta_i^{t,k})\|^2 + \mathbb{E}[\|\xi_{i,t,k}\|^2 \mid \mathcal{F}_{i,t,k}]\\
&\le \sigma_{\mathrm{sgd}}^2 + (1+C_{\mathrm{sgd}})\,\|\nabla f_{i,\boldsymbol{\theta}^t}(\theta_i^{t,k})\|^2,
\end{aligned}
$$

which proves (8). By L-smoothness and the local update,

$$
f_{i,\boldsymbol{\theta}^t}(\theta_i^{t,k+1}) \le f_{i,\boldsymbol{\theta}^t}(\theta_i^{t,k}) - \eta_i\langle\nabla f_{i,\boldsymbol{\theta}^t}(\theta_i^{t,k}), \widetilde{\nabla} f_{i,\boldsymbol{\theta}^t}(\theta_i^{t,k})\rangle + \frac{L\eta_i^2}{2}\|\widetilde{\nabla} f_{i,\boldsymbol{\theta}^t}(\theta_i^{t,k})\|^2.
$$

Taking conditional expectation on the both sides of the inequality above and using

$$
\mathbb{E}[\langle\nabla f_{i,\boldsymbol{\theta}^t}(\theta_i^{t,k}), \widetilde{\nabla} f_{i,\boldsymbol{\theta}^t}(\theta_i^{t,k})\rangle \mid \mathcal{F}_{i,t,k}] = \|\nabla f_{i,\boldsymbol{\theta}^t}(\theta_i^{t,k})\|^2
$$

together with (8), we have (9). □

Lemma 2. *Define*

$$
\Delta G_i^t := G_i^{t+1}(\theta_i^{t+1,0}) - G_i^t(\theta_i^{t+1,0}). \tag{10}
$$

Then

$$
\sum_{t=0}^{T-1}\sum_{k=0}^{K_i-1} a_{i,t,k} \le \frac{\mathbb{E}f_i^0(\theta_i^{0,0}) - f_{\inf}}{C_{\eta_i}} + \frac{TK_iL\eta_i^2\sigma_{sgd}^2}{2C_{\eta_i}} + \frac{(1-\rho_i)}{C_{\eta_i}}\sum_{t=0}^{T-1}\mathbb{E}[\Delta G_i^t]. \tag{11}
$$

Consequently,

$$
S \le A_{\boldsymbol{K}} + \frac{1}{NT}\sum_{i=1}^{N}\frac{(1-\rho_i)}{K_i C_{\eta_i}}\sum_{t=0}^{T-1}\mathbb{E}[\Delta G_i^t]. \tag{12}
$$

Proof. Taking total expectation in (9), summing over $k = 0, \dots, K_i - 1$, and using $\theta_i^{t,K_i} = \theta_i^{t+1,0}$ gives

$$
C_{\eta_i}\sum_{k=0}^{K_i-1} a_{i,t,k} \le \mathbb{E}f_{i,\boldsymbol{\theta}^t}(\theta_i^{t,0}) - \mathbb{E}f_{i,\boldsymbol{\theta}^t}(\theta_i^{t+1,0}) + \frac{K_iL\eta_i^2\sigma_{\mathrm{sgd}}^2}{2}. \tag{13}
$$

Moreover, we have

$$\begin{aligned}
&\sum_{t=0}^{T-1}\big[f_{i,\boldsymbol{\theta}^t}(\theta_i^{t,0}) - f_{i,\boldsymbol{\theta}^t}(\theta_i^{t+1,0})\big] \\
&\quad= f_i^0(\theta_i^{0,0}) - f_{i,\boldsymbol{\theta}^t}(\theta_i^{T,0}) + \sum_{t=0}^{T-1}\big[f_{i,\boldsymbol{\theta}^{t+1}}(\theta_i^{t+1,0}) - f_{i,\boldsymbol{\theta}^t}(\theta_i^{t+1,0})\big] \\
&\quad= f_i^0(\theta_i^{0,0}) - f_{i,\boldsymbol{\theta}^t}(\theta_i^{T,0}) + (1-\rho_i)\sum_{t=0}^{T-1}\Delta G_i^t. \qquad (14)
\end{aligned}$$

We then sum (13) over t, substitute (14), and use $f_{i,\boldsymbol{\theta}^t}(\theta_i^{T,0}) \geq f_{\inf}$. This proves (11). Finally, we divide the inequality for node i by TK_i and average over i to obtain (12). □

Lemma 3. *Under Assumptions 4-5,*

$$\mathbb{E}[\Delta G_i^t] \leq L_d L_h \sum_j w_{ij}\mathbb{E}\|\theta_j^{t+1,0} - \theta_j^{t,0}\|. \qquad (15)$$

Proof.

$$\begin{aligned}
\Delta G_i^t &= \frac{1}{|\widehat{\mathcal{D}}^{\mathrm{re}}|}\sum_{x\in\widehat{\mathcal{D}}^{\mathrm{re}}} d\Big(h_{i,\theta_i^{t+1,0}}(x), a_i^{t+1}(x)\Big) - \frac{1}{|\widehat{\mathcal{D}}^{\mathrm{re}}|}\sum_{x\in\widehat{\mathcal{D}}^{\mathrm{re}}} d\Big(h_{i,\theta_i^{t+1,0}}(x), a_i^{t}(x)\Big) \\
&\leq \frac{L_d}{|\widehat{\mathcal{D}}^{\mathrm{re}}|}\sum_{x\in\widehat{\mathcal{D}}^{\mathrm{re}}} \big\|a_i^{t+1}(x) - a_i^t(x)\big\| \\
&\leq \frac{L_d}{|\widehat{\mathcal{D}}^{\mathrm{re}}|}\sum_{x\in\widehat{\mathcal{D}}^{\mathrm{re}}} \left\|\sum_j w_{ij}\big[h_{j,\theta_j^{t+1,0}}(x) - h_{j,\theta_j^{t,0}}(x)\big]\right\| \\
&\leq \frac{L_d}{|\widehat{\mathcal{D}}^{\mathrm{re}}|}\sum_{x\in\widehat{\mathcal{D}}^{\mathrm{re}}}\sum_j w_{ij}\|h_{j,\theta_j^{t+1,0}}(x) - h_{j,\theta_j^{t,0}}(x)\| \\
&\leq L_d L_h \sum_j w_{ij}\|\theta_j^{t+1,0} - \theta_j^{t,0}\|.
\end{aligned}$$

Taking expectation proves (15). □

Combining Lemmas 2 and 3, and then reordering the finite sums over recipients i and teachers j, we have

$$S \leq A_{\boldsymbol{K}} + \frac{L_d L_h}{NT}\sum_{j=1}^{N} d_j \sum_{t=0}^{T-1}\mathbb{E}\|\theta_j^{t+1,0} - \theta_j^{t,0}\|. \qquad (16)$$

Lemma 4. *Let*

$$V_{j,t} := \sum_{k=0}^{K_j-1} a_{j,t,k}. \qquad (17)$$

Then

$$\mathbb{E}\|\theta_j^{t+1,0} - \theta_j^{t,0}\| \leq \eta_j\Big[\sqrt{K_j V_{j,t}} + \sqrt{K_j\sigma_{sgd}^2 + C_{sgd}V_{j,t}}\Big]. \qquad (18)$$

Consequently,

$$S \leq A_{\boldsymbol{K}} + B\sqrt{S} + \tilde{B}\sqrt{\sigma_{sgd}^2 + C_{sgd}S}. \qquad (19)$$

Proof. First, we utilize the orthogonality among local-step noises similar to Karimireddy et al. (2020). For local indices $r < s$, $\xi_{j,t,r}$ is $\mathcal{F}_{j,t,s}$-measurable. Therefore,

$$\mathbb{E}\langle\xi_{j,t,r}, \xi_{j,t,s}\rangle = \mathbb{E}\left[\mathbb{E}\left[\langle\xi_{j,t,r}, \xi_{j,t,s}\rangle|\mathcal{F}_{j,t,s}\right]\right] = \mathbb{E}[\langle\xi_{j,t,r}, \mathbb{E}[\xi_{j,t,s} \mid \mathcal{F}_{j,t,s}]\rangle] = 0.$$

It thus follows from Assumption 2 that

$$\mathbb{E}\left\|\sum_{k=0}^{K_j-1}\xi_{j,t,k}\right\|^2 = \sum_{k=0}^{K_j-1}\mathbb{E}\|\xi_{j,t,k}\|^2 \le K_j\sigma_{\text{sgd}}^2 + C_{\text{sgd}}V_{j,t}. \tag{20}$$

The local recursion gives

$$\theta_j^{t+1,0} - \theta_j^{t,0} = -\eta_j \sum_{k=0}^{K_j-1}\widetilde{\nabla} f_{j,\boldsymbol{\theta}^t}(\theta_j^{t,k}). \tag{21}$$

Using (21), the Minkowski inequality and Cauchy–Schwarz inequality, we have

$$\begin{aligned}
\mathbb{E}\|\theta_j^{t+1,0} - \theta_j^{t,0}\| &\le \eta_j\sqrt{\mathbb{E}\left\|\sum_{k=0}^{K_j-1}\nabla f_{j,\boldsymbol{\theta}^t}(\theta_j^{t,k})\right\|^2} + \eta_j\sqrt{\mathbb{E}\left\|\sum_{k=0}^{K_j-1}\xi_{j,t,k}\right\|^2} \\
&\le \eta_j\sum_{k=0}^{K_j-1}\sqrt{a_{j,t,k}} + \eta_j\sqrt{K_j\sigma_{\text{sgd}}^2 + C_{\text{sgd}}V_{j,t}} \\
&\le \eta_j\left[\sqrt{K_jV_{j,t}} + \sqrt{K_j\sigma_{\text{sgd}}^2 + C_{\text{sgd}}V_{j,t}}\right].
\end{aligned}$$

This proves (18). Denote $s_i := \frac{1}{TK_i}\sum_{t=0}^{T-1}\sum_{k=0}^{K_i-1}a_{i,t,k}$. Because $\sum_{t=0}^{T-1}V_{j,t} = TK_js_j$, using the Cauchy-Schwarz inequality over t gives

$$\sum_{t=0}^{T-1}\sqrt{K_jV_{j,t}} \le TK_j\sqrt{s_j}, \tag{22}$$

$$\sum_{t=0}^{T-1}\sqrt{K_j\sigma_{\text{sgd}}^2 + C_{\text{sgd}}V_{j,t}} \le T\sqrt{K_j}\sqrt{\sigma_{\text{sgd}}^2 + C_{\text{sgd}}s_j}. \tag{23}$$

Substituting (18)-(23) into (16) yields

$$S \le A_{\boldsymbol{K}} + \frac{L_dL_h}{N}\sum_{j=1}^{N}\Gamma_j\sqrt{s_j} + \frac{L_dL_h}{N}\sum_{j=1}^{N}\frac{\Gamma_j}{\sqrt{K_j}}\sqrt{\sigma_{\text{sgd}}^2 + C_{\text{sgd}}s_j}.$$

The Cauchy-Schwarz inequality over j bounds the second term by $B\sqrt{S}$ and the third term by

$$L_dL_h\sqrt{\frac{1}{N}\sum_{j=1}^{N}\frac{\Gamma_j^2}{K_j}}\sqrt{\frac{1}{N}\sum_{j=1}^{N}(\sigma_{\text{sgd}}^2 + C_{\text{sgd}}s_j)} = \tilde{B}\sqrt{\sigma_{\text{sgd}}^2 + C_{\text{sgd}}S}.$$

This proves (19). □

Lemma 5. *If $x, A, B, \tilde{B}, \sigma, C \ge 0$ and*

$$x \le A + B\sqrt{x} + \tilde{B}\sqrt{\sigma^2 + Cx}, \tag{24}$$

then

$$x \le A + Br_0 + \tilde{B}\sqrt{\sigma^2 + Cr_0^2}, \tag{25}$$

where $r_0 := \frac{\widetilde{D}+\sqrt{\widetilde{D}^2+4\widetilde{A}}}{2}$, $\widetilde{A} := A + \tilde{B}\sigma$, and $\widetilde{D} := B + \tilde{B}\sqrt{C}$.

Proof. Firstly, use $\sqrt{\sigma^2 + Cx} \le \sigma + \sqrt{C}\sqrt{x}$ in (24). With $z := \sqrt{x}$, this gives

$$z^2 - \widetilde{D}z - \widetilde{A} \le 0.$$

The nonnegative upper root is $\frac{\widetilde{D}+\sqrt{\widetilde{D}^2+4\widetilde{A}}}{2}$, proving (25). □

Lemma 6. *For all i, t,*

$$\sqrt{a_{i,t,0}} \le \alpha_i \sqrt{\overline{a}_{i,t}} + \beta_i \sqrt{C_{sgd} \overline{a}_{i,t} + \frac{\sigma_{sgd}^2}{2}}, \tag{26}$$

where

$$\overline{a}_{i,t} := \frac{1}{K_i} \sum_{k=0}^{K_i - 1} a_{i,t,k}. \tag{27}$$

Proof. Fix i, t and abbreviate with

$$K := K_i, \quad \eta := \eta_i, \quad g_k := \nabla f_{i,\boldsymbol{\theta}^t}(\theta_i^{t,k}), \quad \xi_k := \xi_{i,t,k}, \quad a_k := a_{i,t,k}, \quad \overline{a} := \overline{a}_{i,t}.$$

Assumption 1 gives

$$\|g_0\| \le \|g_k\| + L\|\theta_i^{t,k} - \theta_i^{t,0}\|.$$

Then the Minkowski inequality gives,

$$\sqrt{a_0} \le \sqrt{a_k} + L\sqrt{\mathbb{E}\|\theta_i^{t,k} - \theta_i^{t,0}\|^2}. \tag{28}$$

The local recursion gives

$$\theta_i^{t,k} - \theta_i^{t,0} = -\eta \sum_{r=0}^{k-1} (g_r + \xi_r).$$

As in (20), we have

$$\mathbb{E}\left\| \sum_{r=0}^{k-1} \xi_r \right\|^2 \le k\sigma_{\text{sgd}}^2 + C_{\text{sgd}} \sum_{r=0}^{k-1} a_r.$$

Hence the Minkowski inequality gives

$$\sqrt{\mathbb{E}\|\theta_i^{t,k} - \theta_i^{t,0}\|^2} \le \eta \sum_{r=0}^{k-1} \sqrt{a_r} + \eta \sqrt{k\sigma_{\text{sgd}}^2 + C_{\text{sgd}} \sum_{r=0}^{k-1} a_r}. \tag{29}$$

We then average (28) over $k = 0, \ldots, K-1$ and substitute (29). The first resulting term from (28) satisfies

$$\frac{1}{K} \sum_{k=0}^{K-1} \sqrt{a_k} \le \sqrt{\overline{a}}. \tag{30}$$

The first resulting term from (29) satisfies

$$\begin{aligned} \frac{1}{K} \sum_{k=0}^{K-1} \sum_{r=0}^{k-1} \sqrt{a_r} &= \frac{1}{K} \sum_{r=0}^{K-2} (K - 1 - r) \sqrt{a_r} \\ &\le \frac{1}{K} \sqrt{\sum_{r=0}^{K-2} (K-1-r)^2} \sqrt{\sum_{r=0}^{K-2} a_r} \\ &\le \sqrt{\frac{(K-1)(2K-1)}{6}} \sqrt{\overline{a}}, \end{aligned} \tag{31}$$

where $\sum_{r=0}^{K-2}(K-1-r)^2 = K(K-1)(2K-1)/6$. The second resulting term from (29) satisfies

$$\begin{aligned}
&\frac{1}{K}\sum_{k=0}^{K-1}\sqrt{k\sigma_{\text{sgd}}^2 + C_{\text{sgd}}\sum_{r=0}^{k-1}a_r} \\
&\leq \frac{\sqrt{K-1}}{K}\sqrt{\sum_{k=1}^{K-1}\left(k\sigma_{\text{sgd}}^2 + C_{\text{sgd}}\sum_{r=0}^{k-1}a_r\right)} \\
&= \frac{\sqrt{K-1}}{K}\sqrt{\frac{K(K-1)}{2}\sigma_{\text{sgd}}^2 + C_{\text{sgd}}\sum_{r=0}^{K-2}(K-1-r)a_r} \\
&\leq \frac{K-1}{\sqrt{K}}\sqrt{\frac{\sigma_{\text{sgd}}^2}{2} + C_{\text{sgd}}\overline{a}}. \qquad (32)
\end{aligned}$$

Combining (30)-(32) proves (26). □

Lemma 7.

$$S_0 \leq \Phi_{\boldsymbol{K}}(S), \qquad (33)$$

where

$$\forall R \geq 0, \Phi_{\boldsymbol{K}}(R) := \min\left\{K_{\max}R, \left[\alpha_*\sqrt{R} + \sqrt{C_{sgd}\beta_*^2 R + \frac{\sigma_{sgd}^2}{2}\beta_{\text{rms}}^2}\right]^2\right\}, \qquad (34)$$

$c_i := \sqrt{\frac{(K_i-1)(2K_i-1)}{6}}, r_i := \frac{K_i-1}{\sqrt{K_i}}, \alpha_i := 1 + L\eta_i c_i, \beta_i := L\eta_i r_i, \alpha_* := \max_i \alpha_i, \beta_* := \max_i \beta_i, \beta_{\text{rms}} := \sqrt{\frac{1}{N}\sum_{i=1}^N \beta_i^2}$, *and* $K_{\max} := \max_i K_i$.

Proof. Applying Lemma 6 elementwise to the NT entries indexed by (i,t) and using the Minkowski inequality, we have

$$\sqrt{S_0} \leq \sqrt{\frac{1}{NT}\sum_{i,t}\alpha_i^2\overline{a}_{i,t}} + \sqrt{\frac{1}{NT}\sum_{i,t}\beta_i^2\left(C_{\text{sgd}}\overline{a}_{i,t} + \frac{\sigma_{\text{sgd}}^2}{2}\right)} \leq \alpha_*\sqrt{S} + \sqrt{C_{\text{sgd}}\beta_*^2 S + \frac{\sigma_{\text{sgd}}^2}{2}\beta_{\text{rms}}^2}.$$

Squaring both sides gives the second branch of (34).

Moreover, the first branch holds since

$$S_0 \leq \frac{1}{NT}\sum_{i=1}^N\sum_{t=0}^{T-1}\sum_{k=0}^{K_i-1}a_{i,t,k} = \frac{1}{N}\sum_{i=1}^N K_i s_i \leq K_{\max}S.$$

□

C.2 Proof of Convergence Theorems

C.2.1 Proof of Theorem 1

Proof. Consider 19 in Lemma 4. Applying Lemma 5 with

$$x = S, \quad A = A_{\boldsymbol{K}}, \quad \sigma = \sigma_{\text{sgd}}, \quad C = C_{\text{sgd}}$$

gives

$$\widetilde{A} = \hat{A}_{\boldsymbol{K}}, \qquad \widetilde{D} = D_{\boldsymbol{K}}, \qquad r_0 = r_{\boldsymbol{K}}.$$

Then (25) yields

$$S \leq A_{\boldsymbol{K}} + Br_{\boldsymbol{K}} + \tilde{B}\sqrt{\sigma_{\text{sgd}}^2 + C_{\text{sgd}}r_{\boldsymbol{K}}^2} = Q.$$

By Lemma 7, $S_0 \leq \Phi_{\boldsymbol{K}}(S)$. Since $\Phi_{\boldsymbol{K}}$ is increasing and $S \leq Q$, we have

$$S_0 \leq \Phi_{\boldsymbol{K}}(S) \leq \Phi_{\boldsymbol{K}}(Q).$$

□

C.2.2 Proof of Theorem 2

Proof. The local recursion gives

$$\theta_j^{t+1,0} - \theta_j^{t,0} = -\eta_j \sum_{k=0}^{K_j-1} \widetilde{\nabla} f_{j,\boldsymbol{\theta}^t}(\theta_j^{t,k}).$$

By Assumption 7 and the Cauchy-Schwarz inequality, we have

$$\mathbb{E}\|\theta_j^{t+1,0} - \theta_j^{t,0}\| \le \eta_j \sum_{k=0}^{K_j-1} \mathbb{E}\|\widetilde{\nabla} f_{j,\boldsymbol{\theta}^t}(\theta_j^{t,k})\| \le K_j \eta_j U. \tag{35}$$

Substituting (35) into inequality (16) yields

$$\begin{aligned} S &\le A_{\boldsymbol{K}} + \frac{L_d L_h}{NT} \sum_{j=1}^{N} d_j \sum_{t=0}^{T-1} K_j \eta_j U \\ &= A_{\boldsymbol{K}} + \frac{L_d L_h U}{N} \sum_{j=1}^{N} K_j \eta_j d_j \\ &= A_{\boldsymbol{K}} + \frac{L_d L_h U}{N} \sum_{j=1}^{N} \Gamma_j \\ &= \tilde{Q}. \end{aligned}$$

Lemma 7 gives $S_0 \le \Phi_{\boldsymbol{K}}(S)$. Since $\Phi_{\boldsymbol{K}}$ is increasing, we have

$$S_0 \le \Phi_{\boldsymbol{K}}(S) \le \Phi_{\boldsymbol{K}}(\tilde{Q}).$$

□

C.2.3 Proof of Theorem 3

Proof. Based on (13) and Assumption 8, we have

$$\begin{aligned} C_{\eta_i} \sum_{k=0}^{K_i-1} a_{i,t,k} &\le \mathbb{E} f_{i,\boldsymbol{\theta}^t}(\theta_i^{t,0}) - \mathbb{E} f_{i,\boldsymbol{\theta}^t}(\theta_i^{t+1,0}) + \frac{K_i L \eta_i^2 \sigma_{\text{sgd}}^2}{2} \\ &\le \mathbb{E} f_{i,\boldsymbol{\theta}^t}(\theta_i^{t,0}) - \mathbb{E} f_{i,\boldsymbol{\theta}^{t+1}}(\theta_i^{t+1,0}) + \tilde{U} + \frac{K_i L \eta_i^2 \sigma_{\text{sgd}}^2}{2}. \end{aligned} \tag{36}$$

Summing (36) over t, using $f_{i,\boldsymbol{\theta}^t}(\theta_i^{T,0}) \ge f_{\inf}$, dividing the inequality for node i by $TK_iC_{\eta_i}$, and average over i, we have

$$S \le A_{\boldsymbol{K}} + \frac{\tilde{U}}{N} \sum_{i=1}^{N} \frac{1}{K_i C_{\eta_i}}.$$

□

C.3 Additional Details for Tightness Comparison between Q and $\tilde{Q}$

C.3.1 Detailed Conditions

We define $\chi_\Gamma := \frac{N^{-1}\sum_{j=1}^N \Gamma_j}{\sqrt{N^{-1}\sum_{j=1}^N \Gamma_j^2}}$, which satisfies $\frac{1}{\sqrt{N}} \le \chi_\Gamma \le 1$ based on the nonnegativity of the Γ_j and the Cauchy-Schwarz inequality.

First, we give the following necessary and sufficient condition.

Lemma 8. *For $B > 0$,*

$$\Phi_{\boldsymbol{K}}(Q) \le \Phi_{\boldsymbol{K}}(\tilde{Q}) \text{ if and only if } U \ge \frac{r_{\boldsymbol{K}} + \frac{\tilde{B}}{B}\sqrt{\sigma_{\mathrm{sgd}}^2 + C_{\mathrm{sgd}} r_{\boldsymbol{K}}^2}}{\chi_\Gamma}. \tag{37}$$

Proof. Note that the two bounds have the common form:

$$Q = A_{\boldsymbol{K}} + B\left(r_{\boldsymbol{K}} + \frac{\tilde{B}}{B}\sqrt{\sigma_{\mathrm{sgd}}^2 + C_{\mathrm{sgd}} r_{\boldsymbol{K}}^2}\right),$$
$$\tilde{Q} = A_{\boldsymbol{K}} + B\chi_\Gamma U.$$

Then the result holds since $\Phi_{\boldsymbol{K}}$ is increasing.

□

Based on Lemma 8, we immediately have the necessary condition below for strict tightness of Theorem 1.

Corollary 1. *Suppose $B > 0$. If $(\tilde{B}/B)\sigma_{\mathrm{sgd}} \ge \chi_\Gamma U$, then $Q \ge \tilde{Q}$. Hence $Q < \tilde{Q}$ requires*

$$\frac{\tilde{B}}{B}\sigma_{\mathrm{sgd}} < \chi_\Gamma U. \tag{38}$$

Proof. The conclusion follows from $r_{\boldsymbol{K}} + \frac{\tilde{B}}{B}\sqrt{\sigma_{\mathrm{sgd}}^2 + C_{\mathrm{sgd}} r_{\boldsymbol{K}}^2} \ge (\tilde{B}/B)\sigma_{\mathrm{sgd}}$ and (37). □

Condition (38) compares the normalized stochastic-noise contribution with the effective uniform-gradient bound $\chi_\Gamma U$. In particular, $\frac{1}{\sqrt{\max_i K_i}} \le \frac{\tilde{B}}{B} \le \frac{1}{\sqrt{\min_i K_i}}$. It is therefore easier to satisfy (38) with smaller stochastic noise, a more conservative U, or more local steps when the remaining coefficients are fixed.

Furthermore, we can established the following sufficient condition, which is copied from the main body:

Theorem 4. *Suppose $B > 0$ and assume $\chi_\Gamma U > (\tilde{B}/B)\sigma_{\mathrm{sgd}}$. Define $r_u := \frac{(\chi_\Gamma U)^2 - \left((\tilde{B}/B)\sigma_{\mathrm{sgd}}\right)^2}{\chi_\Gamma U + (\tilde{B}/B)\sqrt{\sigma_{\mathrm{sgd}}^2 + C_{\mathrm{sgd}}((\chi_\Gamma U)^2 - \left((\tilde{B}/B)\sigma_{\mathrm{sgd}}\right)^2)}}$, and $E_{\max} := \frac{\chi_\Gamma U r_u^2}{(\tilde{B}/B)\sigma_{\mathrm{sgd}} + \left(1 + (\tilde{B}/B)\sqrt{C_{\mathrm{sgd}}}\right) r_u}$. If $\tilde{Q} \le E_{\max}$, then $Q \le \tilde{Q}$, and consequently $\Phi_{\boldsymbol{K}}(Q) \le \Phi_{\boldsymbol{K}}(\tilde{Q})$.*

Proof. For brevity, let

$$A := A_{\mathbf{K}}, \qquad u := \chi_\Gamma U, \qquad \alpha := \frac{\tilde{B}}{B}, \qquad \sigma := \sigma_{\mathrm{sgd}}, \qquad C := C_{\mathrm{sgd}},$$

and define

$$s := \alpha\sigma, \qquad p := 1 + \alpha\sqrt{C}, \qquad q := 1 + C.$$

Recall that $\tilde{Q} = A + uB$ and $Q = A + BH$, where

$$H = r_{\boldsymbol{K}} + \alpha\sqrt{\sigma^2 + C r_{\boldsymbol{K}}^2},$$

and $r_{\boldsymbol{K}}$ is the positive root of

$$r^2 - pBr - (A + sB) = 0.$$

Define

$$h(r) := r + \alpha\sqrt{\sigma^2 + Cr^2}, \qquad r \ge 0.$$

The function h is continuous and strictly increasing. Moreover,

$$h(0) = \alpha\sigma = s < u, \qquad \lim_{r\to\infty} h(r) = \infty.$$

Thus, we can show that r_u is the unique positve root of

$$h(r) - u = 0.$$

Since $B > 0$,

$$Q \le \tilde{Q} \text{ if and only if } H \le u.$$

Because $H = h(r_{\boldsymbol{K}})$ and h is strictly increasing, we have

$$H \le u \text{ if and only if } r_{\boldsymbol{K}} \le r_u.$$

Since $r_{\boldsymbol{K}}$ is the positive root of $r^2 - pBr - (A + sB) = 0$, we have

$$r_{\boldsymbol{K}} \le r_u \text{ if and only if } A + (s + pr_u)B \le r_u^2. \tag{39}$$

Let

$$c := s + pr_u.$$

Using $h(r_u) = u$, we obtain

$$c - u = s + pr_u - u = \alpha \left(\sigma + \sqrt{C}\, r_u - \sqrt{\sigma^2 + C r_u^2} \right) \ge 0,$$

where the last inequality follows from $\sqrt{\sigma^2 + C r_u^2} \le \sigma + \sqrt{C}\, r_u$.

Therefore, for every $E \ge 0$,

$$\max_{\substack{A,B \ge 0 \\ A + uB \le E}} \{A + cB\} = \frac{c}{u} E. \tag{40}$$

This holds since

$$A + cB = \frac{c}{u}(A + uB) - \left(\frac{c}{u} - 1 \right) A \le \frac{c}{u}(A + uB) \le \frac{c}{u} E.$$

The upper bound is attained at

$$A = 0, \qquad B = \frac{E}{u}.$$

Therefore, for any $E \ge 0$, when $u > s$ and $A + uB \le E$, $A + (s + pr_u)B \le r_u^2$ holds uniformly if and only if

$$\frac{c}{u} E \le r_u^2,$$

which is equivalent to

$$E \le \frac{u r_u^2}{c} = E_{\max}.$$

This proves that if $u > s$ and $\tilde{Q} := A + uB \le E \le E_{\max}$, then $Q \le \tilde{Q}$. □

C.3.2 Control on Neighbors' Shifts

We explain in detail how these two bounds control the neighbors' shifts. As shown in Appendix C.2.2 and Lemma 4 in Appendix C.1, for a fixed node j and round t, our argument gives $\mathbb{E}\|\theta_j^{t+1,0} - \theta_j^{t,0}\| \le M_1$, whereas the uniform-gradient argument (Tan et al., 2022) gives the upper bound M_2, where

$$M_1 := K_j \eta_j \left[\sqrt{\frac{1}{K_j} \sum_{k=0}^{K_j - 1} \mathbb{E}\|\nabla f_{j,\boldsymbol{\theta}^t}(\theta_j^{t,k})\|^2} + \frac{1}{\sqrt{K_j}} \sqrt{\sigma_{\mathrm{sgd}}^2 + C_{\mathrm{sgd}} \left(\frac{1}{K_j} \sum_{k=0}^{K_j - 1} \mathbb{E}\|\nabla f_{j,\boldsymbol{\theta}^t}(\theta_j^{t,k})\|^2 \right)} \right],$$

$$M_2 := K_j \eta_j U.$$

As $\frac{1}{K_j} \sum_{k=0}^{K_j - 1} \mathbb{E}\|\nabla f_{j,\boldsymbol{\theta}^t}(\theta_j^{t,k})\|^2$ decreases, M_1 decreases, approaching $\eta_j \sqrt{K_j} \sigma_{\mathrm{sgd}}$, whereas M_2 remains unchanged for fixed U. Thus, when local gradient magnitudes are sufficiently small and U is of the same order as $\sigma_{\mathrm{sgd}} > 0$, the neighbor's shift bound in Theorem 1 can be tighter than that in Theorem 2 by a factor of order $\sqrt{K_j}$.

C.4 Effect of Proportional Increases in Local Step Counts

Consider scaling the local step counts as $K_i(a) := aK_i$ for every node, where $a \geq 1$ and all scaled counts are integers. Hold the remaining algorithm parameters and assumption constants fixed. We write each quantity as a function of a when evaluated at these scaled step counts.

By definition,

$$\Gamma_j(a) = aK_j\eta_j \sum_{i=1}^{N} \frac{(1-\rho_i)w_{ij}}{aK_iC_{\eta_i}} = \Gamma_j(1).$$

Consequently,

$$B(a) = B(1), \qquad \chi_\Gamma(a) = \chi_\Gamma(1), \qquad \tilde{B}(a) = \frac{\tilde{B}(1)}{\sqrt{a}}.$$

Suppose $B(1) > 0$, and define

$$q(a) := \frac{\tilde{B}(a)}{B(a)} = \frac{q(1)}{\sqrt{a}}, \qquad u := \chi_\Gamma(1)U.$$

Assume the necessary condition (38) is satisfied, i.e., $u > q(1)\sigma_{\mathrm{sgd}}$.

Let $r_u(a)$ be the unique positive solution of

$$r_u(a) + q(a)\sqrt{\sigma_{\mathrm{sgd}}^2 + C_{\mathrm{sgd}}r_u(a)^2} = u. \tag{41}$$

For fixed $q(a)$, the left-hand side is increasing in $r_u(a)$. For fixed $r_u(a)$, the left-hand side is increasing in $q(a)$. Since $q(a)$ is decreasing in a, the solution $r_u(a)$ is increasing in a.

We can rewrite the sufficient-condition threshold as

$$E_{\max}(a) = \frac{u\, r_u(a)}{1 + q(a)\sqrt{C_{\mathrm{sgd}}} + q(a)\sigma_{\mathrm{sgd}}/r_u(a)}. \tag{42}$$

Its numerator is increasing in a, while its denominator is decreasing. Therefore, $E_{\max}(a)$ is increasing in a.

Be its definition, $\tilde{Q}(a)$ is decreasing in a. Hence, if $\tilde{Q}(a_1) \leq E_{\max}(a_1)$ for some admissible $a_1 \geq 1$, then every admissible $a_2 \geq a_1$ satisfies

$$\tilde{Q}(a_2) \leq \tilde{Q}(a_1) \leq E_{\max}(a_1) \leq E_{\max}(a_2).$$

Thus, proportional increases in local step counts preserve the condition $\tilde{Q} \leq E_{\max}$ once it holds, and the gap between these two quantities can also be increased.

C.5 Counterexample to a Globally Uniform Gradient Bound

Consider two nodes with $w_{ij} = 1/2$ for $i, j \in \{1, 2\}$ and $\rho_1 = \rho_2 = 1/2$. Let $\mathcal{X} = \{1, 2\}$, $\widehat{\mathcal{D}}_1 = \widehat{\mathcal{D}}_2 = \{(1, 1), (2, 1)\}$, and $\widehat{\mathcal{D}}^{\mathrm{re}} = \{1, 2\}$. For $\theta \in \mathbb{R}$, define

$$r_x(\theta) := e^{-\theta^2 x^2},$$
$$h_{1,\theta}(x) := \begin{pmatrix} r_x(\theta) \\ 1 - r_x(\theta) \end{pmatrix},$$
$$h_{2,\theta}(x) := q = \begin{pmatrix} 1/2 \\ 1/2 \end{pmatrix}.$$

Both model outputs take values in the probability simplex Δ_2. Set the original supervised loss as $\ell(p, e_1) = -\log p_1$ and the distillation loss as $d(p, q) = \|p - q\|_2^2$.

At round t, write

$$b_z^t := \frac{1}{2}r_z(\theta_1^t) + \frac{1}{4} \text{ and } a^t(z) := \frac{1}{2}h_{1,\theta_1^t}(z) + \frac{1}{2}q = \begin{pmatrix} b_z^t \\ 1 - b_z^t \end{pmatrix},$$

where $a^t(z)$ denotes the ensemble teacher. In particular, $b_z^t \in [\frac{1}{4}, \frac{3}{4}]$ when $z \in \{1, 2\}$. The objective of node 1 in round t is therefore

$$f_{1,\boldsymbol{\theta}^t}(\theta) = \frac{1}{4}\sum_{x=1}^{2} \ell\big(h_{1,\theta}(x), e_1\big) + \frac{1}{4}\sum_{z=1}^{2} d\big(h_{1,\theta}(z), a^t(z)\big)$$
$$= \frac{5}{4}\theta^2 + \frac{1}{2}\sum_{z=1}^{2}\big(r_z(\theta) - b_z^t\big)^2. \tag{43}$$

Differentiating with b_z^t fixed gives

$$(f_{1,\boldsymbol{\theta}^t})'(\theta) = \frac{5}{2}\theta + \sum_{z=1}^{2}\big(r_z(\theta) - b_z^t\big)r_z'(\theta)$$
$$= \frac{5}{2}\theta - 2\theta\sum_{z=1}^{2} z^2 r_z(\theta)\big(r_z(\theta) - b_z^t\big). \tag{44}$$

After differentiating, we substitute $\theta = \theta_1^t$, which yields

$$(f_{1,\boldsymbol{\theta}^t})'(\theta_1^t) = \frac{5}{2}\theta_1^t - \theta_1^t\sum_{z=1}^{2} z^2 r_z(\theta_1^t)\left(r_z(\theta_1^t) - \frac{1}{2}\right).$$

We next verify the assumptions. For any $p, q, q' \in \Delta_2$, we have

$$\big|d(p,q) - d(p,q')\big| = \big|\langle q - q', q + q' - 2p\rangle\big| \le 2\sqrt{2}\,\|q - q'\|_2,$$

where we used

$$\|q + q' - 2p\|_2 \le \|q - p\|_2 + \|q' - p\|_2 \le 2\sqrt{2}.$$

Thus, Assumption 4 holds with

$$L_d = 2\sqrt{2}.$$

Furthermore,

$$\left\|\frac{\partial}{\partial\theta}h_{1,\theta}(x)\right\|_2 = 2\sqrt{2}\,|\theta|x^2 e^{-\theta^2x^2} \le \frac{2|x|}{\sqrt{e}} \le \frac{4}{\sqrt{e}}.$$

The mean-value theorem therefore gives

$$\|h_{1,\theta}(x) - h_{1,\theta'}(x)\|_2 \le \frac{4}{\sqrt{e}}|\theta - \theta'|,$$

uniformly over $x \in \mathcal{X}$. Node 2's model is a constant which does not depends on parameters, so Assumption 5 holds with $L_h = 4/\sqrt{e}$ for both models.

To verify smoothness, observe that

$$r_z'(\theta) = -2\theta z^2 e^{-\theta^2z^2}, \qquad r_z''(\theta) = (4\theta^2z^4 - 2z^2)e^{-\theta^2z^2}.$$

Since $\sup_{u\ge0} ue^{-2u} = \frac{1}{2e}$ and $\sup_{u\ge0}|(4u-2)e^{-u}| = 2$, we have

$$|r_z'(\theta)|^2 \le \frac{2z^2}{e}, \text{ and } |r_z''(\theta)| \le 2z^2.$$

Since $r_z(\theta) \in [0,1]$ and $b_z^t \in [\frac{1}{4}, \frac{3}{4}]$, we have $|r_z(\theta) - b_z^t| \le \frac{3}{4}$ for any θ and t. Then differentiating (44) gives

$$|(f_{1,\boldsymbol{\theta}^t})''(\theta)| = \left|\frac{5}{2} + \sum_{z=1}^{2}\big((r_z'(\theta))^2 + (r_z(\theta) - b_z^t)r_z''(\theta)\big)\right|$$
$$\le \frac{5}{2} + \sum_{z=1}^{2}\left(\frac{2}{e} + \frac{3}{2}\right)z^2 = 10 + \frac{10}{e}.$$

Hence Assumption 1 holds with $L = 10 + \frac{10}{e}$. The objective of node 2 is independent of its optimization parameter, so its gradient is zero and the same smoothness bound applies.

Both objectives are nonnegative, so Assumption 3 holds with $f_{\inf} = 0$. Use full-batch gradients at both nodes. Then Assumption 2 holds with $\sigma_{\mathrm{sgd}} = C_{\mathrm{sgd}} = 0$. Choose any stepsizes $0 < \eta_i < 2/L$, then Assumption 6 holds.

Note that

$$\left|(f_{1,\boldsymbol{\theta}^t})'(\theta) - \frac{5}{2}\theta\right| \le \frac{3}{2}\sum_{z=1}^{2}|\theta| z^2 e^{-\theta^2 z^2} \le \frac{3}{2\sqrt{2e}}\sum_{z=1}^{2} z = \frac{9}{2\sqrt{2e}},$$

where we used $\sup_{u\ge 0} u e^{-u^2} = \frac{1}{\sqrt{2e}}$. Consequently,

$$\left|\widetilde{\nabla} f_{1,\boldsymbol{\theta}^t}(\theta)\right| = \left|(f_{1,\boldsymbol{\theta}^t})'(\theta)\right| \ge \frac{5}{2}|\theta| - \frac{9}{2\sqrt{2e}}.$$

For every prescribed $U < \infty$, choosing $|\theta_1^0| > \frac{2}{5}(U + \frac{9}{2\sqrt{2e}})$ gives $\mathbb{E}|\widetilde{\nabla} f_1^0(\theta_1^0)| > U$, where all of the constants verified above remain unchanged. Thus Assumption 7 does not hold.

C.6 Detailed Comparison Between Theorem 1 and Theorem 3

Comparison of assumptions. Theorem 3 replaces Assumptions 4 and 5 with Assumption 8. In our classification setting, all ensemble labels lie in the probability simplex. Therefore Assumption 4 implies

$$\left|f_{i,\boldsymbol{\theta}^{t+1}}(\theta) - f_{i,\boldsymbol{\theta}^t}(\theta)\right| \le \sqrt{2}(1-\rho_i)L_d.$$

Consequently, the uniform drift assumption holds with $\tilde{U} = \sqrt{2}L_d \max_i(1-\rho_i)$. The converse implication need not hold: an objective may remain unchanged across rounds without satisfying either Lipschitz condition. The advantage of Theorem 1 therefore lies in obtaining a potentially sharper drift estimate through explicit control of the neighbors' updates.

Comparison of the bounds. Suppose the assumptions of both theorems hold. Theorem 1 gives a bound on S_0 no larger than that of Theorem 3 if and only if

$$B r_{\boldsymbol{K}} + \tilde{B}\sqrt{\sigma_{\mathrm{sgd}}^2 + C_{\mathrm{sgd}} r_{\boldsymbol{K}}^2} \le \frac{\tilde{U}}{N}\sum_{i=1}^{N}\frac{1}{K_i C_{\eta_i}}. \tag{45}$$

For fixed other quantities, increasing T reduces the initial objective-gap contribution to $A_{\boldsymbol{K}}$ and hence decreases $r_{\boldsymbol{K}}$, while the right-hand side remains unchanged. Reducing σ_{sgd}, with the other quantities fixed, also decreases the left-hand side, both directly and through $A_{\boldsymbol{K}}$ and $r_{\boldsymbol{K}}$. Smaller effective influence coefficients Γ_j reduce B, $\tilde{B}$, and consequently $r_{\boldsymbol{K}}$, thereby reducing the left-hand side. All these effects can make (45) easier to satisfy for a fixed uniform drift bound $\tilde{U}$.

How objective drift is controlled. Assumptions 4 and 5 give

$$|f_{i,\boldsymbol{\theta}^{t+1}}(\theta) - f_{i,\boldsymbol{\theta}^t}(\theta)| \le (1-\rho_i)\, L_d L_h \sum_j w_{ij}\|\theta_j^{t+1,0} - \theta_j^{t,0}\|.$$

The proof of Theorem 1 controls the expectation of the right-hand side above through the neighbors' local updates, rather than requiring a uniform constant for all nodes, rounds, or parameter values. Its resulting drift estimate can decrease as local gradient magnitudes decrease. In contrast, the stated bound in Theorem 3 retains the same drift penalty given the fixed $\tilde{U}$. This distinction explains the potential improvement achieved by Theorem 1.

C.7 Lemmas and Proofs for Generalization Bound

In this subsection, we show the Lemmas and proofs needed for Theorem 5. For brevity of mathematical expressions, we define:

$$\hat{R}_i(\theta) := \frac{1}{\left|\hat{\mathcal{D}}_i\right|} \sum_{(x,y)\in\hat{\mathcal{D}}_i} \sup_{A\in\mathcal{B}_\mathcal{Y}} \left|\mathcal{P}^{(i)}_{Y|X=x}(A) - \mathcal{Q}_{i,\theta,X}(A)\right|,$$

$$K_i\left(\boldsymbol{\theta}\right) := \mathbb{E}_{X\sim\mathcal{P}^{\text{re}}_X} \text{TV}\left(\mathcal{Q}_{i,\theta,X}, \sum_j w_{ij}\mathcal{Q}_{j,\theta_j}(\cdot|X)\right),$$

$$\hat{K}_i\left(\boldsymbol{\theta}\right) := \frac{1}{|\hat{\mathcal{D}}^{\text{re}}|} \sum_{x\in\hat{\mathcal{D}}^{\text{re}}} \text{TV}\left(\mathcal{Q}_{i,\theta_i,x}, \sum_j w_{ij}\mathcal{Q}_{j,\theta_j,x}\right).$$

Here $\boldsymbol{\theta} := \{\theta_l\}_l$. Recall $\hat{\theta}_i$ is the output model, $b_i := \text{TV}(\mathcal{P}^{(i)}_X, \mathcal{P}^{\text{re}}_X) = \sup_{A\in\mathcal{B}_\mathcal{X}} \left|\mathcal{P}^{(i)}_X(A) - \mathcal{P}^{\text{re}}_X(A)\right|$, where $\mathcal{B}_\mathcal{X}$ is the Borel sigma-field on $\mathcal{X}$ and $A_{ij} = \text{TV}(\mathcal{P}^{(i)}_X, \mathcal{P}^{(j)}_X) = \sup_{A\in\mathcal{B}_\mathcal{X}} \left|\mathcal{P}^{(i)}_X(A) - \mathcal{P}^{(j)}_X(A)\right|$.

Lemma 9. *For all i, we have*

$$R_i(\hat{\theta}_i) \le K_i\left(\hat{\boldsymbol{\theta}}\right) + \sum_j w_{ij}R_j(\hat{\theta}_j) + 2\sum_j w_{ij}A_{ij} + \sum_j w_{ij}b_j. \tag{46}$$

Proof.

$$\begin{aligned}
R_i(\hat{\theta}_i) &= \mathbb{E}_{X\sim\mathcal{P}^{(i)}_X}\left[\text{TV}\left(\mathcal{P}_{Y|X}, \mathcal{Q}_{i,\hat{\theta}_i,X}\right)\right] \\
&\le \mathbb{E}_{X\sim\mathcal{P}^{(i)}_X}\left[\text{TV}\left(\mathcal{P}_{Y|X}, \sum_j w_{ij}\mathcal{Q}_{j,\hat{\theta}_j,X}\right)\right] + \mathbb{E}_{X\sim\mathcal{P}^{(i)}_X}\left[\text{TV}\left(\mathcal{Q}_{i,\hat{\theta}_i,X}, \sum_j w_{ij}\mathcal{Q}_{j,\hat{\theta}_j,X}\right)\right],
\end{aligned}$$

which holds since $\sup_{a,b} a+b \le \sup_a a + \sup_b b$.

For $\mathbb{E}_{X\sim\mathcal{P}^{(i)}_X}\left[\text{TV}\left(\mathcal{Q}_{i,\hat{\theta}_i,X}, \sum_j w_{ij}\mathcal{Q}_{j,\hat{\theta}_j,X}\right)\right]$, we have

$$\begin{aligned}
&\mathbb{E}_{X\sim\mathcal{P}^{(i)}_X}\left[\text{TV}\left(\mathcal{Q}_{i,\hat{\theta}_i,X}, \sum_j w_{ij}\mathcal{Q}_{j,\hat{\theta}_j,X}\right)\right] \le \mathbb{E}_{X\sim\mathcal{P}^{\text{re}}_X}\left[\text{TV}\left(\mathcal{Q}_{i,\hat{\theta}_i,X}, \sum_j w_{ij}\mathcal{Q}_{j,\hat{\theta}_j,X}\right)\right] + \text{TV}(\mathcal{P}^{(i)}_X, \mathcal{P}^{\text{re}}_X) \\
&\le \mathbb{E}_{X\sim\mathcal{P}^{\text{re}}_X}\left[\text{TV}\left(\mathcal{Q}_{i,\hat{\theta}_i,X}, \sum_j w_{ij}\mathcal{Q}_{j,\hat{\theta}_j,X}\right)\right] + \text{TV}\left(\mathcal{P}^{(i)}_X, \sum_j w_{ij}\mathcal{P}^{(j)}_X\right) + \text{TV}\left(\sum_j w_{ij}\mathcal{P}^{(j)}_X, \mathcal{P}^{\text{re}}_X\right) \\
&\le \mathbb{E}_{X\sim\mathcal{P}^{\text{re}}_X}\left[\text{TV}\left(\mathcal{Q}_{i,\hat{\theta}_i,X}, \sum_j w_{ij}\mathcal{Q}_{j,\hat{\theta}_j,X}\right)\right] + \sum_j w_{ij}\text{TV}\left(\mathcal{P}^{(i)}_X, \mathcal{P}^{(j)}_X\right) + \sum_j w_{ij}\text{TV}\left(\mathcal{P}^{(j)}_X, \mathcal{P}^{\text{re}}_X\right) \\
&\le K_i\left(\hat{\boldsymbol{\theta}}\right) + \sum_j w_{ij}A_{ij} + \sum_j w_{ij}b_j,
\end{aligned} \tag{47}$$

where the first inequality holds since for any distributions $\mathcal{P}, \mathcal{Q}$ on $\mathcal{X}$, and any $f : \mathcal{X} \to [0, B]$, we have $|\mathbb{E}_{X\sim\mathcal{P}}[f(X)] - \mathbb{E}_{X\sim\mathcal{Q}}[f(X)]| \le B\text{TV}(\mathcal{P}, \mathcal{Q})$.

For $\mathbb{E}_{X\sim\mathcal{P}_X^{(i)}}\left[\text{TV}\left(\mathcal{P}_{Y|X}, \sum_j w_{ij}\mathcal{Q}_{j,\hat{\theta}_j,X}\right)\right]$, we have

$$\begin{aligned}
&\mathbb{E}_{X\sim\mathcal{P}_X^{(i)}}\left[\text{TV}\left(\mathcal{P}_{Y|X}, \sum_j w_{ij}\mathcal{Q}_{j,\hat{\theta}_j,X}\right)\right] \leq \sum_j w_{ij}\mathbb{E}_{X\sim\mathcal{P}_X^{(i)}}\left[\text{TV}\left(\mathcal{P}_{Y|X}, \mathcal{Q}_{j,\hat{\theta}_j,X}\right)\right] \\
&\leq \sum_j w_{ij}\mathbb{E}_{X\sim\mathcal{P}_X^{(j)}}\left[\text{TV}\left(\mathcal{P}_{Y|X}, \mathcal{Q}_{j,\hat{\theta}_j,X}\right)\right] + \sum_j w_{ij}\text{TV}\left(\mathcal{P}_X^{(i)}, \mathcal{P}_X^{(j)}\right) \\
&\leq \sum_j w_{ij}R_j(\hat{\theta}_j) + \sum_j w_{ij}A_{ij}. \qquad (48)
\end{aligned}$$

Combining (47) and (48) proves (46). □

Lemma 10. *For all i, we have*

$$\begin{aligned}
R_i(\hat{\theta}_i) \leq &\rho R_i(\hat{\theta}_i) + (1-\rho)K_j\left(\hat{\boldsymbol{\theta}}\right) \\
&+ (1-\rho)\sum_j w_{ij}\left(\rho R_j(\hat{\theta}_j) + (1-\rho)K_j\left(\hat{\boldsymbol{\theta}}\right)\right) \\
&+ (1-\rho)^2\sum_j w_{ij}\sum_k w_{jk}R_k(\hat{\theta}_k) + (1-\rho)^2\sum_j w_{ij}\sum_k w_{jk}b_k \\
&+ 2(1-\rho)^2\sum_j w_{ij}\sum_k w_{jk}A_{ik} + (4-2\rho)(1-\rho)\sum_j w_{ij}A_{ij} + (1-\rho)\sum_j w_{ij}b_j. \qquad (49)
\end{aligned}$$

Proof. Inserting the right-hand side for $R_j(\hat{\theta}_j)$ back to (46), we have

$$\begin{aligned}
R_i(\hat{\theta}_i) \leq &K_i\left(\hat{\boldsymbol{\theta}}\right) + \sum_j w_{ij}\left(K_j\left(\hat{\boldsymbol{\theta}}\right) + \sum_k w_{jk}R_k(\hat{\theta}_k) + 2\sum_k w_{jk}A_{jk} + \sum_k w_{jk}b_k\right) \\
&+ 2\sum_j w_{ij}A_{ij} + \sum_j w_{ij}b_j \\
&\leq K_i\left(\hat{\boldsymbol{\theta}}\right) + \sum_j w_{ij}\left(K_j\left(\hat{\boldsymbol{\theta}}\right) + \sum_k w_{jk}R_k(\hat{\theta}_k) + \sum_k w_{jk}b_k\right) \\
&+ 2\sum_j w_{ij}\sum_k w_{jk}\left(A_{ij} + A_{ik}\right) + 2\sum_j w_{ij}A_{ij} + \sum_j w_{ij}b_j \\
&= K_i\left(\hat{\boldsymbol{\theta}}\right) + \sum_j w_{ij}\left(K_j\left(\hat{\boldsymbol{\theta}}\right) + \sum_k w_{jk}R_k(\hat{\theta}_k) + \sum_k w_{jk}b_k\right) \\
&+ 2\sum_j w_{ij}\sum_k w_{jk}A_{ik} + 4\sum_j w_{ij}A_{ij} + \sum_j w_{ij}b_j. \qquad (50)
\end{aligned}$$

Multiplying both sides of (46) by ρ, multiplying both sides of (50) by $1-\rho$, and summing two resulting inequalities, we have

$$\begin{aligned}
R_i(\hat{\theta}_i) \leq &K_i\left(\hat{\boldsymbol{\theta}}\right) + \sum_j w_{ij}\left(\rho R_j(\hat{\theta}_j) + (1-\rho)K_j\left(\hat{\boldsymbol{\theta}}\right)\right) \\
&+ (1-\rho)\sum_j w_{ij}\sum_k w_{jk}R_k(\hat{\theta}_k) + (1-\rho)\sum_j w_{ij}\sum_k w_{jk}b_k \\
&+ 2(1-\rho)\sum_j w_{ij}\sum_k w_{jk}A_{ik} + (4-2\rho)\sum_j w_{ij}A_{ij} + \sum_j w_{ij}b_j. \qquad (51)
\end{aligned}$$

Multiplying $1-\rho$ on the both sides of (51), and adding $\rho R_i(\hat{\theta}_i)$ to both sides of the resulting inequality, (49) is proved. □

Further define

$$\begin{aligned}
\phi_i(\theta) &:= \mathbb{E}_{(X,Y)\sim\mathcal{P}^{(i)}}(1-[h_{i,\theta}(X)]_Y),\\
\hat{\phi}_i(\theta) &:= \frac{1}{\left|\hat{\mathcal{D}}_i\right|}\sum_{(x,y)\in\hat{\mathcal{D}}_i}\left(1-[h_{i,\theta}(x)]_y\right),\\
F_i(\boldsymbol{\theta}) &:= \rho\phi_i(\theta_i)+(1-\rho)K_i(\{\theta_j\}_j),\\
\hat{F}_i(\boldsymbol{\theta}) &:= \rho\hat{\phi}_i(\theta_i)+(1-\rho)\hat{K}_i(\{\theta_j\}_j),\\
g_i\left(\boldsymbol{\theta},\{\hat{\mathcal{D}}_l\}_l,\hat{\mathcal{D}}^{\text{re}}\right) &:= (1+(1-\rho)w_{ii})\hat{F}_i(\boldsymbol{\theta})+(1-\rho)\sum_{j\neq i}w_{ij}\hat{F}_j(\{\theta_k\}_k)\\
g_i(\boldsymbol{\theta}) &:= \mathbb{E}_{\{\{\hat{\mathcal{D}}_l\}_l,\hat{\mathcal{D}}^{\text{re}}\}}g_i\left(\boldsymbol{\theta},\{\hat{\mathcal{D}}_l\}_l,\hat{\mathcal{D}}^{\text{re}}\right)\\
&= (1+(1-\rho)w_{ii})F_i(\boldsymbol{\theta})+(1-\rho)\sum_{j\neq i}w_{ij}F_j(\{\theta_k\}_k).
\end{aligned}$$

Lemma 11. *For all i, we have*

$$\rho R_i(\hat{\theta}_i)+(1-\rho)K_i\left(\hat{\boldsymbol{\theta}}\right)+(1-\rho)\sum_j w_{ij}\left(\rho R_j(\hat{\theta}_j)+(1-\rho)K_j\left(\hat{\boldsymbol{\theta}}\right)\right)\leq g_i\left(\hat{\boldsymbol{\theta}}\right). \tag{52}$$

Proof. By the definition of total variation distance, we have for all θ,

$$\begin{aligned}
\text{TV}\left(\mathcal{P}_{Y|X},\mathcal{Q}_{i,\theta,X}\right) &= 1-\sum_{y=1}^{K}\min\{\mathcal{P}_{Y|X}(Y=y),[h_{i,\theta}(x)]_y\}\\
&\leq 1-\sum_{y=1}^{K}\mathcal{P}_{Y|X}(Y=y)\,[h_{i,\theta}(x)]_y = \mathbb{E}_{Y\sim\mathcal{P}^{(i)}_{Y|X}}(1-[h_{i,\theta}(X)]_Y).
\end{aligned}$$

Therefore, we have

$$R_i(\theta)=\mathbb{E}_{X\sim\mathcal{P}^{(i)}_X}\left[\text{TV}\left(\mathcal{P}_{Y|X},\mathcal{Q}_{i,\theta,X}\right)\right]\leq\mathbb{E}_{X\sim\mathcal{P}^{(i)}_X}\mathbb{E}_{Y\sim\mathcal{P}^{(i)}_{Y|X}}(1-[h_{i,\theta}(X)]_Y)=\phi_i(\theta).$$

Then we have

$$\begin{aligned}
&\rho R_i(\hat{\theta}_i)+(1-\rho)K_i\left(\hat{\boldsymbol{\theta}}\right)+(1-\rho)\sum_j w_{ij}\left(\rho R_j(\hat{\theta}_j)+(1-\rho)K_j\left(\hat{\boldsymbol{\theta}}\right)\right)\leq\\
&(1+(1-\rho)w_{ii})F_i\left(\hat{\boldsymbol{\theta}}\right)+(1-\rho)\sum_{j\neq i}w_{ij}F_j\left(\hat{\boldsymbol{\theta}}\right)=g_i\left(\hat{\boldsymbol{\theta}}\right).
\end{aligned}$$

□

Lemma 12. *For any i, $\forall\delta>0$, we have with probability $\geq 1-\delta$ over the randomness of the datasets $\{\hat{\mathcal{D}}_l\}_l$ and $\hat{\mathcal{D}}^{re}$,*

$$\sup_{\{\theta_l\in\Theta_l\}_l}g_i(\boldsymbol{\theta})-g_i\left(\boldsymbol{\theta},\{\hat{\mathcal{D}}_l\}_l,\hat{\mathcal{D}}^{re}\right)\leq\mathbb{E}_{\{\{\hat{\mathcal{D}}_l\}_l,\hat{\mathcal{D}}^{re}\}}\left[\sup_{\{\theta_l\in\Theta_l\}_l}g_i(\boldsymbol{\theta})-g_i\left(\boldsymbol{\theta},\{\hat{\mathcal{D}}_l\}_l,\hat{\mathcal{D}}^{re}\right)\right]+$$
$$\sqrt{\frac{1}{2}\log\left(\frac{1}{\delta}\right)\left(\frac{((1+(1-\rho)w_{ii})\rho)^2}{\left|\hat{\mathcal{D}}_i\right|}+\sum_{j\neq i}\frac{((1-\rho)w_{ij}\rho)^2}{\left|\hat{\mathcal{D}}_j\right|}+\frac{(1-\rho)^2(2-\rho)^2}{|\hat{\mathcal{D}}^{re}|}\right)}.$$

Proof. This result follows from the McDiarmid's inequality. To apply this inequality, we first show the bounded differences property of $\sup_{\{\theta_l\in\Theta_l\}_l}g_i(\boldsymbol{\theta})-g_i\left(\boldsymbol{\theta},\{\hat{\mathcal{D}}_l\}_l,\hat{\mathcal{D}}^{\text{re}}\right)$.

For any i, replace an arbitrary point (x, y) in $\hat{\mathcal{D}}_i$ by $(x', y') \in \mathcal{X} \times \mathcal{Y}$ and denote the resulting dataset by $\hat{\mathcal{D}}'_i$. We have, for any i and any $\boldsymbol{\theta}$,

$$\sup_{\hat{\mathcal{D}}'_i} \left| g_i\left(\boldsymbol{\theta}, \{\hat{\mathcal{D}}_l\}_l, \hat{\mathcal{D}}^{\text{re}}\right) - g_i\left(\boldsymbol{\theta}, \{\hat{\mathcal{D}}'_i\} \cup \{\hat{\mathcal{D}}_l\}_{l\neq i}, \hat{\mathcal{D}}^{\text{re}}\right)\right| =$$
$$\sup_{\hat{\mathcal{D}}'_i} (1 + (1-\rho)w_{ii})\rho \left|\hat{\phi}_i(\theta_i) - \hat{\phi}'_i(\theta_i)\right| \leq \frac{(1 + (1-\rho)w_{ii})\rho}{\left|\hat{\mathcal{D}}_i\right|},$$

where we define $\hat{\phi}'_i(\theta) := \frac{1}{|\hat{\mathcal{D}}_i|} \sum_{(x,y)\in\hat{\mathcal{D}}_i} \left(1 - [h_{i,\theta}(x)]_y\right)$.

Therefore, we have

$$\sup_{\hat{\mathcal{D}}'_i} \left| \sup_{\{\theta_l\in\Theta_l\}_l} \left[g_i(\boldsymbol{\theta}) - g_i\left(\boldsymbol{\theta}, \{\hat{\mathcal{D}}_l\}_l, \hat{\mathcal{D}}^{\text{re}}\right)\right] - \sup_{\{\theta_l\in\Theta_l\}_l} \left[g_i(\boldsymbol{\theta}) - g_i\left(\boldsymbol{\theta}, \{\hat{\mathcal{D}}'_j\} \cup \{\hat{\mathcal{D}}_l\}_{l\neq j}, \hat{\mathcal{D}}^{\text{re}}\right)\right] \right| \leq$$
$$\sup_{\hat{\mathcal{D}}'_i} \left| \sup_{\{\theta_l\in\Theta_l\}_l} \left[g_i\left(\boldsymbol{\theta}, \{\hat{\mathcal{D}}'_j\} \cup \{\hat{\mathcal{D}}_l\}_{l\neq j}, \hat{\mathcal{D}}^{\text{re}}\right) - g_i\left(\boldsymbol{\theta}, \{\hat{\mathcal{D}}_l\}_l, \hat{\mathcal{D}}^{\text{re}}\right)\right] \right| \leq$$
$$\sup_{\hat{\mathcal{D}}'_i} \sup_{\{\theta_l\in\Theta_l\}_l} \left| g_i\left(\boldsymbol{\theta}, \{\hat{\mathcal{D}}'_j\} \cup \{\hat{\mathcal{D}}_l\}_{l\neq j}, \hat{\mathcal{D}}^{\text{re}}\right) - g_i\left(\boldsymbol{\theta}, \{\hat{\mathcal{D}}_l\}_l, \hat{\mathcal{D}}^{\text{re}}\right)\right| \leq \frac{(1 + (1-\rho)w_{ii})\rho}{\left|\hat{\mathcal{D}}_i\right|}. \tag{53}$$

Similarly, for any i, replace an arbitrary point (x, y) in $\hat{\mathcal{D}}_j, \forall j \neq i$ by $(x', y') \in \mathcal{X} \times \mathcal{Y}$ and denote the resulting dataset by $\hat{\mathcal{D}}'_j$. We have, for any i and any $\boldsymbol{\theta}$,

$$\sup_{\hat{\mathcal{D}}'_j} \left| g_i\left(\boldsymbol{\theta}, \{\hat{\mathcal{D}}_l\}_l, \hat{\mathcal{D}}^{\text{re}}\right) - g_i\left(\boldsymbol{\theta}, \{\hat{\mathcal{D}}'_j\} \cup \{\hat{\mathcal{D}}_l\}_{l\neq j}, \hat{\mathcal{D}}^{\text{re}}\right)\right| =$$
$$\sup_{\hat{\mathcal{D}}'_j} (1-\rho)w_{ij}\rho \left|\hat{\phi}_j(\theta_j) - \hat{\phi}'_j(\theta_j)\right| \leq \frac{(1-\rho)w_{ij}\rho}{\left|\hat{\mathcal{D}}_j\right|}.$$

For any i, replacing an arbitrary point x in $\hat{\mathcal{D}}^{\text{re}}$ by $x' \in \mathcal{X}$ and denoting the resulting dataset by $\hat{\mathcal{D}}'$, we have, for any i and any $\boldsymbol{\theta}$,

$$\sup_{\hat{\mathcal{D}}'} \left| g_i\left(\boldsymbol{\theta}, \{\hat{\mathcal{D}}_l\}_l, \hat{\mathcal{D}}^{\text{re}}\right) - g_i\left(\boldsymbol{\theta}, \{\hat{\mathcal{D}}_l\}_l, \hat{\mathcal{D}}'\right)\right| =$$
$$\sup_{\hat{\mathcal{D}}'} \Big| (1-\rho)(1 + (1-\rho)w_{ii}) \left(\hat{K}_i(\{\theta_j\}_j) - \hat{K}'_i(\{\theta_j\}_j)\right) +$$
$$(1-\rho)^2 \sum_{j\neq i} w_{ij} \left(\hat{K}_i(\{\theta_j\}_j) - \hat{K}'_i(\{\theta_j\}_j)\right) \Big| \leq$$
$$\frac{(1-\rho)(1 + (1-\rho)w_{ii})}{|\hat{\mathcal{D}}^{\text{re}}|} + \frac{(1-\rho)^2 \sum_{j\neq i} w_{ij}}{|\hat{\mathcal{D}}^{\text{re}}|} = \frac{(1-\rho)(2-\rho)}{|\hat{\mathcal{D}}^{\text{re}}|}, \forall i,$$

where we define $\hat{K}'_i(\boldsymbol{\theta}) = \frac{1}{|\hat{\mathcal{D}}^{\text{re}}|} \sum_{x\in\hat{\mathcal{D}}'} \text{TV}\left(\mathcal{Q}_{i,\theta_i,x}, \sum_j w_{ij}\mathcal{Q}_{j,\theta_j,x}\right)$.

Similar to (53), we can show the corresponding differences of $\sup_{\{\theta_l\in\Theta_l\}_l} g_i(\boldsymbol{\theta}) - g_i\left(\boldsymbol{\theta}, \{\hat{\mathcal{D}}_l\}_l, \hat{\mathcal{D}}^{\text{re}}\right)$ are also bounded in $\hat{\mathcal{D}}'_j$ and $\hat{\mathcal{D}}'$ cases.

Therefore, based on the McDiarmid's inequality, we have, for any i and any $\bar{\delta} > 0$,

$$P\left(\sup_{\{\theta_l\in\Theta_l\}_l} g_i(\boldsymbol{\theta}) - g_i\left(\boldsymbol{\theta}, \{\hat{\mathcal{D}}_l\}_l, \hat{\mathcal{D}}^{\text{re}}\right) - \mathbb{E}_{\{\{\hat{\mathcal{D}}_l\}_l, \hat{\mathcal{D}}^{\text{re}}\}}\left[\sup_{\{\theta_l\in\Theta_l\}_l} g_i(\boldsymbol{\theta}) - g_i\left(\boldsymbol{\theta}, \{\hat{\mathcal{D}}_l\}_l, \hat{\mathcal{D}}^{\text{re}}\right)\right] \geq \bar{\delta}\right) \leq$$
$$\exp\left(-\frac{2\bar{\delta}^2}{|\hat{\mathcal{D}}_i|\left(\frac{(1+(1-\rho)w_{ii})\rho}{|\hat{\mathcal{D}}_i|}\right)^2 + \sum_{j\neq i} w_{ij}|\hat{\mathcal{D}}_j|\left(\frac{(1-\rho)w_{ij}\rho}{|\hat{\mathcal{D}}_j|}\right)^2 + |\hat{\mathcal{D}}^{\text{re}}|\left(\frac{(1-\rho)(2-\rho)}{|\hat{\mathcal{D}}^{\text{re}}|}\right)^2}\right) =$$
$$\exp\left(-\frac{2\bar{\delta}^2}{\frac{((1+(1-\rho)w_{ii})\rho)^2}{|\hat{\mathcal{D}}_i|} + \sum_{j\neq i}\frac{((1-\rho)w_{ij}\rho)^2}{|\hat{\mathcal{D}}_j|} + \frac{(1-\rho)^2(2-\rho)^2}{|\hat{\mathcal{D}}^{\text{re}}|}}\right).$$

Equivalently, for any i and $\delta > 0$, we have with probability $\geq 1-\delta$,

$$\sup_{\{\theta_l\in\Theta_l\}_l} g_i(\boldsymbol{\theta}) - g_i\left(\boldsymbol{\theta}, \{\hat{\mathcal{D}}_l\}_l, \hat{\mathcal{D}}^{\text{re}}\right) - \mathbb{E}_{\{\{\hat{\mathcal{D}}_l\}_l, \hat{\mathcal{D}}^{\text{re}}\}}\left[\sup_{\{\theta_l\in\Theta_l\}_l} g_i(\boldsymbol{\theta}) - g_i\left(\boldsymbol{\theta}, \{\hat{\mathcal{D}}_l\}_l, \hat{\mathcal{D}}^{\text{re}}\right)\right] <$$
$$\sqrt{\frac{1}{2}\log\left(\frac{1}{\delta}\right)\left(\frac{((1+(1-\rho)w_{ii})\rho)^2}{\left|\hat{\mathcal{D}}_i\right|} + \sum_{j\neq i}\frac{((1-\rho)w_{ij}\rho)^2}{\left|\hat{\mathcal{D}}_j\right|} + \frac{(1-\rho)^2(2-\rho)^2}{|\hat{\mathcal{D}}^{\text{re}}|}\right)}.$$

□

Lemma 13. *For all i, we have*

$$\mathbb{E}_{\{\{\hat{\mathcal{D}}_l\}_l, \hat{\mathcal{D}}^{re}\}}\left[\sup_{\{\theta_l\in\Theta_l\}_l} g_i(\boldsymbol{\theta}) - g_i\left(\boldsymbol{\theta}, \{\hat{\mathcal{D}}_l\}_l, \hat{\mathcal{D}}^{re}\right)\right] \leq$$
$$2\rho(1+(1-\rho)w_{ii})\Re_i + 2\rho(1-\rho)\sum_{j\neq i} w_{ij}\Re_j + (1-\rho)(2-\rho),$$

where $\Re_i := \mathbb{E}_{\hat{\mathcal{D}}_i, \boldsymbol{\sigma}_i} \sup_{\theta_i\in\Theta_i}\left[\frac{1}{|\hat{\mathcal{D}}_i|}\sum_{l=1}^{|\hat{\mathcal{D}}_i|}\sigma_l\left[h_{i,\theta_i}(x_l)\right]_{y_l}\right]$ and $\boldsymbol{\sigma}_i := \{\sigma_1, ..., \sigma_{|\hat{\mathcal{D}}_i|}\}$ is a set of i.i.d. random variables following the Rademacher distribution.

Proof.

$$\begin{aligned}
&g_i(\boldsymbol{\theta}) - g_i\left(\boldsymbol{\theta}, \{\hat{\mathcal{D}}_l\}_l, \hat{\mathcal{D}}^{\text{re}}\right)\\
=&(1+(1-\rho)w_{ii})\left(F_i(\boldsymbol{\theta}) - \hat{F}_i(\boldsymbol{\theta})\right) + (1-\rho)\sum_{j\neq i} w_{ij}\left(F_j(\{\theta_k\}_k) - \hat{F}_j(\{\theta_k\}_k)\right)\\
=&\rho(1+(1-\rho)w_{ii})\left(\phi_i(\theta_i) - \hat{\phi}_i(\theta_i)\right) + \rho(1-\rho)\sum_{j\neq i} w_{ij}\left(\phi_j(\theta_j) - \hat{\phi}_j(\theta_j)\right)\\
&+(1-\rho)(1+(1-\rho)w_{ii})\left(K_i(\{\theta_l\}_l) - \hat{K}_i(\{\theta_l\}_l)\right) + (1-\rho)^2\sum_{j\neq i} w_{ij}\left(K_j(\{\theta_l\}_l) - \hat{K}_j(\{\theta_l\}_l)\right)\\
\leq&\rho(1+(1-\rho)w_{ii})\left(\phi_i(\theta_i) - \hat{\phi}_i(\theta_i)\right) + \rho(1-\rho)\sum_{j\neq i} w_{ij}\left(\phi_j(\theta_j) - \hat{\phi}_j(\theta_j)\right)\\
&+(1-\rho)(1+(1-\rho)w_{ii}) + (1-\rho)^2\sum_{j\neq i} w_{ij}\\
\leq&\rho(1+(1-\rho)w_{ii})\left(\phi_i(\theta_i) - \hat{\phi}_i(\theta_i)\right) + \rho(1-\rho)\sum_{j\neq i} w_{ij}\left(\phi_j(\theta_j) - \hat{\phi}_j(\theta_j)\right) + (1-\rho)(2-\rho).
\end{aligned}$$

Therefore, we have

$$\mathbb{E}_{\{\{\hat{\mathcal{D}}_l\}_l,\hat{\mathcal{D}}^{\text{re}}\}} \sup_{\{\theta_l\in\Theta_l\}_l} \left[g_i\left(\boldsymbol{\theta}\right) - g_i\left(\boldsymbol{\theta}, \{\hat{\mathcal{D}}_l\}_l, \hat{\mathcal{D}}^{\text{re}}\right)\right] \leq \rho(1+(1-\rho)w_{ii})\mathbb{E}_{\hat{\mathcal{D}}_i} \sup_{\theta_i\in\Theta_i} \left(\phi_i(\theta_i) - \hat{\phi}_i(\theta_i)\right)$$
$$+ \rho(1-\rho)\sum_{j\neq i} w_{ij}\mathbb{E}_{\hat{\mathcal{D}}_j} \sup_{\theta_j\in\Theta_j} \left(\phi_j(\theta_j) - \hat{\phi}_j(\theta_j)\right) + (1-\rho)(2-\rho). \tag{54}$$

Next, we bound the right-hand side of (54) using the corresponding Rademacher complexity, where the procedure is similar to that in Konstantinov & Lampert (2019).

For any i, defining $\bar{\mathcal{D}}_i := \{(\bar{x}_l, \bar{y}_l)\}_{l=1}^{|\hat{\mathcal{D}}_i|}$, which is a set of i.i.d. data points following $\mathcal{P}^{(i)}$, we have

$$\mathbb{E}_{\hat{\mathcal{D}}_i} \sup_{\theta_i\in\Theta_i} \left(\phi_i(\theta_i) - \hat{\phi}_i(\theta_i)\right) = \mathbb{E}_{\hat{\mathcal{D}}_i} \sup_{\theta_i\in\Theta_i} \left[\mathbb{E}_{(X,Y)\sim\mathcal{P}^{(i)}}(1-[h_{i,\theta_i}(X)]_Y) - \frac{1}{\left|\hat{\mathcal{D}}_i\right|}\sum_{l=1}^{|\hat{\mathcal{D}}_i|}\left(1-[h_{i,\theta_i}(x_l)]_{y_l}\right)\right]$$
$$\leq \mathbb{E}_{\hat{\mathcal{D}}_i,\bar{\mathcal{D}}_i} \sup_{\theta_i\in\Theta_i} \left[\frac{1}{\left|\hat{\mathcal{D}}_i\right|}\sum_{l=1}^{|\hat{\mathcal{D}}_i|}\left(1-[h_{i,\theta_i}(\bar{x}_l)]_{\bar{y}_l}\right) - \frac{1}{\left|\hat{\mathcal{D}}_i\right|}\sum_{l=1}^{|\hat{\mathcal{D}}_i|}\left(1-[h_{i,\theta_i}(x_l)]_{y_l}\right)\right]$$
$$= \mathbb{E}_{\hat{\mathcal{D}}_i,\bar{\mathcal{D}}_i,\boldsymbol{\sigma}_i} \sup_{\theta_i\in\Theta_i} \left[\frac{1}{\left|\hat{\mathcal{D}}_i\right|}\sum_{l=1}^{|\hat{\mathcal{D}}_i|}\sigma_l\left[\left(1-[h_{i,\theta_i}(\bar{x}_l)]_{\bar{y}_l}\right) - \left(1-[h_{i,\theta_i}(x_l)]_{y_l}\right)\right]\right]$$
$$= \mathbb{E}_{\hat{\mathcal{D}}_i,\bar{\mathcal{D}}_i,\boldsymbol{\sigma}_i} \sup_{\theta_i\in\Theta_i} \left[\frac{1}{\left|\hat{\mathcal{D}}_i\right|}\sum_{l=1}^{|\hat{\mathcal{D}}_i|}\sigma_l\left([h_{i,\theta_i}(x_l)]_{y_l} - [h_{i,\theta_i}(\bar{x}_l)]_{\bar{y}_l}\right)\right]$$
$$\leq \mathbb{E}_{\hat{\mathcal{D}}_i,\boldsymbol{\sigma}_i} \sup_{\theta_i\in\Theta_i} \left[\frac{1}{\left|\hat{\mathcal{D}}_i\right|}\sum_{l=1}^{|\hat{\mathcal{D}}_i|}\sigma_l\left[h_{i,\theta_i}(x_l)\right]_{y_l}\right] + \mathbb{E}_{\bar{\mathcal{D}}_i,\boldsymbol{\sigma}_i} \sup_{\theta_i\in\Theta_i} \left[\frac{1}{\left|\hat{\mathcal{D}}_i\right|}\sum_{l=1}^{|\hat{\mathcal{D}}_i|}(-\sigma_l)\left[h_{i,\theta_i}(\bar{x}_l)\right]_{\bar{y}_l}\right]$$
$$= 2\mathbb{E}_{\hat{\mathcal{D}}_i,\boldsymbol{\sigma}_i} \sup_{\theta_i\in\Theta_i} \left[\frac{1}{\left|\hat{\mathcal{D}}_i\right|}\sum_{l=1}^{|\hat{\mathcal{D}}_i|}\sigma_l\left[h_{i,\theta_i}(x_l)\right]_{y_l}\right] = 2\Re_i,$$

where $\boldsymbol{\sigma}_i := \{\sigma_1, ..., \sigma_{|\hat{\mathcal{D}}_i|}\}$ is a set of i.i.d. random variables following the Rademacher distribution.

We thus have

$$\mathbb{E}_{\{\{\hat{\mathcal{D}}_l\}_l,\hat{\mathcal{D}}^{\text{re}}\}} \sup_{\{\theta_l\in\Theta_l\}_l} \left[g_i\left(\boldsymbol{\theta}\right) - g_i\left(\boldsymbol{\theta}, \{\hat{\mathcal{D}}_l\}_l, \hat{\mathcal{D}}^{\text{re}}\right)\right] \leq$$
$$2\rho(1+(1-\rho)w_{ii})\Re_i + 2\rho(1-\rho)\sum_{j\neq i} w_{ij}\Re_j + (1-\rho)(2-\rho).$$

□

Lemma 14. *For any i, $\forall\delta > 0$, we have with probability $\geq 1-\delta$,*

$$R_i(\hat{\theta}_i) \leq g_i\left(\hat{\boldsymbol{\theta}}, \{\hat{\mathcal{D}}_l\}_l, \hat{\mathcal{D}}^{re}\right) + 2\rho(1+(1-\rho)w_{ii})\Re_i + 2\rho(1-\rho)\sum_{j\neq i} w_{ij}\Re_j + (1-\rho)(2-\rho) +$$
$$\sqrt{\frac{1}{2}\log\left(\frac{1}{\delta}\right)\left(\frac{\left((1+(1-\rho)w_{ii})\rho\right)^2}{\left|\hat{\mathcal{D}}_i\right|} + \sum_{j\neq i}\frac{\left((1-\rho)w_{ij}\rho\right)^2}{\left|\hat{\mathcal{D}}_j\right|} + \frac{(1-\rho)^2(2-\rho)^2}{|\hat{\mathcal{D}}^{re}|}\right)}$$
$$+ (1-\rho)^2\sum_j w_{ij}\sum_k w_{jk}R_k(\hat{\theta}_k) + (1-\rho)^2\sum_j w_{ij}\sum_k w_{jk}b_k$$
$$+ 2(1-\rho)^2\sum_j w_{ij}\sum_k w_{jk}A_{ik} + (4-2\rho)(1-\rho)\sum_j w_{ij}A_{ij} + (1-\rho)\sum_j w_{ij}b_j.$$

Proof. This result is a direct consequence of 10-13. □

Lemma 15. *When ℓ and d are KL divergence, we have*

$$\hat{F}_i\left(\hat{\boldsymbol{\theta}}\right) \leq f_{i,\hat{\boldsymbol{\theta}}}\left(\hat{\theta}_i\right) + \frac{(1-\rho)M}{4}. \tag{55}$$

When ℓ is KL divergence and d is squared Euclidean norm, we have

$$\hat{F}_i\left(\hat{\boldsymbol{\theta}}\right) \leq f_{i,\hat{\boldsymbol{\theta}}}\left(\hat{\theta}_i\right) + \frac{1-\rho}{8}. \tag{56}$$

Proof. Based on (52), we have

$$\rho R_i(\hat{\theta}_i) + (1-\rho)K_j\left(\hat{\boldsymbol{\theta}}\right) + (1-\rho)\sum_j w_{ij}\left(\rho R_j(\hat{\theta}_j) + (1-\rho)K_j\left(\hat{\boldsymbol{\theta}}\right)\right) \leq g_i\left(\hat{\boldsymbol{\theta}}\right)$$

$$\leq g_i\left(\hat{\boldsymbol{\theta}}, \{\hat{\mathcal{D}}_i'\} \cup \{\hat{\mathcal{D}}_l\}_{l\neq i}, \hat{\mathcal{D}}^{\text{re}}\right) + \sup_{\{\theta_l \in \Theta_l\}_l} g_i\left(\boldsymbol{\theta}\right) - g_i\left(\boldsymbol{\theta}, \{\hat{\mathcal{D}}_l\}_l, \hat{\mathcal{D}}^{\text{re}}\right).$$

Next, we bound $g_i\left(\hat{\boldsymbol{\theta}}, \{\hat{\mathcal{D}}_i'\} \cup \{\hat{\mathcal{D}}_l\}_{l\neq i}, \hat{\mathcal{D}}^{\text{re}}\right)$ by the empirical loss that we can control.

Since $1-u \leq -\log u, \forall u \in (0,1]$, we have

$$\hat{\phi}_i(\hat{\theta}_i) = \frac{1}{\left|\hat{\mathcal{D}}_i\right|}\sum_{(x,y)\in\hat{\mathcal{D}}_i}\left(1 - \left[h_{i,\hat{\theta}_i}(x)\right]_y\right) \leq \frac{1}{\left|\hat{\mathcal{D}}_i\right|}\sum_{(x,y)\in\hat{\mathcal{D}}_i}\left(-\log\left[h_{i,\hat{\theta}_i}(x)\right]_y\right).$$

On the other hand, based on the Pinsker's inequality,

$$\hat{K}_i\left(\hat{\boldsymbol{\theta}}\right) = \frac{1}{|\hat{\mathcal{D}}^{\text{re}}|}\sum_{x\in\hat{\mathcal{D}}^{\text{re}}}\text{TV}\left(\mathcal{Q}_{i,\hat{\theta}_i,x}, w_{ii}\mathcal{Q}_{i,\hat{\theta}_i,x} + \sum_{j\neq i} w_{ij}\mathcal{Q}_{j,\hat{\theta}_j,x}\right)$$

$$\leq \sqrt{\frac{1}{2|\hat{\mathcal{D}}^{\text{re}}|}\sum_{x\in\hat{\mathcal{D}}^{\text{re}}}\text{KL}\left(\mathcal{Q}_{i,\hat{\theta}_i,x} || w_{ii}\mathcal{Q}_{i,\hat{\theta}_i,x} + \sum_{j\neq i} w_{ij}\mathcal{Q}_{j,\hat{\theta}_j,x}\right)}$$

$$= \sqrt{\frac{1}{2|\hat{\mathcal{D}}^{\text{re}}|}\sum_{x\in\hat{\mathcal{D}}^{\text{re}}}\sum_{y=1}^{M}\left[\sum_j w_{ij}h_{j,\hat{\theta}_j}(x)\right]_y \log\frac{\left[\sum_j w_{ij}h_{j,\hat{\theta}_j}(x)\right]_y}{\left[h_{i,\hat{\theta}_i}(x)\right]_y}},$$

where KL is the KL-divergence.

Consider scalars $a, b, c \geq 0$. Let $s = \rho a + (1-\rho)b$. Then

$$\rho a + (1-\rho)c\sqrt{b} = s - (1-\rho)b + (1-\rho)c\sqrt{b} = s + \frac{(1-\rho)c^2}{4} - (1-\rho)\left(\sqrt{b} - \frac{c}{2}\right)^2 \leq s + \frac{(1-\rho)c^2}{4}. \tag{57}$$

Letting $c = \frac{1}{\sqrt{2}}$ in (57), we have

$$\hat{F}_i\left(\hat{\boldsymbol{\theta}}\right) = \rho\hat{\phi}_i(\theta) + (1-\rho)\hat{K}_i\left(\theta_j\}_j\right) \leq$$

$$\frac{\rho}{\left|\hat{\mathcal{D}}_i\right|}\sum_{(x,y)\in\hat{\mathcal{D}}_i}\left(-\log\left[h_{i,\hat{\theta}_i}(x)\right]_y\right) + (1-\rho)\sqrt{\frac{1}{2|\hat{\mathcal{D}}^{\text{re}}|}\sum_{x\in\hat{\mathcal{D}}^{\text{re}}}\sum_{y=1}^{M}\left[\sum_j w_{ij}h_{j,\hat{\theta}_j}(x)\right]_y \log\frac{\left[\sum_j w_{ij}h_{j,\hat{\theta}_j}(x)\right]_y}{\left[h_{i,\hat{\theta}_i}(x)\right]_y}} \leq$$

$$\frac{\rho}{\left|\hat{\mathcal{D}}_i\right|}\sum_{(x,y)\in\hat{\mathcal{D}}_i}\left(-\log\left[h_{i,\hat{\theta}_i}(x)\right]_y\right) + \frac{(1-\rho)}{|\hat{\mathcal{D}}^{\text{re}}|}\sum_{x\in\hat{\mathcal{D}}^{\text{re}}}\sum_{y=1}^{M}\left[\sum_j w_{ij}h_{j,\hat{\theta}_j}(x)\right]_y \log\frac{\left[\sum_j w_{ij}h_{j,\hat{\theta}_j}(x)\right]_y}{\left[h_{i,\hat{\theta}_i}(x)\right]_y} + \frac{1-\rho}{8} =$$

$$f_{i,\hat{\boldsymbol{\theta}}}\left(\hat{\theta}_i\right) + \frac{1-\rho}{8}.$$

Here the distillation loss function d in $f_{i,\hat{\boldsymbol{\theta}}}\left(\hat{\theta}_i\right)$ is KL divergence. This proves (55).

Similarly,

$$\begin{aligned}\hat{K}_i\left(\hat{\boldsymbol{\theta}}\right) &= \frac{1}{|\hat{\mathcal{D}}^{\text{re}}|}\sum_{x\in\hat{\mathcal{D}}^{\text{re}}}\sum_{y=1}^{M}\left|\left[h_{i,\hat{\theta}_i}(x)\right]_y - \left[\sum_j w_{ij}h_{j,\hat{\theta}_j}(x)\right]_y\right| \\ &\leq \frac{\sqrt{M}}{|\hat{\mathcal{D}}^{\text{re}}|}\sum_{x\in\hat{\mathcal{D}}^{\text{re}}}\sqrt{\sum_{y=1}^{M}\left|\left[h_{i,\hat{\theta}_i}(x)\right]_y - \left[\sum_j w_{ij}h_{j,\hat{\theta}_j}(x)\right]_y\right|^2}, \\ &\leq \sqrt{M}\sqrt{\frac{1}{|\hat{\mathcal{D}}^{\text{re}}|}\sum_{x\in\hat{\mathcal{D}}^{\text{re}}}\sum_{y=1}^{M}\left|\left[h_{i,\hat{\theta}_i}(x)\right]_y - \left[\sum_j w_{ij}h_{j,\hat{\theta}_j}(x)\right]_y\right|^2}.\end{aligned}$$

Letting $c=\sqrt{M}$ in (57), we have

$$\hat{F}_i\left(\hat{\boldsymbol{\theta}}\right) \leq f_{i,\hat{\boldsymbol{\theta}}}\left(\hat{\theta}_i\right) + \frac{(1-\rho)M}{4},$$

when d is squared Euclidean norm, which proves (56). □

C.7.1 Proof of Theorem 5

Proof. When Assumptions 9 and 10 hold, based on Lemma 15, we have

$$\begin{aligned}\mathbb{E}g_i\left(\hat{\boldsymbol{\theta}}, \{\hat{\mathcal{D}}_l\}_l, \hat{\mathcal{D}}^{\text{re}}\right) =& (1+(1-\rho)w_{ii})\mathbb{E}\hat{F}_i\left(\hat{\boldsymbol{\theta}}\right) + (1-\rho)\sum_{j\neq i} w_{ij}\mathbb{E}\hat{F}_j\left(\hat{\theta}_j, \{\hat{\theta}_k\}_{k\neq j}\right) \\ \leq& (1+(1-\rho)w_{ii})\mathbb{E}f_{i,\hat{\boldsymbol{\theta}}}\left(\hat{\theta}_i\right) + (1-\rho)\sum_{j\neq i} w_{ij}\mathbb{E}f_{j,\hat{\boldsymbol{\theta}}}\left(\hat{\theta}_i\right) + (2-\rho)r \\ \leq& (1+(1-\rho)w_{ii})\left(\mathbb{E}\left[\frac{1}{2\mu_i}\left\|\nabla f_{i,\hat{\boldsymbol{\theta}}}\left(\hat{\theta}_i\right)\right\|^2\right] + \epsilon_i\right) \\ &+ (1-\rho)\sum_{j\neq i} w_{ij}\left(\mathbb{E}\left[\frac{1}{2\mu_j}\left\|\nabla f_{j,\{\hat{\theta}_l\}_l}\left(\hat{\theta}_j\right)\right\|^2\right] + \epsilon_j\right) + (2-\rho)r. \quad (58)\end{aligned}$$

Further define

$$\begin{aligned} q &:= 1-\rho, \\ \gamma &:= 1-q^2 = \rho(2-\rho), \\ s_j &:= \mathbb{E}\left[\left\|\nabla f_{j,\hat{\boldsymbol{\theta}}}(\hat{\theta}_j)\right\|^2\right], \\ \mathbf{s} &:= (s_1,\ldots,s_N)^\top, \\ D_\mu &:= \text{diag}\left(\frac{1}{2\mu_1},\ldots,\frac{1}{2\mu_N}\right), \\ \boldsymbol{\epsilon} &:= (\epsilon_1,\ldots,\epsilon_N)^\top, \\ \mathbb{E}\mathbf{g} &:= (\mathbb{E}g_1,\ldots,\mathbb{E}g_N)^\top, \text{ where } \mathbb{E}g_i := \mathbb{E}g_i\left(\hat{\boldsymbol{\theta}}, \{\hat{\mathcal{D}}_l\}_l, \hat{\mathcal{D}}^{\text{re}}\right). \end{aligned}$$

Then based on (58), we have

$$\mathbb{E}\mathbf{g} \preceq (I+qW)(D_\mu\mathbf{s} + \boldsymbol{\epsilon} + 2\rho\mathfrak{R}) + (1+q)r\mathbf{1}, \quad (59)$$

where $\preceq$ denotes elementwise inequality.

For brevity of mathematical expressions, denote

$$B_i := 2\sum_j a_{ij}\Re_j + \left[\frac{\log(N/\delta)}{2}\left(\sum_j \frac{a_{ij}^2}{|\hat{\mathcal{D}}_j|} + \frac{q^2(1+q)^2}{|\hat{\mathcal{D}}^{\mathrm{re}}|}\right)\right]^{1/2} + q\sum_j w_{ij}\big[2(1+q)A_{ij} + b_j\big] + q^2\sum_{j,k} w_{ij}w_{jk}\big[2A_{ik} + b_k\big],$$

where $a_{ij} := \rho\big(\mathbf{1}\{i=j\} + qw_{ij}\big)$.

Further denote

$$\begin{aligned}
\boldsymbol{R} &:= (\mathbb{E}R_1(\hat{\theta}_1), \ldots, \mathbb{E}R_N(\hat{\theta}_N))^\top,\\
\boldsymbol{\Re} &:= (\Re_1, \ldots, \Re_N)^\top,\\
\boldsymbol{B} &:= (B_1, \ldots, B_N)^\top,\\
B_i &:= \left[\frac{\log(N/\delta)}{2}\left(\sum_j \frac{a_{ij}^2}{|\hat{\mathcal{D}}_j|} + \frac{q^2(1+q)^2}{|\hat{\mathcal{D}}^{\mathrm{re}}|}\right)\right]^{1/2} + q\sum_j w_{ij}\big[2(1+q)A_{ij} + b_j\big] + q^2\sum_{j,k} w_{ij}w_{jk}\big[2A_{ik} + b_k\big].
\end{aligned}$$

Based on Lemma 14, we have $\forall\delta > 0$, with probability $\geq 1-\delta$,

$$\boldsymbol{R} \preceq q^2W^2\boldsymbol{R} + (I+qW)(D_\mu\mathbf{s} + \boldsymbol{\epsilon} + 2\rho\boldsymbol{\Re}) + \boldsymbol{B} + (1+q)(q+r)\mathbf{1}. \tag{60}$$

Since W is row-stochastic, we have $\|q^2W^2\|_\infty = q^2 < 1$. Consequently, $(I-q^2W^2)^{-1}$ exists and is elementwise nonnegative.

To see this, first observe that W^2 is elementwise nonnegative and

$$W^2\mathbf{1} = W(W\mathbf{1}) = W\mathbf{1} = \mathbf{1}.$$

We then have

$$\|q^2W^2\|_\infty = \max_i \sum_j \big|q^2(W^2)_{ij}\big| = q^2\max_i\sum_j (W^2)_{ij} = q^2 < 1.$$

By submultiplicativity of the induced matrix norm,

$$\|(q^2W^2)^{n+1}\|_\infty \leq \|q^2W^2\|_\infty^{n+1} \leq q^{2(n+1)}.$$

Furthermore, for every integer $n \geq 0$, we have

$$(I-q^2W^2)\sum_{t=0}^{n}(q^2W^2)^t = \left(\sum_{t=0}^{n}(q^2W^2)^t\right)(I-q^2W^2) = I - (q^2W^2)^{n+1}.$$

Since $\|(q^2W^2)^{n+1}\|_\infty \leq q^{2(n+1)} \to 0$ when $n\to\infty$, we have the existence of the inverse

$$(I-q^2W^2)^{-1} = \sum_{t=0}^{\infty}(q^2W^2)^t = \sum_{t=0}^{\infty} q^{2t}W^{2t},$$

which is elementwise nonnegative since every summand is elementwise nonnegative.

Therefore, based on (60), we have

$$\begin{aligned}
\boldsymbol{R} \preceq\ & (I-q^2W^2)^{-1}(I+qW)(D_\mu\mathbf{s} + \boldsymbol{\epsilon} + 2\rho\boldsymbol{\Re})\\
&+ (I-q^2W^2)^{-1}\boldsymbol{B}\\
&+ (1+q)(q+r)(I-q^2W^2)^{-1}\mathbf{1}.
\end{aligned} \tag{61}$$

Since

$$I - q^2W^2 = (I+qW)(I-qW),$$

we have

$$(I - q^2W^2)^{-1}(I + qW) = (I - qW)^{-1}. \tag{62}$$

Moreover, $W^2\mathbf{1} = \mathbf{1}$ implies

$$(I - q^2W^2)\mathbf{1} = (1 - q^2)\mathbf{1} = \gamma\mathbf{1}.$$

Therefore, we have $(I - q^2W^2)^{-1}\mathbf{1} = \frac{1}{\gamma}\mathbf{1}$.

Substituting these identities into (61), we have

$$\boldsymbol{R} \preceq (I - qW)^{-1}(D_\mu \mathbf{s} + \boldsymbol{\epsilon} + 2\rho\mathfrak{R}) + (I - q^2W^2)^{-1}\boldsymbol{B} + \frac{q + r}{\rho}\mathbf{1}. \tag{63}$$

The output selection and Theorem 1 give

$$\frac{1}{N}\sum_{j=1}^{N} s_j \leq \frac{1}{TN}\sum_{t=0}^{T-1}\sum_{j=1}^{N}\mathbb{E}\left\|\nabla f_{j,\boldsymbol{\theta}^t}(\theta_j^t)\right\|^2 \leq \Phi_{\boldsymbol{K}}(Q). \tag{64}$$

Therefore, for every i,

$$\sum_j \frac{\left[(I - (1-\rho)W)^{-1}\right]_{ij}}{2\mu_j} s_j \leq \frac{N\Phi_{\boldsymbol{K}}(Q)}{2}\max_j \frac{\left[(I - (1-\rho)W)^{-1}\right]_{ij}}{\mu_j}. \tag{65}$$

We then have

$$\boldsymbol{R} \preceq (I - qW)^{-1}(\boldsymbol{\epsilon} + 2\rho\mathfrak{R}) + (I - q^2W^2)^{-1}\boldsymbol{B} + \frac{q + r}{\rho}\mathbf{1} + \frac{N\Phi_{\boldsymbol{K}}(Q)}{2}\boldsymbol{\lambda},$$

where $\boldsymbol{\lambda} := (\lambda_i)_{i=1}^N$ and $\lambda_i := \max_j \frac{\left[(I-(1-\rho)W)^{-1}\right]_{ij}}{\mu_j}$. □

C.8 Average Personal Risk Bound and Derivation for (6)

Reuse the notations q and γ in Appendix C.7.1. Define the normalization vectors

$$\begin{aligned}
\mathbf{u} &:= \frac{\mathbf{1}}{N}, \\
\boldsymbol{\nu}^\top &:= \rho\,\mathbf{u}^\top(I - qW)^{-1}, \\
\boldsymbol{\pi}^\top &:= \gamma\,\mathbf{u}^\top(I - q^2W^2)^{-1}.
\end{aligned}$$

Note that

$$\begin{aligned}
(I - qW)^{-1}\mathbf{1} &= \frac{1}{\rho}\mathbf{1}, \\
(I - q^2W^2)^{-1}\mathbf{1} &= \frac{1}{\gamma}\mathbf{1}.
\end{aligned}$$

We then have $\nu_j, \pi_j \geq 0$ and $\sum_j \nu_j = \sum_j \pi_j = 1$.

Multiplying (63) by $\mathbf{u}^\top$ gives

$$\frac{1}{N}\sum_i \mathbb{E}R_i(\hat{\theta}_i) \leq \frac{q + r}{\rho} + \frac{1}{2\rho}\sum_j \frac{\nu_j}{\mu_j}s_j + \frac{1}{\rho}\sum_j \nu_j\epsilon_j + \sum_j 2\nu_j\mathfrak{R}_j + \frac{1}{\gamma}\sum_i \pi_i B_i. \tag{66}$$

Therefore, based on (64), we have

$$\frac{1}{2\rho}\sum_j \frac{\nu_j}{\mu_j}s_j \leq \frac{1}{2\rho}\left(\max_j \frac{\nu_j}{\mu_j}\right)\sum_j s_j \leq \frac{N\Phi_{\boldsymbol{K}}(Q)}{2\rho}\max_j \frac{\nu_j}{\mu_j}. \tag{67}$$

Define the network-and-PL coefficient

$$\kappa_{\rho,\boldsymbol{\mu}}(W) := N\max_j \frac{\nu_j}{\mu_j}. \tag{68}$$

Combining (66) and (67), and substituting $q = 1 - \rho$ and $\gamma = \rho(2 - \rho)$, gives the **average personal risk bound**

$$\frac{1}{N}\sum_i \mathbb{E}R_i(\hat{\theta}_i) \le \frac{1-\rho+r}{\rho} + \frac{\Phi_{\boldsymbol{K}}(Q)}{2\rho}\kappa_{\rho,\boldsymbol{\mu}}(W) + \frac{\sum_j \nu_j\epsilon_j + 2\rho\Re_j}{\rho} + \frac{1}{\rho(2-\rho)}\sum_i \pi_i B_i. \tag{69}$$

However, the averaged bound in (69) is not directly usable as a numerical optimization objective because it depends on generally unavailable quantities, including Lipschitz constants in the optimization-error terms, PL constants, and minimum loss values. We therefore derive a surrogate that makes selected terms independent of W while preserving weight-dependent contributions to guide ensemble-weight selection.

Note that W enters $\Phi_{\boldsymbol{K}}(Q)$ only through $d_j := \sum_i \frac{(1-\rho_i)w_{ij}}{K_i C_{\eta_i}} \le \sum_i \frac{(1-\rho_i)}{K_i C_{\eta_i}}$. Replacing $d_j, \forall j$ in $\Phi_{\boldsymbol{K}}(Q)$ by $\sum_i \frac{(1-\rho_i)}{K_i C_{\eta_i}}$ and denote the resulting quantity $\tilde{\Phi}$. Using $\nu_j \le 1$, we have (6)

$$\frac{1}{N}\sum_i \mathbb{E}R_i(\hat{\theta}_i) \le \frac{(1-\rho)+r}{\rho} + \frac{N\tilde{\Phi}}{2\rho}\max_j \frac{1}{\mu_j} + \frac{1}{\rho}\sum_j \epsilon_j + \sum_j 2\Re_j + \frac{1}{\gamma}\sum_i \pi_i B_i.$$

Now the only remaining W-dependent term is $\sum_{i=1}^N \pi_i B_i$, which is easier to handle.

C.9 Approximation and Estimation Errors of Label-marginal Discrepancies

We bound two sources of error and consider their effect on the generalization bound surrogate: approximating input-distribution TV by label-marginal TV, and estimating label-marginal TV from empirical class proportions.

Population approximation error. Recall

$$A_{ij} = \mathrm{TV}\left(\mathcal{P}_X^{(i)}, \mathcal{P}_X^{(j)}\right), \qquad b_i = \mathrm{TV}\left(\mathcal{P}_X^{(i)}, \mathcal{P}_X^{\mathrm{re}}\right).$$

Their population label-marginal proxies are

$$A_{ij}^Y := \mathrm{TV}\left(\mathcal{P}_Y^{(i)}, \mathcal{P}_Y^{(j)}\right), \qquad b_i^Y := \mathrm{TV}\left(\mathcal{P}_Y^{(i)}, \mathcal{P}_Y^{\mathrm{re}}\right). \tag{70}$$

Write $p_i(y) := \mathcal{P}_Y^{(i)}(y)$ and $p_{\mathrm{re}}(y) := \mathcal{P}_Y^{\mathrm{re}}(y)$, and define

$$\Delta_{ij} := \sum_{y\in\mathcal{Y}} \min\{p_i(y), p_j(y)\}\mathrm{TV}\left(\mathcal{P}_{X|Y=y}^{(i)}, \mathcal{P}_{X|Y=y}^{(j)}\right), \tag{71}$$

$$\Delta_{i,\mathrm{re}} := \sum_{y\in\mathcal{Y}} \min\{p_i(y), p_{\mathrm{re}}(y)\}\mathrm{TV}\left(\mathcal{P}_{X|Y=y}^{(i)}, \mathcal{P}_{X|Y=y}^{\mathrm{re}}\right). \tag{72}$$

Lemma 16 (Population approximation error)**.** *Under the common conditional label distribution assumed above, for every pair of nodes i, j,*

$$0 \le A_{ij} - A_{ij}^Y \le \Delta_{ij}, \qquad 0 \le b_i - b_i^Y \le \Delta_{i,\mathrm{re}}. \tag{73}$$

Proof. Let $p_X^{(i)}$ and $p_X^{(j)}$ denote the densities of $\mathcal{P}_X^{(i)}$ and $\mathcal{P}_X^{(j)}$. We then have

$$p_i(y) - p_j(y) = \int_{\mathcal{X}} P_{Y|X}(y \mid x)\left(p_X^{(i)}(x) - p_X^{(j)}(x)\right)\,dx.$$

Consequently,

$$\begin{aligned} A_{ij}^Y &= \frac{1}{2}\sum_{y\in\mathcal{Y}}\left|\int_{\mathcal{X}} P_{Y|X}(y\mid x)\left(p_X^{(i)}(x)-p_X^{(j)}(x)\right)\,dx\right| \\ &\le \frac{1}{2}\sum_{y\in\mathcal{Y}}\int_{\mathcal{X}} P_{Y|X}(y\mid x)\left|p_X^{(i)}(x)-p_X^{(j)}(x)\right|\,dx \\ &= \frac{1}{2}\int_{\mathcal{X}}\left(\sum_{y\in\mathcal{Y}} P_{Y|X}(y\mid x)\right)\left|p_X^{(i)}(x)-p_X^{(j)}(x)\right|\,dx \\ &= \frac{1}{2}\int_{\mathcal{X}}\left|p_X^{(i)}(x)-p_X^{(j)}(x)\right|\,dx = A_{ij}. \end{aligned} \tag{74}$$

Here we have used the integral triangle inequality, linearity of integration for the finite sum over labels, and $\sum_y P_{Y|X}(y\mid x)=1$.

To prove the upper bound, denote

$$\begin{aligned} m_y &:= \min\{p_i(y), p_j(y)\}, \\ Q_i^y &:= \mathcal{P}_{X|Y=y}^{(i)}, \\ Q_j^y &:= \mathcal{P}_{X|Y=y}^{(j)}. \end{aligned}$$

For any measurable set $E\subseteq\mathcal{X}$,

$$\begin{aligned} \mathcal{P}_X^{(i)}(E)-\mathcal{P}_X^{(j)}(E) = &\sum_y m_y\big(Q_i^y(E)-Q_j^y(E)\big) \\ &+\sum_y\big(p_i(y)-m_y\big)Q_i^y(E)-\sum_y\big(p_j(y)-m_y\big)Q_j^y(E). \end{aligned} \tag{75}$$

For the first term, we have

$$\left|\sum_y m_y Q_i^y(E)-Q_j^y(E)\right| \le \sum_y m_y\left|Q_i^y(E)-Q_j^y(E)\right| \le \Delta_{ij}.$$

Moreover,

$$\begin{aligned} \sum_{y\in\mathcal{Y}} m_y &= \sum_{y\in\mathcal{Y}}\min\{p_i(y),p_j(y)\} \\ &= \frac{1}{2}\sum_{y\in\mathcal{Y}}\left(p_i(y)+p_j(y)-|p_i(y)-p_j(y)|\right) \\ &= \frac{1}{2}\left(\sum_y p_i(y)+\sum_y p_j(y)-\sum_y|p_i(y)-p_j(y)|\right) \\ &= 1-\frac{1}{2}\sum_y|p_i(y)-p_j(y)| \\ &= 1-A_{ij}^Y. \end{aligned}$$

So we have

$$\sum_y\big(p_i(y)-m_y\big)=\sum_y\big(p_j(y)-m_y\big)=A_{ij}^Y.$$

Since $0\le Q_i^y(E),Q_j^y(E)\le 1$, each of the last two terms in (75) belongs to $[0,A_{ij}^Y]$. Their difference therefore has absolute value at most A_{ij}^Y. Hence

$$\left|\mathcal{P}_X^{(i)}(E)-\mathcal{P}_X^{(j)}(E)\right|\le\Delta_{ij}+A_{ij}^Y.$$

Taking the supremum over measurable E yields $A_{ij} \le A_{ij}^Y + \Delta_{ij}$. Together with (74), this proves the first claim.

The reference-distribution claim follows by the same argument, replacing $\mathcal{P}^{(j)}$ with $\mathcal{P}^{\mathrm{re}}$. □

The approximation error is thus small when the conditional input-distribution discrepancies vanish.

Effect of population approximation on $\sum_{i=1}^N \pi_i B_i$. Recall $\tilde{\boldsymbol{B}} := (\tilde{B}_i)_{i=1}^N$ is obtained from $\boldsymbol{B} := (B_i)_{i=1}^N$ in (4) by replacing A_{ij} with A_{ij}^Y and b_i with b_i^Y, while leaving all other terms unchanged. Denoting $\Delta_A := \max_{i,j} \Delta_{ij}$ and $\Delta_B := \max_i \Delta_{i,\mathrm{re}}$, for every row-stochastic W, we have

$$\begin{aligned} 0 \le B_i - \tilde{B}_i &\le (1-\rho) \sum_j w_{ij} \big[2(2-\rho)\Delta_A + \Delta_B\big] \\ &\quad + (1-\rho)^2 \sum_{j,k} w_{ij} w_{jk} \big[2\Delta_A + \Delta_B\big] \\ &= 2(1-\rho)(3-2\rho)\Delta_A + (1-\rho)(2-\rho)\Delta_B. \end{aligned} \tag{76}$$

The equality uses $\sum_j w_{ij} = 1$ and $\sum_{j,k} w_{ij} w_{jk} = 1$.

As shown in Appendix C.7.1, for $0 < \rho \le 1$, the inverse $\big(I - (1-\rho)^2 W^2\big)^{-1}$ is elementwise nonnegative and satisfies

$$\big(I - (1-\rho)^2 W^2\big)^{-1} \mathbf{1} = \frac{1}{\rho(2-\rho)} \mathbf{1}.$$

Multiplying (76) by this inverse gives

$$\mathbf{0} \preceq \big(I - (1-\rho)^2 W^2\big)^{-1} \big(\boldsymbol{B} - \tilde{\boldsymbol{B}}\big) \preceq \epsilon_{\mathrm{app}} \mathbf{1}, \tag{77}$$

where

$$\epsilon_{\mathrm{app}} := \frac{2(1-\rho)(3-2\rho)}{\rho(2-\rho)} \Delta_A + \frac{1-\rho}{\rho} \Delta_B. \tag{78}$$

Averaging (77) also yields

$$0 \le \frac{\mathbf{1}^\top}{N} \big(I - (1-\rho)^2 W^2\big)^{-1} \big(\boldsymbol{B} - \tilde{\boldsymbol{B}}\big) \le \epsilon_{\mathrm{app}}. \tag{79}$$

Replacing these label-marginal discrepancies with empirical class proportions introduces an additional finite-sample estimation error.

Finite-sample estimation error. The empirical class proportions are

$$\begin{aligned} \widehat{p}_i(y) &:= \frac{1}{|\hat{\mathcal{D}}_i|} \sum_{(x,l) \in \hat{\mathcal{D}}_i} \mathbf{1}\{l = y\}, \\ \widehat{p}_{\mathrm{re}}(y) &:= \frac{1}{|\hat{\mathcal{D}}^{\mathrm{re}}|} \sum_{a=1}^{|\hat{\mathcal{D}}^{\mathrm{re}}|} \mathbf{1}\{Y_a^{\mathrm{re}} = y\}. \end{aligned} \tag{80}$$

Here Y_a^{re} denotes the label associated with the a-th reference input. These labels are used only to define the aggregate counts, which need not be disclosed individually or used as supervision during distillation. The implemented discrepancies are

$$\widehat{A}_{ij}^Y := \frac{1}{2} \sum_y |\widehat{p}_i(y) - \widehat{p}_j(y)|, \qquad \widehat{b}_i^Y := \frac{1}{2} \sum_y |\widehat{p}_i(y) - \widehat{p}_{\mathrm{re}}(y)|. \tag{81}$$

For $0 < \delta_{\mathrm{est}} < 1$, define

$$\beta_i := \min\left\{1, \sqrt{\frac{\log\big(2(N+1)(2^M-2)/\delta_{\mathrm{est}}\big)}{2|\hat{\mathcal{D}}_i|}}\right\}, \tag{82}$$

$$\beta_{\mathrm{re}} := \min\left\{1, \sqrt{\frac{\log\big(2(N+1)(2^M-2)/\delta_{\mathrm{est}}\big)}{2|\hat{\mathcal{D}}^{\mathrm{re}}|}}\right\}. \tag{83}$$

Lemma 17. *With probability at least* $1-\delta_{\mathrm{est}}$*, simultaneously for all* i, j*,*

$$|\widehat{A}^Y_{ij} - A^Y_{ij}| \le \min\{1, \beta_i + \beta_j\}, \tag{84}$$

$$|\widehat{b}^Y_i - b^Y_i| \le \min\{1, \beta_i + \beta_{\mathrm{re}}\}. \tag{85}$$

Proof. Consider any one of the $N+1$ label distributions p and its empirical distribution $\widehat{p}$ based on n labels. For any $S \subset \mathcal{Y}$, Hoeffding's inequality for the independent Bernoulli variables $\mathbf{1}\{Y^{\mathrm{re}}_a \in S\}$ and $\mathbf{1}\{l \in S\}$ gives, for $t > 0$,

$$\Pr(|\widehat{p}(S) - p(S)| \ge t) \le 2e^{-2nt^2}.$$

The empty and full sets have zero discrepancy. A union bound over the other $2^M - 2$ subsets gives $\forall t > 0$

$$\Pr(\mathrm{TV}(\widehat{p}, p) \ge t) \le (2^M - 2)2e^{-2nt^2}. \tag{86}$$

Taking $t = \sqrt{\log(2(N+1)(2^M-2)/\delta_{\mathrm{est}})/(2n)}$ makes the right-hand side $\delta_{\mathrm{est}}/(N+1)$. If this threshold is at least one, use the deterministic uniform $\mathrm{TV}(\widehat{p}, p) \le 1$ instead. A union bound over the N local distributions and the reference distribution therefore gives the event

$$\cap_{i=1}^N \{\mathrm{TV}(\widehat{p}_i, p_i) \le \beta_i\} \cap \{\mathrm{TV}(\widehat{p}_{\mathrm{re}}, p_{\mathrm{re}}) \le \beta_{\mathrm{re}}\}, \tag{87}$$

with probability at least $1 - \delta_{\mathrm{est}}$.

Since

$$\begin{aligned}\widehat{A}^Y_{ij} &= \mathrm{TV}(\widehat{p}_i, \widehat{p}_j) \le \mathrm{TV}(\widehat{p}_i, p_i) + \mathrm{TV}(p_i, p_j) + \mathrm{TV}(p_j, \widehat{p}_j) \\ &= \mathrm{TV}(\widehat{p}_i, p_i) + A^Y_{ij} + \mathrm{TV}(\widehat{p}_j, p_j),\end{aligned}$$

and

$$\begin{aligned}A^Y_{ij} &= \mathrm{TV}(p_i, p_j) \le \mathrm{TV}(p_i, \widehat{p}_i) + \mathrm{TV}(\widehat{p}_i, \widehat{p}_j) + \mathrm{TV}(\widehat{p}_j, p_j) \\ &= \mathrm{TV}(\widehat{p}_i, p_i) + \widehat{A}^Y_{ij} + \mathrm{TV}(\widehat{p}_j, p_j).\end{aligned}$$

we have $\left|\widehat{A}^Y_{ij} - A^Y_{ij}\right| \le \mathrm{TV}(\widehat{p}_i, p_i) + \mathrm{TV}(\widehat{p}_j, p_j)$ and similarly, $|\widehat{b}^Y_i - b^Y_i| \le \mathrm{TV}(\widehat{p}_i, p_i) + \mathrm{TV}(\widehat{p}_{\mathrm{re}}, p_{\mathrm{re}})$.

On the event shown in (87), we have

$$\begin{aligned}|\widehat{A}^Y_{ij} - A^Y_{ij}| &\le \mathrm{TV}(\widehat{p}_i, p_i) + \mathrm{TV}(\widehat{p}_j, p_j) \le \beta_i + \beta_j, \\ |\widehat{b}^Y_i - b^Y_i| &\le \mathrm{TV}(\widehat{p}_i, p_i) + \mathrm{TV}(\widehat{p}_{\mathrm{re}}, p_{\mathrm{re}}) \le \beta_i + \beta_{\mathrm{re}}.\end{aligned}$$

The differences are also bounded by one, which proves the lemma. □

Effect of estimation error on $\sum_{i=1}^N \pi_i B_i$. Let $\widehat{\boldsymbol{B}} := (\widehat{B}_i)_{i=1}^N$ be obtained from $\tilde{\boldsymbol{B}}$ by replacing A^Y_{ij} and b^Y_i with $\widehat{A}^Y_{ij}$ and $\widehat{b}^Y_i$, respectively. All other terms are unchanged. Denote

$$\hat{\Delta}_A := \min\{1, 2\max_i \beta_i\}, \qquad \hat{\Delta}_b := \min\{1, \max_i \beta_i + \beta_{\mathrm{re}}\}. \tag{88}$$

On the event defined in (87), $|\widehat{A}^Y_{ij} - A^Y_{ij}| \le \hat{\Delta}_A$ and $|\widehat{b}^Y_i - b^Y_i| \le \hat{\Delta}_b$ for every i, j. For every row-stochastic W, this gives

$$\begin{aligned}|\tilde{B}_i - \widehat{B}_i| &\le (1-\rho)\sum_j w_{ij}\left[2(2-\rho)\hat{\Delta}_A + \hat{\Delta}_b\right] \\ &\quad + (1-\rho)^2 \sum_{j,k} w_{ij}w_{jk}\left[2\hat{\Delta}_A + \hat{\Delta}_b\right] \\ &= 2(1-\rho)(3-2\rho)\hat{\Delta}_A + (1-\rho)(2-\rho)\hat{\Delta}_b.\end{aligned} \tag{89}$$

Unlike population approximation, the estimation error can have either sign. Using nonnegativity of the inverse and its row sums, we have

$$\left\|\left(I - (1-\rho)^2 W^2\right)^{-1}\left(\tilde{\boldsymbol{B}} - \widehat{\boldsymbol{B}}\right)\right\|_\infty \le \epsilon_{\mathrm{est}}, \tag{90}$$

where the norm on the left is the vector maximum norm and

$$\epsilon_{\mathrm{est}} := \frac{2(1-\rho)(3-2\rho)}{\rho(2-\rho)}\hat{\Delta}_A + \frac{1-\rho}{\rho}\hat{\Delta}_b. \tag{91}$$

Combined effect on $\sum_{i=1}^N \pi_i B_i$. Decompose

$$\boldsymbol{B} - \widehat{\boldsymbol{B}} = (\boldsymbol{B} - \tilde{\boldsymbol{B}}) + (\tilde{\boldsymbol{B}} - \widehat{\boldsymbol{B}}).$$

Combining (77) and (90) yields, on the event defined in (87),

$$-\epsilon_{\text{est}}\mathbf{1} \preceq \big(I - (1-\rho)^2 W^2\big)^{-1}\big(\boldsymbol{B} - \widehat{\boldsymbol{B}}\big) \leq (\epsilon_{\text{app}} + \epsilon_{\text{est}})\mathbf{1}. \tag{92}$$

In particular, the averaged comparison is

$$-\epsilon_{\text{est}} \leq \frac{\mathbf{1}^\top}{N}\big(I - (1-\rho)^2 W^2\big)^{-1}\big(\boldsymbol{B} - \widehat{\boldsymbol{B}}\big) \leq \epsilon_{\text{app}} + \epsilon_{\text{est}}. \tag{93}$$

The total error is

$$\epsilon_{\text{app}} + \epsilon_{\text{est}} = \frac{2(1-\rho)(3-2\rho)}{\rho(2-\rho)}(\Delta_A + \hat{\Delta}_A) + \frac{1-\rho}{\rho}(\Delta_B + \hat{\Delta}_b), \tag{94}$$

which can be small when the conditional input-distribution discrepancies are small or the dataset sizes are large.

Recall the normalization vector

$$\boldsymbol{\pi}^\top := \rho(2-\rho)\frac{\mathbf{1}^\top}{N}\big(I - (1-\rho)^2 W^2\big)^{-1}.$$

Therefore, the corresponding comparison for the weight-selection objective in P_1 is given by

$$-\rho(2-\rho)\epsilon_{\text{est}} \leq \sum_i \pi_i B_i - \sum_i \pi_i \widehat{B}_i \leq \rho(2-\rho)(\epsilon_{\text{app}} + \epsilon_{\text{est}}). \tag{95}$$

Generalization guarantee using $\widehat{\boldsymbol{B}}$. Consider the intersect of the event of Theorem 5 and that in (87). Let $0 < \delta, \delta_{\text{est}} < 1$ with $\delta + \delta_{\text{est}} < 1$. Based on (5) and (92), we have with probability at least $1 - \delta - \delta_{\text{est}}$ over the datasets $\{\hat{\mathcal{D}}_l\}_{l=1}^N$ and $\hat{\mathcal{D}}^{\text{re}}$,

$$\begin{aligned}\boldsymbol{R} \preceq\ & \big(I - (1-\rho)W\big)^{-1}(D_\mu \mathbf{s} + \boldsymbol{\epsilon}) \\ & + \big(I - (1-\rho)^2 W^2\big)^{-1}\widehat{\boldsymbol{B}} \\ & + \left(\frac{1-\rho+r}{\rho} + \epsilon_{\text{app}} + \epsilon_{\text{est}}\right)\mathbf{1}.\end{aligned} \tag{96}$$

D Additional experimental details and results

D.1 Data Partition, Metric, and Training Protocol

All datasets have ten classes. Denote the number of dominant classes by d. We consider five groups $g \in \{0, \ldots, 4\}$, and define the dominant class set for group g: $D_g = \{2g, 2g+1, \ldots, 2g+d-1\} \bmod 10$. Each node belongs to one of these five groups, with four nodes per group. For $d < 10$, the prescribed probability of class c is

$$\pi_{g,c} = \begin{cases} 6/(5d+10), & c \in D_g, \\ 1/(5d+10), & c \notin D_g. \end{cases} \tag{97}$$

For $d = 10$, all class probabilities are $1/10$. MNIST and FashionMNIST each contain 60,000 official training images and 10,000 test images; CIFAR-10 contains 50,000 training images and 10,000 test images. We first construct a candidate pool for each node from the official training split according to $\pi_{g,c}$, where the number of points of class c is approximately the pool size times $\pi_{g,c}$. For MNIST or FashionMNIST, each node has 3,000 candidates. For CIFAR-10, the candidate pool contains 1,000 images at each of the 15 smaller nodes and 3,000 at each of the five larger nodes. From its pool, a smaller node selects 500 private images for its training set $\hat{\mathcal{D}}_i$ and a larger node selects 2,500 for its $\hat{\mathcal{D}}_j$, using integer class quotas based on $\pi_{g,c}$. Each node then samples 500 distinct images from

Table 2: Training settings. Within a dataset, all ensemble weight design methods use the same model training settings.

Setting	MNIST / FashionMNIST	CIFAR-10
Model	LeNet, 44,426 parameters	CNN, 1,116,970 parameters
Optimizer	AdamW	SGD, Nesterov momentum 0.9
Learning rate	5×10^{-4}	10^{-2}
Weight decay	10^{-4}	5×10^{-4}
Private/reference batch	256 / 128	128 / 128
Local updates per round	1	1
Rounds	15,000 / 25,000	15,000
KD loss function	Cross entropy	$\frac{1}{2}\|p-q\|_2^2$

the unused portion of its pool for the shared 10,000-point reference set. Any candidates left over are unused. The official test split is reserved for evaluation.

As discussed in Appendix 3.2, GBOW uses the empirical private class proportions $\widehat{p}_{i,c}$, and the empirical reference class proportions $\widehat{p}_{\mathrm{re},c}$, where $\widehat{p}_{i,c}$ is the number of private points of class c at node i divided by $|\hat{\mathcal{D}}_i|$, and $\widehat{p}_{\mathrm{re},c}$ is the number of reference points of class c divided by 10,000.

Let $A_{ic}(t)$ be its accuracy on class c of the test set at round t. The reported system score is

$$A(t) = \frac{100}{20}\sum_{i=1}^{20}\sum_{c=1}^{10}\widehat{p}_{i,c}A_{ic}(t). \tag{98}$$

Each node i is evaluated on the full test set and its personal test accuracy is also weighted by $\sum_{c=1}^{10}\widehat{p}_{i,c}A_{ic}(t)$. A run's score is the mean of its final 1,000 saved values of $A(t)$, i.e., rounds 14,001-15,000 or 24,001-25,000. All uncertainty estimates use independent seeds as the unit of replication.

The used LeNet has two 5×5 convolutions with 6 and 16 output channels, ReLU and 2×2 average pooling, then fully connected widths 120, 84, and 10. Dropout with probability 0.5 is enabled in its hidden classifier layers during private training. The CIFAR-10 CNN has four 3×3 convolutions with widths 32, 32, 64, 64, ReLU activations, max pooling after each pair, and fully connected widths 256 and 10, with dropout 0.3. The training setting details are summarised in Table 2.

D.2 Ensemble Weight Design Methods

Define $\mathcal{S}_i = \{i\}\cup\mathcal{N}_i$ for node i's feasible node sources, including itself, and $h_j(x)$ for node j's probability predictions. In all design methods, all nodes use the same tradeoff factor ρ.

GBOW (ours). As discussed in Appendix 3.2, GBOW minimizes $\min_W \sum_{i=1}^N \pi_i\hat{B}_i$ under constraints (7), using empirical class proportions and training set sizes, and the reference class marginal. In terms of the initial point, we extract some terms from $\hat{B}_i$ and solve the resulting problem for each node i:

$$\min_{\{w_{ij}\}_j}\left[\frac{\log(N/\delta)}{2}\left(\frac{\rho^2\big(1+(1-\rho)w_{ii}\big)^2}{|\hat{\mathcal{D}}_i|}+\rho^2(1-\rho)^2\sum_{j\neq i}\frac{w_{ij}^2}{|\hat{\mathcal{D}}_j|}+\frac{(1-\rho)^2(2-\rho)^2}{|\hat{\mathcal{D}}^{\mathrm{re}}|}\right)\right]^{1/2}$$
$$+(1-\rho)\sum_j w_{ij}\big[2(2-\rho)\mathrm{TV}(\widehat{p}_i,\widehat{p}_j)+\mathrm{TV}(\widehat{p}_j,\widehat{p}_{\mathrm{re}})\big]$$
$$\text{s.t.}\quad w_{ij}\geq0,\forall j,\quad \sum_j w_{ij}=1,\ \text{and } w_{ij}=0 \text{ for disallowed edges.}$$

This problem is convex in $\{w_{ij}\}_j$ for any i, which is solved by the solver MOSEK called with its default settings. Then at most 500 projected-gradient iterations are used for $\hat{P}_1$, which has the

full objective $\min_W \sum_{i=1}^N \pi_i \hat{B}_i$ under constraints in (7). The initial step is 0.1, with maximum 1, minimum 10^{-12}, backtracking factor 0.5, Armijo constant 10^{-4}, iterate tolerance 10^{-8}, objective tolerance 10^{-10}, and $\delta = 0.01$.

UNIFO and TSIZE. The equal-weight ensemble rule (Park et al., 2023) uses $w_{ij} = 1/|\mathcal{S}_i|$. The sample-size rule (Park et al., 2024) uses $w_{ij} = |\hat{\mathcal{D}}_j| / \sum_{k\in\mathcal{S}_i} |\hat{\mathcal{D}}_k|$. The original methods aggregate predictions from all nodes through a server. We adapt their weighting rules to partially connected graphs by restricting the sources to $\mathcal{S}_i$ and normalizing over that set.

ENTRO and VARIA. The entropy-confidence rule (Deng et al., 2023) is

$$w_{ij}(x) = \frac{\exp[-H(h_j(x))]}{\sum_{k\in\mathcal{S}_i} \exp[-H(h_k(x))]}, \text{ where } H(h_j(x)) = -\sum_c [h_j(x)]_c \log[h_j(x)]_c. \tag{99}$$

Following the variance-based confidence weighting of Cho et al. (2022), we use

$$w_{ij}(x) = \frac{V(h_j(x))}{\sum_{k\in\mathcal{S}_i} V(h_k(x))}, \qquad V(h_j(x)) = \frac{1}{C}\sum_{c=1}^{C} \left([h_j(x)]_c - \frac{1}{C}\right)^2, \quad C = 10, \tag{100}$$

If all variances are zero, we assign uniform weights over $\mathcal{S}_i$.

CORRE. We adapt the prediction-vector (CG-logit) component of Shi et al. (2023). For each reference point x, let $r_{ij}(x)$ be the Pearson correlation across classes between node i's probability vector and that of node j. The weight is given by

$$w_{ij}(x) = \frac{1}{|\mathcal{S}_i| - 1}\left[1 - \frac{\exp(r_{ij}(x))}{\sum_{k\in\mathcal{S}_i} \exp(r_{ik}(x))}\right]. \tag{101}$$

The original paper also uses distillation among intermediate features. Our version omits the intermediate-feature branch. Thus CORRE is an adaptation of its weight design rule, not an exact reproduction of the full CG-MTKD system.

COMET. Following the clustered knowledge-transfer idea of Cho et al. (2023), a coordinator clusters flattened node-prediction matrices by K-means, and each node uses its nearest centroid as a frozen teacher for the current round. The implemented number of clusters is three. We use all nodes to provide predictions in each round.

SELEC. This benchmark adapts the selective-sharing idea of Shao et al. (2024). A KuLSIF density-ratio model is fitted using 5% of each private set, with the remaining examples used for calibration. It uses 250 auxiliary samples, a Gaussian kernel function with width five, regularization $250^{-1/2}$, and a node acceptance threshold at the 0.25 calibration quantile. Accepted node predictions are mixed with normalized uniform source weights; points with no accepted source are excluded from distillation. The implemented server score is $2(1-\max_c[a]_c)$ with acceptance threshold two, where a denotes a combined probability prediction after node filtering. SELEC requires a coordinator in this experiment and is only evaluated on the complete graph.

D.3 Full Complete-Graph Comparisons

In this subsection, we provide the detailed results under the complete graph and homogeneous model architecture for three datasets.

Figure 1 presents additional results not shown in Table 1, covering all methods from the same set of experiments.

Tables 3-5 give every method's best mean with a 95% interval. The corresponding ρ values are shown immediately below each table. Bold entries identify the highest mean within each partition.

Figures 2-4 show the observed accuracy curves for all five partitions, together with pointwise 95% confidence intervals. In each figure, all methods share the same $\rho = 1$ endpoint, which represents local-only training. As shown in these figures, GBOW obtains the highest accuracy and remains

Table 3: Homogeneous FashionMNIST: best observed averaged test accuracy (%), followed by selected ρ.

Method	$d = 2$	$d = 4$	$d = 6$	$d = 8$	$d = 10$
GBOW (ours)	**86.13 ± 0.08**	**86.37 ± 0.09**	**85.39 ± 0.08**	**83.47 ± 0.10**	**82.44 ± 0.10**
UNIFO	84.53 ± 0.08	84.29 ± 0.14	82.97 ± 0.06	81.59 ± 0.13	81.59 ± 0.10
ENTRO	84.57 ± 0.09	84.33 ± 0.14	83.03 ± 0.05	81.68 ± 0.11	81.68 ± 0.11
VARIA	84.60 ± 0.09	84.32 ± 0.16	82.99 ± 0.06	81.70 ± 0.09	81.71 ± 0.11
COMET	85.10 ± 0.08	85.08 ± 0.11	83.75 ± 0.06	81.99 ± 0.11	81.58 ± 0.10
TSIZE	84.61 ± 0.08	84.35 ± 0.14	83.07 ± 0.08	81.80 ± 0.10	82.16 ± 0.09
SELEC	84.53 ± 0.10	84.28 ± 0.08	82.95 ± 0.07	81.52 ± 0.14	81.56 ± 0.13
CORRE	84.50 ± 0.07	84.30 ± 0.15	82.95 ± 0.07	81.59 ± 0.13	81.59 ± 0.09

Method	$d = 2$	$d = 4$	$d = 6$	$d = 8$	$d = 10$
GBOW (ours)	0.4	0.3	0.2	0.3	0.3
UNIFO	0.95	0.95	0.95	0.85	0.4
ENTRO	0.95	0.95	0.95	0.9	0.5
VARIA	0.95	0.95	0.95	0.9	0.5
COMET	0.85	0.8	0.8	0.8	0.4
TSIZE	0.95	0.95	0.95	0.7	0.3
SELEC	0.95	0.95	0.95	0.85	0.4
CORRE	0.95	0.95	0.95	0.85	0.5

Table 4: MNIST: best observed averaged test accuracy (%), followed by selected ρ.

Method	$d = 2$	$d = 4$	$d = 6$	$d = 8$	$d = 10$
GBOW (ours)	**98.28 ± 0.03**	**98.25 ± 0.02**	**98.19 ± 0.03**	98.00 ± 0.02	**98.05 ± 0.06**
UNIFO	98.11 ± 0.03	98.07 ± 0.03	97.99 ± 0.03	97.95 ± 0.02	97.90 ± 0.06
ENTRO	98.17 ± 0.03	98.12 ± 0.03	98.02 ± 0.03	97.94 ± 0.02	97.87 ± 0.05
VARIA	98.15 ± 0.03	98.12 ± 0.03	98.00 ± 0.03	97.90 ± 0.02	97.83 ± 0.05
COMET	98.19 ± 0.04	98.20 ± 0.02	98.08 ± 0.03	97.95 ± 0.02	97.90 ± 0.05
TSIZE	98.18 ± 0.03	98.15 ± 0.03	98.08 ± 0.03	**98.05 ± 0.02**	98.02 ± 0.04
SELEC	98.06 ± 0.03	98.02 ± 0.03	97.93 ± 0.03	97.89 ± 0.02	97.87 ± 0.05
CORRE	98.12 ± 0.04	98.07 ± 0.03	97.98 ± 0.03	97.95 ± 0.02	97.90 ± 0.05

Method	$d = 2$	$d = 4$	$d = 6$	$d = 8$	$d = 10$
GBOW (ours)	0.3	0.3	0.2	0.3	0.2
UNIFO	0.8	0.7	0.6	0.3	0.2
ENTRO	0.7	0.7	0.6	0.3	0.2
VARIA	0.7	0.7	0.5	0.3	0.3
COMET	0.6	0.6	0.5	0.3	0.2
TSIZE	0.7	0.8	0.6	0.3	0.2
SELEC	0.7	0.8	0.5	0.3	0.2
CORRE	0.8	0.8	0.6	0.2	0.2

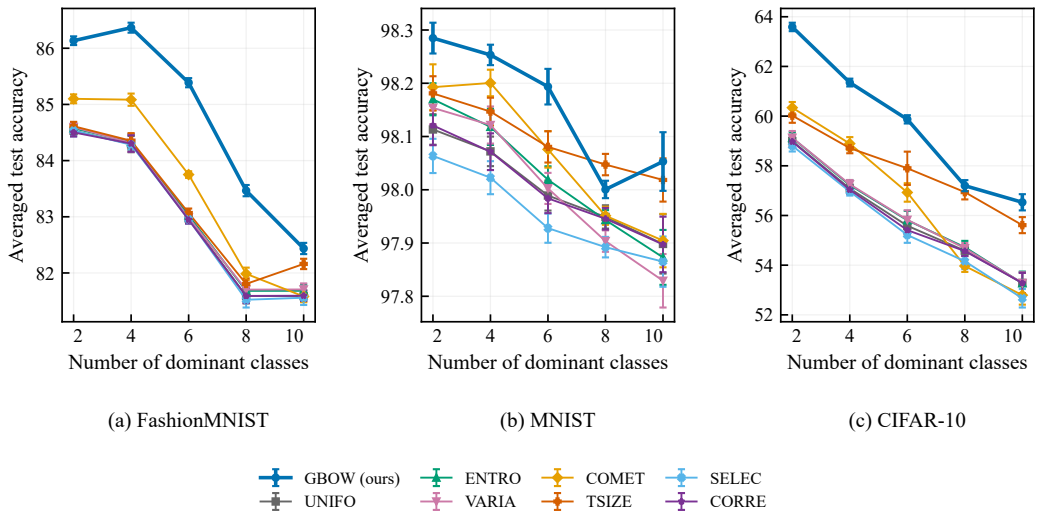


Figure 1: Best-over-ρ averaged test accuracy versus the number of dominant classes on complete graphs.

competitive across a broad range of ρ. Unlike the baseline methods considered here, GBOW explicitly utilize ρ for its weight optimization, and the resulting ensemble weights can better reflect the trade-off between original training loss and KD loss.

D.4 Heterogeneous Architectures on FashionMNIST

In this experiment, we consider heterogeneous model architectures across nodes. Each node independently samples one of five LeNet-family architectures, including the LeNet used in Appendix D.3. The five architectures have 44,426, 11,578, 9,328, 6,616, and 1,466 parameters. The communication graph is complete. Data sizes, loss, and 25,000-round budget match the homogeneous FashionMNIST experiment in Appendix D.3. As shown in Table 6 and Figure 5, GBOW has the highest observed best-over-ρ mean across all five partitions $d = 2, 4, 6, 8, 10$.

D.5 Effect of Communication Radius

In this experiment, we place 20 nodes independently and uniformly in a unit square and connect pairs whose Euclidean distance does not exceed R. We compare $R \in \{0.5, 0.7, 0.9\}$ with a separate complete-graph setting, using homogeneous LeNet models, FashionMNIST, five dominant-class settings, and the twelve-value ρ grid. The results are shown in Figure 6. Since COMET and SELEC require complete communication, they are evaluated only on the complete graph and appear as single endpoint markers. They are also excluded from comparisons of gains across communication settings.

After each method selects its own ρ, GBOW has the highest observed mean among the six graph-compatible methods in all 20 communication–partition settings. Its margin over the best observed alternative ranges from 0.125 to 2.319 percentage points. From $R = 0.5$ to complete communication, GBOW gains 0.620, 0.655, 0.805, 0.645, and 0.209 points at $d = 2, 4, 6, 8, 10$, respectively. These gains exceed those of every individual alternative in every partition. Each alternative's best-over-ρ mean varies by less than 0.137 points across the four communication settings within a partition. Thus, GBOW shows a larger empirical benefit for a denser communication network.

As shown in Table 7, this overall improvement is not strictly monotonic through the complete endpoint, and the pattern suggests that most of the observed connectivity benefit of GBOW has already been obtained by $R = 0.9$.

Table 5: CIFAR-10: best observed averaged test accuracy (%), followed by selected ρ.

Method	$d = 2$	$d = 4$	$d = 6$	$d = 8$	$d = 10$
GBOW (ours)	**63.59 ± 0.17**	**61.35 ± 0.16**	**59.87 ± 0.17**	**57.20 ± 0.22**	**56.53 ± 0.33**
UNIFO	58.99 ± 0.26	57.09 ± 0.18	55.59 ± 0.34	54.58 ± 0.21	53.29 ± 0.41
ENTRO	59.12 ± 0.23	57.22 ± 0.20	55.81 ± 0.38	54.72 ± 0.25	53.27 ± 0.47
VARIA	59.13 ± 0.27	57.26 ± 0.17	55.83 ± 0.38	54.68 ± 0.21	53.28 ± 0.43
COMET	60.34 ± 0.23	58.89 ± 0.26	56.93 ± 0.37	53.97 ± 0.24	52.79 ± 0.38
TSIZE	60.02 ± 0.29	58.71 ± 0.21	57.90 ± 0.67	56.94 ± 0.29	55.61 ± 0.32
SELEC	58.81 ± 0.23	56.97 ± 0.16	55.22 ± 0.32	54.16 ± 0.31	52.67 ± 0.37
CORRE	58.97 ± 0.20	57.04 ± 0.16	55.41 ± 0.18	54.57 ± 0.20	53.28 ± 0.41

Method	$d = 2$	$d = 4$	$d = 6$	$d = 8$	$d = 10$
GBOW (ours)	0.85	0.3	0.3	0.8	0.5
UNIFO	0.95	0.95	0.85	0.85	0.8
ENTRO	0.95	0.95	0.9	0.85	0.8
VARIA	0.95	0.95	0.9	0.85	0.85
COMET	0.95	0.9	0.8	0.85	0.7
TSIZE	0.8	0.7	0.7	0.7	0.6
SELEC	0.95	0.95	0.85	0.8	0.8
CORRE	0.95	0.95	0.85	0.85	0.8

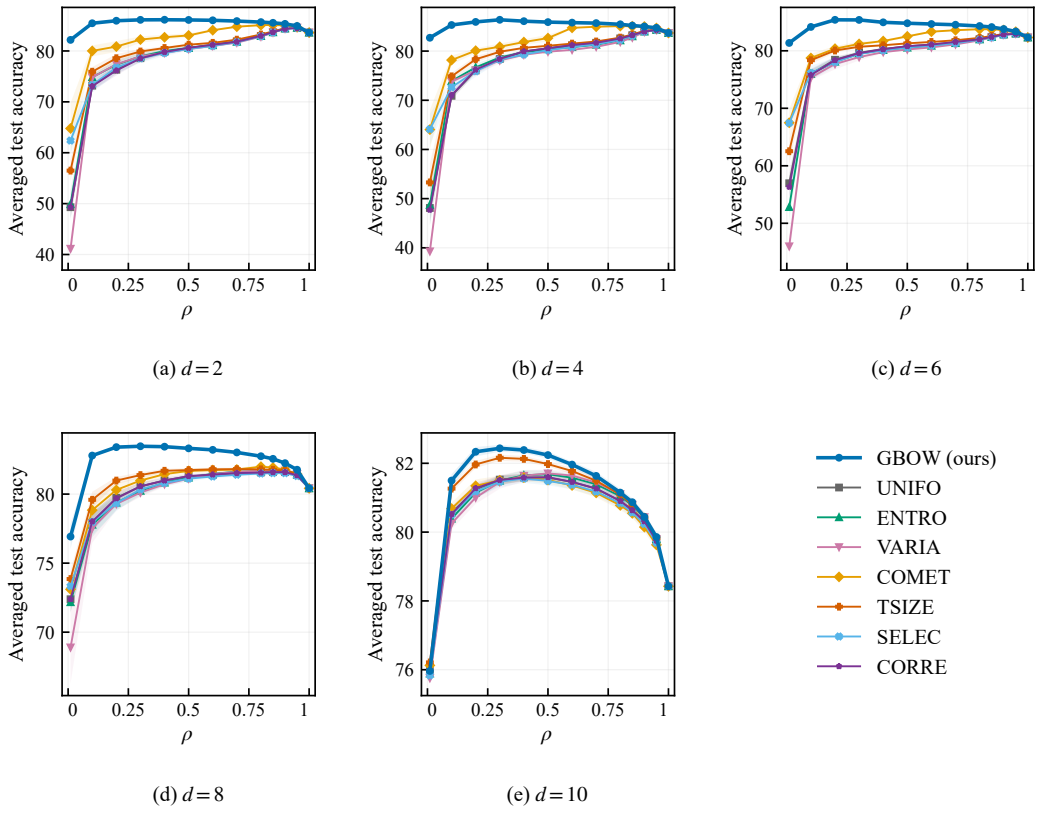


Figure 2: Homogeneous FashionMNIST averaged test accuracy versus ρ. Panels (a)-(e) correspond to $d = 2, 4, 6, 8, 10$. Bands are pointwise 95% confidence intervals.

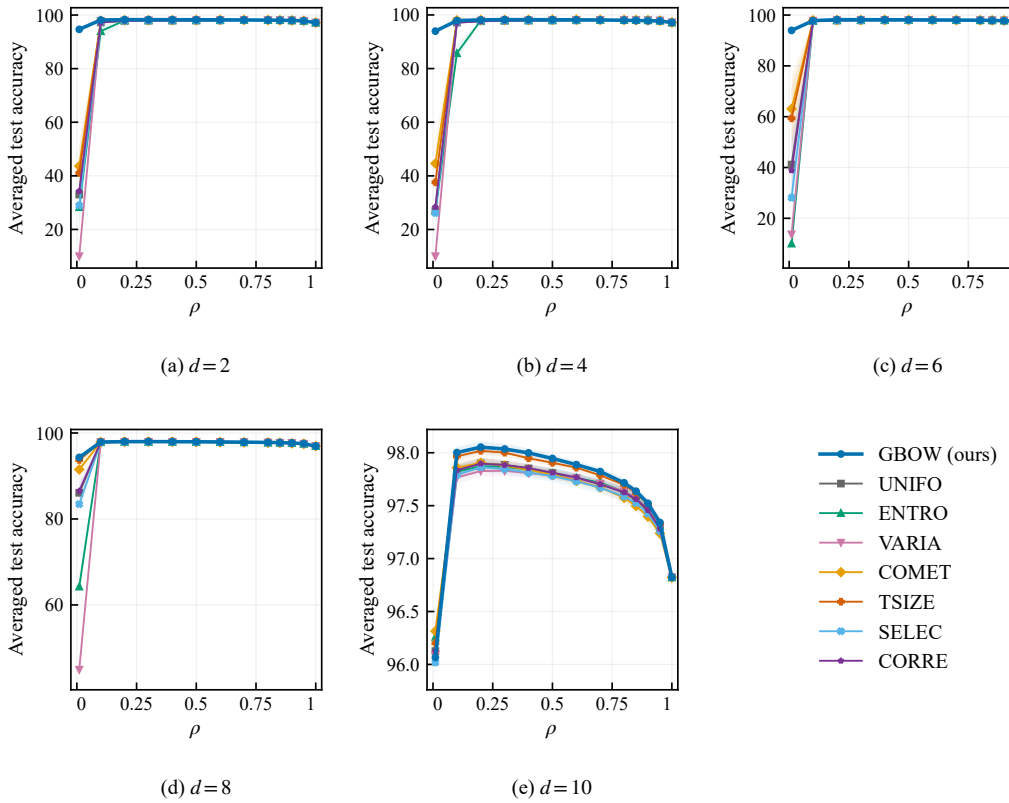


Figure 3: MNIST averaged test accuracy versus ρ. Panels (a)-(e) correspond to $d = 2, 4, 6, 8, 10$. Bands are pointwise 95% confidence intervals.

Table 6: Heterogeneous FashionMNIST: best averaged test accuracy (%) with 95% intervals, followed by selected ρ.

Method	$d = 2$	$d = 4$	$d = 6$	$d = 8$	$d = 10$
GBOW (ours)	**83.11 ± 0.30**	**83.20 ± 0.27**	**81.55 ± 0.38**	**79.01 ± 0.28**	**77.03 ± 0.38**
UNIFO	82.49 ± 0.33	82.43 ± 0.20	80.72 ± 0.32	78.56 ± 0.27	76.86 ± 0.28
ENTRO	82.55 ± 0.32	82.41 ± 0.21	80.74 ± 0.31	78.60 ± 0.26	76.94 ± 0.32
VARIA	82.57 ± 0.33	82.40 ± 0.22	80.73 ± 0.31	78.59 ± 0.26	76.98 ± 0.31
COMET	82.85 ± 0.32	82.75 ± 0.21	81.15 ± 0.25	78.75 ± 0.23	76.92 ± 0.36
TSIZE	82.54 ± 0.33	82.41 ± 0.19	80.84 ± 0.29	78.60 ± 0.28	76.97 ± 0.36
SELEC	82.49 ± 0.33	82.33 ± 0.20	80.71 ± 0.32	78.45 ± 0.26	76.84 ± 0.29
CORRE	82.48 ± 0.32	82.40 ± 0.18	80.74 ± 0.30	78.55 ± 0.29	76.89 ± 0.33

Method	$d = 2$	$d = 4$	$d = 6$	$d = 8$	$d = 10$
GBOW (ours)	0.7	0.3	0.3	0.6	0.6
UNIFO	0.9	0.9	0.9	0.85	0.5
ENTRO	0.9	0.9	0.9	0.8	0.6
VARIA	0.9	0.9	0.9	0.85	0.6
COMET	0.8	0.8	0.8	0.8	0.5
TSIZE	0.9	0.9	0.9	0.8	0.6
SELEC	0.9	0.9	0.9	0.8	0.6
CORRE	0.9	0.9	0.9	0.8	0.6

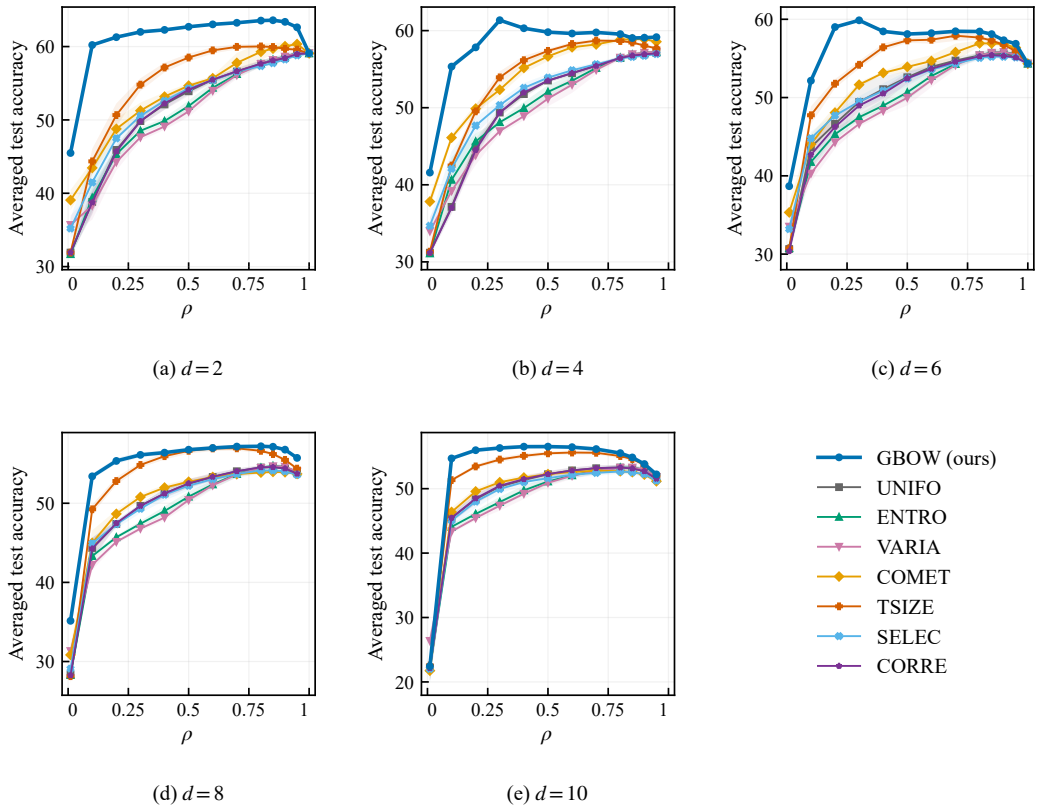


Figure 4: CIFAR-10 averaged test accuracy versus ρ. Panels (a)-(e) correspond to $d = 2, 4, 6, 8, 10$. Bands are pointwise 95% confidence intervals.

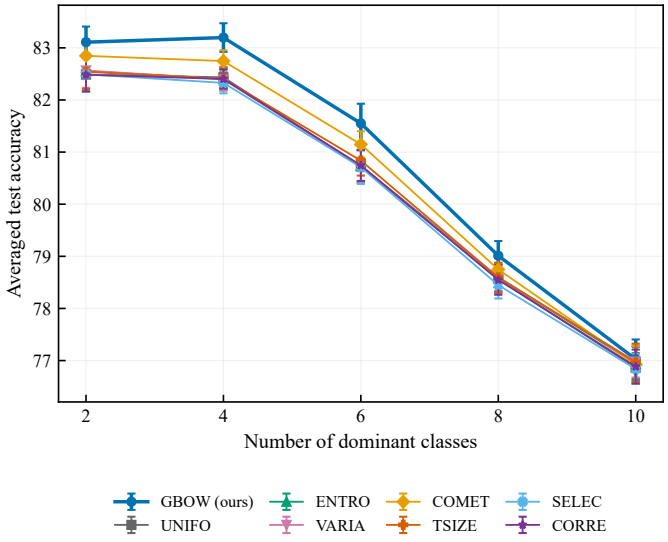


Figure 5: Heterogeneous FashionMNIST, best-over-ρ averaged test accuracy versus the number of dominant classes.

Table 7: GBOW's averaged test accuracy (%) at its best observed ρ. Bold indicates the highest mean in each row.

d	$R = 0.5$	$R = 0.7$	$R = 0.9$	Complete
2	85.51 ± 0.30	86.00 ± 0.17	$\mathbf{86.18 \pm 0.25}$	86.13 ± 0.08
4	85.71 ± 0.38	86.17 ± 0.12	$\mathbf{86.37 \pm 0.23}$	86.37 ± 0.09
6	84.58 ± 0.31	85.15 ± 0.10	85.33 ± 0.13	$\mathbf{85.39 \pm 0.08}$
8	82.82 ± 0.35	83.38 ± 0.16	$\mathbf{83.52 \pm 0.11}$	83.47 ± 0.10
10	82.23 ± 0.41	82.37 ± 0.20	82.38 ± 0.26	$\mathbf{82.44 \pm 0.10}$

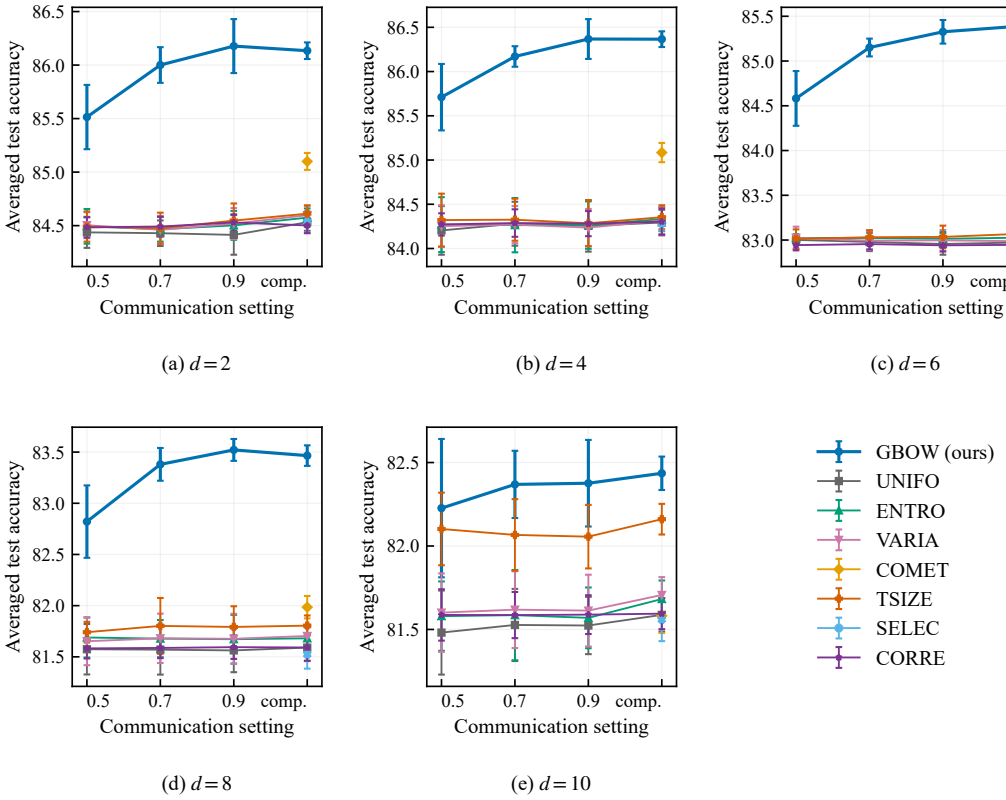


Figure 6: Best-over-ρ averaged test accuracy versus communication setting on FashionMNIST. Panels (a)–(e) correspond to $d = 2, 4, 6, 8, 10$. The rightmost category, `comp.`, is the complete graph. COMET and SELEC appear only as single complete-graph points.