Finite-Sample and Communication-Efficient Networked Information Aggregation

Mohammadhossein Bateni
Google Research
bateni@google.com

Zahra Hadizadeh
University of California, Irvine
zhadizad@uci.edu

MohammadTaghi Hajiaghayi
University of Maryland
hajiagha@umd.edu

Mahdi JafariRaviz
University of Maryland
mahdij@umd.edu

Shayan Taherijam
University of California, Irvine
staherij@uci.edu

Abstract

Building on SODA 2026 work Kearns et al. [2026], we study learning in directed acyclic graphs in which each agent observes only a subset of the input coordinates and may use the predictions of its parents as additional features. In the finite-sample protocol, the message sent along each edge is the agent's prediction vector on a shared sample, so the communication cost is linear in the sample size m. In this paper we study this model under communication constraints. First, we revisit the finite-sample analysis of the full-message protocol. We identify a gap in the finite-sample proof of Kearns et al. [2026] and develop a different argument that establishes a finite-sample population-risk guarantee along paths satisfying deterministic M-coverage. Second, we introduce a shared-sketch protocol in which each agent sends a k-dimensional linear sketch of its prediction vector instead of the full m-dimensional message, where $k \ll m$. We analyze this protocol through a reduction to least squares on the compressed sample together with an oblivious subspace embedding argument. The resulting bound separates statistics from communication: the statistical term is still governed by the original sample size m, while the additional error due to compression is controlled by the sketch dimension k. For Gaussian sketches, we obtain an explicit high-probability population-risk bound, showing that row subsampling is not the only possible communication-statistics tradeoff for networked information aggregation.

1 Introduction

In many distributed learning problems, no single agent observes all of the input features. Instead, each agent sees only a subset of the coordinates and can communicate only with nearby agents in a network. A basic question is whether such local communication is enough to recover the performance of a learner that has access to all features at once.

Kearns et al. [2026] study this question for linear regression on directed acyclic graphs. In their model, agent i observes a subset $S_i \subseteq [d]$ of the coordinates and may use the predictions sent by its parents. Their main analysis is population-based: each agent solves a least-squares problem with respect to the underlying distribution, and along a path they show that if every M consecutive agents together observe all d features, then the network can compete with the best predictor that uses all coordinates.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Our focus is the communication needed to retain this aggregation benefit on a finite shared sample. The model is motivated by vertically partitioned learning on aligned examples and stacked prediction systems that pass upstream fitted values as downstream features.

It is important to be clear about what is communicated in this model. The message passed by an agent is not just a coefficient vector. Rather, the message is the agent's prediction itself, which later agents may use as an additional feature in their own least-squares problems. In a finite table, this corresponds to passing the full column of fitted values. In the population analysis, the analogous message is the prediction as an object defined over the data distribution rather than a fixed finite-dimensional vector. This is one reason why communication is not directly quantified in the population model.

Appendix C of Kearns et al. [2026] considers a finite-sample version of this model. A sample

$$D_m = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\}$$

is shared across the network, each agent still sees only its own feature subset S_i, and the least-squares problem at each node is solved on these sampled rows. In this setting, the message sent by an agent is its prediction vector on the m sampled rows. Thus the recursive learning rule is the same in spirit, but communication is now represented by an explicit vector in $\mathbb{R}^m$.

The finite-sample proof in Appendix C has a gap. The issue is that the proof uses an empirical-risk-minimization comparison between two predictors that are produced by different recursive processes. More precisely, in the notation of Kearns et al. [2026], $\hat{y}_k^{\mathrm{emp}}$ denotes the predictor produced at node k by the empirical recursive process, while $\hat{y}_k$ denotes the predictor produced at node k by the population recursive process. The proof writes $\widehat{R}(\cdot)$ for the empirical squared loss on the sampled rows and uses the inequality

$$\widehat{R}(\hat{y}_k^{\mathrm{emp}}) \le \widehat{R}(\hat{y}_k).$$

The stated justification is that $\hat{y}_k^{\mathrm{emp}}$ is the empirical-risk minimizer at node k. However, this minimization argument would only apply if $\hat{y}_k^{\mathrm{emp}}$ and $\hat{y}_k$ were feasible predictors for the same empirical least-squares problem.

In the networked recursion, this is not generally the case. At node p_t, the least-squares problem is formed from the local sample columns X_{S_t} together with the prediction vectors received from the parents of p_t. In the empirical process, these parent inputs are the previously learned empirical predictions. In the population process, they are the previously learned population predictions. Therefore the two recursive processes generally generate different parent inputs, so the two least-squares problems are built from different linear spans. Consequently, $\hat{y}_k^{\mathrm{emp}}$ is not minimizing the empirical objective over a class that necessarily contains $\hat{y}_k$, and the displayed inequality does not follow as stated.

Finite-sample analysis. Our first contribution is to address the gap in the finite-sample analysis described above. We give a finite-sample population-risk guarantee for the full-message protocol without comparing the sample-based recursive process to a separate population-level recursive process. The proof works entirely on the realized sample D_m: we first prove an empirical path theorem comparing the network output to the empirical all-columns benchmark, that is, the predictor a single learner would obtain from the same sample if it observed all d features. We then apply the uniform generalization bound of Kearns et al. [2026, Theorem C.2] to the selected network predictor and to the population all-columns benchmark. This gives the desired population-risk guarantee while avoiding the invalid sample-versus-population recursive comparison.

The finite-sample model also makes the communication cost explicit. If communication is limited to k numbers per message, a direct approach is to choose k public sample rows and run the finite-sample protocol only on those rows. This reduces communication, but it also reduces the effective sample size from m to k. A different approach is to keep the full m-sample data set, but send only a k-dimensional linear sketch of each m-dimensional prediction vector.

Shared-sketch protocol. Our second contribution is a protocol using a single public sketch matrix $S \in \mathbb{R}^{k \times m}$, drawn independently of the sample. The difficulty is recursive: each agent fits using its parents' sketched predictions, so its predictor depends on S through earlier regressions. By linearity, every implicit full-sample prediction remains a combination of the original feature columns, so all residuals lie in $V := \mathrm{span}\{X_1, \ldots, X_d, y\}$. Conditional on the sample, this subspace is fixed independently of S and has dimension at most $d + 1$. A single oblivious subspace embedding

therefore approximately preserves the empirical squared losses of all linear predictors simultaneously, including those produced by the recursive protocol. Thus, for fixed d, embedding distortion, and failure probability, the required sketch dimension does not grow with network depth.

We also compare the shared-sketch protocol with the row-sampling baseline. Row sampling uses k both as the communication budget and as the statistical sample size, leading to a term of order $\widetilde{O}(d/\sqrt{k})$. Sketching instead uses all m samples and pays a separate compression cost. For Gaussian sketches, the leading k-dependent term is of order $O(\mathcal{R}(f_A^*)\sqrt{d/k})$. Thus sketching is especially useful when the all-columns benchmark has small risk. Even in the worst case, our boundedness assumption gives $\mathcal{R}(f_A^*) \le Y_{\max}^2$, so the sketching term is at most $\widetilde{O}(\sqrt{d/k})$ when $Y_{\max}$ is treated as a constant.

The remainder of the paper is organized as follows. Section 2 introduces the model, the finite-sample protocol, and the all-columns benchmarks. Section 3 proves the finite-sample guarantee for the full-message protocol. Section 4 introduces the shared-sketch protocol, proves its population-risk guarantee, and compares the resulting bound with row subsampling.

1.1 Related Work

Work on information aggregation in networks goes back to the literature on opinion dynamics and social learning. Classical models such as DeGroot dynamics and related heuristic update rules study when repeated local interaction leads to consensus or accurate aggregation of dispersed information [DeGroot, 1974, Bala and Goyal, 1998, Ellison and Fudenberg, 1993, Golub and Jackson, 2010, Akbarpour et al., 2017]. Another major line of work studies Bayesian learning and agreement on networks, including agreement theorems, herding and informational cascades, and conditions under which local interaction does or does not aggregate private information efficiently [Aumann, 1976, Geanakoplos and Polemarchakis, 1982, Aaronson, 2005, Banerjee, 1992, Bikhchandani et al., 1992, Acemoglu et al., 2011, Mossel et al., 2012, 2020, Jadbabaie et al., 2012, Shahrampour and Jadbabaie, 2013, Frongillo et al., 2023, Kong and Schoenebeck, 2023]. These models typically focus on scalar states, actions, or beliefs, rather than supervised learning with high-dimensional feature vectors distributed across agents.

The work most directly related to ours is Kearns et al. [2026], which introduced the DAG-based learning model that we build on. A closely related recent work, Bateni et al. [2026a], studies the same framework for binary classification. Recent work also establishes optimal population aggregation rates [Bateni et al., 2026c] and studies network design for exact aggregation [Bateni et al., 2026b]. Our paper focuses on a different direction: the finite-sample and communication-limited aspects of the squared-loss regression model. In particular, we give a finite-sample population-risk guarantee for the full-message protocol and introduce a shared-sketch protocol that compresses the prediction vectors sent across edges.

Our setting is also related to feature-partitioned collaborative learning, especially vertical federated learning and split learning. In vertical federated learning, different parties hold different feature blocks for the same examples and communicate intermediate quantities in order to train a joint predictor [Yang et al., 2023, Khan et al., 2022, Fu et al., 2022, Castiglia et al., 2022, 2023, Sun et al., 2023, Valdeira et al., 2025]. These works are close in motivation because they also study learning with distributed features under communication constraints. The difference is that they focus on federated training protocols, privacy-oriented system models, or split neural architectures, whereas our model studies information aggregation through a directed network of learners that sequentially pass predictions to their descendants.

The sketching part of our paper is related to randomized numerical linear algebra and sketched least squares. Random projections and oblivious subspace embeddings have been used extensively to compress least-squares problems while preserving their geometry [Clarkson and Woodruff, 2013, Woodruff, 2014, Pilanci and Wainwright, 2014a,b]. Distributed sketching has also been studied as a way to reduce communication in large-scale optimization [Bartan and Pilanci, 2023]. In contrast to these centralized or standard distributed optimization settings, our sketches are used inside a recursive networked prediction process.

A related systems-level perspective comes from the literature on multi-agent and agentic AI workflows. Recent surveys describe agent workflows as directed graphs or structured computation graphs in which multiple agents exchange intermediate outputs and coordinate over a multi-step task [Yu et al.,

2025, Bei et al., 2025, Yue et al., 2026]. This viewpoint is conceptually related to our model, but that literature focuses on LLM-based planning, tool use, memory, and workflow design, whereas our paper studies statistical guarantees for feature-partitioned supervised learning.

2 Preliminaries

Let $\mathcal{D}$ be a distribution over feature-label pairs $(x, y) \in \mathbb{R}^d \times \mathbb{R}$. For a predictor $f : \mathbb{R}^d \to \mathbb{R}$, define its population squared loss by

$$\mathcal{R}(f) := \mathbb{E}_{(x,y)\sim\mathcal{D}}\big[(f(x) - y)^2\big]. \tag{1}$$

Let $D_m = \{(x^{(r)}, y^{(r)})\}_{r=1}^m$ be m i.i.d. samples from $\mathcal{D}$. We write $X \in \mathbb{R}^{m\times d}$ for the sample feature matrix and $y \in \mathbb{R}^m$ for the vector of sampled labels. For a linear predictor $f_\beta(x) = \beta^\top x$, define

$$\widehat{\mathcal{R}}_m(f_\beta) := \frac{1}{m}\,\|X\beta - y\|_2^2\,. \tag{2}$$

For a fitted vector $u \in \mathbb{R}^m$, we also write $\widehat{\mathcal{R}}_m(u) := \frac{1}{m}\,\|u - y\|_2^2$.

2.1 Network and finite-sample protocol

Agents are vertices of a directed acyclic graph. Each agent i observes a subset $S_i \subseteq [d]$ of the original coordinates. If there is an edge $j \to i$ we call j the *parent* of i and denote the set of all parents of i by $\mathrm{Pa}(i)$. In the full-message finite-sample protocol, once the parent fitted vectors have been computed, agent i solves empirical least squares over its own observed sample columns and the fitted vectors received from its parents:

$$(\widehat{w}_i, \widehat{a}_i) \in \arg\min_{w,a} \frac{1}{m}\left\| X_{S_i} w + \sum_{j\in\mathrm{Pa}(i)} a_j u_j - y \right\|_2^2 .$$

The fitted vector sent by agent i on the training sample is $u_i := X_{S_i}\widehat{w}_i + \sum_{j\in\mathrm{Pa}(i)} (\widehat{a}_i)_j u_j$. Equivalently, at a point x, agent i's predictor is the recursively defined linear function

$$\widehat{f}_i(x) = \widehat{w}_i^\top x_{S_i} + \sum_{j\in\mathrm{Pa}(i)} (\widehat{a}_i)_j \widehat{f}_j(x).$$

We focus on a directed path $P = (p_1, p_2, \ldots, p_D)$, where $p_{t-1} \in \mathrm{Pa}(p_t) \quad$ for $t = 2, \ldots, D$. For notational simplicity, along this path we write $u_t := u_{p_t}$, $\widehat{f}_t := \widehat{f}_{p_t}$, and $S_t := S_{p_t}$.

2.2 Coverage under random feature allocation

Assumption 2.1 (Deterministic M-coverage)**.** We assume every contiguous block of M agents on the path observe the full set of features from X. More formally, we have $\bigcup_{t=i}^{i+M-1} S_t = [d]$ for every $i = 1, \ldots, D - M + 1$.

Our proofs use coverage only on the disjoint blocks $\{(q-1)M + 1, \ldots, qM\}$, $q = 1, \ldots, \lfloor D/M \rfloor$.

The random feature allocation model of Kearns et al. [2026, Section 3.3] gives a sufficient condition for Assumption 2.1 to hold with high probability.

Corollary 2.2 (Coverage under random feature allocation)**.** *Fix a path $P = (p_1, \ldots, p_D)$ independently of the feature allocation. Let $p, \delta_{\mathrm{cover}} \in (0, 1)$, and let M be an integer with $1 \le M \le D$. Suppose each feature $\ell \in [d]$ belongs to S_t with probability p, independently over all agents on the path and all features. If*

$$M \ge \frac{\log(dD/\delta_{\mathrm{cover}})}{-\log(1-p)},$$

then Assumption 2.1 holds with probability at least $1 - \delta_{\mathrm{cover}}$ over the feature allocation.

The proof is given in Appendix A.4. For fixed p, the sufficient coverage window grows only logarithmically in d, D, and $1/\delta_{\mathrm{cover}}$, provided that this window fits within the path. Thus the deterministic coverage condition holds with high probability under this random allocation model. Coverage may remain restrictive for sparse or dependent feature allocations. On the coverage event, the results in Sections 3 and 4 apply subject to their remaining assumptions.

2.3 All-columns benchmarks

Fix a radius $A > 0$. The population all-columns benchmark is

$$\beta_A^* \in \arg\min_{\|\beta\|_1 \le A} \mathcal{R}(f_\beta), \qquad f_A^*(x) := (\beta_A^*)^\top x.$$

That is, the predictor that a single agent would learn from observing the distribution $\mathcal{D}$. The empirical all-columns benchmark is the predictor that a single agent would learn from the same finite sample if it observed all d columns:

$$\widehat{\beta}_A \in \arg\min_{\|\beta\|_1 \le A} \widehat{\mathcal{R}}_m(f_\beta), \qquad \widehat{f}_A(x) := \widehat{\beta}_A^\top x.$$

The finite-sample proof below first compares the network output to $\widehat{f}_A$ on the same sample. Only after this purely empirical comparison do we transfer to the population benchmark f_A^*.

Assumption 2.3 (Boundedness and covariance)**.** There are constants $X_{\max}, Y_{\max}, \lambda_0 > 0$ such that

(i) $\|x\|_\infty \le X_{\max}$ almost surely;

(ii) $|y| \le Y_{\max}$ almost surely;

(iii) the population covariance matrix $\Sigma := \mathbb{E}[xx^\top]$ satisfies $\lambda_{\min}(\Sigma) \ge \lambda_0$.

Boundedness controls the empirical and generalization bounds, while the covariance condition provides coefficient-norm control for the population-risk guarantee. The empirical path argument does not require the covariance condition; relaxing boundedness or replacing covariance-based norm control would require new concentration or norm-control arguments.

3 Finite-sample analysis

This section proves a population-risk guarantee for the full-message finite-sample protocol from Section 2. We first state the uniform generalization bound used to pass from empirical risk to population risk. We then state the empirical path theorem, which is proved in Appendix A. Combining these two results with the norm-control lemma from Appendix A.3 gives the main theorem of this section.

For $\Lambda > 0$, let

$$\mathcal{H}_\Lambda := \{x \mapsto \beta^\top x : \|\beta\|_1 \le \Lambda\}.$$

The following theorem follows from the proof of Kearns et al. [2026, Theorem C.2].

Theorem 3.1 (Generalization bound for ℓ_1-bounded linear regression)**.** *Fix $\Lambda > 0$. Assume $\|x\|_\infty \le X_{\max}$ and $|y| \le Y_{\max}$. Let D_m be a training set of m i.i.d. samples from $\mathcal{D}$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$ over the draw of D_m, for all $f \in \mathcal{H}_\Lambda$,*

$$\left|\mathcal{R}(f) - \widehat{\mathcal{R}}_m(f)\right| \le \varepsilon_{\mathrm{C2}}(m, d, \Lambda, \delta),$$

where

$$\varepsilon_{\mathrm{C2}}(m, d, \Lambda, \delta) := C(\Lambda X_{\max} + Y_{\max})\Lambda X_{\max}\sqrt{\frac{\log(2d)}{m}} + (\Lambda X_{\max} + Y_{\max})^2\sqrt{\frac{\log(C'/\delta)}{2m}},$$

and $C, C' > 0$ are universal constants.

The quantity $\varepsilon_{\mathrm{C2}}(m, d, \Lambda, \delta)$ controls the difference between empirical risk and population risk uniformly over $\mathcal{H}_\Lambda$. We keep both terms in this bound. In particular, after substituting $\Lambda_{\mathrm{net}} = \sqrt{\frac{2d}{\lambda_0}}Y_{\max}$, the second term is not lower order in general.

The next theorem is the empirical part of the finite-sample argument. It compares the network output to the empirical all-columns benchmark $\widehat{f}_A$ on the same realized sample. Its proof is deferred to Appendix A.

Theorem 3.2 (Empirical path theorem under M-coverage)**.** *Let $K := \lfloor D/M \rfloor$, and define $\Gamma_{\mathrm{path}} := 2AX_{\max}Y_{\max}\sqrt{\frac{M}{K}}$. Partition the first KM agents on the path into K disjoint blocks of length M. Under Assumption 2.1, there is a block endpoint b^* such that*

$$\widehat{\mathcal{R}}_m(u_{b^*}) \le \widehat{\mathcal{R}}_m(\widehat{f}_A) + \Gamma_{\mathrm{path}}.$$

The same bound holds for every later node on the path.

Using Theorem 3.1, Theorem 3.2, and the norm-control lemma from Appendix A.3, we obtain the main result of this section.

Theorem 3.3 (Finite-sample theorem for the full-message protocol)**.** *Assume Assumption 2.1 and Assumption 2.3. Let $K := \lfloor D/M \rfloor, \Gamma_{\mathrm{path}} := 2AX_{\max}Y_{\max}\sqrt{\frac{M}{K}}, \Lambda_{\mathrm{net}} := \sqrt{\frac{2d}{\lambda_0}}Y_{\max}$, and let $\Lambda := \max\{A, \Lambda_{\mathrm{net}}\}$. On the joint event that*

(i) $\lambda_{\min}(m^{-1}X^\top X) \ge \lambda_0/2$, and

(ii) the uniform-convergence event of Theorem 3.1 holds for $\mathcal{H}_\Lambda$,

there exists a block endpoint b^* on the path such that the network predictor $\widehat{f}_{b^*}$ satisfies

$$\mathcal{R}(\widehat{f}_{b^*}) \le \mathcal{R}(f_A^*) + \Gamma_{\mathrm{path}} + 2\varepsilon_{\mathrm{C2}}(m, d, \Lambda, \delta).$$

The same population-risk bound holds for every later node on the path.

The bound separates the aggregation term Γ_{path} from the statistical term $2\varepsilon_{\mathrm{C2}}(m, d, \Lambda, \delta)$, which accounts for passing from empirical to population risk. Empirical risk is nonincreasing along the path, so some block has a small total risk decrease. Together with the block length, this controls changes in its fitted values. Coverage and local orthogonality then make the endpoint residual approximately orthogonal to every feature, yielding the comparison with the all-columns benchmark. For fixed $A, X_{\max}, Y_{\max}, \lambda_0$, the two terms scale as $O(M/\sqrt{D})$ and $\widetilde{O}(d/\sqrt{m})$, respectively. Thus the bound gives a small excess-risk guarantee when D is large relative to M^2 and m is large enough to make the statistical term small; increasing depth alone does not eliminate the statistical term.

Proof. By Theorem 3.2, there is a block endpoint b^* such that

$$\widehat{\mathcal{R}}_m(u_{b^*}) \le \widehat{\mathcal{R}}_m(\widehat{f}_A) + \Gamma_{\mathrm{path}}.$$

By Lemma A.7, we have $\widehat{f}_{b^*} \in \mathcal{H}_\Lambda$. Also $\widehat{f}_A, f_A^* \in \mathcal{H}_A \subseteq \mathcal{H}_\Lambda$.

Applying Theorem 3.1 to $\widehat{f}_{b^*}$,

$$\mathcal{R}(\widehat{f}_{b^*}) \le \widehat{\mathcal{R}}_m(u_{b^*}) + \varepsilon_{\mathrm{C2}}(m, d, \Lambda, \delta).$$

Using the empirical path bound,

$$\mathcal{R}(\widehat{f}_{b^*}) \le \widehat{\mathcal{R}}_m(\widehat{f}_A) + \Gamma_{\mathrm{path}} + \varepsilon_{\mathrm{C2}}(m, d, \Lambda, \delta).$$

Since $\widehat{f}_A$ minimizes empirical risk over $\mathcal{H}_A$ and $f_A^* \in \mathcal{H}_A$,

$$\widehat{\mathcal{R}}_m(\widehat{f}_A) \le \widehat{\mathcal{R}}_m(f_A^*).$$

Applying Theorem 3.1 to f_A^*, using $\mathcal{H}_A \subseteq \mathcal{H}_\Lambda$, gives

$$\widehat{\mathcal{R}}_m(f_A^*) \le \mathcal{R}(f_A^*) + \varepsilon_{\mathrm{C2}}(m, d, \Lambda, \delta).$$

Combining the last three displays,

$$\mathcal{R}(\widehat{f}_{b^*}) \le \mathcal{R}(f_A^*) + \Gamma_{\mathrm{path}} + 2\varepsilon_{\mathrm{C2}}(m, d, \Lambda, \delta).$$

For any later node $t \ge b^*$, Theorem 3.2 gives the same empirical bound for u_t, and Lemma A.7 gives $\widehat{f}_t \in \mathcal{H}_\Lambda$. Therefore the same argument applies to $\widehat{f}_t$. □

4 Shared-sketch protocol

In this section we introduce the sketching protocol. The training set is still the same m-sample data set used in Section 3. The difference is only in communication: instead of sending a full m-dimensional prediction vector, each agent sends a k-dimensional sketch of that vector, where $k \ll m$.

The main point is that the sample size and the communication size are now separated. The sample size m controls the usual statistical generalization error, while the sketch dimension k controls how much information is lost by compression. If the sketch preserves the geometry of the relevant residual subspace, then the sketched protocol satisfies a population-risk guarantee similar to the finite-sample theorem, but with an additional sketching distortion term.

4.1 Protocol and risks

Throughout this section, $S \in \mathbb{R}^{k \times m}$ is a public sketch matrix drawn independently of the sample. For $v \in \mathbb{R}^m$, write $\widetilde{v} := Sv$. Once the sketched parent messages $\{\widetilde{u}_j^{\mathrm{sk}} : j \in \mathrm{Pa}(i)\}$ are available, agent i solves the sketched least-squares problem

$$(\widetilde{w}_i, \widetilde{a}_i) \in \operatorname*{arg\,min}_{w,a} \frac{1}{m} \left\| SX_{S_i} w + \sum_{j \in \mathrm{Pa}(i)} a_j \widetilde{u}_j^{\mathrm{sk}} - Sy \right\|_2^2 .$$

The agent then broadcasts only the k-dimensional vector

$$\widetilde{u}_i^{\mathrm{sk}} := SX_{S_i} \widetilde{w}_i + \sum_{j \in \mathrm{Pa}(i)} (\widetilde{a}_i)_j \widetilde{u}_j^{\mathrm{sk}} \in \mathbb{R}^k.$$

For the analysis, we also define the corresponding implicit full-sample fitted vector

$$u_i^{\mathrm{sk}} := X_{S_i} \widetilde{w}_i + \sum_{j \in \mathrm{Pa}(i)} (\widetilde{a}_i)_j u_j^{\mathrm{sk}} \in \mathbb{R}^m.$$

By construction, $\widetilde{u}_i^{\mathrm{sk}} = S u_i^{\mathrm{sk}}$. The vector u_i^{sk} is never transmitted. It is only a proof device that lets us view each k-dimensional message as the sketch of a full m-dimensional prediction vector.

Definition 4.1 (Sketched empirical risk)**.** For $u \in \mathbb{R}^m$, define $\widetilde{\mathcal{R}}_{m,S}(u) := \frac{1}{m} \|S(u - y)\|_2^2$. For a linear predictor $f_\beta(x) = \beta^\top x$, define $\widetilde{\mathcal{R}}_{m,S}(f_\beta) := \frac{1}{m} \|S(X\beta - y)\|_2^2$. If $\widetilde{u} = Su$, then $\widetilde{\mathcal{R}}_{m,S}(u) = \frac{1}{m} \|\widetilde{u} - Sy\|_2^2$. Thus the sketched empirical risk can be computed from the sketched message $\widetilde{u}$ alone.

Along the fixed path $P = (p_1, \ldots, p_D)$, we abbreviate $u_t^{\mathrm{sk}} := u_{p_t}^{\mathrm{sk}}, \widetilde{u}_t^{\mathrm{sk}} := S u_t^{\mathrm{sk}}, S_t := S_{p_t}$.

The all-columns sketched benchmark is

$$\widetilde{\beta}_A \in \operatorname*{arg\,min}_{\|\beta\|_1 \le A} \widetilde{\mathcal{R}}_{m,S}(f_\beta), \qquad \widetilde{f}_A(x) := \widetilde{\beta}_A^\top x.$$

This is the predictor that a centralized learner would obtain if it saw all d features but minimized the sketched loss instead of the usual empirical loss.

4.2 Subspace embeddings

Let

$$V := \mathrm{span}\{X_1, \ldots, X_d, y\} \subseteq \mathbb{R}^m,$$

where X_ℓ is column ℓ of X. Then $\dim(V) \le d + 1$. This subspace contains every linear prediction $X\beta$, and also every residual $X\beta - y$. In this section the data matrix X and label vector y are fixed, so V is a fixed subspace of $\mathbb{R}^m$. The randomness in the embedding statements below is only over the draw of the sketch matrix S.

Definition 4.2 (Oblivious subspace embedding)**.** For $\varepsilon \in (0, 1)$ and a subspace $W \subseteq \mathbb{R}^m$, the matrix S is an (ε, W)-*embedding* if for every $v \in W$,

$$(1 - \varepsilon) \|v\|_2^2 \le \|Sv\|_2^2 \le (1 + \varepsilon) \|v\|_2^2 .$$

An (ε, W)-embedding approximately preserves the Euclidean length of every vector in W. Since our risks are squared residual norms, embedding the residual subspace is exactly what is needed to compare empirical risk and sketched empirical risk.

Lemma 4.3 (Sketch preserves residual norms)**.** *Let $\mathcal{V} \subseteq \mathbb{R}^m$, and suppose S is an (ε, W)-embedding for some $W \supseteq \operatorname{span}(\mathcal{V} \cup \{y\})$. Then for every $u \in \operatorname{span}(\mathcal{V})$,*

$$(1-\varepsilon)\widehat{\mathcal{R}}_m(u) \le \widetilde{\mathcal{R}}_{m,S}(u) \le (1+\varepsilon)\widehat{\mathcal{R}}_m(u).$$

Proof. Since $u \in \operatorname{span}(\mathcal{V})$, the residual $u - y$ lies in W. Apply Definition 4.2 to $u - y$ and divide by m. □

The implicit predictors produced by the sketched protocol remain linear functions of the original feature columns, by the same induction used for Lemma A.1. Hence all vectors u_t^{sk} lie in $\operatorname{span}\{X_1, \dots, X_d\} \subseteq V$. Therefore an (ε, V)-embedding is enough for the whole sketching analysis.

A standard way to obtain such an embedding is to use a Gaussian sketch: each entry of $S \in \mathbb{R}^{k \times m}$ is drawn independently from $N(0, 1/k)$. The next lemma is the only random-sketch fact we need. It is a direct consequence of the standard Gaussian oblivious-subspace embedding theorem; see, for example, Woodruff [2014, Section 2.1] for subspace embeddings and Vershynin [2012, Theorem 5.39] for the Gaussian singular-value estimate behind the proof. The general subspace embedding statement and proof are deferred to Appendix B.

Lemma 4.4 (Gaussian embedding for the residual subspace)**.** *Fix $X \in \mathbb{R}^{m \times d}$ and $y \in \mathbb{R}^m$, and let*

$$V = \operatorname{span}\{X_1, \dots, X_d, y\}.$$

Let $S \in \mathbb{R}^{k \times m}$ have i.i.d. entries $N(0, 1/k)$. There is a universal constant $C_{\mathrm{ose}} > 0$ such that, if

$$k \ge C_{\mathrm{ose}} \, \frac{d + 1 + \log(1/\delta_{\mathrm{ose}})}{\varepsilon^2},$$

then S is an (ε, V)-embedding with probability at least $1 - \delta_{\mathrm{ose}}$ over the draw of S.

In the deterministic theorem below, we simply assume that S is an (ε, V)-embedding. Lemma 4.4 shows that this event holds with high probability for Gaussian sketches provided

$$k \gtrsim \frac{d + \log(1/\delta_{\mathrm{ose}})}{\varepsilon^2}.$$

4.3 Main guarantees for the shared-sketch protocol

The proof follows the same idea as the finite-sample proof. We need two facts: sketched risk should be close enough to population risk, and along the path there should be a node whose sketched risk is close to the sketched all-columns benchmark.

The first theorem converts sketched empirical risk into population risk, and also converts population risk back into sketched empirical risk. Its proof is deferred to Appendix B.

Theorem 4.5 (Sketched generalization bound)**.** *Fix $\Lambda > 0$ and assume Assumption 2.3(i)–(ii). Suppose the uniform-convergence event of Theorem 3.1 holds for $\mathcal{H}_\Lambda$, and that S is an (ε, V)-embedding with $\varepsilon \in (0, 1)$. Then, simultaneously for all $f \in \mathcal{H}_\Lambda$,*

$$\mathcal{R}(f) \le \frac{1}{1-\varepsilon}\, \widetilde{\mathcal{R}}_{m,S}(f) + \varepsilon_{\mathrm{C2}}(m, d, \Lambda, \delta), \tag{3}$$

$$\widetilde{\mathcal{R}}_{m,S}(f) \le (1+\varepsilon)\, \mathcal{R}(f) + (1+\varepsilon)\, \varepsilon_{\mathrm{C2}}(m, d, \Lambda, \delta). \tag{4}$$

The second theorem is the sketched analogue of the empirical path theorem. It says that some block endpoint has sketched risk close to the best all-columns sketched predictor. Its proof is also deferred to Appendix B.

Theorem 4.6 (Sketched empirical path theorem)**.** *Assume Assumption 2.1. Let $K := \lfloor D/M \rfloor$ and partition the first KM nodes of the path into K disjoint blocks of length M. Let $\Gamma_{\mathrm{path}} :=$*

$2AX_{\max}Y_{\max}\sqrt{\frac{M}{K}}$. *On the event that S is an (ε, V)-embedding, there is a block endpoint b^* such that*

$$\widetilde{\mathcal{R}}_{m,S}(u^{\mathrm{sk}}_{b^*}) \le \widetilde{\mathcal{R}}_{m,S}(\widetilde{f}_A) + (1+\varepsilon)\Gamma_{\mathrm{path}}.$$

The same bound holds with $\widetilde{f}_A$ replaced by f^*_A.

Combining Theorem 4.5 and Theorem 4.6 gives the main theorem for the shared-sketch protocol.

Theorem 4.7 (Finite-sample theorem for the shared-sketch protocol)**.** *Assume Assumption 2.1 and Assumption 2.3. Let $K := \lfloor D/M \rfloor, \Gamma_{\mathrm{path}} := 2AX_{\max}Y_{\max}\sqrt{\frac{M}{K}}$. Define $\Lambda_{\mathrm{sk}}(\varepsilon) := \sqrt{\frac{2d}{\lambda_0}}\left(1+\sqrt{\frac{1+\varepsilon}{1-\varepsilon}}\right)Y_{\max}$, and let $\Lambda := \max\{A, \Lambda_{\mathrm{sk}}(\varepsilon)\}$. On the joint event that*

(i) S is an (ε, V)-embedding,

(ii) $\lambda_{\min}(m^{-1}X^\top X) \ge \lambda_0/2$, and

(iii) the uniform-convergence event of Theorem 3.1 holds for $\mathcal{H}_\Lambda$,

there exists a block endpoint b^* such that the implicit linear predictor $\widetilde{f}_{b^*}(x) := \widetilde{\beta}^\top_{b^*}x$ satisfies

$$\mathcal{R}(\widetilde{f}_{b^*}) \le \frac{1+\varepsilon}{1-\varepsilon}\,\mathcal{R}(f^*_A) + \frac{(1+\varepsilon)\,\Gamma_{\mathrm{path}}}{1-\varepsilon} + \frac{2}{1-\varepsilon}\,\varepsilon_{\mathrm{C2}}(m,d,\Lambda,\delta).$$

The same bound holds for every later node $t \ge b^*$ on the path.

Remark 4.8 (Multiplicative versus additive form)*.* The leading sketching correction in Theorem 4.7 scales with $\mathcal{R}(f^*_A)$, the risk of the benchmark predictor. If we instead used a purely additive bound, the sketching error would scale with the worst-case quantity $(\Lambda X_{\max} + Y_{\max})^2$. The multiplicative form is therefore at least as good in the worst case and can be much better when f^*_A has small risk.

Proof of Theorem 4.7. By Theorem 4.6, there is a block endpoint b^* such that

$$\widetilde{\mathcal{R}}_{m,S}(u^{\mathrm{sk}}_{b^*}) \le \widetilde{\mathcal{R}}_{m,S}(f^*_A) + (1+\varepsilon)\Gamma_{\mathrm{path}}.$$

By Lemma B.2,

$$u^{\mathrm{sk}}_{b^*} = X\widetilde{\beta}_{b^*}, \qquad \left\|\widetilde{\beta}_{b^*}\right\|_1 \le \Lambda_{\mathrm{sk}}(\varepsilon) \le \Lambda.$$

Hence $\widetilde{f}_{b^*} \in \mathcal{H}_\Lambda$. Also $f^*_A \in \mathcal{H}_A \subseteq \mathcal{H}_\Lambda$, so Theorem 4.5 applies to both predictors.

Since

$$\widetilde{\mathcal{R}}_{m,S}(\widetilde{f}_{b^*}) = \widetilde{\mathcal{R}}_{m,S}(u^{\mathrm{sk}}_{b^*}),$$

Theorem 4.5 gives

$$\mathcal{R}(\widetilde{f}_{b^*}) \le \frac{1}{1-\varepsilon}\widetilde{\mathcal{R}}_{m,S}(u^{\mathrm{sk}}_{b^*}) + \varepsilon_{\mathrm{C2}}(m,d,\Lambda,\delta).$$

Using the sketched path bound,

$$\mathcal{R}(\widetilde{f}_{b^*}) \le \frac{1}{1-\varepsilon}\Big(\widetilde{\mathcal{R}}_{m,S}(f^*_A) + (1+\varepsilon)\Gamma_{\mathrm{path}}\Big) + \varepsilon_{\mathrm{C2}}(m,d,\Lambda,\delta).$$

Applying Theorem 4.5 again to f^*_A,

$$\widetilde{\mathcal{R}}_{m,S}(f^*_A) \le (1+\varepsilon)\mathcal{R}(f^*_A) + (1+\varepsilon)\varepsilon_{\mathrm{C2}}(m,d,\Lambda,\delta).$$

Substituting,

$$\mathcal{R}(\widetilde{f}_{b^*}) \le \frac{1+\varepsilon}{1-\varepsilon}\mathcal{R}(f^*_A) + \frac{(1+\varepsilon)\Gamma_{\mathrm{path}}}{1-\varepsilon} + \frac{2}{1-\varepsilon}\varepsilon_{\mathrm{C2}}(m,d,\Lambda,\delta).$$

For any later node $t \ge b^*$, Lemma B.2 gives $\widetilde{f}_t \in \mathcal{H}_\Lambda$, and sketched monotonicity gives

$$\widetilde{\mathcal{R}}_{m,S}(u^{\mathrm{sk}}_t) \le \widetilde{\mathcal{R}}_{m,S}(u^{\mathrm{sk}}_{b^*}).$$

The same argument therefore gives the same population-risk bound for $\widetilde{f}_t$. □

4.4 Comparison with row subsampling

We briefly compare the shared-sketch protocol with row subsampling at the same per-edge communication budget k. In row subsampling, agents keep k public rows and run the finite-sample protocol on those rows. Thus the communicated vector has length k, but the effective sample size in the generalization term is also k. On the analogous covariance and uniform-convergence events for the k-row sample, Theorem 3.3 gives

$$\mathcal{R}(\widehat{f}_{b^*}) - \mathcal{R}(f_A^*) \le O(\Gamma_{\mathrm{path}}) + \widetilde{O}\left(\frac{d}{\sqrt{k}}\right),$$

where $X_{\max}, Y_{\max}$, and λ_0 are treated as constants.

For Gaussian sketches, Lemma 4.4 and Theorem 4.7 give, for

$$\varepsilon \asymp \sqrt{\frac{d + \log(1/\delta_{\mathrm{ose}})}{k}},$$

the bound

$$\mathcal{R}(\widetilde{f}_{b^*}) - \mathcal{R}(f_A^*) \le O(\Gamma_{\mathrm{path}}) + \widetilde{O}\left(\frac{d}{\sqrt{m}}\right) + O\left(\mathcal{R}(f_A^*)\sqrt{\frac{d + \log(1/\delta_{\mathrm{ose}})}{k}}\right),$$

up to lower-order terms. The key difference is that row sampling pays a statistical price at sample size k, while sketching keeps the statistical term at sample size m and pays a separate compression cost. Subject to the embedding requirement, a useful budget makes the compression term no larger than Γ_{path}, provided the statistical term is already small.

In many natural regimes, the all-columns benchmark risk $\mathcal{R}(f_A^*)$ is small. For example, this happens when the best all-features linear predictor is nearly realizable, or when the noise around the best all-features predictor is low. In such regimes, the leading k-dependent term in the sketched bound, $O\left(\mathcal{R}(f_A^*)\sqrt{\frac{d+\log(1/\delta_{\mathrm{ose}})}{k}}\right)$, is much smaller than the row-sampling term $\widetilde{O}\left(\frac{d}{\sqrt{k}}\right)$. Thus, in the typical regime where the all-columns benchmark is accurate, the sketch bound is expected to be better than the row-sampling bound. When $\mathcal{R}(f_A^*)$ is large, matching this benchmark need not imply small absolute prediction error.

Even in the worst case, since $0 \in \mathcal{H}_A$, $\mathcal{R}(f_A^*) \le \mathcal{R}(0) = \mathbb{E}[y^2] \le Y_{\max}^2$. Therefore, when $Y_{\max}$ is treated as a constant, the leading k-dependent term in the sketched bound is at most $\widetilde{O}\left(\sqrt{\frac{d}{k}}\right)$, whereas row sampling gives $\widetilde{O}\left(\frac{d}{\sqrt{k}}\right)$. So, at the level of the stated upper bounds, sketching has better leading dependence on d and k, while also separating the finite-sample term $\widetilde{O}(d/\sqrt{m})$ from the communication-dependent term. See Appendix B.5 for a more detailed comparison.

The data-generating relationship need not be linear; the benchmark and local learners are linear. Extending the protocol to adaptive nonlinear learners remains future work, since their residuals need not lie in one fixed low-dimensional subspace.

Acknowledgments and Disclosure of Funding

The work is partially supported by DARPA expMath, ONR MURI 2024 award on Algorithms, Learning, and Game Theory, Army-Research Laboratory (ARL) grant W911NF2410052, NSF AF:Small grants 2218678, 2114269, 2347322. The authors declare no competing interests.

References

Scott Aaronson. The complexity of agreement. In *Proceedings of the Thirty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '05, page 634–643, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1581139608. doi: 10.1145/1060590.1060686. URL https://doi.org/10.1145/1060590.1060686.

Daron Acemoglu, Munther A. Dahleh, Ilan Lobel, and Asuman Ozdaglar. Bayesian learning in social networks. *The Review of Economic Studies*, 78(4):1201–1236, October 2011. doi: 10.1093/restud/rdr004. URL https://doi.org/10.1093/restud/rdr004.

Mohammad Akbarpour, Amin Saberi, and Ali Shameli. Information aggregation in overlapping generations, September 2017. URL https://ssrn.com/abstract=3035178. Available at SSRN.

Robert J. Aumann. Agreeing to disagree. *The Annals of Statistics*, 4(6):1236–1239, November 1976. doi: 10.1214/aos/1176343654. URL https://doi.org/10.1214/aos/1176343654.

Venkatesh Bala and Sanjeev Goyal. Learning from neighbours. *The Review of Economic Studies*, 65(3):595–621, 1998. ISSN 00346527, 1467937X. URL http://www.jstor.org/stable/2566940.

Abhijit Banerjee. A simple model of herd behavior. *The Quarterly Journal of Economics*, 107 (3):797–817, 1992. URL https://EconPapers.repec.org/RePEc:oup:qjecon:v:107:y:1992:i:3:p:797-817.

Burak Bartan and Mert Pilanci. Distributed sketching for randomized optimization: Exact characterization, concentration, and lower bounds. *IEEE Transactions on Information Theory*, 69(6): 3850–3879, 2023.

MohammadHossein Bateni, Zahra Hadizadeh, MohammadTaghi Hajiaghayi, Mahdi JafariRaviz, and Shayan Taherijam. Networked information aggregation for binary classification, 2026a. URL https://arxiv.org/abs/2605.01082.

MohammadHossein Bateni, Zahra Hadizadeh, MohammadTaghi Hajiaghayi, Mahdi JafariRaviz, and Shayan Taherijam. Optimal networks for agentic information aggregation, 2026b. URL https://arxiv.org/abs/2609.35537.

MohammadHossein Bateni, Zahra Hadizadeh, MohammadTaghi Hajiaghayi, Mahdi JafariRaviz, and Shayan Taherijam. Optimal rates for agentic networked information aggregation, 2026c. URL https://arxiv.org/abs/2609.05318.

Yuanchen Bei, Weizhi Zhang, Siwen Wang, Weizhi Chen, Sheng Zhou, Hao Chen, Yong Li, Jiajun Bu, Shirui Pan, Yizhou Yu, Irwin King, Fakhri Karray, and Philip S. Yu. Graphs meet ai agents: Taxonomy, progress, and future opportunities, 2025. URL https://arxiv.org/abs/2506.18019.

Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5):992–1026, 1992. ISSN 00223808, 1537534X. URL http://www.jstor.org/stable/2138632.

Timothy Castiglia, Yi Zhou, Shiqiang Wang, Swanand Kadhe, Nathalie Baracaldo, and Stacy Patterson. LESS-VFL: Communication-efficient feature selection for vertical federated learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 3757–3781. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/castiglia23a.html.

Timothy J Castiglia, Anirban Das, Shiqiang Wang, and Stacy Patterson. Compressed-VFL: Communication-efficient learning with vertically partitioned data. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2738–2766. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/castiglia22a.html.

Kenneth L. Clarkson and David P. Woodruff. Low rank approximation and regression in input sparsity time. In *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, STOC '13, page 81–90, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450320290. doi: 10.1145/2488608.2488620. URL https://doi.org/10.1145/2488608.2488620.

Morris H. DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69(345): 118–121, 1974. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2285509.

Glenn Ellison and Drew Fudenberg. Rules of thumb for social learning. *Journal of Political Economy*, 101(4):612–643, 1993. ISSN 00223808, 1537534X. URL http://www.jstor.org/stable/2138741.

Rafael Frongillo, Eric Neyman, and Bo Waggoner. Agreement implies accuracy for substitutable signals. In *Proceedings of the 24th ACM Conference on Economics and Computation*, EC '23, page 702–733, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701047. doi: 10.1145/3580507.3597679. URL https://doi.org/10.1145/3580507.3597679.

Fangcheng Fu, Xupeng Miao, Jiawei Jiang, Huanran Xue, and Bin Cui. Towards communication-efficient vertical federated learning training via cache-enabled local updates. *Proceedings of the VLDB Endowment*, 15(10):2111–2120, 2022. doi: 10.14778/3547305.3547316.

John D. Geanakoplos and Heraklis M. Polemarchakis. We can't disagree forever. *Journal of Economic Theory*, 28(1):192–200, 1982. ISSN 0022-0531. doi: 10.1016/0022-0531(82)90099-0. URL https://doi.org/10.1016/0022-0531(82)90099-0.

Benjamin Golub and Matthew O. Jackson. Naïve learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, 2(1):112–49, February 2010. doi: 10.1257/mic.2.1.112. URL https://www.aeaweb.org/articles?id=10.1257/mic.2.1.112.

Ali Jadbabaie, Pooya Molavi, Alvaro Sandroni, and Alireza Tahbaz-Salehi. Non-bayesian social learning. *Games and Economic Behavior*, 76(1):210–225, 2012. ISSN 0899-8256. doi: 10.1016/j.geb.2012.06.001. URL https://doi.org/10.1016/j.geb.2012.06.001.

Michael Kearns, Aaron Roth, and Emily Ryu. Networked information aggregation via machine learning. In *Proceedings of the 2026 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 4799–4845. SIAM, 2026.

A. Khan, M. ten Thij, and A. Wilbik. Communication-efficient vertical federated learning. *Algorithms*, 15(8):273, 2022. doi: 10.3390/a15080273. URL https://doi.org/10.3390/a15080273.

Yuqing Kong and Grant Schoenebeck. False Consensus, Information Theory, and Prediction Markets. In Yael Tauman Kalai, editor, *14th Innovations in Theoretical Computer Science Conference (ITCS 2023)*, volume 251 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 81:1–81:23, Dagstuhl, Germany, 2023. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-263-1. doi: 10.4230/LIPIcs.ITCS.2023.81. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ITCS.2023.81.

Elchanan Mossel, Allan Sly, and Omer Tamuz. Strategic learning and the topology of social networks. *CoRR*, abs/1209.5527, 2012. URL http://arxiv.org/abs/1209.5527.

Elchanan Mossel, Manuel Mueller-Frank, Allan Sly, and Omer Tamuz. Social learning equilibria. *Econometrica*, 88(3):1235–1267, 2020. ISSN 0012-9682. doi: 10.3982/ecta16465. URL http://dx.doi.org/10.3982/ECTA16465.

Mert Pilanci and Martin J. Wainwright. Randomized sketches of convex programs with sharp guarantees. *CoRR*, abs/1404.7203, 2014a. URL http://arxiv.org/abs/1404.7203.

Mert Pilanci and Martin J. Wainwright. Iterative hessian sketch: Fast and accurate solution approximation for constrained least-squares, 2014b. URL https://arxiv.org/abs/1411.0347.

Shahin Shahrampour and Ali Jadbabaie. Exponentially fast parameter estimation in networks using distributed dual averaging. In *52nd IEEE Conference on Decision and Control*, pages 6196–6201, 2013. doi: 10.1109/CDC.2013.6760868.

Jingwei Sun, Ziyue Xu, Dong Yang, Vishwesh Nath, Wenqi Li, Can Zhao, Daguang Xu, Yiran Chen, and Holger R Roth. Communication-efficient vertical federated learning with limited overlapping samples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5203–5212, 2023.

Pedro Valdeira, João Xavier, Cláudia Soares, and Yuejie Chi. Communication-efficient vertical federated learning via compressed error feedback. *IEEE Transactions on Signal Processing*, 2025.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices., 2012.

David P. Woodruff. Sketching as a tool for numerical linear algebra. *CoRR*, abs/1411.4357, 2014. URL `http://arxiv.org/abs/1411.4357`.

Liu Yang, Di Chai, Junxue Zhang, Yilun Jin, Leye Wang, Hao Liu, Han Tian, Qian Xu, and Kai Chen. A survey on vertical federated learning: From a layered perspective, 2023. URL `https://arxiv.org/abs/2304.01829`.

Chaojia Yu, Zihan Cheng, Hanwen Cui, Yishuo Gao, Zexu Luo, Yijin Wang, Hangbin Zheng, and Yong Zhao. A survey on agent workflow–status and future. In *2025 8th International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pages 770–781. IEEE, 2025.

Ling Yue, Kushal Raj Bhandari, Ching-Yun Ko, Dhaval Patel, Shuxin Lin, Nianjun Zhou, Jianxi Gao, Pin-Yu Chen, and Shaowu Pan. From static templates to dynamic runtime graphs: A survey of workflow optimization for llm agents. *arXiv preprint arXiv:2603.22386*, 2026.

A Deferred proofs for the finite-sample analysis

This appendix proves Theorem 3.2 and Lemma A.7. These results are deterministic statements about the realized sample D_m. The final subsection proves Corollary 2.2, which concerns random feature allocation.

A.1 Least-squares identities

We first collect the basic identities used in the empirical path proof.

Lemma A.1 (Linearity of sequential empirical predictors)**.** *For every agent p_t on the path, there exists $\widehat{\beta}_t \in \mathbb{R}^d$ such that*

$$u_t = X\widehat{\beta}_t.$$

Equivalently, the implicit predictor $\widehat{f}_t(x) = \widehat{\beta}_t^\top x$ is a linear function of the original d features.

Proof. Proceed by induction over the topological order of the graph. At a source node, u_i is a linear combination of the columns of X_{S_i}. At a non-source node,

$$u_i = X_{S_i}\widehat{w}_i + \sum_{j \in \mathrm{Pa}(i)} (\widehat{a}_i)_j u_j.$$

By induction, each parent vector u_j is a linear combination of the columns of X. Therefore u_i is also a linear combination of the columns of X. □

For each path node p_t, define

$$Z_t := \big[\, X_{S_t} \quad u_j : j \in \mathrm{Pa}(p_t) \,\big], \qquad r_t := y - u_t.$$

Lemma A.2 (Empirical orthogonality)**.** *For every t,*

$$Z_t^\top r_t = 0.$$

Consequently,

(i) for every coordinate $\ell \in S_t$, $X_\ell^\top r_t = 0$;

(ii) for $t \geq 2$, $u_{t-1}^\top r_t = 0$;

(iii) $u_t^\top r_t = 0$.

Proof. By definition, $u_t = Z_t\widehat{\theta}_t$, where $\widehat{\theta}_t$ minimizes

$$\frac{1}{m}\left\|Z_t\theta - y\right\|_2^2.$$

The normal equations give

$$Z_t^\top(Z_t\widehat{\theta}_t - y) = 0,$$

which is equivalent to $Z_t^\top r_t = 0$. The three consequences follow because X_ℓ, u_{t-1}, and u_t all lie in $\mathrm{span}(Z_t)$ whenever the corresponding statement applies. □

Lemma A.3 (Empirical monotonicity along the path)**.** *For every $t \geq 2$,*

$$\widehat{\mathcal{R}}_m(u_t) \leq \widehat{\mathcal{R}}_m(u_{t-1}).$$

Proof. Agent p_t can copy p_{t-1} by placing coefficient 1 on u_{t-1} and coefficient 0 on all other parent predictions and local features. Since u_t is chosen by least squares, its empirical risk cannot be larger. □

Lemma A.4 (Empirical Pythagorean identity)**.** *For every $t \geq 2$,*

$$\frac{1}{m}\left\|u_t - u_{t-1}\right\|_2^2 = \widehat{\mathcal{R}}_m(u_{t-1}) - \widehat{\mathcal{R}}_m(u_t).$$

Proof. Since $y - u_{t-1} = r_t + (u_t - u_{t-1})$,

$$\left\|y - u_{t-1}\right\|_2^2 = \left\|r_t\right\|_2^2 + \left\|u_t - u_{t-1}\right\|_2^2 + 2\left\langle r_t, u_t - u_{t-1}\right\rangle.$$

Both u_t and u_{t-1} lie in $\mathrm{span}(Z_t)$, so Lemma A.2 gives $\langle r_t, u_t - u_{t-1}\rangle = 0$. Dividing by m proves the claim. □

Lemma A.5 (Empirical squared-loss decomposition)**.** *For any $u, g \in \mathbb{R}^m$, with $r := y - u$,*

$$\widehat{\mathcal{R}}_m(u) = \widehat{\mathcal{R}}_m(g) + \frac{2}{m}\left\langle g, r\right\rangle - \frac{2}{m}\left\langle u, r\right\rangle - \frac{1}{m}\left\|u - g\right\|_2^2.$$

Proof. Since $y - g = (y - u) + (u - g) = r + (u - g)$,

$$\left\|y - g\right\|_2^2 = \left\|r\right\|_2^2 + 2\left\langle r, u - g\right\rangle + \left\|u - g\right\|_2^2.$$

Dividing by m and rearranging gives

$$\widehat{\mathcal{R}}_m(u) = \widehat{\mathcal{R}}_m(g) - \frac{2}{m}\left\langle r, u - g\right\rangle - \frac{1}{m}\left\|u - g\right\|_2^2.$$

Expanding $\langle r, u - g\rangle = \langle u, r\rangle - \langle g, r\rangle$ proves the identity. □

A.2 The empirical block theorem and proof of the empirical path theorem

For a contiguous block $B = [a, b]$ of nodes on the path, define

$$\mathcal{U}_B := \{X_\ell : \ell \in \textstyle\bigcup_{t=a}^{b} S_t\} \cup \{u_j : j \in \mathrm{Pa}(p_t),\ t \in [a, b]\}.$$

For a representation $g = \sum_{v\in\mathcal{U}_B} \alpha_v v$, define

$$\left\|\alpha\right\|_{1,B} := \sum_{v\in\mathcal{U}_B} |\alpha_v| \frac{\left\|v\right\|_2}{\sqrt{m}}.$$

Theorem A.6 (Empirical block theorem)**.** *Let $B = [a, b] \subseteq \{1, \ldots, D\}$ be a contiguous block on the path, and let*

$$\Delta_B := \widehat{\mathcal{R}}_m(u_a) - \widehat{\mathcal{R}}_m(u_b) \geq 0.$$

Let $g = \sum_{v\in\mathcal{U}_B} \alpha_v v$. Then

$$\widehat{\mathcal{R}}_m(u_b) \leq \widehat{\mathcal{R}}_m(g) + 2\left\|\alpha\right\|_{1,B}\sqrt{|B|\Delta_B}.$$

Proof. Fix $v \in \mathcal{U}_B$. By definition, there is $t = t(v) \in [a, b]$ such that v is either a local feature column available at p_t or a parent prediction available at p_t. Hence $v \in \operatorname{span}(Z_t)$, and Lemma A.2 gives

$$v^\top (y - u_t) = 0.$$

Therefore

$$\frac{1}{m} v^\top (y - u_b) = \frac{1}{m} v^\top (u_t - u_b).$$

By Cauchy–Schwarz,

$$\left| \frac{1}{m} v^\top (y - u_b) \right| \le \frac{\|v\|_2}{\sqrt{m}} \cdot \frac{\|u_b - u_t\|_2}{\sqrt{m}}.$$

Using

$$u_b - u_t = \sum_{s=t+1}^{b} (u_s - u_{s-1}),$$

the triangle inequality, Lemma A.4, and Cauchy–Schwarz give

$$\begin{aligned}
\frac{\|u_b - u_t\|_2}{\sqrt{m}} &\le \sum_{s=t+1}^{b} \frac{\|u_s - u_{s-1}\|_2}{\sqrt{m}} \\
&= \sum_{s=t+1}^{b} \sqrt{\widehat{\mathcal{R}}_m(u_{s-1}) - \widehat{\mathcal{R}}_m(u_s)} \\
&\le \sqrt{(b-t) \sum_{s=a+1}^{b} \left(\widehat{\mathcal{R}}_m(u_{s-1}) - \widehat{\mathcal{R}}_m(u_s) \right)} \\
&\le \sqrt{|B| \Delta_B}.
\end{aligned}$$

Thus

$$\left| \frac{1}{m} v^\top (y - u_b) \right| \le \frac{\|v\|_2}{\sqrt{m}} \sqrt{|B| \Delta_B}.$$

Multiplying by $|\alpha_v|$ and summing over $v \in \mathcal{U}_B$,

$$\left| \frac{1}{m} \langle g, y - u_b \rangle \right| \le \|\alpha\|_{1,B} \sqrt{|B| \Delta_B}.$$

Now apply Lemma A.5 with $u = u_b$. By Lemma A.2, $u_b^\top (y - u_b) = 0$, and the term $-m^{-1} \|u_b - g\|_2^2$ is nonpositive. Therefore

$$\widehat{\mathcal{R}}_m(u_b) \le \widehat{\mathcal{R}}_m(g) + 2 \|\alpha\|_{1,B} \sqrt{|B| \Delta_B}.$$

□

Proof of Theorem 3.2. By Lemma A.3, the empirical risks are nonincreasing. Hence the total drop across the K disjoint blocks is at most

$$\widehat{\mathcal{R}}_m(u_1) \le \widehat{\mathcal{R}}_m(0) \le Y_{\max}^2.$$

Therefore one block $B^* = [a^*, b^*]$ satisfies

$$\Delta_{B^*} \le \frac{Y_{\max}^2}{K}.$$

By Assumption 2.1, the features in B^* cover all of $[d]$. Thus

$$\widehat{f}_A = X \widehat{\beta}_A, \qquad \left\| \widehat{\beta}_A \right\|_1 \le A,$$

is representable in $\mathcal{U}_{B^*}$ using only raw feature columns. For each raw feature column X_ℓ,

$$\frac{\|X_\ell\|_2}{\sqrt{m}} \le X_{\max}.$$

Hence the representation of $\widehat{f}_A$ satisfies

$$\left\|\widehat{\beta}_A\right\|_{1,B^*} \le AX_{\max}.$$

Applying Theorem A.6 with $g = \widehat{f}_A$ and $|B^*| = M$, we get

$$\widehat{\mathcal{R}}_m(u_{b^*}) \le \widehat{\mathcal{R}}_m(\widehat{f}_A) + 2AX_{\max}Y_{\max}\sqrt{\frac{M}{K}}.$$

The bound for later nodes follows from Lemma A.3. □

A.3 Norm control

To apply Theorem 3.1 to the network predictors, we need one common ℓ_1-ball that contains all predictors produced along the path. The next lemma gives such a bound under the empirical covariance event.

Lemma A.7 (ℓ_1-norm control)**.** *Assume*

$$\lambda_{\min}(m^{-1}X^\top X) \ge \lambda_0/2.$$

Then every network predictor on the path admits a representation $u_t = X\widehat{\beta}_t$ with

$$\left\|\widehat{\beta}_t\right\|_2 \le \sqrt{\frac{2}{\lambda_0}}\,Y_{\max}, \qquad \left\|\widehat{\beta}_t\right\|_1 \le \Lambda_{\text{net}} := \sqrt{\frac{2d}{\lambda_0}}\,Y_{\max}.$$

Proof. By Lemma A.1, write $u_t = X\widehat{\beta}_t$. Since u_t is the orthogonal projection of y onto $\operatorname{span}(Z_t)$,

$$\|u_t\|_2 \le \|y\|_2 .$$

Therefore

$$\widehat{\beta}_t^\top \left(\frac{1}{m}X^\top X\right)\widehat{\beta}_t = \frac{1}{m}\|u_t\|_2^2 \le \frac{1}{m}\|y\|_2^2 \le Y_{\max}^2.$$

The empirical covariance lower bound gives

$$\left\|\widehat{\beta}_t\right\|_2 \le \sqrt{\frac{2}{\lambda_0}}\,Y_{\max}.$$

The ℓ_1-bound follows from

$$\left\|\widehat{\beta}_t\right\|_1 \le \sqrt{d}\,\left\|\widehat{\beta}_t\right\|_2 .$$

□

A.4 Proof of the random-coverage corollary

Proof of Corollary 2.2. Fix a feature $\ell \in [d]$ and a block of M consecutive agents on P. By independence across agents, the probability that none of these agents observes ℓ is $(1-p)^M$. There are $D - M + 1$ such blocks and d features. A union bound over all feature–block pairs gives

$$\begin{aligned}\Pr[\text{Assumption 2.1 fails on } P] &\le d(D - M + 1)(1-p)^M \\ &\le dD(1-p)^M \\ &\le \delta_{\text{cover}},\end{aligned}$$

where the last inequality follows from the stated lower bound on M. The union bound does not require independence between overlapping blocks. □

B Deferred proofs for the shared-sketch protocol

This appendix contains the deferred proofs for Section 4.

B.1 Gaussian subspace embeddings

We first record the standard Gaussian oblivious-subspace-embedding statement used in Lemma 4.4.

Lemma B.1 (Gaussian sketches give subspace embeddings)**.** *Let $W \subseteq \mathbb{R}^m$ be a fixed subspace with $\dim(W) = r$. Let $S \in \mathbb{R}^{k\times m}$ have i.i.d. entries $N(0, 1/k)$. There is a universal constant $C_{\mathrm{ose}} > 0$ such that, if*

$$k \geq C_{\mathrm{ose}} \frac{r + \log(1/\delta_{\mathrm{ose}})}{\varepsilon^2},$$

then S is an (ε, W)-embedding with probability at least $1 - \delta_{\mathrm{ose}}$ over the draw of S.

Proof. Let $Q \in \mathbb{R}^{m\times r}$ have orthonormal columns spanning W. Every $v \in W$ can be written as $v = Qz$ for some $z \in \mathbb{R}^r$, and $\|v\|_2 = \|z\|_2$. Therefore S is an (ε, W)-embedding if and only if

$$(1-\varepsilon)\|z\|_2^2 \leq \|SQz\|_2^2 \leq (1+\varepsilon)\|z\|_2^2 \qquad \text{for all } z \in \mathbb{R}^r.$$

Equivalently,

$$\left\|Q^\top S^\top SQ - I_r\right\|_{\mathrm{op}} \leq \varepsilon.$$

Since Q is fixed and has orthonormal columns, SQ is a $k \times r$ Gaussian matrix with i.i.d. entries $N(0, 1/k)$. The standard non-asymptotic singular-value bound for Gaussian matrices gives

$$\left\|Q^\top S^\top SQ - I_r\right\|_{\mathrm{op}} \leq \varepsilon$$

with probability at least $1 - \delta_{\mathrm{ose}}$, provided

$$k \geq C_{\mathrm{ose}} \frac{r + \log(1/\delta_{\mathrm{ose}})}{\varepsilon^2}.$$

This proves the claim. □

Proof of Lemma 4.4. The subspace

$$V = \mathrm{span}\{X_1, \ldots, X_d, y\}$$

is fixed and has dimension at most $d + 1$. Apply Lemma B.1 with $W = V$ and $r = d + 1$. □

B.2 Proof of the sketched generalization theorem

Proof of Theorem 4.5. Fix $f_\beta \in \mathcal{H}_\Lambda$. Since $V = \mathrm{span}\{X_1, \ldots, X_d, y\}$, we have $X\beta - y \in V$. Hence, by Lemma 4.3,

$$(1-\varepsilon)\widehat{\mathcal{R}}_m(f_\beta) \leq \widetilde{\mathcal{R}}_{m,S}(f_\beta) \leq (1+\varepsilon)\widehat{\mathcal{R}}_m(f_\beta). \tag{5}$$

On the uniform-convergence event of Theorem 3.1,

$$\left|\mathcal{R}(f_\beta) - \widehat{\mathcal{R}}_m(f_\beta)\right| \leq \varepsilon_{\mathrm{C2}}(m, d, \Lambda, \delta). \tag{6}$$

By (6), $\mathcal{R}(f_\beta) \leq \widehat{\mathcal{R}}_m(f_\beta) + \varepsilon_{\mathrm{C2}}(m, d, \Lambda, \delta)$. By the left inequality in (5), $\widehat{\mathcal{R}}_m(f_\beta) \leq \frac{1}{1-\varepsilon}\widetilde{\mathcal{R}}_{m,S}(f_\beta)$. Therefore $\mathcal{R}(f_\beta) \leq \frac{1}{1-\varepsilon}\widetilde{\mathcal{R}}_{m,S}(f_\beta) + \varepsilon_{\mathrm{C2}}(m, d, \Lambda, \delta)$, which proves (3).

Similarly, by the right inequality in (5), $\widetilde{\mathcal{R}}_{m,S}(f_\beta) \leq (1+\varepsilon)\widehat{\mathcal{R}}_m(f_\beta)$. Using (6) again, $\widehat{\mathcal{R}}_m(f_\beta) \leq \mathcal{R}(f_\beta) + \varepsilon_{\mathrm{C2}}(m, d, \Lambda, \delta)$. Thus $\widetilde{\mathcal{R}}_{m,S}(f_\beta) \leq (1+\varepsilon)\mathcal{R}(f_\beta) + (1+\varepsilon)\varepsilon_{\mathrm{C2}}(m, d, \Lambda, \delta)$, which proves (4). Since both events hold uniformly over $f_\beta \in \mathcal{H}_\Lambda$, the two inequalities hold simultaneously for all $f \in \mathcal{H}_\Lambda$. □

B.3 Proof of the sketched empirical path theorem

The sketched protocol is the empirical least-squares protocol applied to the compressed data matrix SX and compressed label vector Sy. Therefore the proof of Theorem A.6 applies verbatim after replacing every vector v by its sketch Sv, and after measuring dictionary size by the normalized sketched norm $\|Sv\|_2/\sqrt{m}$.

For a block $B = [a, b]$, define

$$\Delta_B^{\mathrm{sk}} := \widetilde{\mathcal{R}}_{m,S}(u_a^{\mathrm{sk}}) - \widetilde{\mathcal{R}}_{m,S}(u_b^{\mathrm{sk}}).$$

Let

$$\mathcal{U}_B^{\mathrm{sk}} := \{X_\ell : \ell \in \textstyle\bigcup_{t=a}^{b} S_t\} \cup \{u_j^{\mathrm{sk}} : j \in \mathrm{Pa}(p_t),\ t \in [a, b]\}.$$

If $g = \sum_{v \in \mathcal{U}_B^{\mathrm{sk}}} \alpha_v v$, define

$$\|\alpha\|_{1,S,B} := \sum_{v \in \mathcal{U}_B^{\mathrm{sk}}} |\alpha_v| \, \frac{\|Sv\|_2}{\sqrt{m}}.$$

The sketched version of Theorem A.6 is

$$\widetilde{\mathcal{R}}_{m,S}(u_b^{\mathrm{sk}}) \le \widetilde{\mathcal{R}}_{m,S}(g) + 2\,\|\alpha\|_{1,S,B}\,\sqrt{|B|\Delta_B^{\mathrm{sk}}}. \tag{7}$$

This is the same block theorem applied in sketch space.

Proof of Theorem 4.6. Partition the first KM nodes into K disjoint blocks of length M. The sketched risk is nonincreasing along the path, since each node can copy the sketched prediction of its path predecessor. Hence

$$\sum_{\text{blocks } B} \Delta_B^{\mathrm{sk}} \le \widetilde{\mathcal{R}}_{m,S}(u_1^{\mathrm{sk}}) \le \widetilde{\mathcal{R}}_{m,S}(0) = \frac{1}{m}\,\|Sy\|_2^2\,.$$

Since $y \in V$ and S is an (ε, V)-embedding,

$$\frac{1}{m}\,\|Sy\|_2^2 \le (1+\varepsilon)\frac{1}{m}\,\|y\|_2^2 \le (1+\varepsilon)Y_{\max}^2.$$

Thus there exists a block $B^* = [a^*, b^*]$ such that

$$\Delta_{B^*}^{\mathrm{sk}} \le \frac{(1+\varepsilon)Y_{\max}^2}{K}.$$

By Assumption 2.1, the features in B^* cover all of $[d]$. Therefore

$$\widetilde{f}_A = X\widetilde{\beta}_A, \qquad \left\|\widetilde{\beta}_A\right\|_1 \le A,$$

is representable in $\mathcal{U}_{B^*}^{\mathrm{sk}}$ using only raw feature columns. For every feature column X_ℓ, the embedding event gives

$$\frac{\|SX_\ell\|_2}{\sqrt{m}} \le \sqrt{1+\varepsilon}\,\frac{\|X_\ell\|_2}{\sqrt{m}} \le \sqrt{1+\varepsilon}\,X_{\max}.$$

Hence

$$\left\|\widetilde{\beta}_A\right\|_{1,S,B^*} \le \sqrt{1+\varepsilon}\,AX_{\max}.$$

Applying (7) with $g = \widetilde{f}_A$ and $|B^*| = M$, we get

$$\begin{aligned}\widetilde{\mathcal{R}}_{m,S}(u_{b^*}^{\mathrm{sk}}) &\le \widetilde{\mathcal{R}}_{m,S}(\widetilde{f}_A) + 2\sqrt{1+\varepsilon}\,AX_{\max}\sqrt{M\cdot\frac{(1+\varepsilon)Y_{\max}^2}{K}} \\ &= \widetilde{\mathcal{R}}_{m,S}(\widetilde{f}_A) + (1+\varepsilon)\Gamma_{\mathrm{path}}.\end{aligned}$$

Finally, since $\widetilde{f}_A$ minimizes $\widetilde{\mathcal{R}}_{m,S}$ over $\mathcal{H}_A$ and $f_A^* \in \mathcal{H}_A$,

$$\widetilde{\mathcal{R}}_{m,S}(\widetilde{f}_A) \le \widetilde{\mathcal{R}}_{m,S}(f_A^*).$$

Thus the same bound also holds with $\widetilde{f}_A$ replaced by f_A^*. □

B.4 Linearity, monotonicity, and norm control

To apply Theorem 4.5 to the output of the sketched protocol, we need to know that the implicit sketched predictors are linear predictors with bounded ℓ_1-norm. We also use the monotonicity of sketched risk along the path.

Lemma B.2 (Linearity, monotonicity, and norm control for sketched predictors)**.** *On the joint event that S is an (ε, V)-embedding and*

$$\lambda_{\min}(m^{-1}X^\top X) \geq \lambda_0/2,$$

every implicit sketched predictor on the path admits a representation

$$u_t^{\mathrm{sk}} = X\widetilde{\beta}_t$$

with

$$\left\|\widetilde{\beta}_t\right\|_1 \leq \Lambda_{\mathrm{sk}}(\varepsilon) := \sqrt{\frac{2d}{\lambda_0}}\left(1+\sqrt{\frac{1+\varepsilon}{1-\varepsilon}}\right) Y_{\max}.$$

Moreover, the sketched empirical risk is nonincreasing along the path:

$$\widetilde{\mathcal{R}}_{m,S}(u_t^{\mathrm{sk}}) \leq \widetilde{\mathcal{R}}_{m,S}(u_{t-1}^{\mathrm{sk}}) \qquad \textit{for every } t \geq 2.$$

Proof. First we prove linearity. The proof is the same induction as in Lemma A.1. At a source node, the implicit sketched predictor is a linear combination of columns of X_{S_i}. At a non-source node,

$$u_i^{\mathrm{sk}} = X_{S_i}\widetilde{w}_i + \sum_{j\in \mathrm{Pa}(i)} (\widetilde{a}_i)_j u_j^{\mathrm{sk}}.$$

By induction, each parent predictor u_j^{sk} is a linear combination of columns of X. Therefore u_i^{sk} is also a linear combination of columns of X. Hence for every path node t, there exists $\widetilde{\beta}_t \in \mathbb{R}^d$ such that

$$u_t^{\mathrm{sk}} = X\widetilde{\beta}_t.$$

Next we prove sketched monotonicity. Since $p_{t-1} \in \mathrm{Pa}(p_t)$, agent p_t can copy the sketched prediction of its path predecessor by placing coefficient 1 on Su_{t-1}^{sk} and coefficient 0 on all other inputs. Since p_t minimizes the sketched least-squares objective, this feasible choice implies

$$\widetilde{\mathcal{R}}_{m,S}(u_t^{\mathrm{sk}}) \leq \widetilde{\mathcal{R}}_{m,S}(u_{t-1}^{\mathrm{sk}}).$$

It remains to prove the coefficient bound. Since $u_t^{\mathrm{sk}} \in \mathrm{span}\{X_1, \ldots, X_d\} \subseteq V$, the residual $y - u_t^{\mathrm{sk}}$ lies in V. By Lemma 4.3,

$$(1-\varepsilon)\widehat{\mathcal{R}}_m(u_t^{\mathrm{sk}}) \leq \widetilde{\mathcal{R}}_{m,S}(u_t^{\mathrm{sk}}).$$

By sketched monotonicity and the fact that the zero predictor is feasible at the first node,

$$\widetilde{\mathcal{R}}_{m,S}(u_t^{\mathrm{sk}}) \leq \widetilde{\mathcal{R}}_{m,S}(0) = \frac{1}{m}\left\|Sy\right\|_2^2.$$

Since $y \in V$ and S is an (ε, V)-embedding,

$$\frac{1}{m}\left\|Sy\right\|_2^2 \leq (1+\varepsilon)\frac{1}{m}\left\|y\right\|_2^2 \leq (1+\varepsilon)Y_{\max}^2.$$

Therefore

$$\frac{1}{\sqrt{m}}\left\|y - u_t^{\mathrm{sk}}\right\|_2 \leq \sqrt{\frac{1+\varepsilon}{1-\varepsilon}}\, Y_{\max}.$$

By the triangle inequality,

$$\frac{1}{\sqrt{m}}\left\|u_t^{\mathrm{sk}}\right\|_2 \leq \left(1+\sqrt{\frac{1+\varepsilon}{1-\varepsilon}}\right) Y_{\max}.$$

Since $u_t^{\mathrm{sk}} = X\widetilde{\beta}_t$,

$$\widetilde{\beta}_t^\top\left(\frac{1}{m}X^\top X\right)\widetilde{\beta}_t = \frac{1}{m}\left\|u_t^{\mathrm{sk}}\right\|_2^2.$$

Using

$$\lambda_{\min}(m^{-1}X^\top X) \geq \lambda_0/2,$$

we get

$$\left\|\widetilde{\beta}_t\right\|_2 \leq \sqrt{\frac{2}{\lambda_0}}\left(1+\sqrt{\frac{1+\varepsilon}{1-\varepsilon}}\right)Y_{\max}.$$

Finally,

$$\left\|\widetilde{\beta}_t\right\|_1 \leq \sqrt{d}\left\|\widetilde{\beta}_t\right\|_2 \leq \sqrt{\frac{2d}{\lambda_0}}\left(1+\sqrt{\frac{1+\varepsilon}{1-\varepsilon}}\right)Y_{\max} = \Lambda_{\mathrm{sk}}(\varepsilon).$$

□

B.5 Detailed comparison with row subsampling

We give the detailed comparison between row subsampling and Gaussian sketching. The main text states the leading-order conclusion; here we spell out the bounds.

First consider the row-sampling baseline. In this protocol, agents choose k public rows and run the full-message finite-sample protocol on only those rows. Thus the message length is k, but the sample size in the uniform-convergence term is also k. On the analogous empirical covariance and uniform-convergence events for the k-row sample, Theorem 3.3 gives

$$\mathcal{R}(\widehat{f}_{b^*}) - \mathcal{R}(f_A^*) \leq \Gamma_{\mathrm{path}} + 2\varepsilon_{\mathrm{C2}}(k, d, \bar{\Lambda}_{\mathrm{net}}, \delta),$$

where

$$\bar{\Lambda}_{\mathrm{net}} := \max\left\{A, \sqrt{\frac{2d}{\lambda_0}}Y_{\max}\right\}.$$

By definition,

$$\varepsilon_{\mathrm{C2}}(k, d, \bar{\Lambda}_{\mathrm{net}}, \delta) = C(\bar{\Lambda}_{\mathrm{net}}X_{\max}+Y_{\max})\bar{\Lambda}_{\mathrm{net}}X_{\max}\sqrt{\frac{\log(2d)}{k}}+(\bar{\Lambda}_{\mathrm{net}}X_{\max}+Y_{\max})^2\sqrt{\frac{\log(C'/\delta)}{2k}}.$$

If $A, X_{\max}, Y_{\max}$, and λ_0 are treated as fixed constants, then

$$\bar{\Lambda}_{\mathrm{net}} = \Theta(\sqrt{d}).$$

Hence the first term is

$$\widetilde{O}\left(\frac{d}{\sqrt{k}}\right),$$

and the second term is

$$\widetilde{O}\left(\frac{d}{\sqrt{k}}\right).$$

Thus the row-sampling guarantee has the leading-order form

$$\mathcal{R}(\widehat{f}_{b^*}) - \mathcal{R}(f_A^*) \leq O(\Gamma_{\mathrm{path}}) + \widetilde{O}\left(\frac{d}{\sqrt{k}}\right).$$

Now consider Gaussian sketches. By Lemma 4.4, if

$$k \gtrsim \frac{d+\log(1/\delta_{\mathrm{ose}})}{\varepsilon^2},$$

then S is an (ε, V)-embedding with probability at least $1-\delta_{\mathrm{ose}}$. For $\varepsilon \leq 1/2$,

$$\sqrt{\frac{1+\varepsilon}{1-\varepsilon}} \leq \sqrt{3}.$$

Therefore

$$\Lambda_{\mathrm{sk}}(\varepsilon) \leq \sqrt{\frac{2d}{\lambda_0}}(1+\sqrt{3})Y_{\max}.$$

It is enough to use the fixed radius

$$\bar{\Lambda}_{\mathrm{sk}} := \max\left\{A,\ \sqrt{\frac{2d}{\lambda_0}}(1+\sqrt{3})Y_{\max}\right\}.$$

Theorem 4.7 gives

$$\mathcal{R}(\widetilde{f}_{b^*}) \le \frac{1+\varepsilon}{1-\varepsilon}\mathcal{R}(f_A^*) + \frac{(1+\varepsilon)\Gamma_{\mathrm{path}}}{1-\varepsilon} + \frac{2}{1-\varepsilon}\varepsilon_{\mathrm{C2}}(m,d,\bar{\Lambda}_{\mathrm{sk}},\delta).$$

Subtracting $\mathcal{R}(f_A^*)$ and using

$$\frac{1+\varepsilon}{1-\varepsilon} = 1+O(\varepsilon), \qquad \frac{1}{1-\varepsilon} = O(1) \qquad \text{for } \varepsilon \le 1/2,$$

we get

$$\begin{aligned}\mathcal{R}(\widetilde{f}_{b^*}) - \mathcal{R}(f_A^*) \le{}& O(\Gamma_{\mathrm{path}}) + O\big(\varepsilon_{\mathrm{C2}}(m,d,\bar{\Lambda}_{\mathrm{sk}},\delta)\big)\\ &+ O\big(\varepsilon\big[\mathcal{R}(f_A^*) + \Gamma_{\mathrm{path}} + \varepsilon_{\mathrm{C2}}(m,d,\bar{\Lambda}_{\mathrm{sk}},\delta)\big]\big).\end{aligned}$$

Taking

$$\varepsilon \asymp \sqrt{\frac{d+\log(1/\delta_{\mathrm{ose}})}{k}},$$

this becomes

$$\begin{aligned}\mathcal{R}(\widetilde{f}_{b^*}) - \mathcal{R}(f_A^*) \le{}& O(\Gamma_{\mathrm{path}}) + O\big(\varepsilon_{\mathrm{C2}}(m,d,\bar{\Lambda}_{\mathrm{sk}},\delta)\big)\\ &+ O\left(\big[\mathcal{R}(f_A^*) + \Gamma_{\mathrm{path}} + \varepsilon_{\mathrm{C2}}(m,d,\bar{\Lambda}_{\mathrm{sk}},\delta)\big]\sqrt{\frac{d+\log(1/\delta_{\mathrm{ose}})}{k}}\right).\end{aligned}$$

Since $\bar{\Lambda}_{\mathrm{sk}} = \Theta(\sqrt{d})$ under the same constant-parameter convention,

$$\varepsilon_{\mathrm{C2}}(m,d,\bar{\Lambda}_{\mathrm{sk}},\delta) = \widetilde{O}\left(\frac{d}{\sqrt{m}}\right).$$

Thus the sketched guarantee can be summarized as

$$\mathcal{R}(\widetilde{f}_{b^*}) - \mathcal{R}(f_A^*) \le O(\Gamma_{\mathrm{path}}) + \widetilde{O}\left(\frac{d}{\sqrt{m}}\right) + O\left(\mathcal{R}(f_A^*)\sqrt{\frac{d+\log(1/\delta_{\mathrm{ose}})}{k}}\right),$$

up to lower-order terms in which Γ_{path} and $\varepsilon_{\mathrm{C2}}(m,d,\bar{\Lambda}_{\mathrm{sk}},\delta)$ are also multiplied by the sketching distortion.

This highlights the difference between the two protocols. Row sampling uses k both as the communication budget and as the statistical sample size, leading to the term

$$\widetilde{O}\left(\frac{d}{\sqrt{k}}\right).$$

Sketching uses all m samples and pays a separate compression cost. The leading k-dependent term is scaled by the benchmark risk $\mathcal{R}(f_A^*)$. In many natural regimes this risk is small, for example when the all-features linear benchmark is nearly realizable or when the noise level around the best all-features predictor is low.

Even in the worst case, the benchmark risk is bounded by the risk of the zero predictor. Since $0 \in \mathcal{H}_A$,

$$\mathcal{R}(f_A^*) \le \mathcal{R}(0) = \mathbb{E}[y^2] \le Y_{\max}^2.$$

Therefore, when $Y_{\max}$ is treated as a constant, the leading k-dependent term in the sketched bound is at most

$$\widetilde{O}\left(\sqrt{\frac{d}{k}}\right).$$

Thus, at the level of leading-order dimension and communication dependence, sketching replaces the row-sampling term

$$\widetilde{O}\left(\frac{d}{\sqrt{k}}\right)$$

by

$$\widetilde{O}\left(\sqrt{\frac{d}{k}}\right),$$

while keeping the finite-sample term separate as

$$\widetilde{O}\left(\frac{d}{\sqrt{m}}\right).$$

This comparison is at the level of upper bounds; constants, logarithmic factors, and lower-order terms can still matter for finite values of d, m, k.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All claims are accurate and proven in the text.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We provide all assumptions and discuss under what conditions our bounds would be an improvement.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All proofs are provided.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: The paper does not include any experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [N/A]

Justification: The paper does not include any experiments.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: The paper does not include any experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: The paper does not include any experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: The paper does not include any experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We conform to all points in NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: Our work is foundational research and we see no direct path to any negative applications.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not include any experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not include any experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.