Quantum ring all-reduce: communication and privacy advantages for distributed learning

Maria Gragera Garces[1] Lirandë Pira[2]

[1]University of Edinburgh, UK [2]Centre for Quantum Technologies, National University of Singapore, Singapore

Abstract

We show that a quantum communication layer enables provable advantages throughout the distributed learning pipeline, regardless of the model's nature (classical or quantum). Starting from ring all-reduce (the universal primitive for large-scale gradient aggregation) we prove that pre-shared entanglement reduces per-link *online* communication by a provably optimal factor of two and enables information-theoretic private aggregation impossible for any classical protocol. Beyond bandwidth and security, the same quantum setup gives advantages directly relevant to training: for margin-based alignment testing ($\textsc{GapIP}_\tau$), the quantum advantage is *quadratic* ($\widetilde{O}(\tau^{-1}\log P)$ qubits vs. $\widetilde{O}(\tau^{-2})$ bits); and for sign-consistency auditing against private parameter matchings ($\textsc{TieAudit}_\varepsilon$), it is *exponential* ($O(\varepsilon^{-2}\log P)$ qubits vs. $\Omega(\sqrt{P})$ bits).

1 Introduction

Ring all-reduce is the workhorse of large-scale distributed training [1–3]: N workers in a ring exchange gradient chunks until every worker holds the global mean $\bar{g} = \frac{1}{N}\sum_j g_j$, at a total cost of $2Pb$ bits per link (P parameters, b bits precision). At the scale of modern models, interconnect bandwidth is the dominant bottleneck, making any reduction in communication very valuable for training at scale.

We show that a quantum communication layer gives provable advantages in this setting, with no changes to the model, optimiser, or data. Pre-shared EPR pairs and superdense coding halve per-link *online* bandwidth optimally (a critical-path saving. (Section 2), and the same entanglement enables information-theoretic private aggregation that is classically impossible. The same quantum setup also opens up advantages directly relevant to training: gradient conflict detection tasks that arise after the all-reduce in heterogeneous training admit sharp quantum-classical separations, quadratic for margin alignment testing and exponential for relational sign auditing, purely from the structure of what clients need to compute. Related work [4, 5] requires quantum data or studies state certification; our advantages hold for standard stochastic gradient descent (SGD). A full version of this work, including proofs, can be found at `https://arxiv.org/abs/2606.20344` (Appendix).

2 Quantum ring all-reduce

Each adjacent pair pre-shares $\lceil Pb/2 \rceil$ EPR pairs before training begins, input-independently. During the all-reduce, superdense coding [6] lets every m-bit message cross a ring link in $\lceil m/2 \rceil$ qubits. The round structure and gradient arithmetic are unchanged; the *online* qubit cost per link is $Pb/2 + o(Pb)$; the entanglement is distributed beforehand, off the critical path, so total channel use matches the classical baseline.

Figure 1 shows this superdense-coding protocol worked through for $N = 3$.

Optimality: Any cut separating node sets A and B must convey $Pb - O(\log N)$ bits of entropy about $\sum_{j\in B} g_j$; pre-shared entanglement is input-independent and carries none of it. The entanglement-assisted Holevo bound caps each qubit at 2 classical bits, so $Pb/2 - O(\log N)$ qubits must cross every cut. The protocol is optimal.

GHZ as the right resource: Phase-kickback offers an equivalent formulation: each node j encodes its gradient coordinate as a phase $e^{-i\theta_j Z_j/2}$ on a shared N-qubit state; X-basis measurement and parity recovery yield $\Phi = \sum_j \theta_j$. The quantum Fisher information for any N-qubit state satisfies $F_Q \le \left(\sum_j w_j\right)^2$, with equality iff the state is GHZ-type. GHZ states are, in this setting, the optimal probes, shot-for-shot, for any linear gradient aggregate.

Security: Entanglement pre-sharing (before any gradient exists) is the key to information-theoretic privacy that no classical protocol can achieve under the same wiretapped-setup assumptions. We can construct a Verified GHZ Aggregation (VGA) protocol that certifies GHZ copies via a random stabilizer test and then runs phase-kickback, achieving, conditional on verification, composable ε-secure aggregation at a $2\times$ GHZ-copy overhead and no trusted-channel assumption.

3 Quantum advantage in distributed learning

After the all-reduce, the server must communicate the aggregated signal back to clients. In heterogeneous or adversarially distributed settings, clients cannot simply apply the global update blindly: gradient conflicts impede

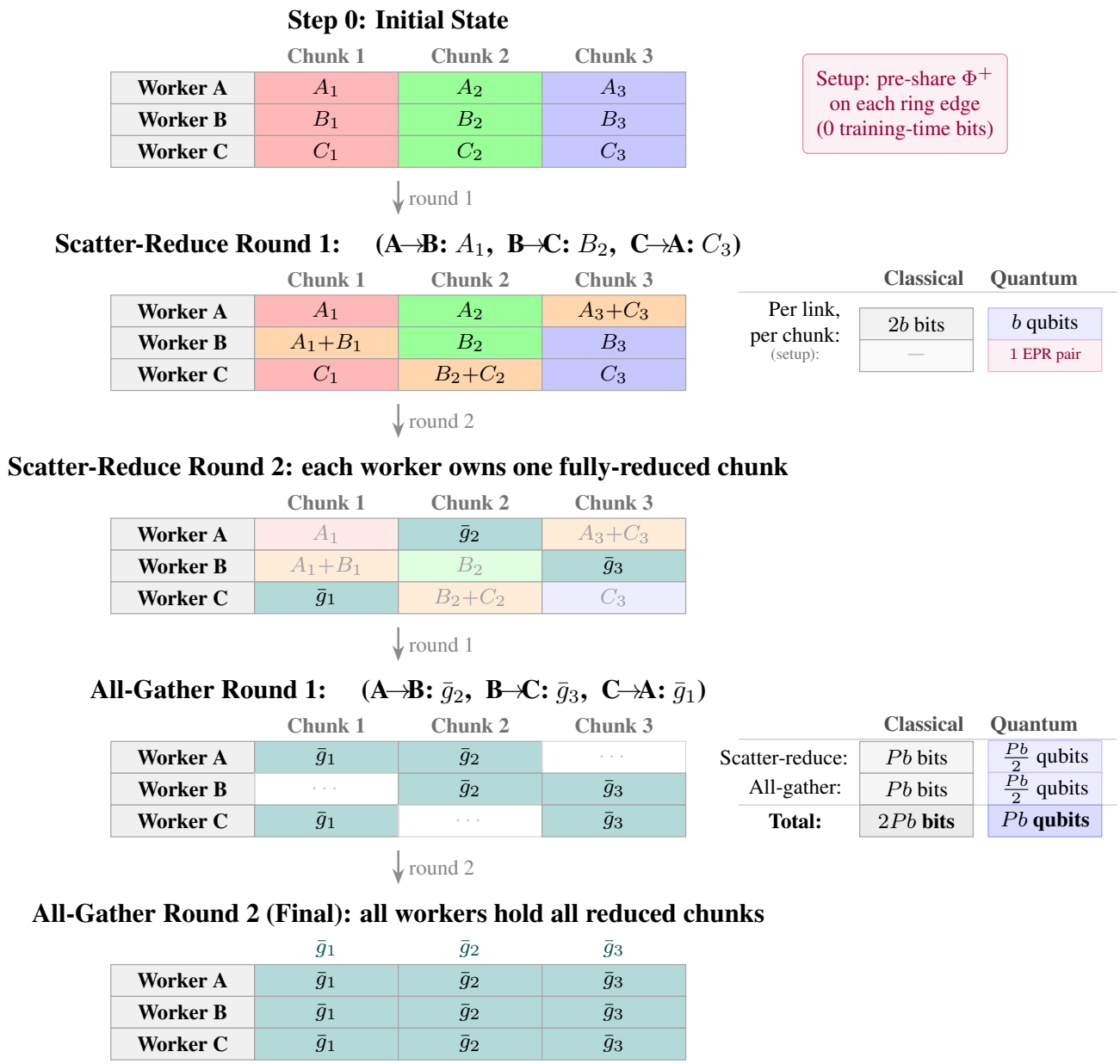

Figure 1: Quantum ring all-reduce for $N = 3$ workers, each holding a gradient partitioned into 3 equal chunks. *Orange*: partial sum accumulated so far.
Teal: fully reduced chunk $\bar{g}_k = \frac{1}{N}\sum_j g_{j,k}$.
Scatter-reduce ($N - 1 = 2$ rounds, clockwise): each worker sends one chunk per round; after round 2 every worker holds exactly one fully-reduced chunk (one third of the gradient).
All-gather ($N - 1 = 2$ rounds): each worker broadcasts its reduced chunk; after round 2 every worker holds all chunks.
Right panels: superdense coding [7] allows $2b$ gradient bits to cross a ring link using only b qubits, given a pre-shared EPR pair Φ^+ loaded in setup; the online qubit cost is Pb versus $2Pb$ bits classically, a factor-of-two saving that meets the entanglement-assisted Holevo lower bound.

convergence [8, 9], and each client needs to assess whether the update is compatible with its local data or model structure before applying it. Transmitting full gradients for this audit costs $O(Pb)$ bits per client and is infeasible at scale. We ask which conflict-detection tasks become possible with a quantum channel that would be intractable classically at the same bandwidth.

3.1 GAPIP: margin conflict detection (quadratic advantage)

Two workers hold unit-norm gradients $g_j, g_k \in \mathbb{R}^P$. Given $|\langle g_j, g_k\rangle| \geq \tau$, they must decide *aligned* ($\geq \tau$) vs. *conflicted* ($\leq -\tau$). The quantum protocol maps this to phase estimation on the Grover operator over the two gradient states, achieving $\widetilde{O}(\tau^{-1}\log P)$ qubits via Heisenberg scaling [10]. Classically, a Johnson–Lindenstrauss sketch gives $\widetilde{O}(\tau^{-2})$ bits, and a Gap-Hamming reduction [11] shows this is tight. The separation is

$$\text{Quantum: } \widetilde{O}(\tau^{-1}\log P) \text{ qubits}, \qquad \text{Classical: } \widetilde{\Theta}(\min(\tau^{-2}, P)) \text{ bits}.$$

The advantage is quadratic and tight on both sides [12]; it requires interaction since any single-round protocol is shot-noise limited at $\widetilde{\Omega}(\tau^{-2})$. Concretely, a quantum channel of width $\widetilde{O}(\tau^{-1}\log P)$ qubits makes pairwise conflict detection feasible at scale, enabling conflict-aware update rules [8] that would be too expensive to run classically at scale.

3.2 TIEAUDIT: sign-consistency auditing (exponential advantage)

After signSGD [13], the server holds $s \in \{\pm 1\}^P$. Each client k holds a private perfect matching M_k on $[P]$ (parameter pairs tied by weight sharing, factorisation, or quantisation) and wants to estimate

$$f_k = \frac{|\{(i,j) \in M_k : s_i \neq s_j\}|}{|M_k|}$$

to additive $\pm\varepsilon$: a high f_k signals the global update is structurally incompatible with the client's model. This is an instance of Boolean Hidden Matching [14].

The server sends $O(\varepsilon^{-2})$ copies of the $\lceil \log_2 P \rceil$-qubit phase state $|s\rangle = P^{-1/2}\sum_c s_c|c\rangle$. The client measures each copy in the matching basis $\{(|i\rangle \pm |j\rangle)/\sqrt{2}\}_{(i,j)\in M_k}$: the wrong-parity amplitude vanishes exactly, so each shot directly samples the conflict fraction. The separation is

$$\text{Quantum: } O(\varepsilon^{-2}\log P) \text{ qubits}, \qquad \text{Classical: } \Omega(\sqrt{P}) \text{ bits}.$$

No classical sketch of $o(\sqrt{P})$ bits can serve all matchings: the client's structure is revealed only after the server commits, so no fixed message can pre-adapt to it. The quantum state encodes all P signs simultaneously and lets the client's measurement post-select the relevant pair. To our knowledge, this is the **first exponential one-way quantum–classical separation instantiated in a gradient-derived statistic** rather than an abstract communication problem; the separation follows because the relational structure of the task (not numerical precision) is what resists classical compression. For a client whose model has $P = 10^8$ parameters, the quantum cost is $O(\varepsilon^{-2} \cdot 27)$ qubits while the classical cost is $\Omega(10^4)$ bits, a gap large enough to determine whether the audit is feasible at all. A client that can compute f_k can use it to trigger adaptive weight untying or local gradient correction when the global update is structurally misaligned with its model.

4 Conclusion

Ring all-reduce is the primitive through which every gradient in large-scale training passes. A quantum communication layer halves its bandwidth optimally, secures it information-theoretically, and exposes downstream communication tasks with provable quantum-classical separations no classical protocol can match, all without touching the model or the data. This shows that quantum networks could be useful for distributed learning regardless of the underlying nature of the model (quantum/classical). Full proofs for all these claims can be found at `https://arxiv.org/abs/2606.20344` (Appendix).

More broadly, our results suggest a hierarchy of quantum advantages for distributed learning. When the task is to communicate the value of a global aggregate, quantum communication is constrained by information-theoretic limits and yields at most constant-factor improvements. When the task is to certify the process by which the aggregate was generated, quantum resources enable privacy guarantees unavailable to classical protocols under the same assumptions. When the task is to evaluate relational properties of the aggregate relative to private local structure, exponential communication advantages become possible. Understanding which learning primitives fall into each of these regimes is, in our view, a promising direction for future work.

References

[1] Pitch Patarasuk and Xin Yuan. Bandwidth optimal all-reduce algorithms for clusters of workstations. *Journal of Parallel and Distributed Computing*, 69(2):117–124, 2009.

[2] Alexander Sergeev and Mike Del Balso. Horovod: fast and easy distributed deep learning in TensorFlow. *arXiv preprint arXiv:1802.05799*, 2018.

[3] Shen Li, Yanli Zhao, Rohan Varma, Omkar Salpekar, Pieter Noordhuis, Teng Li, Adam Paszke, Jeff Smith, Brian Vaughan, Pritam Luo, and Soumith Chintala. PyTorch distributed: Experiences on accelerating data parallel training. *Proceedings of the VLDB Endowment*, 13(12):3005–3018, 2020.

[4] Dar Gilboa, Hagay Michaeli, Daniel Soudry, and Jarrod R. McClean. Exponential quantum communication advantage in distributed inference and learning. In *Advances in Neural Information Processing Systems*, volume 37, pages 30425–30473, 2024. arXiv:2310.07136.

[5] Mina Doosti, Ryan Sweke, and Chirag Wadhwa. Distributed quantum property testing with communication constraints. *arXiv preprint arXiv:2604.05962*, 2026.

[6] Charles H. Bennett and Stephen J. Wiesner. Communication via one- and two-particle operators on Einstein-Podolsky-Rosen states. *Physical Review Letters*, 69(20):2881–2884, 1992.

[7] Charles H Bennett and Stephen J Wiesner. Communication via one-and two-particle operators on einstein-podolsky-rosen states. *Physical review letters*, 69(20):2881, 1992.

[8] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. arXiv:2001.06782.

[9] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 2020.

[10] Gilles Brassard, Peter Høyer, Michele Mosca, and Alain Tapp. Quantum amplitude amplification and estimation. In *Quantum Computation and Quantum Information: A Millennium Volume*, volume 305 of *AMS Contemporary Mathematics*, pages 53–74. American Mathematical Society, 2002. arXiv:quant-ph/0005055.

[11] Amit Chakrabarti and Oded Regev. An optimal lower bound on the communication complexity of Gap-Hamming-Distance. *SIAM Journal on Computing*, 41(5):1299–1317, 2012.

[12] Alexander A. Razborov. Quantum communication complexity of symmetric predicates. *Izvestiya: Mathematics*, 67(1):145–159, 2003. arXiv:quant-ph/0204025.

[13] Jeremy Bernstein, Jiawei Zhao, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD with majority vote is communication efficient and fault tolerant. In *International Conference on Learning Representations (ICLR)*, 2019. arXiv:1810.05291.

[14] Dmitry Gavinsky, Julia Kempe, Iordanis Kerenidis, Ran Raz, and Ronald de Wolf. Exponential separations for one-way quantum communication complexity, with applications to cryptography. *SIAM Journal on Computing*, 38(5):1695–1708, 2008. Earlier version in STOC 2007, pp. 516–525.

A Attached paper

Quantum ring all-reduce: communication and privacy advantages for distributed learning

María Gragera Garcés[1] and Lirandë Pira[2]

[1]University of Edinburgh, Edinburgh, United Kingdom
[2]Centre for Quantum Technologies, National University of Singapore, Singapore

Machine learning models have scaled to unprecedented sizes, making training across distributed devices the de facto standard in the field. In this work, we explore how quantum communications can make distributed training both more communication-efficient and information-theoretically private, for both classical and quantum learning models. Ring all-reduce is the foundational communication primitive for large-scale distributed training. We present a quantum version that reduces per-link *online* communication by a provably optimal factor of two (a saving on the synchronisation critical path; total lifetime channel use is unchanged) using pre-shared entanglement and superdense coding, without requiring the learning model or gradient computation to change. Beyond bandwidth, the primitive enables privacy guarantees that are information-theoretically impossible for any classical protocol, achieving composable ε-secure aggregation, via verified entanglement, at a $2\times$ overhead in GHZ copies. Our hybrid quantum-classical communication architecture yields simultaneous communication and security advantages for large scale distributed training, regardless of whether the learning itself is quantum or classical. Finally, we characterise quantum advantages in gradient conflict detection for server-to-client communication under bandwidth constraints, a setting that arises after ring all-reduce is completed, when full gradient broadcast to external clients is infeasible due to bandwidth constraints. Two variants of the problem admit qualitatively different separations. For margin-based alignment testing ($\textsc{GapIP}_\tau$), the quantum advantage is quadratic in the margin parameter: $\widetilde{O}(\tau^{-1}\log P)$ qubits versus $\widetilde{O}(\min(\tau^{-2}, P))$ bits classically. For sign-consistency auditing against a private parameter matching ($\textsc{TieAudit}_\varepsilon$), the advantage represents an exponential separation in communication complexity: classical one-way communication requires $\Omega(\sqrt{P})$ bits whereas $O(\varepsilon^{-2}\log P)$ qubits suffice.

María Gragera Garcés: m.gragera.garces@ed.ac.uk
Lirandë Pira: lpira@nus.edu.sg

Contents

1 Introduction **3**

2 Related work **4**
2.1 Gradient compression and ring all-reduce . . . 4
2.2 Quantum advantages in distributed machine learning . . . 4
2.3 Quantum secure aggregation . . . 4
2.4 Communication complexity . . . 5

3 The quantum ring all-reduce primitive **5**
3.1 Setup: pre-sharing entanglement . . . 5
3.2 Online: superdense-coding all-reduce . . . 5
3.3 GHZ phase encoding: an equivalent formulation and its optimality . . . 6

4 A secure quantum ring all-reduce protocol **8**
4.1 Model and definitions . . . 8
4.2 Classical impossibility . . . 9
4.3 Verified GHZ aggregation . . . 10
4.4 Multi-round privacy . . . 11

5 Quantum advantage in gradient conflict detection under bandwidth constraints **12**
5.1 Problem definitions . . . 13
5.2 GapIP . . . 13
5.3 TieAudit_ϵ . . . 15

6 Conclusion and outlook **16**

7 Acknowledgments **17**

A Superdense coding primer **19**
A.1 The protocol . . . 19
A.2 Why this gives the factor of two . . . 19

B Communication-complexity primer **20**
B.1 The model . . . 20
B.2 One-way versus interactive . . . 20
B.3 Randomness and shared resources . . . 20
B.4 Qubits are not free . . . 21
B.5 Separations, with examples . . . 21

C Boolean Hidden Matching primer **21**
C.1 The problem . . . 22
C.2 Why it is hard classically . . . 22
C.3 Why a small quantum message suffices . . . 22
C.4 From Boolean Hidden Matching to TieAudit . . . 22

D GapIP_τ Quantum protocol **23**
D.1 Encoding . . . 23
D.2 Local reflections . . . 23
D.3 Distributed Grover operator . . . 23
D.4 Estimation and decision . . . 23

1 Introduction

Training a competitive machine learning model today requires processing datasets of billions of samples, optimising hundreds of billions of parameters, and executing arithmetic operations that would take enormous resources on a single processor. The gap between what a single device can deliver and what training demands has grown faster than hardware alone can close. As model sizes, datasets, and computational demands continue to grow, training is performed across clusters of devices that must repeatedly communicate in order to perform as a single learning system [3]. This suggests that industry-scale machine learning is by default distributed, and thus, modern distributed training is dominated by gradient communication.

Ring all-reduce is a standard communication primitive for distributed training at scale [27]. This algorithm works by arranging N workers in a ring, each holding a local gradient $g_j \in \mathbb{R}^P$ computed on its local data; in each round every worker exchanges one message with each ring neighbour, and after $\lfloor N/2 \rfloor$ rounds every worker holds the global mean $\bar{g} = \frac{1}{N}\sum_j g_j$ to b bits of precision per coordinate (Pb bits total). The primitive underpins large-scale machine learning, and more specifically deep learning systems across industry, implemented in frameworks such as Horovod [31], PyTorch DDP [23], and NCCL (NVIDIA Collective Communications Library).

In this work we explore how quantum technologies can reduce the resources necessary to perform this primitive. One principle organises what follows: the size of the quantum advantage is governed by *what one asks of the all-reduce aggregate*. Transporting its value buys only a Holevo-bounded constant factor; certifying how it was produced buys a qualitative privacy separation, classically impossible at any communication cost; and querying a privately chosen relation over it can buy an exponential separation. These regimes coexist within a single GHZ-based protocol, leaving the learning model untouched. The first of them, the bandwidth saving, is a baseline rather than the headline. Bandwidth saving relocates online communication off the critical path but does not reduce total channel use (made more precise in Section 3), and it is the substrate on which the privacy and conflict-detection results build. The main contributions of this work are threefold.

First, we propose a superdense-coding strategy that exploits pre-shared EPR pairs to replace each classical message of m bits with $\lceil m/2 \rceil$ qubits, reducing the online communication across every ring link by exactly a factor of two. This is provably optimal: a cut argument shows at least $Pb - O(\log N)$ classical bits of information must cross every cut in the online phase of the ring, and the entanglement-assisted capacity of a qubit channel is at most two classical bits even with unlimited entanglement, so the factor-of-two is the ceiling of what entanglement can buy. The round structure and latency are unchanged: no quantum protocol, regardless of pre-shared entanglement, can reduce the blocking latency below $\lfloor N/2 \rfloor$ post-barrier rounds.

Second, we present the security guarantees a quantum ring all-reduce offers. In a model where every channel (including the setup phase) is wiretapped, classical protocols cannot achieve information-theoretically private aggregation at *any* communication cost, while our Verified GHZ Aggregation (VGA) protocol achieves it at a $2\times$ overhead in GHZ copies, regardless of whether the learning itself is quantum or classical.

Third, we study two variants of gradient conflict detection for communication-limited settings where broadcasting full gradient vectors is infeasible. In distributed training, the server often broadcasts a compressed global signal that each client must interpret against its own private model structure: the compressed message alone does not tell the client whether the proposed update is compatible with its internal constraints. In the first variant (GapIP), two workers check peer-to-peer whether their unit-normalised gradients are aligned or conflicting up to a margin τ, without exchanging full vectors: the quantum communication cost is $\widetilde{O}(\tau^{-1}\log P)$ versus $\widetilde{O}(\min(\tau^{-2}, P))$ classically, a tight quadratic separation. The second variant, TieAudit, arises in the signSGD setting where the server's compressed global signal is the sign-gradient $s \in \{\pm 1\}^P$, one bit per parameter. Each client holds a private matching M_k of parameter-index pairs that are structurally coupled in its local model (via weight sharing, factorisation, or quantisation); the matching acts as a private key for checking whether the global update direction is internally consistent. Concretely, the client wants to estimate what fraction of its tied pairs $(i, j) \in M_k$ receive conflicting global signs $s_i \neq s_j$: a high fraction signals that the server's update is structurally incompatible with the client's model and should be locally corrected. The quantum advantage here is exponential: classical one-way communication requires $\Omega(\sqrt{P})$ bits regardless of approximation accuracy, while

$O(\varepsilon^{-2} \log P)$ qubits suffice.

The quantum-communication results we draw on here — such as superdense coding, GHZ-based metrology, the impossibility of secret-key agreement from independent randomness, and Boolean Hidden Matching — are each individually standard. Our aim is not to promote any particular approach, but to identify the communication layer of distributed learning, a setting of clear practical importance, as a regime in which these results yield concrete advantages. These advantages arise from the communication layer alone: unlike Gilboa et al. [17], whose exponential separation stems from the nodes holding quantum data, our inputs and outputs can be entirely classical, as the magnitude of our advantage claims is governed by the type of question asked of the ring all-reduce aggregate. Moving the aggregate's value buys a Holevo-bounded constant factor; certifying the setup that produced it buys a qualitative privacy separation, impossible to achieve classically at any communication cost; and extracting a privately chosen relation over it can buy an exponential separation. A single GHZ-based communication protocol spans all three regimes while leaving the learning model untouched: this is the central message of our work.

This paper is organised as follows. Section 2 presents related work including the ring all-reduce architecture and existing literature regarding quantum advantage in distributed machine learning. Section 3 introduces the quantum ring all-reduce primitive in both its raw and GHZ phase encoding forms. Section 4 demonstrates the security advantages of a GHZ phase encoding ring all-reduce architecture protocol. Finally, Section 5 presents the gradient conflict settings that might arise from ring all-reduce and how quantum architectures can enable quadratic and exponential advantages. Conclusions and future outlooks can be found in Section 6.

2 Related work

2.1 Gradient compression and ring all-reduce

Ring all-reduce is the bandwidth-optimal topology for gradient aggregation in data-parallel training [27], and its implementation in Horovod [31] and PyTorch DDP [23] makes it the de facto standard for large-scale deep learning. Gradient compression methods, including signSGD [6] and its fault-tolerant majority-vote variant [7], reduce per-link bandwidth at the cost of introducing sign-aggregation artifacts; TieAudit formalises one such artifact as a detectable relational statistic. The gradient conflict literature [34, 19] motivates GapIP: both PCGrad and SCAFFOLD identify conflicting updates as a source of convergence degradation, but neither addresses the communication cost of detecting conflicts under bandwidth constraints.

2.2 Quantum advantages in distributed machine learning

Gilboa et al. [17] show an exponential quantum communication advantage for distributed inference when nodes hold *quantum* data; the advantage follows from the exponential classical simulation cost of quantum states rather than from the communication primitive itself. Our work is complementary: inputs are classical gradient vectors and the quantum advantage arises entirely from the communication layer (superdense coding for bandwidth, phase encoding aggregation for security, and one-way quantum messages for gradient conflict detection). Doosti et al. [14] study distributed quantum state certification under one-way communication constraints and obtain a quadratic quantum-over-classical advantage via quantum random compression. Their setting shares structural parallels with our GapIP and TieAudit problems, both exploit one-way quantum advantages in relational statistics, but addresses state certification to a central node rather than gradient aggregation for learning and conflict detection in a ring. Pira and Ferrie [28] survey distributed quantum neural networks, focusing on architectures rather than communication primitives, and identify quantum generalisations of collective communication protocols as an open challenge.

2.3 Quantum secure aggregation

Classical private aggregation relies on pairwise masking over trusted side channels; quantum proposals extend this via gradient hiding [21], blind quantum computation [22, 35], and Byzantine-

robust quantum aggregation [26]. A common assumption across this line is that entanglement or keys are distributed over a channel not visible to the adversary. Theorem 2 establishes that the trusted-channel assumption in classical schemes [8] is not merely a convenience but a necessity: no classical protocol achieves information-theoretic privacy when every channel is observed. Quantumly the barrier is lifted: Proposition 1 gives per-coordinate δ-security conditional on verification, and composing this additively across coordinates and rounds (Lemma 3) yields, for any $\varepsilon > 0$, composable ε-secure aggregation at sufficiently many verified GHZ copies, at a $2\times$ overhead in GHZ copies for verification and with no trusted-channel assumption.

2.4 Communication complexity

The lower bounds in Section 5 draw on two results from one-way communication complexity. The Gap-Hamming Distance lower bound of Chakrabarti and Regev [12] gives the tight $\widetilde{\Omega}(\tau^{-2})$ classical lower bound for GapIP; the quantum lower bound uses Razborov's symmetric-predicate technique [29]. The exponential separation for TieAudit reduces to the Boolean Hidden Matching lower bound of Gavinsky et al. [16], which established the first exponential separation between quantum and classical one-way communication complexity; Bar-Yossef, Jayram, and Kerenidis [2] give an alternative construction. Our results instantiate these separations in concrete gradient statistics rather than abstractly defined communication problems.

3 The quantum ring all-reduce primitive

The goal of the quantum ring all-reduce is to compute $\bar{g} = \frac{1}{N}\sum_j g_j$ with minimum qubit communication while preserving the round structure of the classical primitive. The protocol has two phases: a *setup* phase in which entanglement is distributed before any gradient exists, and an *online* phase in which each training step runs the all-reduce using that entanglement.

3.1 Setup: pre-sharing entanglement

Before training begins, every pair of adjacent nodes $(j, j+1)$ pre-shares $\lceil Pb/2 \rceil$ EPR pairs $|\Phi^+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$, one pair per two classical bits of gradient that will cross that link. This resource is *input-independent*: it can be generated, certified, and distributed before any training data or model gradient is seen, which is both a practical convenience and the key property the security analysis exploits (see Section 4).

3.2 Online: superdense-coding all-reduce

The bandwidth advantage of replacing classical messages with qubits is fundamentally capped at $2\times$ by the entanglement-assisted Holevo capacity; the protocol below meets this bound exactly.

The classical ring all-reduce proceeds in two sub-phases: a *scatter-reduce*, in which each node accumulates one P/N-dimensional partial sum by passing it clockwise around the ring, followed by an *all-gather*, in which the completed mean is disseminated counter-clockwise. Each sub-phase requires $N-1$ rounds, and a total of Pb classical bits cross every cut of the ring.

A straightforward quantum version would replace each classical message of m bits with $\lceil m/2 \rceil$ qubits. Node j encodes two classical bits into one qubit by applying one of the four Pauli operations $\{I, X, iY, Z\}$ to its half of a pre-shared EPR pair; the receiving node recovers both bits via a Bell measurement (superdense coding [5]). The round structure, the all-reduce schedule, and the gradient recovery arithmetic are identical to the classical case; only the physical carriers change. The online qubit cost across every link is therefore $\frac{Pb}{2}(1 + o(1))$, a factor of two reduction.

Remark 1 (Resource accounting across rounds)**.** *Superdense coding consumes entanglement: the Bell measurement that decodes each message destroys the EPR pair it uses, so the pairs are not reusable and a fresh supply must be distributed for every training round. Because entanglement cannot be created across a cut by local operations and classical communication, distributing one EPR pair across a link costs one qubit channel-use across that link, and each online qubit consumes exactly one such pair. Counting both phases, the total number of carriers crossing every link*

therefore equals the classical baseline: consistent with the Holevo bound, superdense coding does not reduce the total *communication, it halves only the* online *portion. The factor of two is thus a relocation of communication off the critical path rather than a net reduction in lifetime channel use, and this is the operationally relevant quantity. The setup phase is input-independent and can be scheduled during computation or on a dedicated entanglement-distribution layer, whereas the online phase runs inside the all-reduce barrier that blocks all N workers. In the communication-bound regime typical of large-scale distributed training, where barrier bandwidth on the critical path is the scarce resource, halving the online volume is the saving that matters even though total channel use is unchanged; an ideal setup for our architecture.*

The factor is tight: a cut separating any node set A from B must convey the Shannon entropy of $\sum_{j\in B} g_j$ equals $Pb - O(\log N)$ bits (the sum of uniform b-bit words loses only carry overhead), and pre-shared entanglement is input-independent so by the no-communication theorem it carries none of that entropy on its own. Each qubit crossing the cut conveys at most $C_E = 2$ classical bits even with unlimited entanglement, the entanglement-assisted Holevo bound [4]. Hence at least $\frac{Pb}{2} - O(\log N)$ qubits must cross every cut in the online phase: the protocol meets the lower bound.

Figure 1 shows this superdense-coding protocol worked through for $N = 3$. The construction so far is a direct application of superdense coding, included because it targets the defining cost of ring all-reduce, its per-link bandwidth; it is the simplest component of the paper, and its value is as a substrate rather than a result in its own right. Section 3.3 recasts the same all-reduce in terms of a shared GHZ state rather than EPR pairs, and it is this reformulation carries the conceptual content of what follows: the optimal-probe characterisation of Algorithm 1, the security guarantees of Section 4, and the structure exploited by the conflict-detection separations of Section 5.

3.3 GHZ phase encoding: an equivalent formulation and its optimality

An alternative formulation encodes gradients directly in the phases of a shared GHZ state rather than compressing classical bit strings. This view is the basis of the security protocol in Section 4, and it also yields the sharpest statement of *why* GHZ is the right resource.

Definition 1 (Local phase encoding)**.** *Each node j holds one qubit of a shared N-qubit pure state $|\psi\rangle$ and applies $U_j(\theta_j) = e^{-i\theta_j Z_j/2}$, where θ_j is a b-bit encoding of its gradient coordinate. Writing $H = \frac{1}{2}\sum_j w_j Z_j$ for weights $w_j > 0$ (the equal-weight case $w_j = 1$ encodes a simple sum), the post-encoding state along the direction $\theta_j \equiv \vartheta$ is $e^{-i\vartheta H}|\psi\rangle$. Every node measures in the X basis, broadcasts its outcome bit, and all nodes recover the aggregate $\Phi = \sum_j \theta_j$ by parity.*

Definition 2 (Quantum Fisher information [11])**.** *For the family $|\psi_\vartheta\rangle = e^{-i\vartheta H}|\psi\rangle$, the quantum Fisher information is $F_Q = 4\,\mathrm{Var}_\psi(H)$. Every locally unbiased estimator of Φ from ν independent copies obeys $\mathrm{Var}(\hat{\Phi}) \geq 1/(\nu F_Q)$. F_Q bounds the information extractable by* any *measurement strategy, including adaptive LOCC with classical broadcast.*

Theorem 1 (GHZ uniquely maximises linear-aggregate precision)**.** *Fix the resource model of Definition 1: a single N-qubit pure probe state $|\psi\rangle$[1], encoding restricted to local phase rotations $U_j(\theta_j)$ (no entangling gates during encoding, no ancillas beyond the N encoding qubits), and ν independent, identically prepared, unentangled copies of the protocol. For every such $|\psi\rangle$ and weights $w_j > 0$,*

$$F_Q = 4\,\mathrm{Var}_\psi(H) \leq \Big(\sum_j w_j\Big)^2,$$

with equality if and only if $|\psi\rangle = \frac{1}{\sqrt{2}}(|0^N\rangle + e^{i\varphi}|1^N\rangle)$, a GHZ-type state. Any amplitude γ on a non-extremal eigenspace of H with eigenvalue λ_χ incurs the quantitative penalty $F_Q = \Big(\sum_j w_j\Big)^2\Big(1 - |\gamma|^2\big(1 - \lambda_\chi^2/\lambda_{\max}^2\big)\Big) - 4|\gamma|^4\lambda_\chi^2$.

[1] That is, $\rho = |\psi\rangle\langle\psi|$ is rank one, not a statistical mixture. This restriction matters: F_Q is convex in ρ, so any mixed probe has F_Q no larger than the best pure state in its decomposition, and the GHZ-uniqueness claim would need restating in terms of ρ if mixed probes were allowed.

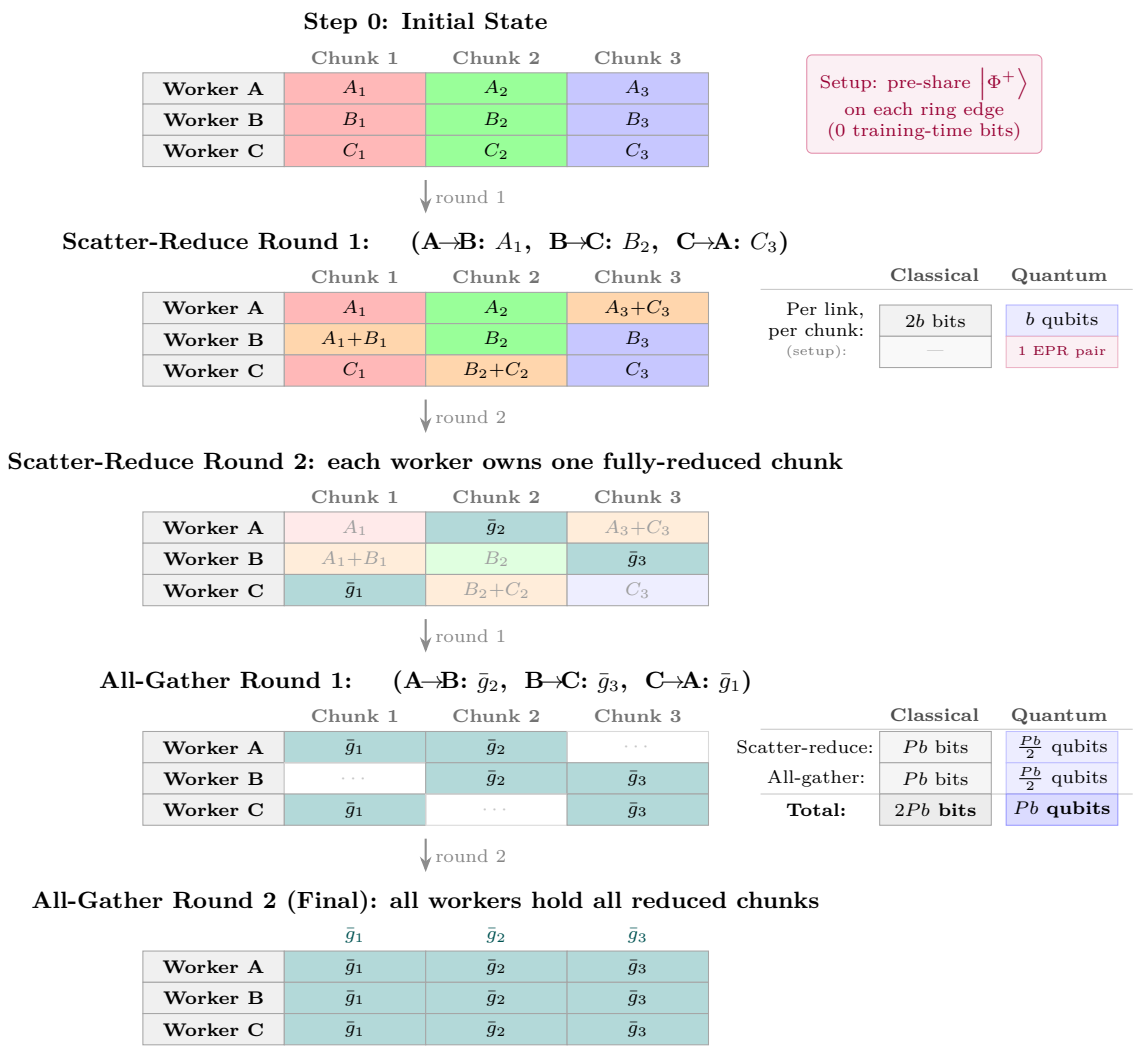

Figure 1: Quantum ring all-reduce for $N = 3$ workers, each holding a gradient partitioned into 3 equal chunks. *Orange*: partial sum accumulated so far.
Teal: fully reduced chunk $\bar{g}_k = \frac{1}{N}\sum_j g_{j,k}$.
Scatter-reduce ($N - 1 = 2$ rounds, clockwise): each worker sends one chunk per round; after round 2 every worker holds exactly one fully-reduced chunk (one third of the gradient).
All-gather ($N - 1 = 2$ rounds): each worker broadcasts its reduced chunk; after round 2 every worker holds all chunks.
Right panels: superdense coding [5] allows $2b$ gradient bits to cross a ring link using only b qubits, given a pre-shared EPR pair $|\Phi^+\rangle$ loaded in setup; the online qubit cost is Pb versus $2Pb$ bits classically, a factor-of-two saving that meets the entanglement-assisted Holevo lower bound.

Algorithm 1 Quantum all-reduce: GHZ phase encoding formulation

Require: Each node $j \in \{1, \ldots, N\}$ holds a gradient coordinate $\theta_j \in [0, 2\pi)$ encoded to b-bit precision. *AND* Each node j holds a weight $w_j > 0$. *AND* A fresh N-qubit GHZ state $|\psi\rangle = \frac{1}{\sqrt{2}}(|0^N\rangle + |1^N\rangle)$ is supplied to the ring before each of the b rounds below.

Output: $\hat{\Phi}$, an estimate of $\Phi = \sum_j w_j \theta_j \bmod 2\pi$ accurate to b bits, identical at every node

```
for k = b−1, b−2, ..., 0 do
    Distribute one fresh copy of |ψ⟩, one qubit per node
    for each node j in parallel do
        θ_j^(k) ← w_j θ_j · 2^k (mod 2π)
        Apply U_j(θ_j^(k)) = e^(−i θ_j^(k) Z_j/2) to its half of |ψ⟩
        Measure local qubit in the X basis → outcome m_j^(k) ∈ {0,1}
        Broadcast m_j^(k) to all other nodes                    ▷ N − 1 classical bits
    end for
    Every node computes the parity bit p^(k) ← ⊕_j m_j^(k)
end for
Every node combines (p^(b−1), ..., p^(0)) into an estimate Φ̂ return Φ̂
```

Proof. The spectrum of $H = \frac{1}{2}\sum_j w_j Z_j$ lies in $[\lambda_{\min}, \lambda_{\max}] = [-\frac{1}{2}\sum w_j, \frac{1}{2}\sum w_j]$. Popoviciu's variance inequality gives $\mathrm{Var}_\psi(H) \leq \left(\frac{\lambda_{\max}-\lambda_{\min}}{2}\right)^2$ with equality iff the spectral distribution of $|\psi\rangle$ puts mass $\frac{1}{2}$ on each extremum. Since all $w_j > 0$, the extreme eigenspaces are one-dimensional ($|0^N\rangle$ and $|1^N\rangle$), forcing $|\psi\rangle$ to have the GHZ form up to a relative phase. The penalty formula follows by direct variance computation on the three-component decomposition. □

Reaching aggregate precision τ costs $\nu \geq 1/(\tau^2 F_Q)$ encoded copies, so GHZ-type states are the unique optimal probe shot-for-shot among all N-qubit states for any linear aggregate $\sum_j w_j \theta_j$. This optimality is specific to linear functionals; for nonlinear functionals of the gradients, no single probe state is universally optimal and performance becomes instance- and structure-dependent.

Remark 2 (Saturation of the Quantum Fisher Information bound by the parity protocol)**.** *Theorem 1 bounds F_Q, which by the quantum Cramér–Rao bound of Definition 1 is itself only a lower bound on the variance of any estimator; attaining it in general requires the optimal (symmetric-logarithmic-derivative) measurement, not the concrete X-basis, parity-readout scheme of Definition 1. For the GHZ state these coincide: the SLD measurement for $H = \frac{1}{2}\sum_j w_j Z_j$ on $\frac{1}{\sqrt{2}}(|0^N\rangle + e^{i\varphi}|1^N\rangle)$ is exactly the transversal X-basis measurement with parity readout, so the protocol of Definition 1 saturates the bound of Theorem 1 with no further assumptions. This is what upgrades the theorem's claim from "GHZ uniquely maximises F_Q" to "GHZ is the unique optimal probe* for the stated protocol*, shot-for-shot."*

4 A secure quantum ring all-reduce protocol

The quantum communication advantage of Section 3 uses entanglement as a resource established before any gradient exists. That entanglement pre-sharing enables a *information-theoretic privacy* guarantee, that cannot be achieved by any classical protocol in a setting where an adversary observes every channel, including the setup phase. Building on the GHZ phase encoding formulation of Section 3.3, we propose the Verified GHZ Aggregation (VGA) protocol, which achieves *information-theoretic private* aggregation at a cost of twice as many GHZ copies as the unverified phase encoding protocol.

4.1 Model and definitions

Eve is read-only on the authenticated classical channels (she sees every classical symbol but cannot modify it) and may act arbitrarily on the quantum systems in transit, including tampering; such tampering is caught, except with small probability, by the verification step of Section 4. By

no-cloning [33] she cannot copy a quantum system in transit without disturbing it, so any such attempt is caught by verification rather than yielding a silent copy. Nodes hold independent private randomness but begin with no pre-shared secrets. Leakage about node j's gradient g_j is measured as the mutual information $I(g_j; V_E)$, where V_E is Eve's total accumulated view. Eve's goal is to learn as much as possible about individual nodes' gradients beyond what the aggregate output $\bar{g}$ already discloses. In distributed learning this is a concrete threat: gradient values have been shown to allow reconstruction of private training data and membership inference [35]; our metric $I(g_j; V_E)$ measures exactly how much Eve's view of the protocol exceeds what the aggregate alone reveals.

Definition 3 (Wiretapped-setup model)**.** *This model runs the same N-node ring topology and scatter-reduce / all-gather structure as Section 3, but with no prior correlations between nodes. Every channel (classical and quantum) in both the setup phase (quantum state distribution) and the online phase is exposed to an adversary Eve, who is read-only on the authenticated classical channels and may act arbitrarily (including tampering) on the quantum systems in transit. Classical channels are authenticated. Nodes hold independent private randomness. Leakage is $I(g_j; V_E)$, where V_E is Eve's total view.*

To state security precisely we use composable simulation security: a protocol is *secure* if whatever Eve learns from participating, she could equally have learned by interacting with a simulator that sees only the aggregate output $\bar{g}$ and nothing else, meaning the real protocol leaks no information that the output alone would not already reveal.

Definition 4 (ε-secure aggregation (composable) [25])**.** *A protocol is ε-secure if for every adversary there exists a simulator interacting only with an ideal functionality, which collects inputs, returns $\bar{g}$ (or aborts), and reveals nothing else, such that the joint state of honest outputs and the adversary's view in the real protocol is within trace distance ε of the corresponding state in the simulated ideal execution. Trace-distance security composes additively.*

The trace-distance bound ε upper-bounds the distinguishing advantage of any measurement between the real and ideal experiments.

4.2 Classical impossibility

Theorem 2 (No classical IT privacy under wiretapped setup)**.** *Any classical protocol in Definition 3 that computes $\bar{g}$ with error probability ϵ admits an input distribution under which*

$$I(g_j; T_{\text{full}}) \;\geq\; (1-\epsilon)\, H(g_j) - h(\epsilon),$$

where $H(\cdot)$ denotes Shannon entropy, $h(\cdot)$ binary entropy, and T_{full} is the full protocol transcript. Vanishing leakage is therefore unachievable at any communication cost.

Proof. We reduce to secret-key agreement. Set $g_j = K$ uniformly random and all other nodes run honestly on input 0, so the honest output is $\bar{g} = K/N$, from which any other node k recovers K exactly. By Fano's inequality, $H(K \mid T_{\text{full}}, R_k) \leq h(\epsilon) + \epsilon\, H(K)$, where R_k is node k's private randomness. If the transcript satisfied $I(K; T_{\text{full}}) \leq \ell$, then nodes j and k (starting from independent randomness) could agree, via authenticated public discussion over the already-transmitted transcript, on a key of entropy at least $H(K) - \ell - h(\epsilon) - \epsilon\, H(K)$ hidden from any passive observer. The secret-key capacity of independent sources over public authenticated channels is zero [24, 1]; hence $\ell \geq (1-\epsilon)H(K) - h(\epsilon)$. □

This closes, in the strongest form, the "pre-shared one-time pad" objection: classical masking schemes [8] achieve information-theoretically private aggregation if pads can be delivered secretly, and Theorem 2 shows that a fundamental information-theoretic barrier when every channel is observed.

4.3 Verified GHZ aggregation

The VGA protocol builds on the phase encoding formulation of Definition 1. The challenge of this setup is that Eve can observe the GHZ distribution step and may tamper with the states en route. The VGA protocol handles this attack by executing a random stabilizer test before any gradient is encoded: surviving copies are certified close to genuine GHZ states, and because certification happens before inputs arrive, abort decisions are input-independent, mirroring QKD protocols, where a random subset of transmitted systems is sacrificed for parameter estimation before any key is extracted.

Protocol VGA (Verified GHZ Aggregation), per gradient coordinate

Step 1 (Distribution) A designated node prepares $2t$ copies of $|\mathrm{GHZ}_N\rangle = \frac{1}{\sqrt{2}}(|0^N\rangle + |1^N\rangle)$ and routes one qubit of each copy to every other node around the ring.
Step 2 (Verification) A public random string selects t copies and, for each, announces a uniformly random nontrivial stabilizer $g \in \mathcal{S} \setminus \{I\}$ of the GHZ group $\mathcal{S}$. Every such g is a tensor product of single-qubit Paulis, so all nodes measure locally and broadcast outcomes. *Abort* unless every tested g yields $+1$.
Step 3 (Aggregation) On the surviving t copies, run phase encoding aggregation (Definition 1): node j applies $e^{-i\theta_j Z_j/2}$ to its share of each copy, all nodes measure in the X basis and broadcast the outcome bit, and the aggregate $\Phi = \sum_j \theta_j$ is recovered by parity.

We establish security in two layers. On exact GHZ states, every classical bit broadcast during readout is simulatable from the aggregate alone, so the protocol leaks nothing beyond its output (Lemmas 1 and 2). Verification then certifies that the accepted states are close to exact GHZ, and security degrades only gracefully in that distance (Proposition 1); the closeness itself is bounded by the sampling analysis that follows.

Lemma 1 (Per-copy privacy, including the broadcast readout)**.** *For exact GHZ states, the joint view of any coalition $S \subsetneq [N]$ (their qubits, any applied deviations, Eve's residual system, and* all classical bits broadcast during the readout*) is simulatable from the coalition's own inputs and the honest-sum $\Phi_H = \sum_{j \notin S} \theta_j$. Individual gradients θ_j for $j \notin S$ are never revealed beyond what Φ_H implies.*

Proof. After honest encoding the shared state is $\frac{1}{\sqrt{2}}(|0^N\rangle + e^{i\Phi_H} e^{i\phi_S} |1^N\rangle)$, where ϕ_S is the phase applied by the coalition (known to the simulator). The distribution of the honest nodes' X-measurement outcomes, conditioned on any coalition measurement with outcome ω, depends on the state only through the relative phase $\Phi_H + \phi_S$:

$$\Pr\Big[\bigoplus_{j \notin S} b_j = 0 \;\Big|\; \text{coalition outcome } \omega\Big] = \tfrac{1}{2}\big(1 + \mathrm{Re}\; e^{i(\Phi_H + \phi_S)} c_\omega\big),$$

where c_ω depends only on the coalition's (known) measurement. The simulator samples honest outcome bits from this distribution; individual θ_j never appear beyond their contribution to Φ_H. □

Lemma 2 (Ideal-GHZ security, exact)**.** *On exact GHZ copies, VGA realises the ideal sums-only functionality of Definition 4 exactly: there is a simulator interacting only with the ideal functionality (which returns $\Phi = \sum_j \theta_j$ or aborts) whose output is identical to the real execution. In particular $\varepsilon = 0$.*

Proof. Distribution and verification are input-independent, so the simulator runs them against the real Eve, reproducing the accepted state and the abort decision identically. Eve's residual view is the broadcast readout together with her side system; this is the case $S = \varnothing$ of Lemma 1 (a purely external, passive Eve controls no inputs), for which $\Phi_H = \sum_j \theta_j = \Phi$, so the readout is simulatable exactly from Φ. Composing over the t surviving copies gives an exact simulator; hence $\varepsilon = 0$. □

Proposition 1 (Conditional robust security of VGA)**.** *Condition on acceptance, and suppose the joint state $\rho_{C\bar{T}E}$ of the t surviving copies (honest register C) and Eve's system E satisfies*

$$\left\|\rho_{C\bar{T}E} - \tilde{\rho}_{C\bar{T}E}\right\|_1 \;\leq\; \delta,$$

where $\tilde{\rho}$ has honest register equal to t exact GHZ copies $|\mathrm{GHZ}_N\rangle^{\otimes t}$ (Eve may hold arbitrary side information). Then VGA is δ-secure per coordinate in the sense of Definition 4.

Proof. The remaining protocol (input encoding $U_j(\theta_j)$ on each honest share, X-basis measurement, broadcast, and Eve's processing) is a fixed channel Λ_θ depending on the inputs but not on the probe state. By Lemma 2, $\Lambda_\theta(\tilde{\rho})$ equals the ideal simulator output $\mathsf{S}(\Phi)$. Since trace distance is non-increasing under the common channel Λ_θ,

$$\left\|\Lambda_\theta(\rho) - \mathsf{S}(\Phi)\right\|_1 = \left\|\Lambda_\theta(\rho) - \Lambda_\theta(\tilde{\rho})\right\|_1 \leq \left\|\rho - \tilde{\rho}\right\|_1 \leq \delta.$$

Abort is input-independent and reproduced by the simulator, and trace-distance security composes additively, giving δ-security per coordinate. □

The verification of Step 2 is a Bouman–Fehr sampling strategy [9] in the GHZ-stabiliser eigenbasis; a uniformly random nontrivial stabiliser flags a corrupted copy ($\vec{\sigma} \neq 0$) with probability $\geq \frac{1}{2}$. Read at the level of the purification, acceptance certifies (except with probability $\varepsilon_{\mathrm{samp}}$) that the surviving ensemble is within trace distance $\delta = O\big(\sqrt{\varepsilon_{\mathrm{samp}}} + \sqrt{\beta}\big)$ of an ideal-GHZ register, with $\beta = O\big(\log(1/\varepsilon_{\mathrm{samp}})/t\big)$. Substituting this δ into Proposition 1 yields the security claim, with overhead a factor 2 in GHZ copies and $O(1)$ broadcast bits per tested copy.

We emphasise that this certification step is asymptotic. Three refinements remain for deployment-grade constants:

1. The $\frac{1}{2}$ detection probability requires an imperfect-detection Bouman–Fehr bound, or stabiliser amplification, to fix the constant in β;
2. The entropic-uncertainty/decoupling conversion [32, 30] is used here in its asymptotic, leading-order form; the explicit finite-key correction terms, which would convert it into a security parameter at finite copy count t, are left to future work;
3. A noise-tolerant acceptance threshold replacing the perfect-pass rule remains to be established.

None of these affects Proposition 1 or the classical impossibility of Theorem 2; they bound only the achievable δ.

4.4 Multi-round privacy

In practice, trainings with ring all-reduce are often completed in T rounds. After observing T consecutive aggregate sums, an adversary might try to piece together information about individual node gradients that no single sum would reveal alone. The theorem below shows our protocol is safe against this type of attack: the coalition's full multi-round view is provably no more informative than the sequence of sums itself, VGA's per-round privacy extends to T rounds at no extra cost.

Theorem 3 (View reduction across rounds)**.** *For T rounds of ideal VGA (exact GHZ each round) against a static coalition $S \subsetneq [N]$, there exists a simulator that, given only the coalition's own inputs and the sequence of honest sums $\{\Phi_H^{(r)}\}_{r=1}^T$, reproduces the coalition's joint view (qubits, applied deviations, and all broadcast bits across all rounds) exactly. Consequently,*

$$I\big(\{g_j^{(r)}\}_{j\notin S,\, r\leq T};\ \mathrm{view}_S\big) \;=\; I\big(\{g_j^{(r)}\}_{j\notin S,\, r\leq T};\ \{\Phi_H^{(r)}\}_{r\leq T}\big).$$

Multi-round leakage collapses, with no loss, to the information in the sequence of aggregates.

Proof. We proceed by induction on T.

Base case ($T = 1$): This is Lemma 1.

Inductive step: Assume the claim holds for $T-1$ rounds: there exists a simulator Sim_{T-1} that reconstructs the coalition's view for rounds $1,\ldots,T-1$ from the coalition's inputs and $\{\Phi_H^{(r)}\}_{r<T}$ alone. For round T, the only cross-round information is *public*: the model iterate $w^{(T-1)}$ is a deterministic function of past broadcasts, already present in Sim_{T-1}'s output. Conditioned on this shared public state, round T is a fresh independent run of VGA encoding $\Phi_H^{(T)}$; Lemma 1 supplies a simulator $\mathrm{Sim}_1^{(T)}$ for this single round from $\Phi_H^{(T)}$ alone. Composing Sim_{T-1} with $\mathrm{Sim}_1^{(T)}$ gives the joint simulator for all T rounds. The mutual information identity follows: exact simulation means the real and ideal joint distributions coincide, so by the data-processing inequality $I(\text{hidden gradients}; \text{view}_S) = I(\text{hidden gradients}; \{\Phi_H^{(r)}\})$. □

Theorem 3 reduces the question of "What does running VGA for T rounds reveal?" to the question "What does the sequence of sums reveal?". For Gaussian inputs with a static personal component μ_j contributing through the *same* linear functional $M = \sum_j \mu_j$ every round, data processing gives $I(\mu_1; \{\Phi_H^{(r)}\}_{r\le T}) \le \frac{1}{2}\ln\frac{N}{N-1}$, a bound that is independent of T. In regimes where gradients genuinely change across rounds, the sums-only reduction makes Gaussian differential privacy efficient: nodes clip and add local noise inside their phase encoding, the N independent noise contributions pool in the aggregate, and the released sum has $\sqrt{N}$ times less per-node noise than local differential privacy (DP) [18, 15] (a framework that bounds how much any single node's data can shift the output distribution, with *local* DP meaning each node adds calibrated noise before any sharing).

The following lemma extends these guarantees from the ideal oracle to the real VGA protocol against a fully adaptive adversary.

Lemma 3 (Adaptive chaining)**.** *Let Π be T-round real VGA and $\Pi^\star$ the ideal sums-only oracle. If each round of Π, against any adversary adaptive on the public history, is ε_r-close in trace distance to the corresponding ideal round, then Π and $\Pi^\star$ are $\left(\sum_{r=1}^T \varepsilon_r\right)$-close in trace distance, even against a fully adaptive cross-round adversary.*

Proof. Define a sequence of *hybrid* experiments (a hybrid argument interpolates between two protocols one step at a time, bounding their distance via the triangle inequality over consecutive swaps) $H_0 = \Pi, \ldots, H_T = \Pi^\star$, where H_m runs the ideal round for $r \le m$ and the real round for $r > m$. Adjacent hybrids differ only in round $m{+}1$; conditioned on the identical public history up to round m, they are the real vs. ideal single round and therefore ε_{m+1}-close. Trace distance is non-increasing under the common CPTP continuation (remaining rounds plus Eve's processing), so $\|H_m - H_{m+1}\|_1 \le \varepsilon_{m+1}$. The triangle inequality over $m = 0,\ldots,T{-}1$ gives $\|\Pi - \Pi^\star\|_1 \le \sum_r \varepsilon_r$. Adaptivity is handled because the hybrid swap in round $m{+}1$ is conditioned on the shared public history, generated identically by both H_m and H_{m+1} for all earlier rounds. □

Together, Theorem 3 and lemma 3 show that T-round VGA leaks no more than the sequence of gradient sums it releases: the ideal multi-round view collapses exactly to the sums (Theorem 3), and real VGA is within $\sum_r \varepsilon_r$ in trace distance of that ideal (Lemma 3), with ε_r the per-round bound supplied by Proposition 1.

5 Quantum advantage in gradient conflict detection under bandwidth constraints

In distributed training, the ring all-reduce is not the end of the communication story. Once workers have aggregated their gradients, a server or parameter node typically distills the result into a compact signal and broadcasts it back to clients, each of which must interpret that signal against its own private model structure. The bandwidth available for this downstream broadcast can often be far smaller than what the ring itself uses: clients may be edge devices or be connected over a low-rate link. How much useful information a bandwidth-limited message can carry, and whether quantum communication offers an advantage, is the question we will now explore.

In large-scale distributed training with heterogeneous or adversarial data distributions, *gradient conflicts* are a central obstacle [34, 19]: two workers' gradient vectors may point in nearly opposite

directions, and naively averaging them cancels signal or slows convergence. Detecting such conflicts after the all-reduce allows practitioners to selectively re-weight or drop conflicting updates, an idea that underlies conflict-aware federated learning algorithms and untying/refactoring schedules in over-parameterised networks.

The baseline for conflict detection is to broadcast all gradients: $O(Pb)$ bits per worker. The question is how much cheaper this can be made with quantum resources. This section characterises the quantum-classical separation for two natural variants of the problem, which arise from two different ways of formalising "gradient conflict":

- *Margin conflict detection (*$\textsc{GapIP}_\tau$*):* two workers, each holding a unit-normalised gradient, test whether the signed inner product of their gradients falls above or below a threshold τ. The quantum advantage is **quadratic**, tight on both sides.
- *Relational sign audit (*$\textsc{TieAudit}_\epsilon$*):* a server holds the global sign-gradient produced by signSGD; each client holds a private matching of parameter pairs (tied weights) and wants the fraction of its ties on which the global update is sign-inconsistent. The quantum advantage is **exponential** in $\log P$, arising from the hidden-matching structure of the relational task.

5.1 Problem definitions

Throughout this section, $P \in \mathbb{Z}_{>0}$ is the number of model parameters, $\tau \in (0, 1)$ a margin threshold, and $\epsilon \in (0, \frac{1}{2})$ an additive estimation error. Communication complexity is one-way (client to server) unless stated otherwise. We work with symmetric resources: both quantum and classical protocols may use input-independent pre-shared correlations (EPR pairs or shared randomness respectively). Readers new to one-way communication complexity and shared-resource models may consult Section B for a self-contained primer; Section C covers the Boolean Hidden Matching problem on which Theorem 5 rests.

Definition 5 (Margin conflict detection: $\textsc{GapIP}_\tau$)**.** *Two workers hold unit-norm gradients $g_j, g_k \in \mathbb{R}^P$, $\|g_j\| = \|g_k\| = 1$. Given the promise that $|\langle g_j, g_k\rangle| \geq \tau$, decide whether $\langle g_j, g_k\rangle \geq \tau$ (*aligned*) or $\langle g_j, g_k\rangle \leq -\tau$ (*conflicted*), with error probability at most $\frac{1}{3}$. Communication is from one worker to the other (one-way) or interactive.*

Definition 6 (Relational sign statistics: $\textsc{TieAudit}_\epsilon$)**.** *A server holds the global sign-gradient $s \in \{\pm 1\}^P$, where $s_c = \mathrm{sign}([\bar{g}]_c)$ is the coordinate-wise sign of the all-reduce output from signSGD [7]. Each client k holds a private perfect matching M_k on the coordinate set $[P] = \{1, \ldots, P\}$: a set of $P/2$ disjoint index pairs (i, j) induced by parameter ties (e.g. shared weights in a convolutional layer, factor pairs in a matrix factorisation, or quantisation bins). A pair $(i, j) \in M_k$ is* sign-inconsistent *if $s_i \neq s_j$: the global update disagrees on the tied coordinates. The task is to estimate the fraction of conflicted pairs*

$$f_k \;=\; \frac{|\{(i,j) \in M_k : s_i \neq s_j\}|}{|M_k|}$$

to additive $\pm\epsilon$, given a one-way message from server to client. This is an instance of the Boolean Hidden Matching *problem [16]: the server holds a bit string $x \in \{0, 1\}^P$, the receiver holds a perfect matching M on $[P]$ and a label vector $w \in \{0, 1\}^{P/2}$ with the promise that the edge-parity vector $(x_i \oplus x_j)_{(i,j)\in M}$ equals w or $\bar{w}$; the task is to decide which. Setting $x_c = \mathbf{1}[s_c = +1]$ identifies the two problems. A self-contained account of Boolean Hidden Matching is given in Section C.*

5.2 $\textsc{GapIP}$

The $\textsc{GapIP}$ problem can be described as deciding in which side of $\pi/2$ the angle between g_j and g_k falls on (to within margin τ). These types of angle decisions are precisely the use-case for amplitude estimation: namely, encoding each gradient into a quantum state, applying a Grover reflection product, and reading off the resulting angle without reconstructing the vectors. Resolving an angle to precision $\Theta(\tau)$ takes $O(1/\tau)$ reflections, independent of P; only the cost of a single round-trip carries the $\log P$ dependence.

Classically the angle is also defined, but there is no classical mechanism analogous to coherent phase estimation for resolving it: any estimator built from samples or random projections is bound by the standard $1/\sqrt{n}$ concentration rate, so precision τ forces $n = \Theta(\tau^{-2})$ samples, by Johnson–Lindenstrauss [13]. This is the same shot-noise-versus-Heisenberg contrast as Theorem 1: amplitude estimation reaches τ precision in $O(1/\tau)$ rounds where classical sampling needs $O(1/\tau^2)$.

We formalize this separation below.

Theorem 4 (Gap Inner Product separation)**.** *With symmetric resources:*

1. Quantum upper bound: *$\textsc{GapIP}_\tau$ is solvable with $\widetilde{O}(\tau^{-1}\log P)$ qubits of interactive communication. See D for the protocol implementation.*

2. Classical bounds: *The randomised communication complexity satisfies $R(\textsc{GapIP}_\tau) = \widetilde{\Theta}(\min(\tau^{-2}, P))$ bits.*

3. Quantum lower bound: *$Q(\textsc{GapIP}_\tau) = \widetilde{\Omega}(\tau^{-1})$ qubits. The separation is quadratic and the quantum protocol of (1) is optimal up to log factors.*

Proof. 1. *Quantum upper bound:* Augment each unit vector to remove the sign ambiguity: set $\hat{u} = (1, g_j)/\sqrt{2}$ and $\hat{v} = (1, g_k)/\sqrt{2}$ so that $\langle \hat{u}, \hat{v}\rangle = \frac{1+\langle g_j, g_k\rangle}{2} \in [0, 1]$ is a monotone re-scaling of the signed inner product. The two-party Grover operator $G = (2\,|\hat{u}\rangle\langle\hat{u}| - I)(2\,|\hat{v}\rangle\langle\hat{v}| - I)$ rotates by angle $\theta = \arccos(\langle\hat{u}, \hat{v}\rangle)$; recovering $\langle g_j, g_k\rangle$ to precision $\pm\tau$ is equivalent to estimating θ to precision $\Theta(\tau)$. Kitaev phase estimation on G [10] achieves this with $O(1/\tau)$ applications of G; each application requires one round-trip of $O(\log P)$ qubits. Total: $\widetilde{O}(\tau^{-1}\log P)$ qubits.

2. *Classical bounds:*

   - *Upper:* draw a random matrix $R \in \mathbb{R}^{k\times P}$ with i.i.d. $\mathcal{N}(0, 1/k)$ entries (shared randomness so all parties hold the same R). One worker broadcasts Rg_j (a k-dimensional sketch). By the distributional Johnson-Lindenstrauss lemma, for $k = O(\tau^{-2}\log(1/\delta))$, $\langle Rg_j, Rg_k\rangle$ approximates $\langle g_j, g_k\rangle$ to $\pm\tau/2$ with probability $1-\delta$. Total: $O(\tau^{-2})$ words, i.e. $\widetilde{O}(\tau^{-2})$ bits. The trivial upper bound of $O(P)$ bits (send the whole vector) gives $R(\textsc{GapIP}_\tau) = \widetilde{O}(\min(\tau^{-2}, P))$.
   - *Lower:* Gap-Hamming Distance (GHD_n) on n bits reduces to $\textsc{GapIP}_\tau$ with $n = 4/\tau^2$: encode Alice's string as $u = (-1)^x/\sqrt{n} \in \mathbb{R}^n$, so that gap τ in inner product corresponds exactly to the Hamming gap in GHD. By [12], $R(\mathrm{GHD}_n) = \Omega(n)$, giving $R(\textsc{GapIP}_\tau) = \Omega(\tau^{-2})$.

3. *Quantum lower bound:* The same embedding into GHD_n with $n = 4/\tau^2$ gives $Q(\textsc{GapIP}_\tau) \geq Q(\mathrm{GHD}_{4/\tau^2})$. The quantum lower bound $Q(\mathrm{GHD}_n) = \Omega(\sqrt{n})$ follows from a blow-up reduction from set-intersection [29] (repeat each coordinate $\sqrt{n}$ times; Razborov's symmetric-predicate bound applies at the threshold jump). Hence $Q(\textsc{GapIP}_\tau) \geq \Omega(\sqrt{4/\tau^2}) = \Omega(1/\tau)$, matching the upper bound of (1) to log factors.

□

Remark 3 (Scope of the quadratic advantage)**.** *The quantum protocol wins when $P > 1/\tau$: otherwise the classical $O(P)$-bit upper bound matches the quantum $O(1/\tau)$ bound and there is no separation. For N workers, classical conflict detection broadcasts one sketch per worker ($\widetilde{O}(N/\tau^2)$ total), while the quantum protocol is inherently pairwise ($\widetilde{O}(N^2/\tau)$ total rounds); quantum wins in total communication only when $N \lesssim 1/\tau$. Finally, the Heisenberg scaling in (1) requires interaction: any one-round protocol estimating a state overlap to $\pm\tau$ is shot-noise limited at $\widetilde{\Omega}(\tau^{-2})$ copies [10], so the quadratic communication advantage and $O(1)$-round operation are mutually exclusive.*

5.3 TieAudit$_\epsilon$

The margin result shows quantum is faster for *numerical* conflict scores, by a quadratic factor. The exponential separation below does not come from precision, but the relational nature of the task. The main distinction is between a *value* and a *relation*. A value task asks for a number (e.g. $\langle g_j, g_k \rangle$) that tolerates additive error and can be compressed by sketching or sampling. A relational task asks for one valid fact drawn from a family that the *receiver* privately chooses; no fixed short message can serve all possible matchings, because the receiver's choice is made *after* the sender commits. This is why TieAudit resists classical compression while admitting a small quantum message: the sender's phase state encodes all P signs simultaneously in its amplitudes, and *any* matching the receiver holds can extract a pair's parity in a single measurement, because the wrong-parity amplitude is exactly zero.

Theorem 5 (Tie-audit separation)**.** *TieAudit$_\epsilon$ has one-way communication complexity*

$$\textit{Classical:}\ \ \Omega(\sqrt{P}), \qquad \textit{Quantum:}\ \ O(\epsilon^{-2}\log P).$$

Proof. 1. *Quantum upper bound:* The server prepares and sends $O(\epsilon^{-2})$ independent copies of the $\lceil \log_2 P \rceil$-qubit phase state

$$|s\rangle \;=\; \frac{1}{\sqrt{P}} \sum_{c=1}^{P} s_c \, |c\rangle \,,$$

a coherent superposition over parameter indices with sign-gradient amplitudes. A client holding matching M_k measures each copy in the *matching basis* $\big\{(|i\rangle + |j\rangle)/\sqrt{2},\ (|i\rangle - |j\rangle)/\sqrt{2}\big\}_{(i,j)\in M_k}$. This measurement collapses to a uniformly random pair $(i,j) \in M_k$ with outcome $+$ if $s_i = s_j$ and $-$ if $s_i \neq s_j$: the wrong-parity amplitude is exactly zero, because $s_i \neq s_j$ implies $\langle i| + \langle j|$ and $s_i\,|i\rangle + s_j\,|j\rangle$ are orthogonal. Averaging $O(\epsilon^{-2})$ outcome bits estimates f_k to additive $\pm\epsilon$ by Chernoff bounds. Total: $O(\epsilon^{-2}\log P)$ qubits.

2. *Classical lower bound:* Any estimator of f_k to additive $\pm\frac{1}{4}$ distinguishes the two promise branches of Boolean Hidden Matching: the fractions $\mathrm{wt}(w)/(P/2)$ and $1 - \mathrm{wt}(w)/(P/2)$ differ by at least $\frac{1}{2}$ for balanced w, so an additive-$\frac{1}{4}$ estimate identifies the branch. By the classical lower bound for Boolean Hidden Matching [16], any classical one-way protocol for this task requires $\Omega(\sqrt{P})$ bits.

□

It is important to note that the exponential separation concerns communication required to estimate a relational statistic of the aggregated gradient; **it does not** by itself **imply** an exponential **improvement in training performance**.

Remark 4 (Why relational tasks resist compression)**.** *No classical sketch of $o(\sqrt{P})$ bits can serve all possible client matchings, because the matching M_k is chosen after the server's message is fixed. A Johnson-Lindenstrauss sketch[2] preserves inner products and norms but not the parity of a specific coordinate pair under an unanticipated matching. The quantum state $|s\rangle$ succeeds precisely because it encodes all P signs in its amplitude structure and lets the receiver's choice of measurement basis post-select onto the relevant pair without any prior knowledge of the matching.*

Remark 5 (Practical scope)**.** *Theorem 5 establishes the first exponential one-way quantum advantage for a statistic of the aggregated gradient in server-to-client gradient communication. Two things to consider alongside the result.*

1. *Training benefit is open: the natural use is an adaptive untying schedule that breaks ties whose conflict fraction remains high across rounds (e.g. splitting a shared weight when the global update consistently disagrees with the tie). Whether consuming the TieAudit statistic improves convergence is plausible but must be demonstrated experimentally.*

[2] A Johnson-Lindenstrauss (JL) sketch compresses a vector $s \in \mathbb{R}^P$ to $k = O(\varepsilon^{-2}\log(1/\delta))$ dimensions via a random linear map $A \in \mathbb{R}^{k\times P}$, with entries drawn i.i.d. $\pm 1/\sqrt{k}$ or Gaussian. The JL lemma guarantees $|\langle As, Au\rangle - \langle s, u\rangle| \le \varepsilon$ for any *fixed* query vector u with probability $1-\delta$. The sketch is therefore useful when the query direction is known at encoding time, but cannot answer queries whose structure (here, the client matching M_k) is revealed only after the message is sent.

2. *Quantum states cannot be broadcast, so serving N clients requires N separate copies of $|s\rangle$: total cost $\Theta(N\epsilon^{-2}\log P)$ qubits versus one classical $O(\sqrt{P})$-bit broadcast. The per-client advantage is exponential; the system-wide quantum advantage holds only when $N \ll \sqrt{P}/(\epsilon^{-2}\log P)$.*

6 Conclusion and outlook

We introduced a quantum version of ring all-reduce and characterised the advantages that quantum communication can provide for distributed learning. Our results carry several qualifications. Our first result is a bandwidth-optimal quantum ring all-reduce primitive that uses pre-shared entanglement and superdense coding to reduce per-link communication by exactly a factor of two while preserving the classical round structure. We proved that this factor is optimal, and no entanglement-assisted protocol can reduce the communication below the entanglement-assisted Holevo limit. In the second result, the bandwidth analysis assumes noiseless quantum channels and perfect EPR pair generation; realistic hardware introduces fidelity loss that erodes the factor-of-two gain and requires a separate noise analysis. The VGA security proof is asymptotic: as noted in Section 4, finite-key constants and a noise-tolerant verification threshold remain to be established. The gradient conflict detection results are worst-case communication complexity separations; whether the quantum advantage materialises in practice depends on the overhead of preparing and measuring the required states, which remains non-trivial on current hardware. Finally, quantum states cannot be broadcast: serving many clients requires independent copies of the quantum message, limiting system-wide advantage to the regime identified in Remark 5.

Several avenues remain open for future work:

- On the learning side, it remains to determine which gradient statistics beyond those studied here admit substantial quantum communication advantages, and whether acting on TieAudit output to drive adaptive model corrections yields measurable convergence improvements.

- Regarding physical realism, every advantage we establish here assumes high-fidelity entanglement and noiseless channels, so real hardware will require entanglement distillation, quantum repeater networks to distribute entanglement across the distances separating workers, and ultimately fault-tolerant entanglement distribution. Part of this is already built into VGA: its verification is a fidelity test, so noise tolerance amounts to setting its acceptance threshold above the hardware noise floor rather than demanding perfect agreement.

- Topology is a another such question: we use the ring because it is the bandwidth-optimal collective for data-parallel training, but the GHZ phase-encoding formulation of Section 3.3 is largely topology-agnostic, as the aggregate $\sum_j \theta_j$ is a global property of the shared GHZ state rather than of any particular interconnect, and it is natural to ask whether the same advantages extend to tree all-reduce, parameter-server architectures, and the hierarchical collectives that combine intra- and inter-node reduction. The per-link superdense-coding saving transfers to any topology directly, whereas the security construction depends on how the GHZ state is distributed and verified.

Settling these questions will require characterising the end-to-end overhead of the protocol in a realistic network architecture, including entanglement generation rates, qubit coherence requirements, and integration with collective communication libraries such as NCCL and Horovod, which will ultimately decide whether the bandwidth saving and the privacy guarantee remain net-positive once the cost of producing the states they consume is counted.

More broadly, our results suggest a hierarchy of quantum advantages for distributed learning. When the task is to communicate the value of a global aggregate, quantum communication is constrained by information-theoretic limits and yields at most constant-factor improvements. When the task is to certify the process by which the aggregate was generated, quantum resources enable privacy guarantees unavailable to classical protocols under the same assumptions. When the task is to evaluate relational properties of the aggregate relative to private local structure, exponential communication advantages become possible. Understanding which learning primitives fall into each of these regimes is, in our view, a promising direction for future work.

7 Acknowledgments

This research is partially supported by the National Research Foundation, Singapore through the National Quantum Office, hosted in A*STAR, under its Centre for Quantum Technologies Funding Initiative (S24Q2d0009). MGG is funded by the EPSRC UK Quantum Technologies Programme under grant EP/T001062/1 and VeriQloud. The authors extend their gratitude to the Centre for Quantum Technologies in Singapore for hosting Hackamonth 2026, where this project was developed.

References

[1] Rudolf Ahlswede and Imre Csiszár. Common randomness in information theory and cryptography. i. secret sharing. *IEEE Transactions on Information Theory*, 39(4):1121–1132, 1993.

[2] Ziv Bar-Yossef, T. S. Jayram, and Iordanis Kerenidis. Exponential separation of quantum and classical one-way communication complexity. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing (STOC)*, pages 128–137, 2004.

[3] Tal Ben-Nun and Torsten Hoefler. Demystifying parallel and distributed deep learning: An in-depth concurrency analysis. *ACM Comput. Surv.*, 52(4), August 2019.

[4] Charles H Bennett, Peter W Shor, John A Smolin, and Ashish V Thapliyal. Entanglement-assisted capacity of a quantum channel and the reverse shannon theorem. *IEEE transactions on Information Theory*, 48(10):2637–2655, 2002.

[5] Charles H Bennett and Stephen J Wiesner. Communication via one-and two-particle operators on einstein-podolsky-rosen states. *Physical review letters*, 69(20):2881, 1992.

[6] Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimisation for non-convex problems. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 560–569. PMLR, 2018.

[7] Jeremy Bernstein, Jiawei Zhao, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD with majority vote is communication efficient and fault tolerant. In *International Conference on Learning Representations (ICLR)*, 2019. arXiv:1810.05291.

[8] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1175–1191, 2017.

[9] Niek J Bouman and Serge Fehr. Sampling in a quantum population, and applications. In *Annual Cryptology Conference*, pages 724–741. Springer, 2010.

[10] Gilles Brassard, Peter Høyer, Michele Mosca, and Alain Tapp. Quantum amplitude amplification and estimation. In *Quantum Computation and Quantum Information: A Millennium Volume*, volume 305 of *AMS Contemporary Mathematics*, pages 53–74. American Mathematical Society, 2002. arXiv:quant-ph/0005055.

[11] Samuel L Braunstein and Carlton M Caves. Statistical distance and the geometry of quantum states. *Physical Review Letters*, 72(22):3439, 1994.

[12] Amit Chakrabarti and Oded Regev. An optimal lower bound on the communication complexity of Gap-Hamming-Distance. *SIAM Journal on Computing*, 41(5):1299–1317, 2012.

[13] Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

[14] Mina Doosti, Ryan Sweke, and Chirag Wadhwa. Distributed quantum property testing with communication constraints. *arXiv preprint arXiv:2604.05962*, 2026.

[15] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference (TCC)*, volume 3876 of *Lecture Notes in Computer Science*, pages 265–284. Springer, 2006.

[16] Dmitry Gavinsky, Julia Kempe, Iordanis Kerenidis, Ran Raz, and Ronald de Wolf. Exponential separations for one-way quantum communication complexity, with applications to cryptography. *SIAM Journal on Computing*, 38(5):1695–1708, 2008. Earlier version in STOC 2007, pp. 516–525.

[17] Dar Gilboa, Hagay Michaeli, Daniel Soudry, and Jarrod R. McClean. Exponential quantum communication advantage in distributed inference and learning. In *Advances in Neural Information Processing Systems*, volume 37, pages 30425–30473, 2024. arXiv:2310.07136.

[18] Peter Kairouz, Ziyu Liu, and Thomas Steinke. The distributed discrete gaussian mechanism for federated learning with secure aggregation. In *International conference on machine learning*, pages 5201–5212. PMLR, 2021.

[19] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 2020.

[20] Alexei Yu. Kitaev, Alexander H. Shen, and Mikhail N. Vyalyi. *Classical and Quantum Computation*, volume 47 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, Rhode Island, 2002.

[21] Changhao Li, Niraj Kumar, Zhixin Song, Shouvanik Chakrabarti, and Marco Pistoia. Privacy-preserving quantum federated learning via gradient hiding. *Quantum Science and Technology*, 9(3):035028, 2024. arXiv:2312.04447.

[22] Changhao Li, Boning Li, Omar Amer, Ruslan Shaydulin, Shouvanik Chakrabarti, Guoqing Wang, Haowei Xu, Hao Tang, Isabel Schoch, Niraj Kumar, Charles Lim, Ju Li, Paola Cappellaro, and Marco Pistoia. Blind quantum machine learning with quantum bipartite correlator. *Physical Review Letters*, 133(12):120602, 2024. arXiv:2310.12893.

[23] Shen Li, Yanli Zhao, Rohan Varma, Omkar Salpekar, Pieter Noordhuis, Teng Li, Adam Paszke, Jeff Smith, Brian Vaughan, Pritam Luo, and Soumith Chintala. PyTorch distributed: Experiences on accelerating data parallel training. *Proceedings of the VLDB Endowment*, 13(12):3005–3018, 2020.

[24] Ueli M Maurer. Secret key agreement by public discussion from common information. *IEEE transactions on information theory*, 39(3):733–742, 1993.

[25] Jörn Müller-Quade and Renato Renner. Composability in quantum cryptography. *New Journal of Physics*, 11(8):085006, 2009.

[26] Arnab Nath and Harsh Kasyap. CQSA: Byzantine-robust clustered quantum secure aggregation in federated learning. *arXiv preprint arXiv:2602.22269*, 2026. FL-AsiaCCS 2026.

[27] Pitch Patarasuk and Xin Yuan. Bandwidth optimal all-reduce algorithms for clusters of workstations. *Journal of Parallel and Distributed Computing*, 69(2):117–124, 2009.

[28] Lirandë Pira and Chris Ferrie. An invitation to distributed quantum neural networks. *Quantum Machine Intelligence*, 5(2):23, 2023.

[29] Alexander A. Razborov. Quantum communication complexity of symmetric predicates. *Izvestiya: Mathematics*, 67(1):145–159, 2003. arXiv:quant-ph/0204025.

[30] Renato Renner. Security of quantum key distribution. *International Journal of Quantum Information*, 6(01):1–127, 2008.

[31] Alexander Sergeev and Mike Del Balso. Horovod: fast and easy distributed deep learning in TensorFlow. *arXiv preprint arXiv:1802.05799*, 2018.

[32] Marco Tomamichel and Renato Renner. Uncertainty relation for smooth entropies. *Physical review letters*, 106(11):110506, 2011.

[33] William K. Wootters and Wojciech H. Zurek. A single quantum cannot be cloned. *Nature*, 299(5886):802–803, 1982.

[34] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. arXiv:2001.06782.

[35] Yichi Zhang, Chao Zhang, Cai Zhang, Lixin Fan, Bei Zeng, and Qiang Yang. Federated learning with quantum secure aggregation. *arXiv preprint arXiv:2207.07444*, 2022.

Appendix

A Superdense coding primer

Superdense coding [5] is the protocol underlying the factor-of-two bandwidth saving of Section 3. It is the mirror image of teleportation: teleportation spends one shared EPR pair and two classical bits to move one qubit, whereas superdense coding spends one shared EPR pair and one transmitted qubit to move two classical bits. We restate it here for readers coming from the distributed-learning side, in the same accessible spirit as the Boolean Hidden Matching primer of Section C.

A.1 The protocol

Alice and Bob share one EPR pair in the state

$$\left|\Phi^+\right\rangle = \frac{1}{\sqrt{2}}\big(|00\rangle + |11\rangle\big),$$

with Alice holding the first qubit and Bob the second. Crucially, this pair is distributed in advance, during the input-independent setup phase, before Alice knows which bits she will send. To transmit two classical bits $(a, b) \in \{0, 1\}^2$, Alice applies one of four Pauli operations to her half *alone* and then sends that single qubit to Bob. The four operations rotate the shared pair onto the four mutually orthogonal Bell states:

(a, b)	Alice applies	resulting Bell state
00	I	$\|\Phi^+\rangle = (\|00\rangle + \|11\rangle)/\sqrt{2}$
01	X	$\|\Psi^+\rangle = (\|01\rangle + \|10\rangle)/\sqrt{2}$
10	Z	$\|\Phi^-\rangle = (\|00\rangle - \|11\rangle)/\sqrt{2}$
11	ZX	$\|\Psi^-\rangle = (\|01\rangle - \|10\rangle)/\sqrt{2}$

Now holding both qubits, Bob performs a Bell measurement. Because the four states are mutually orthogonal, this measurement distinguishes them perfectly, and Bob recovers (a, b) exactly in the noiseless setting.

A.2 Why this gives the factor of two

Only one qubit crosses the channel during the online phase, yet two classical bits are delivered: the entanglement-assisted capacity of a noiseless qubit channel is two classical bits, and superdense coding saturates it [4]. This is exactly the substitution used in Section 3, where each classical message of m bits becomes $\lceil m/2 \rceil$ transmitted qubits on every ring link.

The saving is not a free lunch in total channel use. Each pair consumed online had to be distributed during setup, at a cost of one qubit channel-use across the same link, so the lifetime carrier count matches the classical baseline, consistent with the Holevo bound. What superdense coding changes is the *timing*: the setup traffic is input-independent and can be scheduled off the synchronisation critical path, leaving only half the volume inside the blocking all-reduce barrier. In the communication-bound regime of large-scale training, that relocation is the operationally relevant gain.

Classical systems can also exploit input-independent preprocessing, but pre-shared classical resources do not provide an analogue of superdense coding. The factor-of-two reduction in online communication arises specifically from entanglement-assisted communication.

B Communication-complexity primer

Communication complexity studies how much two parties must exchange in order to jointly compute a function of their separate inputs. It is important because in distributed systems the dominant cost is frequently not local computation but the data moved between machines, and communication complexity makes that cost precise: it yields lower bounds that hold no matter how clever the local processing is. The point is sharp in large-scale machine learning, where training is communication-bound and the exchange of gradients across devices, rather than the arithmetic on any one of them, sets the pace. A lower bound on the bits needed for a coordination task is then a statement that no implementation, present or future, can do better, and a quantum protocol that beats it is exploiting a genuinely non-classical effect. This is the setting for the separations of Section 5, and we collect here the notions our results use.

B.1 The model

Two parties, conventionally Alice and Bob, hold inputs x and y and wish to evaluate a function $f(x, y)$. Local computation is free; the only charged resource is the number of bits, or qubits, that they send to one another. The communication complexity of f is the least communication, over all protocols, that outputs $f(x, y)$ correctly. In our setting the inputs are gradient-derived. In GapIP (Definition 5) Alice holds g_j, Bob holds g_k, and they decide whether the two gradients are aligned or conflicting; the point is that they do so *without* exchanging the gradients themselves, since the naive solution sends a full P-dimensional vector (Pb bits) and GapIP asks how much smaller a message can be while still resolving the alignment question, which is the regime where a client must act on a gradient it cannot afford to receive in full. In TieAudit (Definition 6) the server holds the sign-gradient s and the client holds a private matching. A *matching* in this setting is a pairing of the P coordinate indices into $P/2$ disjoint pairs (i, j), each pair marking two parameters the client's own model treats as coupled, for instance weights shared across a convolutional filter, or two entries forced equal by a factorisation or a quantisation bin. It is *private* because it is fixed by the client's local structure and is never sent to the server, so the server's message must be useful for whatever pairing the client happens to hold.

B.2 One-way versus interactive

A protocol is *one-way* when it consists of a single message from one party to the other, who then produces the answer, and *interactive* when the parties exchange several messages, each able to depend on what was received so far. The distinction is not cosmetic. TieAudit is one-way: the server sends a message and the client decides, with no return channel, mirroring a server that broadcasts to bandwidth-limited clients. GapIP is solved interactively: the phase-estimation protocol of Section D shuttles a register back and forth, and Remark 3 shows that the same precision is unattainable in a single round.

B.3 Randomness and shared resources

If the parties must always be correct the problem is *deterministic*; if they may err with probability at most 1/3, a threshold that repetition can drive down, it is *randomised*. A randomised protocol may use *shared randomness*: a common random string, fixed independently of the inputs, that both parties can read. Shared randomness is a free, input-independent correlation, and it is essential for fairly bounding quantum advantage claims, since a quantum advantage should not originate from the mere presence of randomness but from our ability to exploit quantum-mechanical laws, such as quantum correlations, to outperform classical methods. The classical GapIP protocol of Theorem 4 uses randomness directly: both workers draw the same random projection matrix R from shared randomness and compute the sketches Rg_j and Rg_k, which would not agree if each worker sampled R privately.

The quantum counterpart of shared randomness, in our setting, is *pre-shared entanglement*, a supply of EPR pairs distributed before any input arrives (the multi-party analogue, used by the aggregation protocols of Section 3 and Section 4, is a shared GHZ state). Throughout Section 5 we use *symmetric resources*: classical protocols may use shared randomness and quantum protocols

may use pre-shared entanglement, both input-independent. This is what makes the comparison fair. A gap obtained by giving the quantum protocol a pre-shared correlation while denying the classical protocol any would be uninformative, because it could merely reflect that one side has a shared resource and the other does not. We instead equip each side with the strongest input-independent correlation of its own kind, shared randomness classically and entanglement quantumly, so that any remaining separation is attributable to quantum mechanics itself rather than to an asymmetry in free resources.

B.4 Qubits are not free

A quantum protocol sends qubits in place of bits. One might expect a qubit to carry unboundedly more than a bit, but it cannot: by Holevo's bound a single qubit conveys at most one classical bit of accessible information, and at most two even with unlimited pre-shared entanglement, the latter saturated by superdense coding [5, 4]. This ceiling is precisely the factor of two of Section 3. The advantages of Section 5 therefore cannot come from packing more data onto each carrier. They come from a different mechanism: the quantum message lets the receiver extract, through its own choice of measurement, an answer that no equally short classical message could have encoded in advance. This is also the mechanism behind the broader claim of the paper. The quantum communication layer introduced for the all-reduce in Section 3 does more than complete aggregation at lower online cost: the same layer, by letting receivers choose what to extract, enables capabilities with no classical counterpart at any communication cost, namely the information-theoretic privacy of Section 4 and the conflict-detection separations of the present section.

B.5 Separations, with examples

A *separation* compares the quantum cost $Q(f)$ with the classical cost $R(f)$ as functions of the input size. Two regimes appear in this paper.

A *quadratic* separation has $Q(f) = \widetilde{\Theta}(\sqrt{R(f)})$. The canonical example is Gap-Hamming Distance, in which the two parties decide whether their n-bit strings are close or far in Hamming distance: classically this requires $\Theta(n)$ bits [12], whereas $\widetilde{O}(\sqrt{n})$ qubits suffice, a quadratic gap. GapIP inherits exactly this separation through the reduction in Theorem 4.

An *exponential* separation has the classical cost polynomially large while the quantum cost is only polylogarithmic. The canonical example is Boolean Hidden Matching (Section C), where any classical one-way protocol needs $\Omega(\sqrt{P})$ bits but $O(\log P)$ qubits suffice. TieAudit inherits this separation. To make the gap concrete, a model with $P = 10^9$ parameters forces on the order of 3×10^4 classical bits per client, while the quantum message is a handful of states on $\lceil \log_2 P \rceil \approx 30$ qubits each.

These two examples also display the role of interaction noted in Section 5: the quadratic GapIP advantage needs an interactive protocol, whereas the exponential TieAudit advantage already holds for a single one-way message.

Note 1 (On terminology)**.** *We speak of* separations *rather than* advantages *deliberately. A separation is a proven gap between the quantum and classical communication complexity of a fixed task, resting on a lower bound for the classical side, and it is a worst-case, asymptotic statement about communication alone. The word "advantage" is looser and is often read as an end-to-end practical benefit. Keeping to "separation" makes precise what we claim and what we do not: the quantum protocols provably move less information, but, as Remark 5 stresses, this does not by itself establish a faster or better training procedure. What it does establish is a reduction in communication, the metric our ring all-reduce protocol is built to optimise; which in itself is a plausible route to more efficient training.*

C Boolean Hidden Matching primer

Boolean Hidden Matching is the problem underlying the exponential separation of Theorem 5. It was introduced by Gavinsky et al. [16], with an alternative construction by Bar-Yossef, Jayram, and Kerenidis [2], as the first example of an exponential gap between quantum and classical one-way

communication. We state it here in an accessible manner to explain both why it is hard classically and why a small quantum message solves it, building on the one-way model of Section B.

C.1 The problem

Alice holds a bit string $x \in \{0,1\}^P$. Bob holds a *perfect matching* M on the index set $[P] = \{1, \ldots, P\}$, that is, a set of $P/2$ disjoint pairs (i, j) that together cover every index once, together with a target vector $w \in \{0,1\}^{P/2}$, one bit per pair. The promise is that the *edge-parity vector*

$$\big(x_i \oplus x_j\big)_{(i,j)\in M}$$

equals either w or its bitwise complement $\bar{w}$. From a single message sent by Alice, Bob must decide which. The only thing that matters about x is, for each of Bob's pairs, the parity $x_i \oplus x_j$; but Alice does not know M, so she must commit her message before learning which pairs Bob will read.

For a concrete instance, take $P = 4$ and $x = (0, 0, 1, 1)$. The matching $M = \{(1, 2), (3, 4)\}$ yields parities $(x_1 \oplus x_2,\, x_3 \oplus x_4) = (0, 0)$, while the matching $M' = \{(1, 3), (2, 4)\}$ on the same x yields $(x_1 \oplus x_3,\, x_2 \oplus x_4) = (1, 1)$. The parities Bob extracts depend on the matching he holds, and Alice cannot tailor her message to a matching she has not seen.

C.2 Why it is hard classically

Alice's message is a function of x alone, fixed before M is revealed. To answer correctly for whatever matching Bob holds, it would have to preserve the parity $x_i \oplus x_j$ of an arbitrary disjoint pairing. There are about $P^2/2$ candidate pairs but only $P/2$ in any one matching, chosen adversarially after the message is sent, so any short summary of x, a random sketch or a subset of its bits, will miss the pairs Bob happens to ask about. Making this precise, Gavinsky et al. [16] show that any classical one-way protocol with constant advantage must send $\Omega(\sqrt{P})$ bits. The threshold sits at $\sqrt{P}$ rather than P because a message of that size already starts to cover pairs by a birthday-type coincidence; below it, almost every matching is effectively unseen.

C.3 Why a small quantum message suffices

Alice instead sends the *phase state*

$$|x\rangle \;=\; \frac{1}{\sqrt{P}} \sum_{c=1}^{P} (-1)^{x_c}\, |c\rangle\,,$$

a single register of $\lceil \log_2 P \rceil$ qubits that records all P bits at once in its signs. Bob, holding M, measures in the basis adapted to his matching, namely $\big\{(|i\rangle + |j\rangle)/\sqrt{2},\ (|i\rangle - |j\rangle)/\sqrt{2}\big\}_{(i,j)\in M}$. Restricted to one pair (i, j) the state is $\big((-1)^{x_i}\, |i\rangle + (-1)^{x_j}\, |j\rangle\big)/\sqrt{P}$, and its overlap with $(|i\rangle - |j\rangle)/\sqrt{2}$ is

$$\frac{(-1)^{x_i} - (-1)^{x_j}}{\sqrt{2P}},$$

which vanishes exactly when $x_i = x_j$. The measurement therefore returns a uniformly random pair $(i, j) \in M$ together with its parity, read off with no error: outcome $(|i\rangle + |j\rangle)/\sqrt{2}$ means $x_i \oplus x_j = 0$, and $(|i\rangle - |j\rangle)/\sqrt{2}$ means $x_i \oplus x_j = 1$.

A single measurement decides the problem. If the edge-parity vector is w, the sampled parity agrees with the corresponding bit of w; if it is $\bar{w}$, it disagrees. Comparing one extracted parity against w identifies the branch with constant advantage, using one message of $O(\log P)$ qubits, an exponential saving over the classical $\Omega(\sqrt{P})$.

C.4 From Boolean Hidden Matching to TieAudit

Theorem 5 is the estimation form of this problem. Setting $x_c = \mathbf{1}[s_c = +1]$ turns the sign-gradient into a bit string, so that $x_i \oplus x_j = 1$ precisely when $s_i \neq s_j$, that is, when the tied pair (i, j) is sign-inconsistent. The phase state coincides with the TieAudit state: $(-1)^{x_c} = -s_c$ gives $|x\rangle = -\,|s\rangle$,

the same physical state up to a global sign, and the client's matching-basis measurement is exactly the extraction above. The one difference is the output. Boolean Hidden Matching asks a single yes/no question, w versus $\bar{w}$, which one copy answers, whereas TieAudit asks for the *fraction* f_k of inconsistent pairs to additive accuracy $\pm\epsilon$. Estimating a fraction rather than deciding a binary promise is why TieAudit averages over $O(\epsilon^{-2})$ copies, for $O(\epsilon^{-2}\log P)$ qubits in total, while the $\Omega(\sqrt{P})$ classical bound carries over because an additive-$\frac{1}{4}$ estimate already decides the underlying branch.

D GapIP$_\tau$ Quantum protocol

D.1 Encoding

Worker j augments its unit vector to $\hat{u} = (1, g_j)/\sqrt{2} \in \mathbb{R}^{P+1}$ and locally prepares the amplitude-encoded state $|\hat{u}\rangle = U_j |0\rangle$ on $n = \lceil \log_2(P+1) \rceil = O(\log P)$ qubits, where U_j depends only on g_j; worker k likewise prepares $|\hat{v}\rangle = U_k |0\rangle$ from $\hat{v} = (1, g_k)/\sqrt{2}$. The augmentation does two things: $\langle \hat{u}, \hat{v} \rangle = \frac{1}{2}\big(1 + \langle g_j, g_k \rangle\big) \in [0, 1]$ is nonnegative, hence a *monotone* function of the signed inner product (overlap estimation natively returns the unsigned angle, and the offset restores the sign), and it sends the decision threshold $\langle g_j, g_k \rangle = \pm\tau$ to $\langle \hat{u}, \hat{v} \rangle = \frac{1}{2}(1 \pm \tau)$, in the interior of $[0, 1]$ where arccos is well-conditioned. Only communication is counted; the local cost of U_j (up to $O(P)$ gates) is not charged.

D.2 Local reflections

Each reflection is a local operation on the shared register: $R_{\hat{u}} = 2|\hat{u}\rangle\langle\hat{u}| - I = U_j R_0 U_j^\dagger$ with $R_0 = 2|0\rangle\langle 0| - I$ a multi-controlled phase, applied by worker j alone; symmetrically $R_{\hat{v}} = U_k R_0 U_k^\dagger$ by worker k.

D.3 Distributed Grover operator

The two-party operator $G = R_{\hat{u}} R_{\hat{v}}$ is realised on a single travelling n-qubit register: worker k applies $R_{\hat{v}}$, transmits the register to worker j, who applies $R_{\hat{u}}$ and returns it. One application of G (or of its controlled version, carrying an $O(1)$-qubit phase-estimation ancilla along with the register) therefore costs one round-trip of $O(\log P)$ qubits. Within the plane $\mathrm{span}\{\hat{u}, \hat{v}\}$, G is a rotation by 2θ with $\theta = \arccos\langle \hat{u}, \hat{v} \rangle$; it acts trivially on the orthogonal complement, and the initial state $|\hat{v}\rangle$ lies in the plane.

D.4 Estimation and decision

Iterative (Kitaev) phase estimation [20] on G, run on $|\hat{v}\rangle$, estimates the eigenphase 2θ (hence $\langle \hat{u}, \hat{v} \rangle = \cos\theta$, hence $\langle g_j, g_k \rangle$) to additive $\pm\Theta(\tau)$ using $O(1/\tau)$ applications of G, the Heisenberg rate ($\cos\theta$ is injective on $\theta \in [0, \pi]$, so the eigenphase sign is immaterial). Each application is one round-trip, giving $O(\tau^{-1}\log P)$ qubits; boosting the success probability to $\geq 2/3$ multiplies this by $O(\log(1/\delta))$, for $\widetilde{O}(\tau^{-1}\log P)$ total. The protocol is manifestly interactive ($O(1/\tau)$ sequential round-trips), consistent with Remark 3: no one-round protocol attains the Heisenberg rate.