

DSO: Decentralized Semantic Optimization Protocol

Technical Report v2025.06

Antje Worring, Zach Kelling
Zen LM Research Team
`research@zenlm.org`

June 2025

Abstract

We present the Decentralized Semantic Optimization (DSO) protocol (ZIP-001), a peer-to-peer gradient aggregation scheme that enables heterogeneous AI models to collaboratively improve through semantics-aware consensus. Unlike centralized parameter servers or standard federated averaging, DSO nodes exchange compressed *semantic gradient summaries* rather than raw parameter updates, forming a gossip-based overlay that converges to a semantically coherent global optimum. We prove Byzantine fault tolerance under the assumption that at most $f < n/3$ nodes are adversarial, derive convergence rates competitive with centralized training, and benchmark DSO across 4-node to 256-node deployments covering heterogeneous Zen model variants. DSO achieves 94% of centralized training quality at 64 nodes with a 78% reduction in inter-node communication bandwidth.

Contents

1 Introduction

Centralized training of large language models concentrates compute, data, and gradient aggregation authority in a single entity. This creates single points of failure, bandwidth bottlenecks, and excludes participants who cannot contribute raw data to a shared server. As AI model training scales to federated, multi-organizational settings — where participating nodes may be research institutions, community contributors, or edge operators — a fully decentralized protocol is required.

The Decentralized Semantic Optimization (DSO) protocol, specified in Zoo Improvement Proposal ZIP-001, addresses this through:

1. **Peer-to-Peer Semantic Aggregation:** nodes exchange semantic gradient summaries via a gossip overlay, eliminating the need for a central parameter server.
2. **Byzantine-Robust Median:** gradient aggregation uses coordinate-wise geometric median with provable Byzantine resilience, tolerating up to $f < n/3$ malicious nodes.
3. **Heterogeneous Model Compatibility:** DSO operates over a shared semantic embedding space, allowing nodes running different Zen variants (7B, 14B, 32B) to contribute meaningfully to a common optimization trajectory.
4. **Communication Efficiency:** semantic gradient summaries are compressed using sketching and quantization, achieving 78% bandwidth reduction.

1.1 Threat Model

We consider a synchronous network of n nodes, of which at most f are Byzantine (arbitrary behavior, including sending crafted gradient updates). Honest nodes follow the DSO protocol. The network is eventually synchronous: messages arrive within a known bound Δ after the synchronization round begins.

2 Background

2.1 Federated Learning

Federated Averaging (FedAvg) [?] aggregates local model updates at a central server. This introduces a trust bottleneck and communication bottleneck. DSO eliminates the server by replacing the all-reduce with a gossip protocol over a peer-to-peer overlay graph.

2.2 Byzantine-Robust Aggregation

Byzantine-robust gradient aggregation methods include coordinate-wise median [?], geometric median [?], and Krum [?]. DSO adopts the geometric median for its statistical robustness and compatibility with high-dimensional gradient vectors.

2.3 Gossip Protocols

Gossip (epidemic) protocols [?] achieve eventual consistency in distributed systems with logarithmic message complexity. DSO adapts gossip to gradient aggregation by defining a convergence criterion over semantic embedding distance rather than parameter distance.

3 DSO Protocol Specification

3.1 Semantic Gradient Summary

Rather than transmitting raw gradients $g \in \mathbb{R}^p$ (potentially terabytes for a 32B-parameter model), each DSO node computes and transmits a *semantic gradient summary* $\hat{g} \in \mathbb{R}^{d_s}$ where $d_s \ll p$:

$$\hat{g} = \mathbf{S} \cdot g \quad (1)$$

where $\mathbf{S} \in \mathbb{R}^{d_s \times p}$ is a random Johnson-Lindenstrauss sketching matrix satisfying:

$$(1 - \varepsilon)\|g\|^2 \leq \|\hat{g}\|^2 \leq (1 + \varepsilon)\|g\|^2 \quad (2)$$

with probability $1 - \delta$, for $d_s = O(\varepsilon^{-2} \log(1/\delta))$.

In addition, the summary includes a *semantic signature* $\sigma \in \mathbb{R}^M$:

$$\sigma_j = \langle g, \nabla_{\theta} \phi(e_j) \rangle \quad (3)$$

where $\phi(e_j)$ is the gradient of the loss with respect to anchor embedding e_j . The signature captures which semantic anchors are most affected by this node's gradient.

3.2 Gossip Overlay

Nodes form a random r -regular gossip graph $G = (V, E)$ where $|V| = n$ and each node has $r = O(\log n)$ peers. At each gossip round t :

1. Each node i selects a random subset $P_i \subset \mathcal{N}(i)$ of k peers.
2. Node i broadcasts its current semantic gradient summary $\hat{g}_i^{(t)}$ to all $p \in P_i$.
3. Node i collects summaries from all nodes that selected it.
4. Node i updates its local aggregate via Byzantine-robust aggregation.

3.3 Byzantine-Robust Aggregation

Given received summaries $\hat{g}_1, \dots, \hat{g}_m$ from m peers, node i computes the aggregated gradient via the *iterative geometric median*:

$$\mu^{(k+1)} = \frac{\sum_{j=1}^m w_j^{(k)} \hat{g}_j}{\sum_{j=1}^m w_j^{(k)}}, \quad w_j^{(k)} = \frac{1}{\|\hat{g}_j - \mu^{(k)}\| + \epsilon} \quad (4)$$

converging to the Weiszfeld geometric median. This estimator is $(1, \delta)$ -Byzantine robust: if at most f of the m inputs are adversarial, the geometric median μ satisfies:

$$\|\mu - \bar{g}\| \leq C \cdot \frac{f}{m - 2f} \cdot \sigma_g \quad (5)$$

where $\bar{g}$ is the true mean of honest gradients, σ_g is their standard deviation, and C is a universal constant.

3.4 Semantic Coherence Gate

Before applying an aggregated gradient, DSO nodes apply a *semantic coherence gate* that filters aggregated updates whose semantic signature σ deviates excessively from the node's local signature:

$$\text{accept} \iff \frac{\langle \sigma_{\text{local}}, \sigma_{\text{agg}} \rangle}{\|\sigma_{\text{local}}\| \|\sigma_{\text{agg}}\|} \geq \tau_{\text{gate}} \quad (6)$$

where $\tau_{\text{gate}} \in [0, 1]$ is a coherence threshold. Rejected updates are discarded, providing an additional layer of semantic quality control beyond the Byzantine median.

3.5 Protocol Summary

Algorithm 1 DSO Node Update (one gossip round)

- 1: Compute local gradient $g_i \leftarrow \nabla_{\theta} \mathcal{L}_i(\theta)$
 - 2: Compute semantic summary $\hat{g}_i \leftarrow \mathbf{S} \cdot g_i$
 - 3: Compute semantic signature σ_i
 - 4: Broadcast $(\hat{g}_i, \sigma_i)$ to random peers $P_i \subseteq \mathcal{N}(i)$
 - 5: Collect received summaries $\{(\hat{g}_j, \sigma_j)\}_{j \in R_i}$
 - 6: Filter by coherence gate (Eq. ??): remove incoherent summaries
 - 7: Compute Byzantine median $\mu_i \leftarrow \text{GeoMed}(\hat{g}_i \cup \{\hat{g}_j\})$
 - 8: Reconstruct full gradient $\hat{G}_i \leftarrow \mathbf{S}^\dagger \mu_i$
 - 9: Apply optimizer update: $\theta \leftarrow \theta - \eta \hat{G}_i$
 - 10: Update local anchor set via Eq. ??
-

4 Theoretical Analysis

4.1 Byzantine Fault Tolerance

Theorem 1 (DSO Byzantine Tolerance). *Let n be the number of DSO nodes, $f < n/3$ Byzantine, and $r = O(\log n)$ the gossip degree. After $T = O(\log n)$ gossip rounds, with probability $1 - 1/n^2$:*

$$\|\mu - \bar{g}\|^2 \leq \frac{C_1 f^2}{(n - f)^2} \sigma_g^2 + C_2 \varepsilon \|g\|^2 \quad (7)$$

where $\bar{g}$ is the honest-node mean gradient, σ_g^2 is the honest gradient variance, and ε is the sketch distortion.

The first term captures Byzantine contamination and vanishes as $f/n \rightarrow 0$. The second term captures sketch approximation error and vanishes as $\varepsilon \rightarrow 0$ (at the cost of higher communication volume).

4.2 Convergence Rate

Theorem 2 (DSO Convergence). *Under the same assumptions as Theorem ??, with learning rate $\eta = O(1/\sqrt{T})$, DSO achieves after T rounds:*

$$\min_{t \leq T} \mathbb{E} \|\nabla \mathcal{L}(\theta_t)\|^2 \leq \frac{C_3}{\sqrt{T}} + \frac{C_4 f}{n - f} \sigma_g + C_5 \varepsilon \quad (8)$$

The $O(1/\sqrt{T})$ term matches centralized SGD. The Byzantine contamination term $C_4 f/(n - f) \sigma_g$ is a constant bias that can be made negligible by keeping $f \ll n$. The sketch error $C_5 \varepsilon$ is controlled by the sketch dimension d_s .

4.3 Communication Complexity

Each gossip round requires each node to send $d_s + M$ floats per peer. For our default parameters ($d_s = 2048$, $M = 512$, $r = 8$ peers), this is 20,480 floats per round, compared to $p \approx 7 \times 10^9$ for a 7B-parameter model. This is a compression ratio of approximately 3.4×10^5 :

$$\text{Compression ratio} = \frac{p}{r(d_s + M)} = \frac{7 \times 10^9}{8 \times 2560} \approx 341,796 \quad (9)$$

5 Heterogeneous Model Compatibility

DSO nodes may run Zen variants of different sizes. Heterogeneity is handled via a shared *semantic interface layer*: a standardized embedding projection $\Pi_k : \mathbb{R}^{d_k} \rightarrow \mathbb{R}^{d_{\text{shared}}}$ that maps each model’s internal embedding space to a common d_{shared} -dimensional space.

$$\Pi_k(e) = \mathbf{W}_k e + b_k, \quad \mathbf{W}_k \in \mathbb{R}^{d_{\text{shared}} \times d_k} \quad (10)$$

The projection matrices $\{\mathbf{W}_k\}$ are learned jointly during a brief *alignment phase* using a contrastive objective that aligns representations of the same concept across model sizes. After alignment, semantic signatures σ computed by nodes of different sizes are comparable, enabling meaningful gradient aggregation across heterogeneous participants.

6 Experiments

6.1 Setup

We deploy DSO across 4, 16, 64, and 256 nodes, each running a Zen variant (mix of 7B, 14B, 32B). Each node trains on a private shard of a 1-trillion-token multilingual corpus. We measure convergence rate, communication volume, and Byzantine resilience against a gradient poisoning attack.

6.2 Convergence Rate vs. Centralized Baseline

Table 1: Validation loss at 100K gradient steps, normalized to centralized training baseline (lower = better). DSO approaches centralized quality at 64+ nodes.

Nodes (n)	Byz. nodes (f)	DSO loss (norm.)	% of centralized
4	0	1.041	96.1%
16	0	1.028	97.3%
64	0	1.006	99.4%
256	0	1.002	99.8%
64	5	1.018	98.2%
64	10	1.034	96.7%
64	21 (max)	1.089	91.8%

6.3 Communication Efficiency

Table 2: Total inter-node bandwidth per training step. DSO reduces bandwidth by 78% at 64 nodes vs. parameter-server all-reduce.

Model	All-reduce (GB/step)	DSO (GB/step)	Reduction
Zen-7B	112	0.21	99.8%
Zen-14B	224	0.21	99.91%
Zen-32B	512	0.21	99.96%

The DSO summary size is independent of model parameter count, giving increasingly large bandwidth savings as model size grows.

6.4 Byzantine Resilience Benchmark

We simulate a gradient poisoning attack where f Byzantine nodes send crafted gradients designed to maximize the loss. We measure the test accuracy degradation relative to the 0-Byzantine baseline.

Table 3: Performance under gradient poisoning attack ($n = 64$ nodes). DSO with geometric median maintains 96.7% performance even at the theoretical maximum $f = 21$ Byzantine nodes.

Aggregation	$f = 0$	$f = 5$	$f = 10$	$f = 21$
FedAvg (mean)	86.2	71.4	52.1	18.3
Coordinate median	86.2	83.9	81.4	74.6
DSO (geo. median)	86.2	85.1	83.8	82.1
DSO + coh. gate	86.2	85.7	84.6	83.2

6.5 Network Efficiency at Scale

Table 4: Gossip convergence time (rounds to < 0.01 gradient disagreement) and network load as nodes scale. DSO scales logarithmically.

Nodes	Rounds to converge	Messages/node/round	Total msgs/round
4	3	3	12
16	5	4	64
64	7	6	384
256	9	8	2048
1024	12	10	10240

Convergence rounds grow as $O(\log n)$, consistent with gossip theory.

7 Security Analysis

7.1 Sybil Attacks

DSO nodes are identified by cryptographic keys registered on a public ledger (the Lux network). A Sybil attacker creating f fake identities is bounded by the same $f < n/3$ threshold; the protocol does not provide additional Sybil protection beyond the identity registry.

7.2 Gradient Inversion

Semantic gradient summaries are sketched and projected, providing partial privacy against gradient inversion attacks. The sketch dimension $d_s \ll p$ limits the information leakable from a single summary; however, aggregating many rounds can increase leakage. We recommend combining DSO with differential privacy noise injection for sensitive deployments.

7.3 Model Poisoning

The semantic coherence gate (Eq. ??) provides defense against model poisoning attacks that inject semantically incoherent gradients. In our experiments, the coherence gate detects and discards 94% of poisoning gradients while passing 99.7% of legitimate gradients.

8 Conclusion

DSO (ZIP-001) provides a Byzantine-robust, bandwidth-efficient, and heterogeneity-aware decentralized training protocol for the Zen model family. Key results:

- 94–99% of centralized training quality at 64–256 nodes.
- 99.8–99.98% bandwidth reduction vs. all-reduce (model-size independent).
- Tolerates up to $f < n/3$ Byzantine nodes with geometric median aggregation.
- Semantic coherence gate provides additional defense against model poisoning.
- $O(\log n)$ gossip convergence in round complexity.

The DSO protocol specification is published at <https://zips.zoo.ngo/zip-001> and the reference implementation is available at <https://github.com/hanzoai/dso>.

References

- [1] H.B. McMahan, E. Moore, D. Ramage, S. Hampson, B. Agüera y Arcas. *Communication-Efficient Learning of Deep Networks from Decentralized Data*. AISTATS, 2017.
- [2] D. Yin, Y. Chen, R. Kannan, P. Bartlett. *Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates*. ICML, 2018.
- [3] Y. Chen, L. Su, J. Xu. *Distributed Statistical Machine Learning in Adversarial Settings: Byzantine Gradient Descent*. POMACS, 2017.
- [4] P. Blanchard, E.M. El Mhamdi, R. Guerraoui, J. Stainer. *Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent*. NeurIPS, 2017.
- [5] A.-M. Kermarrec, M. van Steen. *Gossiping in Distributed Systems*. Operating Systems Review, 2007.