HiZO: Hebbian-informed Zeroth-order Optimization for Efficient Fine-tuning

Anonymous authors
Paper under double-blind review

Abstract

Zeroth-order optimization enables language model fine-tuning without backpropagation, but isotropic perturbations do not exploit the task-dependent information contained in forward activations. We introduce Hebbian-informed Zeroth-order Optimization (HiZO), which uses this information to guide the perturbation distribution and improve convergence. HiZO uses the magnitude of centered input–output covariance to allocate perturbation variance across weight coordinates while preserving the total variance budget. The resulting guidance reflects both learned weights and downstream input statistics. These statistics are collected during existing loss evaluations and require no additional model forward passes. We also develop subsampled and separable variants to reduce the computation and storage needed for guidance. For smooth nonconvex objectives, we establish convergence bounds in the directional-derivative limit that characterize the roles of covariance–gradient alignment and estimator variability. A complementary one-layer analysis identifies conditions under which Hebbian covariance provides information about task-gradient magnitudes. Language-model experiments and controlled ablations show faster optimization with Hebbian guidance while retaining comparable or better predictive performance. These findings suggest that forward activations offer a practical source of task-dependent information for improving zeroth-order convergence.

1 Introduction

Backpropagation is the standard approach to fine-tuning pretrained language models (Rumelhart et al., 1986; Devlin et al., 2019), but storing activations for backward computation can be memory-intensive, and non-differentiable objectives prevent direct gradient computation. Parameter-efficient methods such as adapters and LoRA reduce the number of trainable parameters (Houlsby et al., 2019; Hu et al., 2022). Zeroth-order (ZO) optimization offers a complementary approach: it estimates updates from objective evaluations, avoiding backpropagation and accommodating non-differentiable objectives. Memory-efficient ZO (MeZO) enables language-model fine-tuning with a memory footprint close to inference (Malladi et al., 2023). With only a few objective evaluations per update, choosing informative perturbation directions becomes central to optimization efficiency.

A common ZO approach samples isotropic Gaussian directions, assigning equal expected squared perturbation to every trainable parameter (Nesterov & Spokoiny, 2017; Malladi et al., 2023). This sampling rule does not use the structure of the pretrained model or its downstream inputs. Guided Evolutionary Strategies uses surrogate gradients to shape exploration (Maheswaranathan et al., 2019), while sparse parameter selection (Liu et al., 2025), low-rank perturbations (Chen et al., 2025), and curvature estimates (Zhao et al., 2025) exploit different aspects of parameter structure. Activation-guided methods construct low-rank search subspaces during optimization or at initialization (Lin et al., 2026; Dong et al., 2026) (see Appendix A for a broader discussion of related work).

The co-variation of connected neurons provides information at the level of individual weights. These co-activations depend on both the model's learned weights and the task's input distribution. Motivated by the covariance rule in Hebbian learning (Gerstner et al., 2014), we use centered input–output covariance as Hebbian information. It requires only forward activations and can be collected during existing objective evaluations, without additional model passes or backward computation. Rather

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

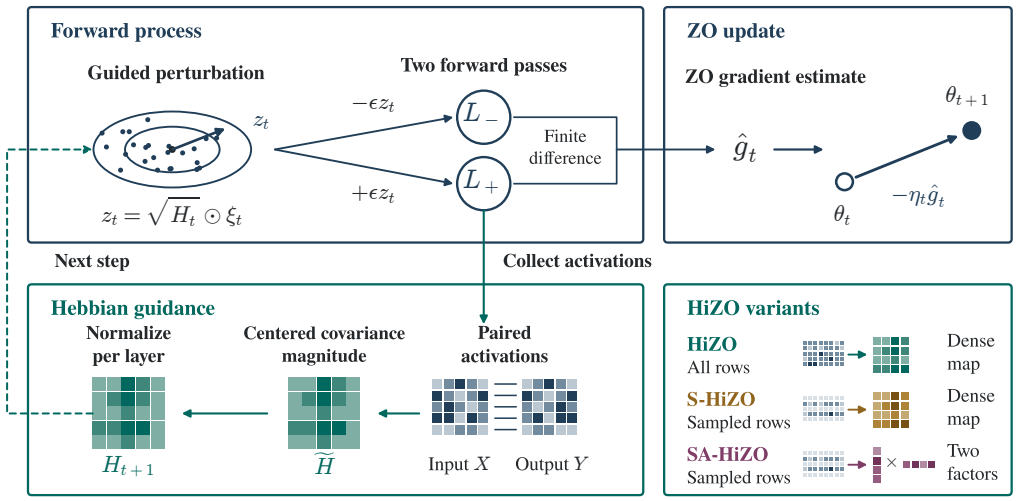


Figure 1: **Overview of HiZO and its variants.** Hebbian statistics guide perturbation sampling. Two forward losses $L_\pm$ yield the ZO estimate $\hat{g}_t$ for the parameter update. The positive evaluation also collects activations to construct H_{t+1}. S-HiZO subsamples activation rows; SA-HiZO additionally stores the variance weights as two factors per layer.

than constructing an activation-based subspace, this statistic describes the joint activity associated with each connection. We therefore ask:

Can Hebbian information help distribute the perturbation budget more effectively across weights?

We introduce **Hebbian-informed Zeroth-order Optimization (HiZO)**, which uses the magnitude of centered input–output covariance to score each connection and guide the sampled directions used in the zeroth-order updates. Layerwise normalization converts these scores into a diagonal Gaussian sampling covariance whose trace equals the number of trainable parameters. This preserves the total expected squared perturbation of isotropic sampling while allocating more variance to higher-scoring coordinates. A standard two-sided finite-difference update then uses the objective difference to determine the update along the sampled direction. Statistics collected during one update guide the next, with two objective evaluations per update (Figure 1).

The usefulness of this allocation depends on its alignment with the downstream gradient and its effect on estimator variability. Our analysis connects these quantities through covariance–gradient alignment and Gaussian fourth moments. For smooth nonconvex objectives, we establish convergence bounds in the directional-derivative limit and identify conditions under which Hebbian guidance improves the bound over isotropic exploration. Further analysis in a stylized setting with a one-layer model explains when Hebbian guidance carries information about gradient magnitudes and satisfies the conditions.

Our main contributions are as follows:

- We propose HiZO, which uses model- and task-dependent co-activation statistics to guide ZO exploration under a fixed perturbation budget. Subsampled and separable variants reduce the computation and persistent storage of these statistics.
- We establish a nonconvex convergence analysis linking alignment to estimator variability, together with sufficient conditions for the Hebbian information to have the useful alignment property in a one-layer model.
- Our evaluations show the benefits of Hebbian guidance through faster optimization and improved performance over isotropic MeZO across multiple fine-tuning settings. Controlled ablations examine score assignment, covariance scaling, and the contributions of weights and input statistics.

2 METHODS

2.1 HIZO

Hebbian-informed Zeroth-order Optimization (HiZO) allocates more perturbation variance to weights whose connected neurons exhibit stronger co-variation on downstream inputs. These forward-pass statistics guide exploration, while loss evaluations determine the update along each sampled direction.

Hebbian information from forward activations. For a layer with weight matrix $W \in \mathbb{R}^{m\times n}$, let $X \in \mathbb{R}^{N\times n}$ and $Y \in \mathbb{R}^{N\times m}$ denote its input and output activations, with rows corresponding to examples or token positions. Motivated by covariance-based Hebbian learning (Sejnowski, 1977; Gerstner et al., 2014), we use the magnitude of centered output–input covariance as an importance score for each connection:

$$\widetilde{H}_{ij} = \left| \frac{1}{N} \sum_{k=1}^{N} (Y_{ki} - \bar{Y}_i)(X_{kj} - \bar{X}_j) \right|, \tag{1}$$

where $\bar{X}_j$ and $\bar{Y}_i$ are the corresponding sample means. The magnitude captures strong joint variation of either sign. For a linear layer before any nonlinearity, the underlying signed covariance equals $W\Sigma_x$, where $\Sigma_x = N^{-1}X_c^\top X_c$ is the empirical input covariance.

From co-variation to exploration. For layer ℓ with m_ℓ outputs, n_ℓ inputs, and $d_\ell = m_\ell n_\ell$ trainable weights, we flatten its scores into a vector $\widetilde{H}^{(\ell)} \in \mathbb{R}^{d_\ell}_{\geq 0}$. For a nonzero score vector, we normalize it within the layer:

$$H^{(\ell)} = \frac{d_\ell \widetilde{H}^{(\ell)}}{\|\widetilde{H}^{(\ell)}\|_1}, \qquad \sum_{e=1}^{d_\ell} H_e^{(\ell)} = d_\ell. \tag{2}$$

For an all-zero score vector, we set $H^{(\ell)} = \mathbf{1}$. We may instead use a mean-preserving affine rescaling to bound the variance weights in $[c, c^{-1}]$, with $c \in (0, 1]$, while preserving the same variance budget (Appendix C).

Let $\theta \in \mathbb{R}^d$ collect all trainable parameters and let $H \in \mathbb{R}^d_{\geq 0}$ collect their variance weights. Trainable parameters without a Hebbian score, such as biases, retain unit variance. Setting $H = \mathbf{1}$ recovers the isotropic perturbation distribution. At step t, HiZO uses the cached variance weights H_t to sample

$$z_t = \sqrt{H_t} \odot \xi_t, \qquad \xi_t \sim \mathcal{N}(0, I_d). \tag{3}$$

Here, I_d is the $d \times d$ identity matrix, the square root is taken elementwise, and $\odot$ denotes elementwise multiplication. The total expected squared norm remains $\mathbb{E}[\|z_t\|_2^2 \mid H_t] = d$.

We use the standard two-sided finite-difference update (Nesterov & Spokoiny, 2017; Malladi et al., 2023):

$$s_t = \frac{F_{B_t}(\theta_t + \epsilon z_t) - F_{B_t}(\theta_t - \epsilon z_t)}{2\epsilon}, \qquad \theta_{t+1} = \theta_t - \eta_t s_t z_t, \tag{4}$$

where F_{B_t} is the minibatch loss, $\epsilon > 0$ is the perturbation scale, and η_t is the learning rate. Both loss evaluations use the same minibatch, with dropout disabled.

We initialize $H_0 = \mathbf{1}$. During the positive evaluation at $\theta_t + \epsilon z_t$, we collect activations and construct H_{t+1} for the next iteration. The current variance weights H_t remain fixed during direction sampling, both loss evaluations, and the parameter update. When parameter changes and perturbations are small and the normalized statistics vary smoothly, the cache can approximate the current activation structure. Each iteration requires two forward loss evaluations plus the computation and storage of Hebbian statistics.

2.2 HIZO VARIANTS

Subsampling for HiZO. HiZO computes its covariance scores using all activation rows. Subsampled HiZO (S-HiZO) reduces this cost by estimating the same statistic from fewer paired input–output rows. Given a row budget $2 \leq q < N$, we sample q rows uniformly with replacement, preserving each input–output pair, and center the activations within the sample.

Subsampling reduces the cost of the covariance matrix product from $O(Nmn)$ to $O(qmn)$. S-HiZO retains a dense score matrix and uses the same normalization and update as HiZO. Taking absolute values and normalizing the sampled covariances yields an approximation to the variance weights computed from all rows.

Separable approximation for HiZO. Separable-approximation HiZO (SA-HiZO) represents each layer's scores with two vectors: one summarizes connection strengths for output neurons, and the other for input neurons. Each weight's score is the product of the corresponding factors. For an $m \times n$ weight matrix, this reduces persistent statistics storage from $O(mn)$ to $O(m+n)$.

We estimate the covariance-magnitude matrix $\widetilde{H} \in \mathbb{R}_{\geq 0}^{m \times n}$ from sampled activation rows, as in S-HiZO. We compute its row and column Euclidean norms and their respective means:

$$r_i = \|\widetilde{H}_{i,:}\|_2, \quad \bar{r} = \frac{1}{m}\sum_{i=1}^{m} r_i, \quad c_j = \|\widetilde{H}_{:,j}\|_2, \quad \bar{c} = \frac{1}{n}\sum_{j=1}^{n} c_j. \tag{5}$$

For $\widetilde{H} \neq 0$, the reconstructed variance weight is $H_{ij}^{\mathrm{SA}} = (r_i/\bar{r})(c_j/\bar{c})$. These weights already have mean one. For an all-zero matrix, we use unit factors. The two normalized vectors define the variance weights for the same two-sided ZO update. In the language-model experiments, we apply the bounded affine rescaling in Appendix C to the reconstructed outer-product scores, giving $H_{ij}^{\mathrm{SA,aff}} = 1 + s(H_{ij}^{\mathrm{SA}} - 1)$, where s is the slope in Eq. (19); constant maps retain unit weights. Caching the two factors and the rescaling scalars preserves $O(m+n)$ persistent storage.

The separable approximation retains aggregate connection strengths while discarding pair-specific co-variation. It compresses the variance weights rather than restricting the sampled direction to a low-rank subspace. Each weight receives independent Gaussian noise scaled by the square root of its reconstructed variance weight. Computing the factors still requires covariance estimation, so the reduction in persistent storage does not imply an equivalent reduction in computation.

Algorithm 1 in the appendix provides the complete procedure for HiZO and its variants. As in MeZO, the sampled direction can be regenerated from its random seed, avoiding storage of a full direction vector (Malladi et al., 2023).

3 Experiments

We evaluate the predictive performance, optimization speed, and computational cost of HiZO and its efficient variants on three language models, then examine Hebbian guidance and its constituent factors through two multilayer perceptron (MLP) ablations.

3.1 Experimental Setup

Models, tasks, and baselines. We evaluate RoBERTa-large (Liu et al., 2019) on SST-2 sentiment classification (Socher et al., 2013) and SNLI natural language inference (Bowman et al., 2015). For the decoder-only models OPT-1.3B (Zhang et al., 2022) and Qwen3-1.7B (Yang et al., 2025), we use SST-2, COPA commonsense causal reasoning (Roemmele et al., 2011), and SQuAD question answering (Rajpurkar et al., 2016). We report accuracy for SST-2, SNLI, and COPA, and token-level F1 for SQuAD, all on a 0–100 scale. We compare HiZO, S-HiZO, and SA-HiZO with MeZO (Malladi et al., 2023) and AGZO (Lin et al., 2026) across all three models. The RoBERTa-large comparison also includes zero-shot prediction and first-order fine-tuning (FT) as references.

Training and evaluation. We follow MeZO's full-parameter fine-tuning setup, sharing data splits, input formatting, trainable parameters, and batch sizes across ZO methods within each model–task pair. RoBERTa-large is trained to convergence, while decoder models use a fixed budget of 10,000 optimizer steps. We report mean accuracy and sample standard deviation over three runs with different seeds for RoBERTa-large. Decoder-model scores are evaluated at checkpoints selected by validation performance within the 10,000-step budget. Appendix D gives the detailed protocol.

Table 1: RoBERTa-large accuracy (%). ZO results are mean ± sample standard deviation over three seeds. Zero-shot and FT entries are reference scores. Bold marks the highest mean among ZO methods.

Method	SST-2	SNLI
Zero-shot	79.0	50.2
FT	91.9	77.5
MeZO	91.7 ± 5.2	70.8 ± 4.2
AGZO	92.7 ± 2.0	73.6 ± 1.4
HiZO	$\mathbf{93.8} \pm 3.2$	$\mathbf{75.7} \pm 1.3$
S-HiZO	92.7 ± 4.1	72.9 ± 4.1
SA-HiZO	93.0 ± 3.9	70.8 ± 2.1

Table 2: Mean paired-run steps-to-target ratios S and mean Hebbian alignment $\overline{A}$ on RoBERTa-large. Target crossings are measured at 500-step evaluation intervals. Alignment averages exclude the first 1,000 optimizer steps. MeZO is the reference for both quantities.

	SST-2		SNLI	
Method	$S\,(\times)$	$\overline{A}$	$S\,(\times)$	$\overline{A}$
MeZO	1.00	1.00	1.00	1.00
HiZO	3.28	3.47	3.42	3.58
S-HiZO	3.40	3.38	3.46	3.43
SA-HiZO	3.28	3.45	3.55	3.53

3.2 Main Results

Accuracy after convergence. HiZO achieves the highest mean ZO accuracy on both RoBERTa-large tasks (Table 1), improving over MeZO by 2.1 percentage points on SST-2 and 4.9 points on SNLI. AGZO also improves on MeZO, but HiZO exceeds AGZO by 1.1 and 2.1 points, respectively. Relative to FT, HiZO has higher mean accuracy on SST-2 and narrows the gap on SNLI.

Performance within 10,000 steps. HiZO's largest decoder-model gains occur on OPT-1.3B, where it improves SST-2 accuracy by 9.4 percentage points and SQuAD F1 by 7.9 points over MeZO, while matching MeZO on COPA (Figure 2). HiZO exceeds AGZO on OPT-1.3B SST-2 and COPA, but trails it by 0.5 F1 points on SQuAD. On Qwen3-1.7B, HiZO exceeds both baselines on all three tasks. Table 5 provides the full results.

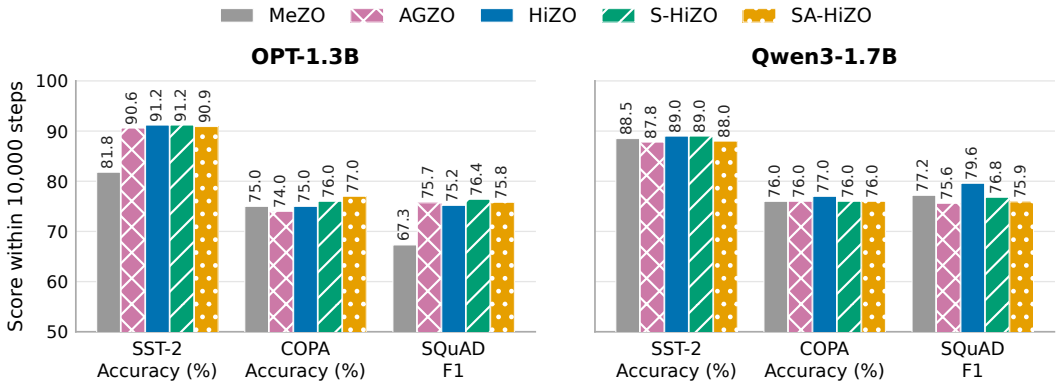


Figure 2: Task scores at validation-selected checkpoints within 10,000 optimizer steps for OPT-1.3B and Qwen3-1.7B. SST-2 and COPA use accuracy (%); SQuAD uses token-level F1. Bars show point estimates; exact values appear in Table 5.

Effect of the efficient variants. S-HiZO retains most of HiZO's predictive gains and achieves the highest OPT-1.3B SQuAD score. SA-HiZO also improves on MeZO on several tasks, including the best OPT-1.3B COPA score. Both approximations underperform MeZO on Qwen3-1.7B SQuAD, indicating that the effect of compressing the guidance depends on the model and task.

Steps to a common validation-loss target. For MeZO and the HiZO variants, we measure empirical convergence speed by the number of optimizer steps needed to reach a shared validation-loss target. For each task, we manually select a target validation loss based on empirical results of MeZO

(Appendix D.2). We define the step speedup as

$$S_{M,r} = \frac{T^{\text{target}}_{\text{MeZO},r}}{T^{\text{target}}_{M,r}}, \tag{6}$$

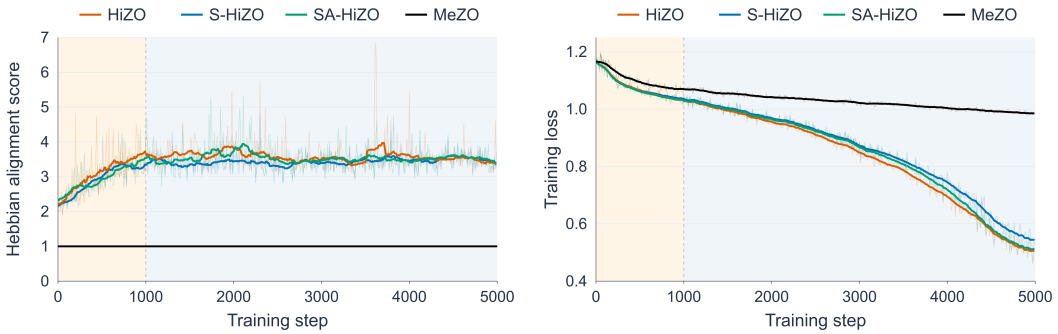
where $T^{\text{target}}_{M,r}$ is the first evaluated step at which method M reaches or falls below the target in run r (Appendix D.2). On RoBERTa-large, the HiZO variants reach these targets in roughly one-third of MeZO's steps (Table 2). Section 3.4 accounts for their per-step cost.

Hebbian alignment. To connect the optimization results with the covariance geometry, we measure

$$A_t = \frac{g_t^\top C_t g_t}{\|g_t\|_2^2}, \qquad \|g_t\|_2 > 0, \tag{7}$$

where g_t is a diagnostic gradient and C_t is the covariance used to sample the next perturbation. With total perturbation variance fixed by $\text{tr}(C_t) = d$, isotropic MeZO has $A_t = 1$. An alignment above one means that more perturbation variance is allocated along the gradient direction. Diagnostic gradients are used for analysis and do not enter the training updates. This diagnostic corresponds to the covariance–gradient alignment in Section 4.1. All HiZO variants have mean alignment above the isotropic reference, alongside faster progress toward the shared validation-loss targets.

Alignment and training dynamics on SNLI. Over the first 5,000 updates of RoBERTa-large on SNLI, the HiZO variants exhibit comparable alignment, consistently above the isotropic reference, and reduce training loss more rapidly than MeZO (Figure 3).



(a) Hebbian alignment.

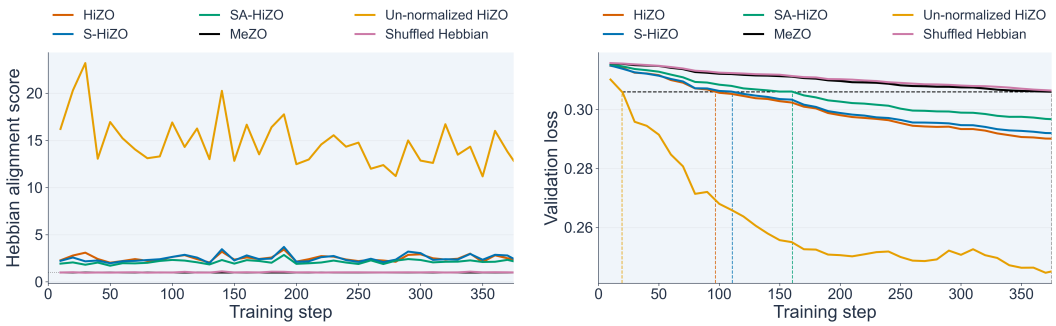
(b) Training loss.

Figure 3: RoBERTa-large on SNLI over the first 5,000 optimizer steps. Thick curves show centered 21-point moving averages; faint curves show raw values. MeZO has isotropic alignment $A = 1$. The dashed divider marks the end of the 1,000-step warm-up phase.

3.3 Ablation Studies

We conduct two ablation studies using an MLP on ten EMNIST classes (Cohen et al., 2017) to examine Hebbian guidance and its constituent factors. Each run consists of one pass over the selected training set (375 optimizer steps). Appendix D.4 provides the detailed setup and results.

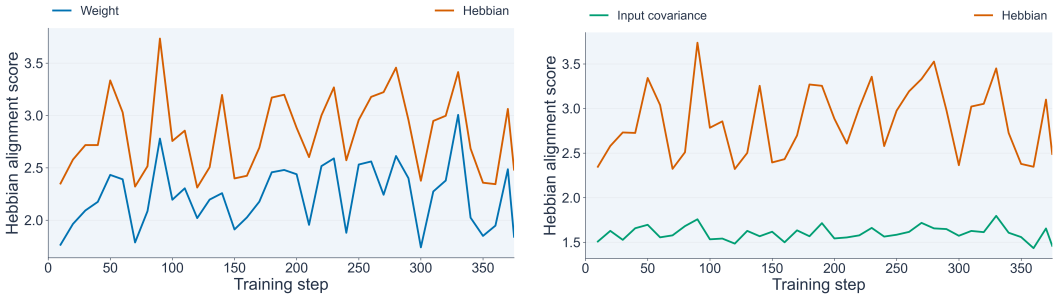
Hebbian structure and the variance budget. Figure 4 compares HiZO and its variants with isotropic MeZO, shuffled Hebbian scores, and unnormalized HiZO. At matched total perturbation variance, HiZO and S-HiZO show similar alignment and faster loss reduction than MeZO, while SA-HiZO retains most of this gain. Shuffling the scores within each layer preserves their distribution but reduces alignment to the isotropic level and produces a loss trajectory close to MeZO. The benefit therefore depends on assigning the scores to the corresponding weight coordinates. Unnormalized HiZO reduces loss faster, but its larger covariance scale also increases the effective learning rate. Matching the total variance isolates the benefit of Hebbian structure from this scale effect (Appendix D.4.1).



(a) Hebbian alignment. (b) Validation loss.

Figure 4: Hebbian-guidance ablation over one epoch (375 steps), using every training example from the selected classes once. All methods except unnormalized HiZO share the same total perturbation variance. Dashed guides mark MeZO's final validation loss and interpolated first crossings.

Weights and input covariance contribute jointly. For a linear layer, the signed output–input covariance factorizes as $B = W\Sigma_x$, where W is the weight matrix and Σ_x is the input covariance. Both weight-only and input-covariance-only guidance yield alignment above one, but full Hebbian guidance has higher alignment at the same checkpoints (Figure 5). Combining learned weights with input statistics therefore captures more gradient-relevant structure than either factor alone.



(a) Weight-only guidance. (b) Input-covariance-only guidance.

Figure 5: Component ablation with normalized guidance. Each panel compares a component's sampling covariance with full Hebbian guidance evaluated along that component's training trajectory. All covariances have matched total variance.

3.4 Computational and Memory Efficiency

Table 3 reports the measured per-step time and peak GPU memory on RoBERTa-large with SST-2. Combining these costs with the steps-to-target ratios gives estimated training-time speedups of $2.19\times$, $2.85\times$, and $2.55\times$ for HiZO, S-HiZO, and SA-HiZO, respectively. S-HiZO provides the largest estimated time saving, while SA-HiZO keeps peak GPU memory close to MeZO's. Appendix E details the measurement protocol, estimation assumptions, and complexity analysis.

Table 3: Computational and memory cost on RoBERTa-large with SST-2 under matched hardware and execution settings. Step time includes statistics collection for the HiZO variants. Peak memory denotes allocated GPU memory.

Metric	MeZO	HiZO	S-HiZO	SA-HiZO
Time per step (s)	0.226	0.339	0.270	0.291
Peak GPU memory (MB)	3075	4779	4463	3079

4 THEORETICAL ANALYSIS

4.1 CONVERGENCE UNDER HEBBIAN COVARIANCE ALIGNMENT

HiZO redistributes a fixed perturbation-variance budget toward connections with stronger co-activation. We first quantify the benefit when this allocation emphasizes coordinates with large task gradients, then give a mechanism by which Hebbian covariance can provide this information. We analyze a fixed, lower-bounded, L-smooth objective $F : \mathbb{R}^d \to \mathbb{R}$, which may be nonconvex. Unsubscripted vector norms are Euclidean. Let $g_t = \nabla F(\theta_t)$ and let $\mathcal{F}_t$ contain the optimization history and cached Hebbian statistics before the current direction is sampled. Write $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_t]$ and $C_t = \mathbb{E}_t[z_t z_t^\top]$, with $\mathbb{E}_t[z_t] = 0$ and $\mathrm{tr}(C_t) = d$, where tr denotes the matrix trace.

Assumption 1 (Hebbian covariance alignment (HCA)). *There is a $\gamma > 0$ such that, almost surely at every analyzed iteration,*

$$g_t^\top C_t g_t \geq (1+\gamma)\|g_t\|^2. \tag{8}$$

For the cached Gaussian sampler in Eq. (3), $C_t = \mathrm{diag}(H_t)$, where diag forms a diagonal matrix from a vector. Layerwise normalization in Eq. (2), together with unit variance for unscored parameters, gives $\sum_i H_{t,i} = d$. Consequently,

$$g_t^\top C_t g_t - \|g_t\|^2 = d\ \mathrm{Cov}_{i\sim\mathrm{Unif}([d])}(H_{t,i}, g_{t,i}^2).$$

Here $[d] = \{1, \ldots, d\}$ and $\mathrm{Unif}([d])$ is the uniform distribution over coordinates. Thus HCA means that the allocated variance is positively associated with squared gradient magnitudes under the same total variance budget. The analyzed segment starts after isotropic initialization or warm-up, where the alignment ratio is one. Define $\Delta = \mathbb{E}[F(\theta_0)] - F_{\inf}$ for the start of this segment, where $F \geq F_{\inf}$.

Theorem 1 (Alignment gain in the fixed-norm idealization). *Suppose HCA holds and $\|z_t\|^2 = d$ almost surely. In the directional-derivative limit, use $\widehat{g}_t = (g_t^\top z_t) z_t$ and $\theta_{t+1} = \theta_t - \widehat{g}_t/(Ld)$. Then*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla F(\theta_t)\|^2 \leq \frac{2Ld\Delta}{(1+\gamma)T}. \tag{9}$$

Why the factor is $1+\gamma$. Let $D_t = g_t^\top C_t g_t$ denote the numerator of the alignment ratio in Eq. (7). The two estimator moments are $\mathbb{E}_t[\widehat{g}_t] = C_t g_t$ and $\mathbb{E}_t\|\widehat{g}_t\|^2 = dD_t$. Smoothness therefore gives

$$\mathbb{E}_t[F(\theta_{t+1})] \leq F(\theta_t) - \frac{D_t}{2Ld} \leq F(\theta_t) - \frac{1+\gamma}{2Ld}\|g_t\|^2. \tag{10}$$

Telescoping proves the theorem. Isotropic fixed-norm sampling has $C_t = I_d$ and bound $2Ld\Delta/T$, so alignment reduces the stationarity-bound constant, and the corresponding sufficient iteration budget, by a factor $1+\gamma$. The order remains $O(T^{-1})$ for the expected squared gradient norm. Appendix F gives the proof and finite-difference correction.

Connection to Gaussian HiZO. A fixed-norm surrogate that preserves HiZO's covariance exactly is $z_t = \sqrt{H_t} \odot u_t$, where the coordinates of u_t are independent uniform signs. HiZO instead uses Gaussian coordinates. For its directional-derivative estimator,

$$\mathbb{E}_t\|\widehat{g}_t\|^2 = dD_t + 2g_t^\top C_t^2 g_t. \tag{11}$$

If $g_t^\top C_t^2 g_t \leq K D_t$ throughout the analyzed segment, the step size $\eta = 1/[L(d+2K)]$ yields

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|g_t\|^2 \leq \frac{2L(d+2K)\Delta}{(1+\gamma)T}. \tag{12}$$

The ratio of the isotropic Gaussian bound constant to HiZO's is $(1+\gamma)(d+2)/(d+2K)$, which gives an improvement when it exceeds one. For the bounded variance weights in Appendix C, $K = c^{-1}$ is valid. With fixed c, the correction factor $(d+2)/(d+2K)$ approaches one as d grows. This connects the method's variance control to its estimator second moment: alignment increases the descent term, while bounded weights control the Gaussian correction. These bounds use exact evaluations of F in the directional-derivative limit; Appendix F.6 incorporates minibatch noise.

4.2 Why Hebbian covariance can carry task-gradient information

We give a sufficient mechanism for HCA in a one-layer model. Let $x \in \mathbb{R}^n$ be centered, $y = Wx + b \in \mathbb{R}^m$, and $\hat{y} = e_c$ be a one-hot class label, where c is the class index and e_c is the cth standard basis vector. We optimize W with fixed b under $F(W) = \frac{1}{2}\mathbb{E}\|y - \hat{y}\|^2$. Thus $y - \hat{y}$ is the output residual, and the sample gradient is $(y - \hat{y})x^\top$.

Assume the input covariance $\Sigma = \mathbb{E}[xx^\top]$ is positive definite, and define the signed Hebbian covariance $B = \mathrm{Cov}(y, x)$ and the target covariance $Q = \mathrm{Cov}(\hat{y}, x)$. Then

$$B = W\Sigma, \qquad G := \nabla_W F = B - Q, \qquad Q_{k,:} = \Pr(c = k)\,\mathbb{E}[x \mid c = k]^\top. \tag{13}$$

This is the centered form of the decomposition into output and target Hebbian products discussed by Melchior & Wiskott (2019). The identity $B = W\Sigma$ is the population counterpart of the activation statistic in Eq. (1): it combines the learned weights with the task-dependent input covariance. The target covariance Q describes the task map that the model is learning.

Proposition 1 (A sufficient mechanism for HCA)**.** *Let $W_\star = Q\Sigma^{-1} \neq 0$ be the least-squares solution and suppose $W = aW_\star$ for a scalar $a > 0$, $a \neq 1$. Set $d = mn$, $h = |\,\mathrm{vec}(B)|$, and use HiZO's variance allocation $C = \mathrm{diag}(dh/\|h\|_1)$. Here* vec *stacks matrix entries into a vector, and $|\cdot|$ acts elementwise. For $g = \mathrm{vec}(G)$,*

$$G = \frac{a-1}{a}B, \qquad \frac{g^\top Cg}{\|g\|^2} = \Gamma(h) := \frac{d\sum_{i=1}^d h_i^3}{(\sum_{i=1}^d h_i)(\sum_{i=1}^d h_i^2)} \geq 1. \tag{14}$$

The inequality is strict whenever the entries of h are not all equal. In this nonuniform case, HCA holds with $\gamma = \Gamma(h) - 1 > 0$.

Interpretation and derivation. The condition $W = aW_\star$ describes a model that has learned the task map up to an overall scale. Since $B = aQ$ and $G = (a-1)Q$, their magnitudes share the same coordinate structure even when their signs are opposite. Substituting this relation into $g^\top Cg$ gives (14). The inequality follows because h_i and h_i^2 are positively associated over coordinates. With equal nonzero scores on exactly s of d coordinates, $\Gamma = d/s$: guidance concentrates exploration on the coordinates that also carry the residual gradient.

This regime has a constructive realization. For $\Sigma = \sigma^2 I_n$, ordinary gradient flow from $W(0) = 0$ gives $W(t) = (1 - e^{-\sigma^2 t})W_\star$ for every $t > 0$. This first-order trajectory illustrates how learning a task map can produce aligned model states, following the correlation-based view of linear-network dynamics in Saxe et al. (2014). The convergence result applies to HiZO when its own trajectory maintains the stated alignment condition.

Appendix G proves the result, quantifies its stability when G is approximately proportional to B, and gives one-hot examples with $\Gamma = 2$ for squared loss and softmax cross-entropy. It also shows how this margin transfers to cached statistics, bounded affine rescaling, and layerwise variance allocation.

5 Conclusions and Discussion

HiZO uses task-dependent centered input–output covariance, collected during existing forward evaluations, to guide zeroth-order exploration under a fixed perturbation-variance budget. Its subsampled and separable variants reduce computation and persistent storage, respectively. Our analysis identifies when Hebbian statistics capture task-gradient magnitudes and improve convergence bounds through covariance–gradient alignment. On RoBERTa-large, the HiZO variants reach shared validation-loss targets in roughly one-third of MeZO's optimizer steps. Full HiZO maintains or improves predictive performance over MeZO across the evaluated settings, with higher mean accuracy than AGZO on both RoBERTa-large tasks. MLP ablations support the importance of matching scores to connections and the joint contribution of weights and input statistics to alignment.

Hebbian guidance could complement other ZO methods by combining task-dependent activation information with sparse exploration, structured perturbations, or curvature estimates. A natural next step is to study how these sources of information interact and when their combination improves convergence enough to offset the additional computation and memory required for guidance.

AI USE STATEMENT

In this work, we used generative AI tools to assist with writing refinement, proof development, section drafting, and numerical verification. All AI-assisted content was carefully reviewed by the authors. We take full responsibility for the final content of this work, including all text, claims, analyses, and other artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

Section 2 and Algorithm 1 describe HiZO and its efficient variants, including the construction of Hebbian guidance and the perturbation and update procedures. Appendix D documents the datasets, task formulations, data splits, hyperparameter search grids, training budgets, checkpoint selection, and ablation settings. It also specifies the alignment diagnostics and validation-loss targets used to measure optimization progress. Appendix E reports the computational and memory measurements and the assumptions used to estimate training-time speedups. The assumptions underlying our theoretical results are stated in Section 4, with detailed derivations and proofs in Appendices F and G. We will publicly release the code for reproducing our experiments.

REFERENCES

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642. Association for Computational Linguistics, 2015. doi: 10.18653/v1/D15-1075. URL https://aclanthology.org/D15-1075/.

Yiming Chen, Yuan Zhang, Liyuan Cao, Kun Yuan, and Zaiwen Wen. Enhancing zeroth-order fine-tuning for language models with low-rank structures. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=9BiVepgmWW.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. EMNIST: an extension of MNIST to handwritten letters, 2017. URL https://arxiv.org/abs/1702.05373.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Xun Dong, Yibo Xu, Naigang Wang, Xin Li, Penghang Yin, and Zi Yang. ZO-Act: Efficient zeroth-order fine-tuning via one-shot activation-informed low-rank subspaces. *arXiv preprint arXiv:2607.01125*, 2026. URL https://arxiv.org/abs/2607.01125.

Wulfram Gerstner, Werner M. Kistler, Richard Naud, and Liam Paninski. *Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition*. Cambridge University Press, 2014. URL https://neuronaldynamics.epfl.ch/.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 2019. URL https://proceedings.mlr.press/v97/houlsby19a.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Adrien Journé, Hector Garcia Rodriguez, Qinghai Guo, and Timoleon Moraitis. Hebbian deep learning without feedback. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=8gd4M-_Rj1.

Dmitry Krotov and John J. Hopfield. Unsupervised learning by competing hidden units. *Proceedings of the National Academy of Sciences*, 116(16):7723–7731, 2019. doi: 10.1073/pnas.1820458116. URL https://arxiv.org/abs/1806.10181.

Wei Lin, Yining Jiang, Qingyu Song, Qiao Xiang, and Hong Xu. AGZO: Activation-guided zeroth-order optimization for LLM fine-tuning. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2601.17261.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. URL https://arxiv.org/abs/1907.11692.

Yong Liu, Zirui Zhu, Chaoyu Gong, Minhao Cheng, Cho-Jui Hsieh, and Yang You. Sparse MeZO: Less parameters for better performance in zeroth-order LLM fine-tuning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2402.15751.

Niru Maheswaranathan, Luke Metz, George Tucker, Dami Choi, and Jascha Sohl-Dickstein. Guided evolutionary strategies: Augmenting random search with surrogate gradients. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4264–4273. PMLR, 2019. URL https://proceedings.mlr.press/v97/maheswaranathan19a.html.

Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D. Lee, Danqi Chen, and Sanjeev Arora. Fine-tuning language models with just forward passes. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2305.17333.

Jan Melchior and Laurenz Wiskott. Hebbian-descent. *arXiv preprint arXiv:1905.10585*, 2019. URL https://arxiv.org/abs/1905.10585.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017. doi: 10.1007/s10208-015-9296-2. URL https://doi.org/10.1007/s10208-015-9296-2.

Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15:267–273, 1982. doi: 10.1007/BF00275687. URL https://doi.org/10.1007/BF00275687.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, 2016. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264/.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, 2011. URL https://asgordon.github.io/publications/AAAI-SPRING11A.PDF.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986. doi: 10.1038/323533a0. URL https://www.nature.com/articles/323533a0.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014. URL https://arxiv.org/abs/1312.6120.

Johannes Schmidt-Hieber and Wouter M. Koolen. Hebbian learning inspired estimation of the linear regression parameters from queries. *arXiv preprint arXiv:2311.03483*, 2023. URL https://arxiv.org/abs/2311.03483.

T. J. Sejnowski. Storing covariance with nonlinearly interacting neurons. *Journal of Mathematical Biology*, 4(4):303–321, 1977. doi: 10.1007/BF00275079. URL https://doi.org/10.1007/BF00275079.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642. Association for Computational Linguistics, 2013. URL https://aclanthology.org/D13-1170/.

James C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37(3):332–341, 1992. doi: 10.1109/9.119632. URL https://www.jhuapl.edu/SPSA/PDF-SPSA/Spall_TAC92.pdf.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022. URL https://arxiv.org/abs/2205.01068.

Yihua Zhang, Pingzhi Li, Junyuan Hong, Jiaxiang Li, Yimeng Zhang, Wenqing Zheng, Pin-Yu Chen, Jason D. Lee, Wotao Yin, Mingyi Hong, Zhangyang Wang, Sijia Liu, and Tianlong Chen. Revisiting zeroth-order optimization for memory-efficient LLM fine-tuning: A benchmark. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 59173–59190. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhang24ad.html.

Yanjun Zhao, Sizhe Dang, Haishan Ye, Guang Dai, Yi Qian, and Ivor W. Tsang. Second-order fine-tuning without pain for LLMs: A hessian informed zeroth-order optimizer. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=bEqI61iBue.

Appendix

A Related Work

A.1 Zeroth-Order Optimization and Language Model Fine-Tuning

Zeroth-order (ZO) optimization uses function evaluations to optimize an objective without access to its derivatives. Finite differences along random perturbations provide a classical approach to gradient estimation (Spall, 1992; Nesterov & Spokoiny, 2017). For an objective $F(\theta)$ with $\theta \in \mathbb{R}^d$, a two-sided Gaussian estimator takes the form

$$\widehat{g}(\theta; z) = \frac{F(\theta + \epsilon z) - F(\theta - \epsilon z)}{2\epsilon} z, \qquad z \sim \mathcal{N}(0, I_d), \tag{15}$$

where $\epsilon > 0$ controls the perturbation scale. MeZO adapts this estimator to language model fine-tuning through in-place parameter perturbations and seed-based regeneration of random directions, with memory use close to inference (Malladi et al., 2023). It supports both full-parameter and parameter-efficient fine-tuning, including LoRA (Hu et al., 2022; Malladi et al., 2023). A broader benchmark examines how optimizer design, model architecture, and fine-tuning scheme affect the trade-off between predictive performance and memory use (Zhang et al., 2024).

Subsequent methods refine the parameters explored or the perturbation distribution. Sparse MeZO restricts perturbations and updates to a selected subset of parameters (Liu et al., 2025). LOZO exploits the low-rank structure of fine-tuning gradients through low-rank perturbations and offers a memory-efficient momentum variant (Chen et al., 2025). HiZOO incorporates diagonal Hessian estimates to account for differences in curvature across parameters (Zhao et al., 2025). More broadly, Guided Evolutionary Strategies allocates additional search variance to a subspace spanned by surrogate gradients (Maheswaranathan et al., 2019).

Forward activations offer another source of guidance. AGZO constructs activation-informed low-rank search subspaces during optimization (Lin et al., 2026), while ZO-Act computes an activation basis once and optimizes coefficient matrices within the resulting fixed subspace (Dong et al., 2026). HiZO uses the magnitudes of centered input–output covariance to allocate perturbation variance across individual weights under a fixed trace budget. These connection-level statistics depend on the model weights and task inputs and are collected during existing forward evaluations. They shape a diagonal Gaussian search covariance rather than defining an activation subspace.

A.2 Hebbian Information and Neural Network Learning

Hebbian learning links changes in synaptic strength to the joint activity of connected neurons (Gerstner et al., 2014). For a layer with input activations x, output activations y, and weight matrix W, where W_{ij} connects input j to output i, the elementary Hebbian rule has the outer-product form

$$\Delta W = \eta y x^\top, \qquad \Delta W_{ij} = \eta y_i x_j, \tag{16}$$

where η is a learning rate. Centering the pre- and postsynaptic activities gives the covariance rule, whose expected update is proportional to

$$C_{yx} = \mathbb{E}\big[(y - \mathbb{E}[y])(x - \mathbb{E}[x])^\top\big]. \tag{17}$$

This statistic captures joint fluctuations around mean activity, removing the contribution of nonzero means (Sejnowski, 1977; Gerstner et al., 2014). Oja's rule adds a stabilizing normalization term and connects linear-neuron learning to principal component extraction (Oja, 1982).

Hebbian mechanisms also support representation learning in artificial neural networks. Competitive Hebbian and anti-Hebbian updates learn hidden features without labels, after which a supervised output layer is trained on the frozen representation (Krotov & Hopfield, 2019). Multilayer SoftHebb extends this approach to deep networks through soft winner-take-all interactions, learning hidden representations without backward error or target signals and evaluating them with a supervised linear classifier (Journé et al., 2023).

Hebbian learning has also been connected to loss-based optimization. For mean-squared error, Hebbian-descent expresses its update as a difference between target-dependent and output-dependent

Hebbian terms (Melchior & Wiskott, 2019). In a loss-query setting, Hebbian-inspired rules estimate linear regression parameters with guarantees on estimation error (Schmidt-Hieber & Koolen, 2023). In HiZO, Hebbian information guides exploration rather than directly specifying weight updates: co-activation statistics allocate perturbation variance, while two-sided loss evaluations determine the update along each sampled direction.

B Algorithm for HiZO and Its Variants

Algorithm 1 HiZO and Its Variants

Input: Initial parameters θ_0; minibatch loss F_B; steps T; learning rates $\{\eta_t\}$; perturbation scale ϵ; variant v; sampling budget $q \geq 2$.
Output: Fine-tuned parameters θ_T.

$H_0 \leftarrow \mathbf{1}$
for $t = 0, \ldots, T-1$ **do**
  Sample minibatch B_t and $\xi_t \sim \mathcal{N}(0, I_d)$
  $z_t \leftarrow \sqrt{H_t} \odot \xi_t$ {Hebbian-guided perturbation}
  $\ell_+ \leftarrow F_{B_t}(\theta_t + \epsilon z_t)$ {Collect paired activations $\{X_\ell, Y_\ell\}$}
  $H_{t+1} \leftarrow \textsc{HebbianWeights}(\{X_\ell, Y_\ell\}, v, q)$
  $\ell_- \leftarrow F_{B_t}(\theta_t - \epsilon z_t)$
  $s_t \leftarrow (\ell_+ - \ell_-)/(2\epsilon)$
  $\theta_{t+1} \leftarrow \theta_t - \eta_t s_t z_t$
end for
Subroutine $\textsc{HebbianWeights}(\{X_\ell, Y_\ell\}, v, q)$
$H \leftarrow \mathbf{1}$ {Default variance for zero or missing scores}
for each guided layer ℓ **do**
  if $v \neq$ HiZO and $\mathrm{rows}(X_\ell) > q$ **then**
    $(X_\ell, Y_\ell) \leftarrow q$ paired rows sampled uniformly with replacement
  end if
  $X_c \leftarrow X_\ell - \mathbf{1}\bar{X}_\ell^\top$, $Y_c \leftarrow Y_\ell - \mathbf{1}\bar{Y}_\ell^\top$ {Center retained rows}
  $A \leftarrow |Y_c^\top X_c| / \mathrm{rows}(X_\ell)$ {Hebbian covariance magnitude}
  if $A \neq 0$ **then**
    if $v =$ SA-HiZO **then**
      $r_i \leftarrow \|A_{i,:}\|_2$, $c_j \leftarrow \|A_{:,j}\|_2$
      $H_\ell \leftarrow \big(r/\,\mathrm{mean}(r), c/\,\mathrm{mean}(c)\big)$ {Store factors of the variance map}
    else
      $H_\ell \leftarrow A/\,\mathrm{mean}(A)$
    end if
  end if
end for
return H

C Variance-preserving rescaling of Hebbian scores

In our larger-model training runs, nonzero Hebbian scores can span up to ten orders of magnitude. Mean normalization preserves their nonzero ratios in exact arithmetic, so a few dominant scores can concentrate the perturbation variance. Finite-precision arithmetic can further affect near-zero scores and their normalization. We therefore bound the variance weights while preserving the total perturbation variance.

A fixed variance budget. For a layer with D trainable weights, let $A = \mathrm{vec}(|N^{-1}Y_c^\top X_c|) \in \mathbb{R}^D_{\geq 0}$ collect the covariance magnitudes in Eq. (1). Here X_c and Y_c contain N paired, centered activation rows, and $D = d_\ell$ in the notation of the main text. With $\mu = D^{-1}\sum_{i=1}^D A_i > 0$, the usual normalization is $H_i = A_i/\mu$. For an all-zero map, we set $H_i = 1$. Sampling $z = \sqrt{H} \odot \xi$, with

$\xi \sim \mathcal{N}(0, I_D)$, gives

$$\mathbb{E}\big[\|\epsilon z\|_2^2 \mid H\big] = \epsilon^2 \sum_{i=1}^{D} H_i = \epsilon^2 D. \tag{18}$$

Thus, mean-one variance weights redistribute the same expected perturbation energy as MeZO. Individual variance weights can nevertheless approach D when one score dominates the map.

Mean-preserving affine rescaling. We use a single parameter $c \in (0, 1]$ to specify the interval $[c, c^{-1}]$. For a finite, nonconstant score map, define $a = \min_i A_i$ and $b = \max_i A_i$, so that $a < \mu < b$, and set

$$s = \min\left\{ \frac{1-c}{\mu - a}, \frac{c^{-1}-1}{b-\mu} \right\}, \qquad H_i = 1 + s(A_i - \mu). \tag{19}$$

For a constant map, including an all-zero map, we set $H_i = 1$. The slope s is the largest nonnegative slope compatible with both bounds. Indeed,

$$\begin{aligned} \frac{1}{D}\sum_{i=1}^{D} H_i &= 1, \\ \min_i H_i &= 1 - s(\mu - a) \geq c, \\ \max_i H_i &= 1 + s(b - \mu) \leq c^{-1}. \end{aligned} \tag{20}$$

For a nonconstant map and $c < 1$, the transformation preserves score ordering and the variance budget in Eq. (18), reaching at least one interval endpoint. Setting $c = 1$ gives $H_i = 1$, recovering isotropic sampling. Mapping both extrema to prescribed endpoints would not generally preserve the mean. The positive lower bound retains exploration even at coordinates with zero Hebbian score.

The resulting variance ratio is bounded by c^{-2}. For example, $c = 0.01$ gives $H_i \in [0.01, 100]$ and a maximum variance ratio of 10^4. The corresponding coordinatewise perturbation standard deviations lie in $[0.1\epsilon, 10\epsilon]$. These constraints regulate the sampling distribution rather than equating HiZO's parameter-update norm with MeZO's.

In our experiments, we use $c = 0.001$, which bounds the variance weights in $[10^{-3}, 10^3]$ to limit extreme values in finite-precision arithmetic. The rescaling preserves the total variance budget, but its effect on convergence depends on the resulting covariance–gradient alignment and estimator variability.

D Experimental Setup and Additional Results

D.1 Language-Model Fine-Tuning

Data and task formulation. We adopt MeZO's data preparation and prompting conventions (Malladi et al., 2023). RoBERTa-large uses the $k = 16$ setting: training and validation each contain 16 examples per class, giving 32 examples per split for SST-2 and 48 for SNLI. Final evaluation uses up to 1,000 examples per task. Decoder-model experiments use up to 1,000 training, 500 validation, and 1,000 evaluation examples. For the smaller COPA dataset, the validation subset contains 100 examples and the training set uses the remaining available training examples. The splits are shared by all methods within a run.

RoBERTa-large predicts verbalized class labels at a masked position. The decoder models score candidate continuations for SST-2 and COPA, and generate answers for SQuAD. We retain MeZO's task-specific templates, target-token loss construction, and decoding procedure. Qwen3-1.7B uses the same task formulation as OPT-1.3B. SST-2, SNLI, and COPA use accuracy, while SQuAD uses token-level F1.

Optimization settings. Table 4 lists the full-parameter ZO settings, following MeZO's model-specific configurations. All ZO methods use the same learning-rate search grid within each model–task pair, with rates selected separately for each method using validation performance. MeZO and the HiZO variants also share the batch size, perturbation scale, and constant learning-rate schedule.

Table 4: Optimization settings for MeZO, HiZO, S-HiZO, and SA-HiZO; AGZO uses the same learning-rate search grid. Decoder settings apply to OPT-1.3B and Qwen3-1.7B. Training duration follows the model-specific protocol used in this work.

Hyperparameter	RoBERTa-large	Decoder models
Trainable parameters	Full model	Full model
Batch size	64	16
Learning-rate grid	$\{10^{-7}, 10^{-6}, 10^{-5}\}$	$\{10^{-7}, 5\times10^{-7}, 10^{-6}\}$
Learning-rate schedule	Constant	Constant
Perturbation scale ϵ	10^{-3}	10^{-3}
Weight decay	0	0
Training duration	To convergence	10K steps
Maximum optimizer steps	100K	10K

RoBERTa-large is trained to convergence within a maximum of 100,000 steps; ZO results are summarized over three runs with different seeds. For OPT-1.3B and Qwen3-1.7B, we use a shorter, fixed budget of 10,000 steps, select checkpoints by validation performance, and report their evaluation scores. The RoBERTa-large FT entries use the corresponding MeZO FT setup. AGZO (Lin et al., 2026) is included in the predictive-performance comparisons on all three language models. The steps-to-target and training-cost comparisons cover MeZO and the HiZO variants.

Hebbian-statistics configuration. HiZO collects centered input–output activation cross-covariances during the positive loss evaluation and uses the resulting guidance at the next optimizer step. Statistics are refreshed at every step. S-HiZO and SA-HiZO sample up to 32 activation rows per guided layer. S-HiZO stores a dense score matrix, whereas SA-HiZO retains row and column summaries. We normalize the sampling covariance to have mean diagonal entry one, using the bounded rescaling in Appendix C with $c = 0.001$ for language models, and disable guidance annealing. For SA-HiZO, this rescaling is applied to the reconstructed outer-product scores. The cached covariance is held fixed for both loss evaluations and the parameter update within each step, so they use the same perturbation direction. Statistics collection reuses an existing forward pass.

D.2 Checkpoint Selection and Alignment

Steps to a common validation-loss target. For each task q, we manually select a single target validation loss $\ell_q^\star$ based on empirical MeZO training runs. For the RoBERTa-large convergence comparison, we use $\ell^\star_{\text{SNLI}} = 0.75$ and $\ell^\star_{\text{SST-2}} = 0.45$. This target remains fixed across methods and runs for that task. Let $\ell_{M,r,q}(t)$ be the validation loss of method M at step t in run r, and let $\mathcal{E}_{M,r,q}$ be its evaluated steps. The first target-reaching checkpoint is

$$T^{\text{target}}_{M,r,q} = \min\{t \in \mathcal{E}_{M,r,q} : \ell_{M,r,q}(t) \le \ell_q^\star\}\,. \tag{21}$$

Equation (6) compares these first-crossing steps for paired runs, with the task index omitted for readability. Ratios are defined when both methods reach the target at positive steps. Validation is evaluated every 500 optimizer steps, which determines the temporal resolution. A speedup of 3.5 means that a method takes approximately $1/3.5$ as many steps as MeZO to reach the same validation-loss threshold. Predictive task scores are reported separately from this convergence measure.

Alignment diagnostics. For the language-model analysis, the diagnostic gradient is evaluated on a fixed training subset and restricted to the ZO trainable coordinates. Each unperturbed checkpoint is paired with the cached covariance that determines its next perturbation. We use the actual covariance of each variant, including its subsampling or separable approximation, in Eq. (7). Gradients are computed offline and do not affect optimization. Alignment is averaged over the common diagnostic checkpoints with nonzero gradient norm from step 1,000 onward, excluding warm-up, and then across runs. Reported speedups are arithmetic means of the paired-run ratios $S_{M,r}$. Table 2 reports the resulting point summaries.

Table 5: Scores at validation-selected checkpoints within 10,000 optimizer steps. SST-2 and COPA entries are accuracy (%), and SQuAD entries are token-level F1. Bold marks the highest score within each model and task, including ties.

Method	SST-2	COPA	SQuAD
	OPT-1.3B		
MeZO	81.8	75.0	67.3
AGZO	90.6	74.0	75.7
HiZO	**91.2**	75.0	75.2
S-HiZO	**91.2**	76.0	**76.4**
SA-HiZO	90.9	**77.0**	75.8
	Qwen3-1.7B		
MeZO	88.5	76.0	77.2
AGZO	87.8	76.0	75.6
HiZO	**89.0**	**77.0**	**79.6**
S-HiZO	**89.0**	76.0	76.8
SA-HiZO	88.0	76.0	75.9

D.3 Decoder-Model Results

Table 5 lists the point estimates shown in Figure 2. Each entry is the evaluation score at a checkpoint selected by validation performance within the 10,000-step budget. On OPT-1.3B, HiZO exceeds AGZO on SST-2 and COPA, while AGZO has a 0.5-point advantage on SQuAD. S-HiZO achieves the highest SQuAD F1, and SA-HiZO achieves the highest COPA accuracy. On Qwen3-1.7B, HiZO exceeds both baselines on all three tasks and ties S-HiZO for the highest SST-2 accuracy.

D.4 MLP Ablation Details

Shared setup. Both ablations start from the same pretrained MLP, with three ReLU hidden layers of width 1,568 and a 47-class output head. We fine-tune on ten classes from EMNIST (Cohen et al., 2017), using the full training data for those classes exactly once. This single epoch comprises 375 optimizer steps. A hyperparameter grid search gives a batch size of 64, learning rate of 5×10^{-5}, and perturbation scale of 10^{-3}. Statistics are refreshed at every update and annealing is disabled. S-HiZO and SA-HiZO sample 32 activation rows per layer.

Alignment and validation loss are logged every 10 updates and at the last step. Alignment uses the current training minibatch gradient over linear weights and the cached sampling covariance for the update. These diagnostic gradients do not affect training. Validation loss is measured on the selected classes from the official EMNIST test split.

D.4.1 Hebbian Structure and Covariance Scale

Normalized and shuffled controls. For a zero-mean perturbation z with covariance C over d coordinates,

$$\mathbb{E}\|\epsilon z\|_2^2 = \epsilon^2 \operatorname{tr}(C). \tag{22}$$

We normalize each guided layer to $\operatorname{tr}(C) = d$, matching its total perturbation variance to MeZO. The shuffled control permutes the normalized scores within each layer at every refresh, preserving their distribution and trace. The unnormalized control retains the score scale, so it is excluded from the fixed-variance comparison.

Why covariance scale acts as learning-rate scaling. For the two-sided estimator

$$\widehat{g}_\epsilon(z) = \frac{F(\theta + \epsilon z) - F(\theta - \epsilon z)}{2\epsilon}\, z, \tag{23}$$

rescaling a perturbation by $\sqrt{\alpha}$ gives the exact identity

$$\widehat{g}_\epsilon(\sqrt{\alpha}\, z) = \alpha\, \widehat{g}_{\epsilon\sqrt{\alpha}}(z), \qquad \alpha > 0. \tag{24}$$

Table 6: Hebbian-guidance ablation. Alignment is averaged over 38 logged checkpoints. A dash indicates that the reference loss was not reached. Unnormalized HiZO uses a different total variance budget.

Method	Mean alignment	Final loss	First crossing	Speedup
MeZO	1.00	0.3060	375	1.00×
HiZO	2.51	0.2901	100	3.75×
S-HiZO	2.51	0.2920	120	3.13×
SA-HiZO	2.12	0.2967	170	2.21×
Shuffled Hebbian	1.01	0.3064	—	—
Unnormalized HiZO	14.80	0.2451	20	18.75×

Thus, multiplying a covariance by α multiplies the estimator scale by α and the finite-difference radius by $\sqrt{\alpha}$. In the local linear regime, with $g = \nabla F(\theta)$, $\mathbb{E}[\widehat{g}_\epsilon(z)] \simeq Cg$, so this produces the same expected update scaling as replacing η by $\alpha\eta$. Layer-dependent covariance scales similarly induce layer-dependent effective learning rates. Meanwhile, the alignment score scales exactly as $A(\alpha C) = \alpha A(C)$. Fixing the trace removes this scale effect from the alignment comparison.

Numerical summary. Table 6 summarizes the runs in Figure 4. The common-loss reference is MeZO's final validation loss, 0.3060. Speedup is 375 divided by the first logged step reaching this threshold. The figure's dashed guides use interpolated crossings, whereas the table uses logged evaluations. This ablation uses the final loss directly, whereas the language-model comparison uses the manually selected task-specific targets in Appendix D.2.

D.4.2 Weight and Input-Covariance Contributions

Factorization and normalization. For a linear layer $y = Wx + b$, centering removes the bias and gives

$$B = \mathbb{E}\big[(y - \mathbb{E}y)(x - \mathbb{E}x)^\top\big] = W\Sigma_x, \qquad \Sigma_x = \mathbb{E}\big[(x - \mathbb{E}x)(x - \mathbb{E}x)^\top\big]. \tag{25}$$

Here B is the signed activation cross-covariance before conversion to nonnegative guidance scores, and Σ_x is the covariance of layer inputs. It differs from the parameter-space sampling covariance C used in the alignment metric. We compare guidance derived from the weights alone, the input covariance alone, and the full Hebbian statistic. Each resulting sampling covariance is normalized to the same trace.

Paired alignment measurements. In each component run, we compare its cached guidance with full Hebbian guidance recomputed at the same diagnostic checkpoint, using the same gradient. The orange curves in Figure 5 are therefore paired diagnostics along the respective component trajectories, rather than a separate HiZO training run. Over 38 logged checkpoints, weight-only guidance has mean alignment 2.23, compared with 2.83 for its paired Hebbian diagnostic. Input-covariance-only guidance has mean alignment 1.60, compared with 2.85 for its paired Hebbian diagnostic. Both factors carry alignment information, and combining them yields higher alignment at a matched variance budget.

Complementarity of the two factors. The mean alignment gains above the isotropic reference are approximately additive: the weight-only and input-covariance-only gains sum to 1.83, close to the full Hebbian gains of 1.83 and 1.85 in the paired diagnostics. This numerical pattern suggests complementary information in the two factors, but does not establish an additive decomposition of alignment or convergence speed.

E Computational and Memory Efficiency

We measure the cost of Hebbian guidance using optimizer-step time and peak GPU memory on RoBERTa-large with SST-2. All methods use the same NVIDIA A800-80GB GPU, model precision, microbatch size, sequence length, trainable parameters, and gradient-accumulation setting. Table 3 summarizes the measurements.

Measurement protocol. Step time includes both loss evaluations, perturbation generation and restoration, Hebbian-statistics collection and normalization, and the parameter update. We exclude validation, alignment diagnostics, and checkpoint writing. We time complete optimizer steps, synchronizing the GPU at the boundaries of the measurement interval. Per-step time is the elapsed time divided by the number of measured updates. Peak memory is the maximum GPU memory allocated during the measured training interval, including model parameters, cached statistics, and temporary buffers. The memory counter is reset immediately before this interval.

HiZO's per-step overhead comes from collecting and processing activation covariances. S-HiZO reduces this additional time by 61.1% through subsampling. SA-HiZO has a slightly higher step cost than S-HiZO but brings peak memory to within 4 MB of MeZO, reflecting the smaller persistent statistics representation.

Relating step counts to training time. Let τ_M denote the measured time per optimizer step. Combining this cost with the SST-2 steps-to-target ratio S_M gives the estimate

$$\widehat{S}_M^{\text{time}} = S_M \frac{\tau_{\text{MeZO}}}{\tau_M}. \tag{26}$$

Using the ratios in Table 2 yields estimated training-time speedups of $2.19\times$, $2.85\times$, and $2.55\times$, respectively. This calculation assumes that the measured mean step cost is representative throughout training and excludes validation, diagnostic, and checkpoint-writing time. These are estimates of training time to the shared validation-loss target, not separately measured end-to-end runtimes.

Computational complexity. Consider a guided weight matrix $W \in \mathbb{R}^{m\times n}$ with N input–output activation rows per statistics-collection pass, and let $q < N$ denote the number of sampled rows. Full HiZO forms the centered cross-covariance $Y_c^\top X_c$ at a cost of $O(Nmn)$, followed by $O(mn)$ score normalization. S-HiZO reduces the matrix-product cost to $O(qmn)$ by sampling activation rows. SA-HiZO forms the same sampled cross-covariance and computes its row and column norms in $O(mn)$ time, so its covariance-estimation cost remains $O(qmn)$. Its factorized score representation reduces storage but does not eliminate this matrix product. Applying guidance to a dense perturbation requires $O(mn)$ operations in all three variants.

Let d be the number of trainable parameters and $\mathcal{C}_F$ the cost of one loss evaluation. Without gradient accumulation, MeZO requires $2\mathcal{C}_F + O(d)$ work per update. Collecting statistics once per update adds

$$O\left(\sum_{\ell\in\mathcal{G}} N_\ell m_\ell n_\ell\right) \quad \text{for HiZO,} \qquad O\left(\sum_{\ell\in\mathcal{G}} q_\ell m_\ell n_\ell\right) \quad \text{for S-HiZO and SA-HiZO,}$$

where $\mathcal{G}$ denotes the guided layers. Statistics are collected during an existing loss evaluation, requiring no additional model forward pass. These expressions describe the dominant matrix products; activation preparation and row selection add implementation-dependent costs. For example, casting all activation rows before sampling costs $O(N(m+n))$. With gradient accumulation, forward and statistics-collection costs scale with the number of times each is performed per optimizer step.

Memory complexity. Beyond model storage and the working memory needed for forward evaluation and perturbation replay, HiZO and S-HiZO retain $O(mn)$ scores per guided matrix. SA-HiZO retains two vectors and a constant number of normalization scalars, reducing persistent score storage to $O(m+n)$. Across guided layers, these costs become

$$O\left(\sum_{\ell\in\mathcal{G}} m_\ell n_\ell\right) \quad \text{and} \quad O\left(\sum_{\ell\in\mathcal{G}} (m_\ell + n_\ell)\right),$$

respectively. Sampling also reduces the centered activation workspace from $O(N(m+n))$ to $O(q(m+n))$, although casting before sampling can still create full-size activation buffers. The current SA-HiZO implementation temporarily materializes an $m \times n$ cross-covariance during collection and a dense guidance map during perturbation generation. Its transient workspace therefore still includes $O(mn)$ buffers. The reduction to $O(m+n)$ concerns persistent scores. Measured peak memory additionally depends on buffer lifetimes, precision, and the rest of the training workload.

F FULL DERIVATION OF THE ALIGNMENT-DEPENDENT CONVERGENCE RATE

F.1 CONDITIONAL SETUP AND THE MEANING OF HCA

Let $F : \mathbb{R}^d \to \mathbb{R}$ be differentiable and lower bounded by $F_{\inf}$, with

$$\|\nabla F(x) - \nabla F(y)\| \le L\|x - y\|, \qquad L > 0. \tag{27}$$

The objective may be nonconvex. Write $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_t]$, where the filtration $\mathcal{F}_t$ includes θ_t and the cached statistics used to sample z_t. The current random seed is fresh conditional on this history. In particular, statistics collected during a current perturbed evaluation may determine C_{t+1}, while C_t remains fixed throughout the current update. Set

$$g_t = \nabla F(\theta_t), \qquad C_t = \mathbb{E}_t[z_t z_t^\top], \qquad D_t = g_t^\top C_t g_t. \tag{28}$$

All directions below have conditional mean zero and covariance trace d. We assume HCA, $D_t \ge (1+\gamma)\|g_t\|^2$, along the analyzed trajectory, and write $\Delta = \mathbb{E}[F(\theta_0)] - F_{\inf} < \infty$. Expectations and moments in the bounds are assumed finite.

For diagonal $C_t = \mathrm{diag}(c_{t,1}, \dots, c_{t,d})$ with trace d, the mean of $c_{t,i}$ over coordinates is one; for HiZO, $c_{t,i} = H_{t,i}$. Hence

$$\begin{aligned} d\ \mathrm{Cov}_{i\sim\mathrm{Unif}([d])}(c_{t,i}, g_{t,i}^2) &= \sum_{i=1}^d c_{t,i} g_{t,i}^2 - \sum_{i=1}^d g_{t,i}^2 \\ &= D_t - \|g_t\|^2. \end{aligned} \tag{29}$$

HCA measures how the sampler allocates its fixed variance budget along the current gradient. For a nonzero gradient, it implies $1 + \gamma \le \lambda_{\max}(C_t) \le d$, where $\lambda_{\max}$ denotes the largest eigenvalue. The condition is imposed along the analyzed trajectory, so the covariance may adapt as the gradient changes.

HiZO initializes its variance weights to one. At such a step, $D_t = \|g_t\|^2$, so a positive uniform γ can start only after this initialization unless $g_t = 0$. Our indexing restarts at an aligned segment. If the restart occurs after s warm-up steps, replace Δ by $\Delta_s = \mathbb{E}[F(\theta_s)] - F_{\inf}$ and add s to the resulting iteration budget. The theorem uses a uniform lower bound on alignment over this segment.

F.2 FIXED-NORM IDEALIZATION AND ITS RELATION TO HIZO

HiZO forms nonnegative layerwise variance weights

$$H_t^{(\ell)} = \frac{d_\ell \widetilde{H}_t^{(\ell)}}{\|\widetilde{H}_t^{(\ell)}\|_1}, \qquad \sum_i H_{t,i}^{(\ell)} = d_\ell, \tag{30}$$

where $\widetilde{H}_t^{(\ell)}$ contains magnitudes of centered output–input covariances. Layers with zero total score and parameters without scores use unit variance. Thus the concatenated weights satisfy $\sum_i H_{t,i} = d$. The Gaussian rule in Eq. (3) is

$$z_t^{\mathrm{G}} = \sqrt{H_t} \odot \xi_t, \qquad \xi_t \sim \mathcal{N}(0, I_d), \qquad C_t^{\mathrm{G}} = \mathrm{diag}(H_t). \tag{31}$$

To isolate covariance alignment from fluctuations in the direction norm, consider the covariance-matched surrogate

$$z_t^{\mathrm{S}} = \sqrt{H_t} \odot u_t, \qquad \Pr(u_{t,i} = 1) = \Pr(u_{t,i} = -1) = \tfrac{1}{2}, \tag{32}$$

with independent coordinates. Since $\mathbb{E}_t[u_{t,i}u_{t,j}]$ is one for $i = j$ and zero otherwise,

$$C_t^{\mathrm{S}} = \mathrm{diag}(H_t) = C_t^{\mathrm{G}}, \qquad \|z_t^{\mathrm{S}}\|^2 = \sum_i H_{t,i} = d. \tag{33}$$

This surrogate retains the cached Hebbian variance allocation and the two-sided update, changing the base noise from Gaussian variables to uniform signs. Its HCA parameter is therefore exactly the one for Gaussian HiZO at the same parameter vector and cache. A uniform HCA condition is imposed separately along each method's trajectory.

F.3 Conditional moments and nonconvex descent

For $\epsilon > 0$, define the two-sided directional difference and update by

$$q_\epsilon(\theta, z) = \frac{F(\theta + \epsilon z) - F(\theta - \epsilon z)}{2\epsilon}, \qquad \widehat{g}_t = q_\epsilon(\theta_t, z_t) z_t, \qquad \theta_{t+1} = \theta_t - \eta \widehat{g}_t. \tag{34}$$

The ideal directional-derivative oracle is defined by $q_0(\theta_t, z_t) = g_t^\top z_t$. This is the limit as $\epsilon \to 0$, and equals q_ϵ for every quadratic objective and every $\epsilon > 0$.

Lemma 1 (Fixed-norm conditional moments)**.** *If $\|z_t\|^2 = d$ almost surely and $\widehat{g}_t = (g_t^\top z_t) z_t$, then*

$$\mathbb{E}_t[\widehat{g}_t] = C_t g_t, \qquad \mathbb{E}_t\|\widehat{g}_t\|^2 = d D_t. \tag{35}$$

Proof. The first identity follows from $\mathbb{E}_t[z_t z_t^\top g_t] = C_t g_t$. For the second,

$$\mathbb{E}_t\|\widehat{g}_t\|^2 = \mathbb{E}_t[(g_t^\top z_t)^2 \|z_t\|^2] = d\, \mathbb{E}_t[(g_t^\top z_t)^2] = d D_t.$$

□

Proof of Theorem 1. Smoothness implies the descent inequality

$$F(\theta_t - \eta \widehat{g}_t) \le F(\theta_t) - \eta g_t^\top \widehat{g}_t + \frac{L\eta^2}{2} \|\widehat{g}_t\|^2. \tag{36}$$

For completeness, it follows by integrating $\nabla F(\theta_t + sv)^\top v$ over $s \in [0, 1]$ and bounding the difference from $g_t^\top v$ by $Ls\|v\|^2$, with $v = -\eta \widehat{g}_t$. Taking conditional expectations and applying Lemma 1,

$$\mathbb{E}_t[F(\theta_{t+1})] \le F(\theta_t) - \left(\eta - \frac{L\eta^2 d}{2}\right) D_t. \tag{37}$$

The coefficient is maximized at $\eta = 1/(Ld)$ and equals $1/(2Ld)$. Using HCA, taking total expectations, and summing yields

$$\begin{aligned} \frac{1+\gamma}{2Ld} \sum_{t=0}^{T-1} \mathbb{E}\|g_t\|^2 &\le \sum_{t=0}^{T-1} \big(\mathbb{E}F(\theta_t) - \mathbb{E}F(\theta_{t+1})\big) \\ &= \mathbb{E}F(\theta_0) - \mathbb{E}F(\theta_T) \le \Delta. \end{aligned} \tag{38}$$

Rearranging proves (9). □

If R is uniform on $\{0, \dots, T-1\}$ and independent of the algorithm, the left-hand side of (9) equals $\mathbb{E}\|\nabla F(\theta_R)\|^2$. Thus a sufficient iteration count for $\mathbb{E}\|\nabla F(\theta_R)\|^2 \le \delta$, with $\delta > 0$, is

$$T_{\mathrm{H}}(\delta) = \left\lceil \frac{2Ld\Delta}{(1+\gamma)\delta} \right\rceil. \tag{39}$$

Isotropic sphere sampling $u = \sqrt{d}\,\xi/\|\xi\|$, with $\xi \sim \mathcal{N}(0, I_d)$, has $\mathbb{E}[uu^\top] = I_d$, as does independent sign sampling. Both give $T_{\mathrm{I}}(\delta) = \lceil 2Ld\Delta/\delta \rceil$ under the same analysis. Before integer rounding, the sufficient budgets differ by $1+\gamma$ for the same objective, starting gap, and oracle. This is a comparison of stationarity bounds. Each finite-radius update uses the two objective evaluations in Eq. (4).

F.4 Finite-radius correction for the fixed-norm surrogate

Proposition 2 (Finite-radius fixed-norm bound)**.** *Under the conditions of Theorem 1, use the actual central difference (34) with $\eta = 1/(Ld)$. Suppose $q_\epsilon(\theta_t, z_t) = g_t^\top z_t + r_t$ and $\mathbb{E}_t[r_t^2] \le \beta_\epsilon^2$. Then*

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\|g_t\|^2 \le \frac{2Ld\Delta}{(1+\gamma)T} + \frac{\beta_\epsilon^2}{1+\gamma}. \tag{40}$$

Under L-smoothness one may take $\beta_\epsilon = L\epsilon d/2$. If F is twice differentiable with ρ-Lipschitz Hessian in operator norm, one may instead take $\beta_\epsilon = \rho\epsilon^2 d^{3/2}/6$.

Proof. Write $a_t = g_t^\top z_t$. Applying (36) pathwise and using $\|z_t\|^2 = d$ gives an exact cancellation:

$$\begin{aligned} F(\theta_{t+1}) &\le F(\theta_t) - \frac{a_t(a_t + r_t)}{Ld} + \frac{(a_t + r_t)^2}{2Ld} \\ &= F(\theta_t) - \frac{a_t^2}{2Ld} + \frac{r_t^2}{2Ld}. \end{aligned} \tag{41}$$

No independence between a_t and r_t is needed. Taking conditional expectations yields $\mathbb{E}_t F(\theta_{t+1}) \le F(\theta_t) - D_t/(2Ld) + \beta_\epsilon^2/(2Ld)$. Apply HCA and telescope to obtain (40).

For the smoothness-only remainder, the fundamental theorem of calculus gives

$$\begin{aligned} r_t &= \frac{1}{2\epsilon} \int_{-\epsilon}^{\epsilon} \big(\nabla F(\theta_t + s z_t) - g_t\big)^\top z_t \, ds, \\ |r_t| &\le \frac{L\|z_t\|^2}{2\epsilon} \int_{-\epsilon}^{\epsilon} |s| \, ds = \frac{L\epsilon d}{2}. \end{aligned} \tag{42}$$

With a ρ-Lipschitz Hessian, the second-order Taylor expansion obeys

$$F(\theta_t + h) = F(\theta_t) + g_t^\top h + \tfrac{1}{2} h^\top \nabla^2 F(\theta_t) h + R_t(h), \qquad |R_t(h)| \le \frac{\rho}{6}\|h\|^3.$$

Subtract the expansions at $h = \epsilon z_t$ and $h = -\epsilon z_t$. The quadratic terms cancel, so $|r_t| \le \rho\epsilon^2\|z_t\|^3/6 = \rho\epsilon^2 d^{3/2}/6$. □

In particular, the Hessian-Lipschitz error floor is $\rho^2\epsilon^4 d^3/[36(1+\gamma)]$ and vanishes without changing the leading constant as $\epsilon \to 0$. For an isotropic fixed-norm baseline, the same derivation gives $2Ld\Delta/T + \beta_\epsilon^2$. Thus the bounds differ by $1+\gamma$ at the same radius and step size when HCA holds along the finite-radius trajectory. The remainder determines the accuracy attainable at a fixed radius.

F.5 Exact Gaussian moments and the HiZO convergence bound

For Gaussian HiZO, $\mathbb{E}_t\|z_t\|^2 = d$ holds, whereas $\|z_t\|^2 = d$ does not hold pathwise. The relevant fourth moment can be computed exactly, including when some variance weights are zero.

Lemma 2 (Gaussian conditional moments)**.** *For $z_t \mid \mathcal{F}_t \sim \mathcal{N}(0, C_t)$ with $\mathrm{tr}(C_t) = d$, the ideal estimator satisfies*

$$\mathbb{E}_t[\widehat{g}_t] = C_t g_t, \qquad \mathbb{E}_t\|\widehat{g}_t\|^2 = dD_t + 2g_t^\top C_t^2 g_t. \tag{43}$$

Proof. Diagonalize $C_t = U \,\mathrm{diag}(\lambda_1, \dots, \lambda_d) U^\top$ and write $v = U^\top g_t$. For independent standard Gaussian coordinates, $\mathbb{E}\xi_i^2 = 1$, $\mathbb{E}\xi_i^4 = 3$, and odd cross terms vanish. Consequently,

$$\begin{aligned} \mathbb{E}_t[(g_t^\top z_t)^2\|z_t\|^2] &= \sum_i \lambda_i v_i^2 \left(3\lambda_i + \sum_{j \ne i} \lambda_j\right) \\ &= \mathrm{tr}(C_t) \sum_i \lambda_i v_i^2 + 2\sum_i \lambda_i^2 v_i^2 = dD_t + 2g_t^\top C_t^2 g_t. \end{aligned} \tag{44}$$

The first moment follows directly from $\mathbb{E}_t[z_t z_t^\top] = C_t$. □

Proposition 3 (Gaussian HiZO in the directional-derivative limit)**.** *Suppose HCA holds and a deterministic constant $K > 0$ satisfies*

$$g_t^\top C_t^2 g_t \le K g_t^\top C_t g_t \tag{45}$$

almost surely throughout the analyzed segment. Then Gaussian HiZO with q_0 and $\eta = 1/[L(d+2K)]$ satisfies (12). One sufficient choice is $K \ge \sup_t \|C_t\|_{\mathrm{op}}$, where $\|\cdot\|_{\mathrm{op}}$ denotes the operator norm.

Proof. Lemma 2 and (36) give

$$\mathbb{E}_t F(\theta_{t+1}) \le F(\theta_t) - \eta D_t + \frac{L\eta^2}{2}(dD_t + 2g_t^\top C_t^2 g_t)$$

$$\leq F(\theta_t) - \left(\eta - \frac{L\eta^2(d+2K)}{2}\right) D_t$$
$$\leq F(\theta_t) - \frac{1+\gamma}{2L(d+2K)}\|g_t\|^2. \tag{46}$$

The last line uses the stated step size and HCA. Taking expectations and telescoping proves the result. The spectral sufficient condition follows from $C_t^2 \preceq \|C_t\|_{\mathrm{op}} C_t$. □

For isotropic Gaussian directions, $C_t = I_d$ and the exact second moment is $(d+2)\|g_t\|^2$. The same proof with $\eta = 1/[L(d+2)]$ gives $2L(d+2)\Delta/T$. The ratio of the isotropic and HiZO bound constants is

$$\mathcal{S}_{\mathrm{G}} = (1+\gamma)\frac{d+2}{d+2K}, \qquad \frac{\mathcal{S}_{\mathrm{G}}}{1+\gamma} = 1 - \frac{2(K-1)}{d+2K}. \tag{47}$$

The HiZO bound is tighter precisely when $(1+\gamma)(d+2) > d+2K$. For $K = o(d)$, $\mathcal{S}_{\mathrm{G}}/(1+\gamma) \to 1$; this is a high-dimensional moment condition in addition to taking the directional-derivative limit. For the diagonal sampler, the directional ratio $g_t^\top C_t^2 g_t / D_t$ is at most $\max_i H_{t,i}$ whenever $D_t > 0$. In particular, a uniform bound on all variance weights supplies a valid dimension-independent K. Trace normalization alone supplies $K = d$, giving the conservative bound $6Ld\Delta/[(1+\gamma)T]$.

The affine rescaling in Appendix C supplies $K = c^{-1}$ by enforcing $H_{t,i} \leq c^{-1}$. It therefore gives a direct way to control this correction while retaining a fixed total variance. Appendix G.5 shows how the same rescaling preserves a positive alignment margin. The Gaussian result here uses q_0; the finite-radius cancellation in Appendix F.4 uses the fixed-norm property.

F.6 Minibatch extension and the oracle setting

The main theorem uses exact evaluations of one fixed objective F. To connect this setting to minibatch HiZO, consider the ideal derivative of the minibatch loss, with $G_t = \nabla F_{B_t}(\theta_t)$ and $\widehat{g}_t = (G_t^\top z_t) z_t$. Assume B_t and the current direction are conditionally independent given $\mathcal{F}_t$, and

$$\mathbb{E}[G_t \mid \mathcal{F}_t] = g_t, \qquad \mathbb{E}[(G_t - g_t)^\top C_t (G_t - g_t) \mid \mathcal{F}_t] \leq v_C^2. \tag{48}$$

The cached H_t is constructed before drawing B_t and z_t, as in Algorithm 1; fresh independent draws therefore give the conditional independence used here. The quantity v_C^2 bounds minibatch gradient noise weighted by the exploration covariance. Let $M = d$ for fixed-norm directions. For Gaussian directions take $M = d + 2K$ with the spectral bound $\|C_t\|_{\mathrm{op}} \leq K$, so the moment bound holds for both gradients and their noise.

Conditional independence and unbiasedness imply

$$\mathbb{E}[\widehat{g}_t \mid \mathcal{F}_t] = C_t g_t,$$
$$\mathbb{E}[\|\widehat{g}_t\|^2 \mid \mathcal{F}_t] \leq M\, \mathbb{E}[G_t^\top C_t G_t \mid \mathcal{F}_t] \leq M(D_t + v_C^2). \tag{49}$$

For $0 < \eta \leq 1/(LM)$, smoothness and HCA therefore yield

$$\mathbb{E}[F(\theta_{t+1}) \mid \mathcal{F}_t] \leq F(\theta_t) - \frac{\eta(1+\gamma)}{2}\|g_t\|^2 + \frac{L\eta^2 M}{2} v_C^2.$$

Telescoping and dividing by $\eta(1+\gamma)T/2$ gives

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|g_t\|^2 \leq \frac{2\Delta}{\eta(1+\gamma)T} + \frac{L\eta M v_C^2}{1+\gamma}. \tag{50}$$

For $v_C^2 > 0$, a step size proportional to $T^{-1/2}$, capped at $1/(LM)$, gives an $O(T^{-1/2})$ bound. With exact gradients and $\eta = 1/(LM)$, the $O(T^{-1})$ result is recovered. Equation (50) connects alignment to minibatch HiZO through both descent and covariance-weighted noise; comparisons between samplers must use their respective M and v_C^2. At a positive finite-difference radius, the remainder must also be bounded for the sampled losses.

G When Hebbian covariance contains task-gradient information

G.1 From the output residual to the exact gradient

Consider a one-layer linear predictor $y = Wx + b$, where $W \in \mathbb{R}^{m\times n}$ and the target is $\hat{y} = e_c \in \mathbb{R}^m$, the standard basis vector for class c. The input distribution is fixed, $\mathbb{E}x = 0$, and $\Sigma = \mathbb{E}[xx^\top] \succ 0$ is finite. The bias b is fixed in this analysis, so the trainable dimension is $d = mn$. Taking $b = 0$ recovers $y = Wx$; taking $b = \mathbb{E}\hat{y}$ fits the target mean without affecting the calculations for W.

Under squared loss, the output residual $y - \hat{y}$ determines the gradient:

$$\ell(W; x, \hat{y}) = \tfrac{1}{2}\|Wx + b - \hat{y}\|^2, \qquad F(W) = \mathbb{E}\ell(W; x, \hat{y}). \tag{51}$$

Differentiation gives

$$\nabla_W \ell = (y - \hat{y})x^\top = yx^\top - \hat{y}x^\top. \tag{52}$$

This decomposition separates the output and target Hebbian products (Melchior & Wiskott, 2019).

Let $\pi = \mathbb{E}\hat{y}$ and define signed cross-covariances by

$$B = \mathbb{E}[(y - \mathbb{E}y)x^\top], \qquad Q = \mathbb{E}[(\hat{y} - \pi)x^\top]. \tag{53}$$

Since $\mathbb{E}y = b$ and $\mathbb{E}x = 0$,

$$B = W\Sigma, \qquad G := \nabla_W F(W) = W\Sigma - Q = B - Q. \tag{54}$$

HiZO estimates $|B|$ from forward activations using Eq. (1). Thus its scores reflect both model weights and the input distribution through $W\Sigma$, while the labeled loss evaluations supply the task objective for the update.

Centering matters for this identity. If the original input x_0 has mean μ, with $y = Wx_0 + b$, then

$$\nabla_W F = \mathrm{Cov}(y, x_0) - \mathrm{Cov}(\hat{y}, x_0) + (\mathbb{E}y - \mathbb{E}\hat{y})\mu^\top. \tag{55}$$

Centering the features or fitting the target mean with the bias ($\mathbb{E}y = \mathbb{E}\hat{y}$) removes the last term. Otherwise this mean-residual contribution is included in the mismatch term considered in Appendix G.4.

G.2 What one-hot labels imply

For an individual example, $\hat{y}x^\top = e_c x^\top$ changes only row c of $yx^\top$. Thus $m - 1$ of the m rows of the sample gradient equal the corresponding rows of the uncentered output Hebbian product. The corresponding magnitudes satisfy

$$\|e_c x^\top\|_{\mathrm{F}} = \|x\|, \qquad \|yx^\top\|_{\mathrm{F}} = \|y\|\,\|x\|. \tag{56}$$

Here $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. HiZO centers the activations and averages their outer products before taking covariance magnitudes. The relevant target contribution is therefore the population quantity Q, determined by class-conditional input means.

For the population statistic, let $\pi_k = \Pr(c = k) > 0$ and $\mu_k = \mathbb{E}[x \mid c = k]$. The kth row of Q is

$$Q_{k,:} = \mathbb{E}[\mathbf{1}_{\{c=k\}}x]^\top = \pi_k \mu_k^\top. \tag{57}$$

Rows for classes of zero probability are zero. Conditional Jensen's inequality gives a useful bound:

$$\begin{aligned}\|Q\|_{\mathrm{F}}^2 &= \sum_k \pi_k^2 \|\mu_k\|^2 \le \pi_{\max} \sum_k \pi_k \|\mu_k\|^2 \\ &\le \pi_{\max}\mathbb{E}\|x\|^2 = \pi_{\max}\,\mathrm{tr}(\Sigma).\end{aligned} \tag{58}$$

Here $\pi_{\max} = \max_k \pi_k$; balanced classes give $\pi_{\max} = 1/m$. When this bound is small relative to $\|B\|_{\mathrm{F}}^2$, it supports $G \approx B$. The next result explains alignment when the model has learned the structure of Q, even if its magnitude is substantial.

G.3 A LEARNED TASK MAP GIVES EXACT MAGNITUDE ALIGNMENT

The least-squares normal equation is $W_\star \Sigma = Q$. Since $\Sigma \succ 0$, the unique minimizer over W is $W_\star = Q\Sigma^{-1}$. This conclusion requires only second moments, without assuming that the conditional mean of the labels is linear in x. Suppose $Q \neq 0$ and the current map is $W = aW_\star$, with $a > 0$ and $a \neq 1$. Then

$$B = aQ, \qquad G = (a-1)Q = \kappa B, \qquad \kappa = \frac{a-1}{a} \neq 0. \tag{59}$$

For $0 < a < 1$, the gradient and signed Hebbian covariance have opposite signs. Their absolute values are proportional, which is the relation needed for allocating exploration variance.

Lemma 3 (Magnitude weighting and alignment)**.** *Let $B \neq 0$, $h = |\operatorname{vec}(B)| \in \mathbb{R}^d_{\geq 0}$, and $C = \operatorname{diag}(dh/\|h\|_1)$. If $G = \kappa B$ with $\kappa \neq 0$, then, for $g = \operatorname{vec}(G)$,*

$$\frac{g^\top C g}{\|g\|^2} = \Gamma(h) = \frac{d\sum_i h_i^3}{(\sum_i h_i)(\sum_i h_i^2)}, \qquad 1 \leq \Gamma(h) \leq \|C\|_{\mathrm{op}} \leq d. \tag{60}$$

Equality $\Gamma(h) = 1$ holds exactly when all h_i are equal.

Proof. The squared coordinate g_i^2 is $\kappa^2 h_i^2$. Therefore

$$g^\top C g = \frac{d\kappa^2 \sum_i h_i^3}{\sum_i h_i}, \qquad \|g\|^2 = \kappa^2 \sum_i h_i^2,$$

which proves the ratio. To establish the lower bound and its equality case, expand

$$d\sum_i h_i^3 - (\sum_i h_i)(\sum_i h_i^2) = \sum_{i<j}(h_i - h_j)^2(h_i + h_j) \geq 0. \tag{61}$$

If two entries differ, their contribution is strictly positive. The upper bounds follow from the Rayleigh quotient and $\operatorname{tr}(C) = d$. □

Combining (59) and Lemma 3 proves Proposition 1. The HCA margin has the explicit form

$$\gamma_\star = \Gamma(h) - 1 = \frac{\sum_{i<j}(h_i - h_j)^2(h_i + h_j)}{(\sum_i h_i)(\sum_i h_i^2)}. \tag{62}$$

If h has exactly s equal nonzero entries, then $\Gamma(h) = d/s$. The gain comes from nonuniform task-relevant magnitudes in the weight coordinates where HiZO allocates variance.

A concrete learning trajectory. Linear-network learning dynamics can be expressed through input and input–target correlations (Saxe et al., 2014, Section 1). For the one-layer squared-loss model, the gradient-flow equation is

$$\dot{W}(t) = Q - W(t)\Sigma. \tag{63}$$

If $\Sigma = \sigma^2 I_n$ and $W(0) = 0$, solving this linear differential equation gives

$$W(t) = (1 - e^{-\sigma^2 t})W_\star, \qquad B(t) = (1 - e^{-\sigma^2 t})Q, \qquad G(t) = -e^{-\sigma^2 t}Q. \tag{64}$$

For every finite $t > 0$, the magnitude lemma applies. The normalized covariance C and Γ are constant along this trajectory because the scale of $|B(t)|$ cancels in normalization. Discrete gradient descent has the same property for $W_0 = 0$ and $0 < \eta\sigma^2 < 1$, since $W_k = [1 - (1 - \eta\sigma^2)^k]W_\star$ for $k \geq 1$. At initialization $B = 0$, HiZO's unit-variance fallback is isotropic. These first-order trajectories illustrate how task learning produces aligned states. The next proposition extends the mechanism to nearby states, allowing departures from exact scalar rescaling.

One-hot example with a strict gain. Let $x = (u, v)^\top$, where u and v are independent uniform signs, and set $\hat{y} = e_1$ for $u = 1$ and $\hat{y} = e_2$ for $u = -1$. Then

$$\Sigma = I_2, \qquad \pi = (1/2, 1/2)^\top, \qquad Q = W_\star = \begin{pmatrix} 1/2 & 0 \\ -1/2 & 0 \end{pmatrix}. \tag{65}$$

For $W = aW_\star$, $0 < a < 1$, and $b = \pi$, the predictor has learned the correct class-dependent direction but its predictions have reduced amplitude. Both B and G are supported on the two entries multiplying u; the two entries multiplying the irrelevant feature v are zero. Thus $d = 4$, $s = 2$, and $\Gamma = 2$, giving HCA with $\gamma = 1$ at these states. The score normalization allocates variance to the task-relevant feature while preserving the total budget.

G.4 Stability under approximate proportionality

Exact rescaling is sufficient, but approximate proportionality can also preserve alignment. The following bound quantifies how much mismatch the alignment margin can tolerate.

Proposition 4 (Alignment with a residual mismatch)**.** *Let $B \neq 0$, $h = |\operatorname{vec}(B)|$, and $C = \operatorname{diag}(dh/\|h\|_1)$, with $\Gamma = \Gamma(h)$ and $\Lambda = \|C\|_{\mathrm{op}}$. Suppose*

$$G = \kappa B + E, \qquad \kappa \neq 0, \qquad \|E\|_{\mathrm{F}} \leq r|\kappa|\|B\|_{\mathrm{F}}, \qquad 0 \leq r < 1. \tag{66}$$

Then $g = \operatorname{vec}(G) \neq 0$ and

$$\frac{g^\top C g}{\|g\|^2} \geq \left(\frac{[\sqrt{\Gamma} - \sqrt{\Lambda}\, r]_+}{1 + r} \right)^2, \qquad [z]_+ = \max\{z, 0\}. \tag{67}$$

In particular, when $\Gamma > 1$, a strictly positive HCA margin is guaranteed if

$$r < \frac{\sqrt{\Gamma} - 1}{\sqrt{\Lambda} + 1}. \tag{68}$$

Proof. Put $q = \kappa \operatorname{vec}(B)$ and $e = \operatorname{vec}(E)$. The reverse triangle inequality gives $\|g\| \geq (1-r)\|q\| > 0$, while $\|g\| \leq (1+r)\|q\|$. Lemma 3 gives $\|C^{1/2}q\| = \sqrt{\Gamma}\|q\|$. Hence

$$\begin{aligned} \|C^{1/2}g\| &\geq [\|C^{1/2}q\| - \|C^{1/2}e\|]_+ \\ &\geq [\sqrt{\Gamma} - \sqrt{\Lambda}\, r]_+ \|q\|. \end{aligned} \tag{69}$$

Squaring and dividing by the upper bound on $\|g\|^2$ proves (67). Condition (68) makes its right-hand side strictly larger than one. □

This result covers two interpretations of the one-layer decomposition. First, when $G = B - Q$, take $\kappa = 1$ and $E = -Q$. The actual relative error is $r = \|Q\|_{\mathrm{F}}/\|B\|_{\mathrm{F}}$. Equation (58) bounds it by $\sqrt{\pi_{\max} \operatorname{tr}(\Sigma)}/\|B\|_{\mathrm{F}}$. If this upper bound meets (68), it certifies a positive alignment margin. Second, for an approximately rescaled task map $W = aW_\star + D$,

$$B = aQ + D\Sigma, \qquad G = \frac{a-1}{a} B + \frac{1}{a} D\Sigma. \tag{70}$$

For $a > 0$, $a \neq 1$, an admissible relative mismatch is

$$r = \frac{\|D\Sigma\|_{\mathrm{F}}}{|a-1|\|B\|_{\mathrm{F}}}, \tag{71}$$

provided it is less than one. Meeting Eq. (68) preserves positive alignment, with model mismatch measured relative to the remaining gradient signal.

G.5 Estimated statistics and layerwise covariance

HiZO constructs its covariance from cached empirical statistics. Let C be the covariance from the exact statistic B, and let $\widehat{C}$ be the covariance actually used for sampling. If $\|\widehat{C} - C\|_{\mathrm{op}} \leq \zeta$, then for every $g \neq 0$,

$$\frac{g^\top \widehat{C} g}{\|g\|^2} \geq \frac{g^\top C g}{\|g\|^2} - \zeta. \tag{72}$$

Indeed, $g^\top(\widehat{C} - C)g \geq -\zeta\|g\|^2$ by the operator-norm bound. Combining this inequality with (67) gives an HCA margin of at least

$$\gamma_{\mathrm{eff}} = \left(\frac{[\sqrt{\Gamma} - \sqrt{\Lambda}\, r]_+}{1 + r} \right)^2 - \zeta - 1, \tag{73}$$

whenever this expression is positive. For diagonal covariances, $\zeta = \max_i |\widehat{H}_i - H_i|$ measures the largest variance-weight error. It can include empirical estimation, the one-step cache delay, collection at the positively perturbed parameters, and the approximations used by S-HiZO or SA-HiZO. The bound shows that these effects preserve HCA when their combined error is smaller than the available alignment margin; it assumes a bound on ζ rather than estimating it from the algorithm.

Bounded variance rescaling. The affine rule in Appendix C also has an exact alignment interpretation. For one block, let h denote its nonzero covariance-magnitude vector, $\mu = d^{-1}\sum_i h_i$, and $C = \mathrm{diag}(h/\mu)$. The rescaled weights are $H_i^{\mathrm{aff}} = 1 + s(h_i - \mu)$, so with $\alpha = s\mu$,

$$C_{\mathrm{aff}} = I_d + \alpha(C - I_d), \qquad \frac{g^\top C_{\mathrm{aff}} g}{\|g\|^2} - 1 = \alpha\left(\frac{g^\top C g}{\|g\|^2} - 1\right). \tag{74}$$

For nonconstant scores and $c < 1$, $\alpha > 0$, so positive alignment is preserved. Under exact proportionality, the margin becomes $\alpha[\Gamma(h) - 1]$. At the same time, $\|C_{\mathrm{aff}}\|_{\mathrm{op}} \le c^{-1}$ controls the Gaussian moment correction. For cached or approximate scores, apply Eq. (72) to the rescaled covariances.

Layerwise aggregation. For J independently perturbed parameter blocks, write $C = \mathrm{blockdiag}(C^{(1)}, \dots, C^{(J)})$ and $g = (g^{(1)}, \dots, g^{(J)})$. With zero-gradient blocks omitted,

$$\frac{g^\top C g}{\|g\|^2} = \sum_\ell \omega_\ell \frac{(g^{(\ell)})^\top C^{(\ell)} g^{(\ell)}}{\|g^{(\ell)}\|^2}, \qquad \omega_\ell = \frac{\|g^{(\ell)}\|^2}{\|g\|^2}. \tag{75}$$

This identity follows by expanding the block quadratic form. If every guided block has margin at least $\gamma_0 > 0$ and all remaining blocks use unit covariance, the global margin is at least $\gamma_0 \sum_{\ell\,\mathrm{guided}} \omega_\ell$. Thus per-layer alignment contributes according to the gradient energy in that layer. With affine rescaling, each layer contributes its rescaled margin through the same gradient-energy weighting. This matches HiZO's layerwise normalization and unit-variance treatment of unscored parameters.

A uniform positive lower bound on the resulting global margin supplies γ in Proposition 3, using the actual cached Gaussian covariance and its moment bound.

G.6 Connection to classification and hidden layers

Alignment depends on the relation between forward covariance and the task gradient; one-hot labels alone are insufficient.

Softmax classification. For logits $u = Wx + b$, probabilities $p = \mathrm{softmax}(u)$, and cross-entropy $\ell = -\hat{y}^\top \log p$, direct differentiation gives

$$\frac{\partial \ell}{\partial u_k} = p_k - \hat{y}_k, \qquad \nabla_W F_{\mathrm{CE}} = \mathbb{E}[(p - \hat{y})x^\top] = \mathrm{Cov}(p, x) - Q \tag{76}$$

for centered x, where $F_{\mathrm{CE}} = \mathbb{E}\ell$. The first identity follows from $\ell = \log\sum_j e^{u_j} - \hat{y}^\top u$ and $\sum_j \hat{y}_j = 1$. This residual form is also discussed by Melchior & Wiskott (2019). HiZO's linear-layer hooks collect logit covariance $B = \mathrm{Cov}(u, x) = W\Sigma$, whereas the gradient involves probability covariance. To relate the actual statistic to a classification gradient, consider the data in (65). Set $W = aQ$, $a > 0$, and $b = 0$, so the logits are $(ax_1/2, -ax_1/2)^\top$. Writing $s(a) = 1/(1 + e^{-a})$, the class-one probability is $s(ax_1)$. For $x_1 = 1$, $(p_1 - \hat{y}_1)x_1 = s(a) - 1$; for $x_1 = -1$, it is $-s(-a) = s(a) - 1$. Independence and centering of x_2 make its expected gradient zero. Thus

$$B = aQ, \qquad G_{\mathrm{CE}} = \begin{pmatrix} s(a) - 1 & 0 \\ 1 - s(a) & 0 \end{pmatrix} = -\frac{2[1 - s(a)]}{a} B. \tag{77}$$

Lemma 3 again gives $\Gamma = 2$ for every finite $a > 0$, using the covariance of the logits actually collected by a linear hook. Population gradient flow preserves this family with $\dot{a} = 2[1 - s(a)]$. Thus forward logit covariance can identify the gradient magnitudes under cross-entropy as well as squared loss.

Hidden-layer guidance. For a hidden linear layer $u^{(\ell)} = W^{(\ell)}x^{(\ell)} + b^{(\ell)}$, the chain rule gives

$$\nabla_{W^{(\ell)}} F = \mathbb{E}[\delta^{(\ell)}(x^{(\ell)})^\top], \qquad \delta^{(\ell)} = \frac{\partial \ell}{\partial u^{(\ell)}}. \tag{78}$$

HiZO collects $B^{(\ell)} = \mathrm{Cov}(u^{(\ell)}, x^{(\ell)}) = W^{(\ell)}\Sigma_{x^{(\ell)}}$ from the same forward pass. A sufficient success condition is that the layer gradient can be written as $G^{(\ell)} = \kappa_\ell B^{(\ell)} + E^{(\ell)}$ with nonuniform covariance magnitudes and mismatch satisfying Eq. (68). Proposition 4 then gives a layerwise margin, and Eq. (75) combines the margins across the network. These are conditions on the actual layer gradients, supplied exactly by the examples above and approximately by the stability bound. They explain how activation-only statistics can guide exploration while HiZO determines updates from loss evaluations, without computing backpropagated residuals.