Trust the Direction, Search the Step: Zero-and-First-Order Methods for LLM Fine-Tuning

Cristian McGee
University of Central Florida
Orlando, FL, USA

El Houcine Bergou
Mohammed VI Polytechnic University
Ben Guerir, Morocco

Aritra Dutta
University of Central Florida
Orlando, FL, USA

Abstract

Step-size selection remains a central challenge in large-scale neural network optimization; conservative steps slow convergence, while aggressive steps can destabilize it. We combine **Z**ero-and-**F**irst-**O**rder optimization (ZFO) and propose a framework that decouples direction selection from step-size. ZFO uses a trusted first-order optimizer to determine the direction and performs zeroth-order evaluations only along this one-dimensional subspace to choose how far to move. Using the current gradient information and two additional objective function evaluations, ZFO instances construct a local model of the objective function along the proposed direction and select a curvature-aware step within a bounded search interval. This yields an adaptive step-selection mechanism that costs less than a full line search. We provide theoretical guarantees to show that shared-sample evaluations produce reliable finite-difference curvature estimates, that the induced local model selects a near-optimal step along the search interval, and that ZFO converges to a neighborhood of a stationary point. Across the evaluated settings, language models and datasets, ZFO frequently improves optimization and final performance relative to fixed-step first-order baselines, with the magnitude and preferred local model depending on the objective. Our code is publicly available at: `https://github.com/nizswan/Zeroth-First-Order-Framework`.

1 Introduction

Gradient-based first-order (FO) optimization methods are the main workhorse of modern deep neural network training [14, 25, 31]. During training, gradient information provides the *direction to move along*, which, coupled with a step size (or learning rate), updates the model parameters. While the gradient direction is generally informative, the step size used along that direction is typically chosen through schedules, heuristics, or hyperparameter tuning [18, 30, 48, 49]. For larger networks, whose local geometry can change rapidly during adaptation, a poorly chosen step size may lead to unstable training [13]. Despite these step-size instabilities, FO methods remain the standard choice for neural network adaptation because the FO gradient provides strong directional guidance.

In contrast, zeroth-order (ZO) methods [5–7, 16, 17, 26, 34, 38] do not compute gradients. They construct approximate gradient updates using only objective function evaluations, often via finite differences along random directions. Although ZO methods reduce computational overhead, gradient approximations are often noisy and high-variance in large-dimensional settings [2, 23]. However, when FO provides the update direction, the high-dimensional problem becomes a univariate step-

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

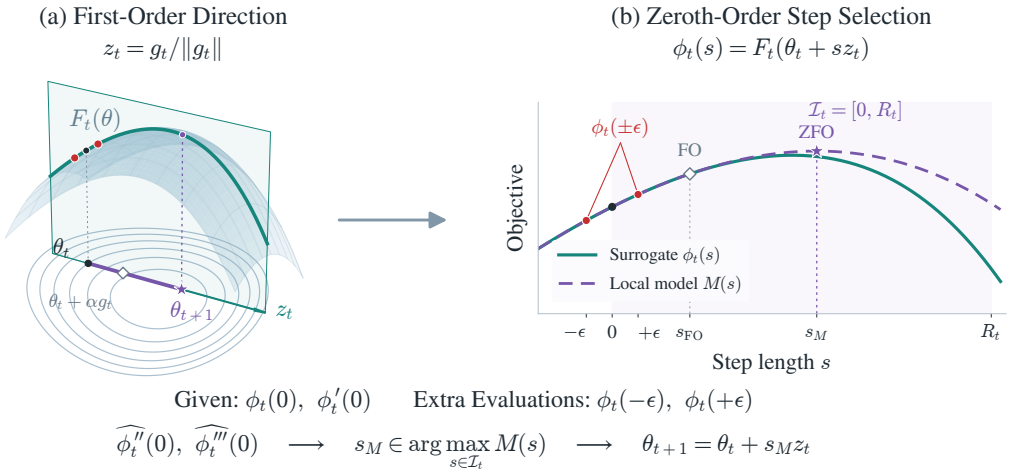

Figure 1: **The Zero-and-First-Order (ZFO) Framework.** *Left:* A first-order optimizer proposes the direction z_t at θ_t, which restricts the objective to the line $\theta_t + sz_t$. *Right:* Along this line, $\theta_t + sz_t$, the forward pass gives $\phi_t(0)$, the backward pass gives $\phi_t'(0)$, and two symmetric probes at $\pm\epsilon$ supply curvature estimates. ZFO fits a local model $M(s)$ and takes the optimized step s_M over the interval $[0, R_t]$; see Algorithm 1.

selection problem; see Figure 1. Consequently, computationally inexpensive ZO evaluations become substantially informative for modeling local function behavior. This raises a natural question:

If FO methods are strong at choosing an update direction, can ZO evaluations be used to decide how far to move along that direction?

We investigate this question by using ZO evaluations not to replace gradients, but to select reliable step sizes. This idea is related to recent work on generalized Newton step selection (GeN) [9], which uses function evaluations along a first-order direction to construct a curvature-informed step. In contrast, ZFO takes a broader view: rather than committing to a particular step construction, we formulate fixed-budget directional step selection as a local modeling problem. This perspective is not restricted to a specific optimization setup; in principle, the framework applies whenever an approximate gradient direction is available, but its appropriate magnitude is uncertain, a pattern that recurs in existing step-size rules [33]. This uncertainty becomes especially relevant in noisy or stochastic optimization settings.

Large language model (LLM) fine-tuning is a natural setting for this hybridization. Modern fine-tuning pipelines construct minibatches of prompt-response examples and optimize a surrogate objective on these fixed examples. Once a minibatch is formed, nearby parameter evaluations can reuse the same data and require only additional forward evaluations of the perturbed model. This makes ZO probes especially useful—not as a replacement for gradients, but as a cheap mechanism for modeling objective variation along a trusted FO direction. In contrast, fully ZO LLM fine-tuning methods use forward evaluations to estimate high-dimensional gradients [26, 34]. ZFO exploits this structure by assigning direction selection to a base FO optimizer, such as AdamW [31] or Muon [25], while using a small number of ZO function evaluations to build a univariate local model of the objective function and select the step size along that direction. Taken together, the ZFO framework acts as a lightweight step-selection wrapper around a standard FO optimizer. This is particularly valuable in LLM fine-tuning, where local geometry can shift rapidly and poorly tuned learning rates can cause unstable or slow training [27, 50].

Problem Setup. Let fine-tuning an LLM be represented by a conditional distribution $\pi_\theta(y \mid x)$, where x denotes an input prompt, y denotes a model response, and $\theta \in \mathbb{R}^d$ are the model parameters. We initialize the model from a pretrained reference model and further optimize it using task-specific feedback. We summarize the main notation used throughout the paper in §D; see Table 15.

At a given iteration t, a FO fine-tuning algorithm constructs a minibatch of information, denoted by S_t. This minibatch contains all quantities needed to define the current training objective. Depending on the method, S_t may include prompts, sampled responses, supervised labels, etc. Conditioned on S_t, the fine-tuning method defines a sampled surrogate objective optimized during the current update

$F_t(\theta) = F(\theta; S_t)$. Different LLM fine-tuning algorithms instantiate this surrogate differently. Our framework does not require a particular choice of F_t, and only assumes that the surrogate can be evaluated on the fixed minibatch S_t.

This fixed-batch viewpoint is standard in LLM fine-tuning [22, 26, 34, 46]. Once S_t is constructed, it is fixed within the update. Evaluating F_t at a nearby parameter value therefore does not require constructing a new batch or collecting new feedback; it only requires re-scoring the same prompt-response pairs under the perturbed model.

A standard fine-tuning step computes a stochastic gradient of the sampled surrogate at the current parameters, and passes it through an optimizer such as AdamW. If we denote the resulting update direction by g_t, the usual first-order gradient ascent update is: $\theta_{t+1} = \theta_t + \alpha g_t$, where $\alpha > 0$ is the learning rate. This update contains two decisions: the direction g_t and the step length controlled by α. The chosen fine-tuning surrogate and the optimizer determine the direction. The step length, however, is usually specified by a learning-rate schedule or hyperparameter tuning [32]. In LLM fine-tuning, this scalar choice can be sensitive; excessively small steps may fail to fully exploit the information, while aggressive steps can destabilize training [51].

We improve the step-length decision while leaving the optimizer direction unchanged and formulate all objectives as maximization problems. Let $g_t \neq 0$, and define the normalized direction

$$z_t = \frac{g_t}{\|g_t\|}. \tag{1}$$

We restrict the current surrogate to the one-dimensional path along this direction:

$$\phi_t(s) = F_t(\theta_t + s z_t), \quad \text{for } s \in \mathbb{R}.$$

The standard optimizer step corresponds to $s_{\mathrm{FO}} = \alpha\|g_t\|$. Thus, choosing the learning rate is equivalent to choosing a scalar displacement along the fixed-batch improvement of ϕ_t.

Rather than relying only on a global learning-rate schedule, we ask whether the current surrogate F_t supports a better step along the direction already selected by the optimizer. We search over a bounded interval $\mathcal{I}_t = [0, R_t], \quad R_t = \beta\alpha\|g_t\|$, where $\beta > 0$ controls the search radius relative to the baseline optimizer step. When $\beta = 1$, the interval ends at the usual first-order update; larger values allow the method to consider longer steps while remaining constrained to the same direction:

$$s_t^\star \approx \arg\max_{s \in \mathcal{I}_t} \phi_t(s). \tag{2}$$

A full line search would evaluate ϕ_t at many candidate values of s, which is costly for LLMs. Instead, we use a fixed-budget step-selection procedure: the current value $\phi_t(0)$ is obtained from the usual forward evaluation of the surrogate on S_t, and the backward pass gives the directional derivative $\phi_t'(0) = \langle \nabla F_t(\theta_t), z_t \rangle$. We then evaluate the same fixed-batch surrogate at two symmetric perturbations along the update direction:

$$\phi_t(\epsilon) = F_t(\theta_t + \epsilon z_t), \quad \phi_t(-\epsilon) = F_t(\theta_t - \epsilon z_t), \text{ for a small } \epsilon > 0.$$

Together, these quantities $\{\phi_t(0), \phi_t'(0), \phi_t(\epsilon), \phi_t(-\epsilon)\}$, define a local model of the surrogate improvement along the optimizer direction. The selected step s_t is obtained by maximizing this model over $\mathcal{I}_t$, and the parameters are updated as $\theta_{t+1} = \theta_t + s_t z_t$. In this formulation, the base LLM fine-tuning algorithm determines the surrogate F_t and the update direction g_t, while our method determines how far to move along that direction; see Figure 1. The procedure is therefore compatible with a broad class of LLM fine-tuning objectives. Due to space limitations, we defer the related work to §A.

Our contributions. We develop ZFO as a general framework that separates first-order direction selection from low-query zeroth-order step selection. This formulation allows different local step-selection models to be developed and analyzed within the same optimization framework. Below, we summarize our main contributions:

(*i*) **A ZFO optimization framework.** We propose the general ZFO framework in which a standard first-order optimizer proposes an update direction, and a small number of zeroth-order evaluations are used to choose an improved step size along this direction. Within this framework, we develop four practical ZFO instances based on Taylor and Padé local models. Whereas GeN represents a specific step-selection rule, ZFO provides a more general fixed-budget directional modeling framework. GeN can be interpreted within the ZFO framework as a particular fixed-budget directional step-selection rule.

(*ii*) **Theoretical justification of adaptive step selection.** Our theoretical analysis of the one-dimensional objective follows from restricting the training objective to the first-order search direction. We show that the finite-difference curvature estimates with common random numbers (CRNs) concentrate around their population counterparts with high probability. We further prove that maximizing the resulting local model yields a step whose objective value is near-optimal within a bounded search region. Finally, we establish a global stationarity guarantee showing that ZFO converges to a stationary neighborhood whose size is controlled by the accumulated step-selection error.

(*iii*) **Empirical validation across scales.** We evaluate ZFO on both controlled low-dimensional problems and LLM fine-tuning tasks. The experiments show that ZFO is robust to the choice of initial learning rate and can match or outperform standard first-order and zeroth-order baselines on most benchmarks. These results demonstrate that lightweight zeroth-order step adaptation can improve the stability and reliability of first-order fine-tuning without incurring the cost of full line search or extensive hyperparameter tuning.

2 Methodology

We fix an iteration t and suppress the subscript t for readability. We use this convention throughout the section.

Local derivative estimates. We recall that the forward-backward pass provides $\phi(0)$ and the directional derivative, $\phi'(0) = \langle \nabla F(\theta), z \rangle$. ZFO adds two symmetric fixed-batch evaluations,

$$\phi(\epsilon) = F(\theta + \epsilon z), \text{ and } \phi(-\epsilon) = F(\theta - \epsilon z).$$

We estimate the second directional derivative by the central finite difference

$$\widehat{\phi''}(0) = \frac{\phi(\epsilon) - 2\phi(0) + \phi(-\epsilon)}{\epsilon^2}.$$

Using the available first-order derivative, the same two probes also yield

$$\widehat{\phi'''}(0) = \frac{3}{\epsilon^3}\left(\phi(\epsilon) - \phi(-\epsilon) - 2\epsilon\phi'(0)\right).$$

For compactness, define: $a = \phi'(0), b = \widehat{\phi''}(0)$, and $c = \widehat{\phi'''}(0)$. ZFO uses these quantities to construct a local approximation of ϕ (see Figure 1), considering both polynomial models derived from Taylor expansions and rational models derived from Padé approximations.

(a) **Taylor models.** The second-order Taylor model is $T_2(s) = \phi(0) + as + \frac{1}{2}bs^2$. Since T_2 is quadratic, the maximizer is obtained by checking the endpoints of I and, when feasible, the stationary point $s = -a/b$. Similarly, the third-order Taylor model is $T_3(s) = \phi(0) + as + \frac{1}{2}bs^2 + \frac{1}{6}cs^3$. The candidate maximizers are the endpoints of I and the real roots inside I of $a + bs + \frac{1}{2}cs^2 = 0$.

(b) **Padé models.** The Padé variants approximate ϕ using rational functions. The second-order Padé model is

$$P_2(s) = \phi(0) + \frac{as}{1 + q_2 s}, \quad q_2 = -\frac{b}{2a},$$

Similarly, the third-order Padé model is

$$P_3(s) = \frac{\phi(0) + (a + \phi(0)q_3)s}{1 + q_3 s + r_3 s^2}, \text{ where } q_3 = \frac{\phi(0)c - 3ab}{3(2a^2 - \phi(0)b)}, \quad r_3 = \frac{3b^2 - 2ac}{6(2a^2 - \phi(0)b)}.$$

If a Padé denominator has a pole inside I, or if its coefficients are ill-defined, ZFO rejects the Padé step and falls back to the quadratic Taylor step. This *safeguard* avoids unstable rational extrapolation near singularities.

Model selection and update. Each ZFO variant corresponds to one model choice from $\{T_2, T_3, P_2, P_3\}$. For a chosen model M in this set, ZFO computes the step by solving

$$s_M \in \arg\max_{s \in I} M(s). \tag{3}$$

All ZFO variants use the same direction, surrogate, and feasible interval. They differ only in how they convert two additional evaluations into a scalar step length. GeN [9] also fits within this abstraction as a specific quadratic ZFO instance. GeN uses two function evaluations along the first-order direction to construct a generalized Newton step. Our instances reuse the directional derivative already available from the backward pass so that the same two probes recover both second- and third-order directional

Algorithm 1 Zero-and-First-Order (ZFO) Framework

Require: θ_0, base optimizer, learning rate α, $\epsilon > 0$, $\beta > 0$, model $M \in \{T_2, T_3, P_2, P_3\}$

for $t = 0, 1, 2, \ldots$ **do**
  Compute the direction g_t using the base optimizer ▷ First-order direction
  Set $z_t = g_t / \|g_t\|$ and $I_t = [0, \beta\alpha\|g_t\|]$
  Compute: $a_t = \phi_t'(0), b_t = \frac{\phi_t(\epsilon) - 2\phi_t(0) + \phi_t(-\epsilon)}{\epsilon^2}, c_t = \frac{3}{\epsilon^3}\left(\phi_t(\epsilon) - \phi_t(-\epsilon) - 2\epsilon a_t\right)$
  Construct the chosen local model M using a_t, b_t, c_t
  if M is a Padé model and has a pole in I_t **then**
    Replace M by T_2
  end if
  Select $s_t = s_M \in \arg\max M(s)$, s.t. $s \in I_t$ ▷ Zeroth-order step selection
  Update $\theta_{t+1} = \theta_t + s_t z_t$ ▷ Parameter update
end for

information and support Taylor and Padé models. GeN uses curvature-based rejection as its safeguard, while the instances studied here optimize over a bounded interval and fall back when a Padé model is ill-defined. ZFO's novelty is the fixed-budget local-model abstraction, its family of Taylor/Padé instantiations, and the associated unified analysis and empirical validation. Algorithm 1 summarizes our framework.

3 Theoretical Analysis

This section provides a theoretical justification for the ZFO step-selection framework. The analysis has three components. First, we show that shared-batch evaluations yield reliable finite-difference estimates of directional curvature; see §3.1 (*i*). Second, conditional on the sampled examples, we show that maximizing the local ZFO model returns a step whose objective value is close to the best value attainable over the search interval; see §3.1 (*ii*). Third, we connect this local step-selection property to a global stationarity guarantee for the ZFO update sequence; see §3.2. We omit the iteration index t when no ambiguity arises.

3.1 Local Convergence Analysis

Shared-batch common-random-numbers (CRN). ZFO evaluates all local objective values on the same minibatch. At a fixed iteration, the examples are sampled once and then reused across the evaluations $\{\phi(-\epsilon), \phi(0), \phi(\epsilon)\}$. After conditioning on these shared examples, ϕ becomes a deterministic function of the scalar step length, s. This conditional determinism is the basis of the local approximation analysis. Let $\tau = \{\tau_1, \ldots, \tau_n\}$ denote the shared sampled examples used under CRN. For each example τ_i, define the per-sample directional objective $\phi_i(s) = F(\theta + sz; \tau_i)$. The empirical shared-batch CRN objective is $\phi(s) = \frac{1}{n}\sum_{i=1}^{n} \phi_i(s)$.

Definition 1 (Population directional objective)**.** *The population counterpart of the empirical CRN objective is $\psi(s) = \mathbb{E}_\tau[\phi_i(s)]$, where the expectation is taken with respect to the distribution used to generate the empirical objective.*

(*i*) **Local smoothness assumptions.** The first group of assumptions ensures that the directional objective admits Taylor expansions with controlled remainders.

Assumption 1 (Pathwise smoothness)**.** *For each example τ_i, the per-sample directional objective ϕ_i is five times continuously differentiable on the interval $[-\epsilon, R]$: $\phi_i \in C^5([-\epsilon, R])$.*

Assumption 2 (Higher-order derivative bounds)**.** *There exist constants $W, M, B, L > 0$ such that, for every i and every $s \in [-\epsilon, R]$, $|\phi_i''(s)| \le W$, $|\phi_i'''(s)| \le M$, $|\phi_i^{(4)}(s)| \le B$, and $|\phi_i^{(5)}(s)| \le L$.*

(*ii*) **Sampling assumptions.** Population concentration is analyzed at a fixed iterate and direction under the independent-example assumptions below. Algorithm 1 reuses its batch across probes, but also computes the direction from that batch; CRN alone does not establish the independence needed by this concentration argument. Thus Theorem 1 is a fixed-direction result, whereas the approximation guarantee in Theorem 2 is pathwise once the batch and direction are fixed. Bounded directional derivatives imply the sub-Gaussian condition by Hoeffding's lemma; see (4).

Assumption 3 (CRN)**.** *For each index i, the same example τ_i is used to evaluate $\{\phi_i(-\epsilon), \phi_i(0), \phi_i(\epsilon)\}$. Moreover, for $i \neq j$, the examples τ_i and τ_j are sampled independently.*

Assumption 4. *(Sub-Gaussian Deviation) For each ϕ_i, the directional deviations at $s = 0$ are sub-Gaussian up to third order, i.e., there exists a constant σ such that, for every $k \in \{0, 1, 2, 3\}$ and all $\lambda \in \mathbb{R}$, $\mathbb{E}[\exp(\lambda[\phi_i^{(k)}(0) - \psi^{(k)}(0)])] \leq \exp(\frac{\lambda^2\sigma^2}{2})$.*

Theorem 1 (CRN curvature concentration)**.** *Suppose Assumptions 1–4 hold. Then, with probability at least $1 - \delta$, $\left|\widehat{\phi''}(0) - \psi''(0)\right| \leq \frac{B\epsilon^2}{12} + \sigma\sqrt{\frac{2\log(4/\delta)}{n}}$, and $\left|\widehat{\phi'''}(0) - \psi'''(0)\right| \leq \frac{L\epsilon^2}{20} + \sigma\sqrt{\frac{2\log(4/\delta)}{n}}$.*

Theorem 1 shows that the finite-difference curvature estimates used by ZFO concentrate around their population counterparts. Each bound contains two terms. The first term is a deterministic finite-difference bias controlled by the perturbation scale ϵ and the relevant higher-order derivative bound. The second term is a stochastic sampling term that decreases at the classic rate $1/\sqrt{n}$. Thus, under CRN, the two additional objective evaluations provide statistically meaningful curvature information.

After conditioning on the shared examples, the empirical objective ϕ is deterministic. Let $s_M \in \arg\max_{s\in[0,R]} T_2(s)$ denote the model-selected step by T_2, and let $s_\phi \in \arg\max_{s\in[0,R]} \phi(s)$ denote the best step for the empirical CRN objective over the same interval.

Theorem 2 (Second-order Taylor approximation bound)**.** *Suppose Assumptions 1 and 2 hold. Then the step selected by maximizing the second-order Taylor model satisfies $\phi(s_M) \geq \phi(s_\phi) - \frac{MR^3}{3} - \frac{B\epsilon^2R^2}{12}$.*

Theorem 2 shows that, once the examples are fixed, ZFO selects a step that is close to the best step available in the search interval. The first error term, $\frac{MR^3}{3}$, is the deterministic Taylor approximation error over the interval. The second error term, $\frac{B\epsilon^2R^2}{12}$, is the error caused by estimating curvature through finite differences. The guarantee is strongest when the search radius R is moderate, the perturbation scale ϵ is small, and the directional objective has limited higher-order variation. The same proof strategy extends to the third-order Taylor model. For the Padé variants, the local model is rational rather than polynomial. These models can better capture nonlinear curvature changes over a finite interval. Their approximation guarantees follow from the same maximizer-comparison argument, replacing the Taylor remainder by a rational residual term. We defer the analysis of these variants to §B.1.

3.2 Global Convergence Analysis

We now connect the local analysis to the true optimization objective J. Recall that ZFO optimizes on the surrogate F_t, whose expectation under the current sampling distribution is $\bar{F}_t(\theta + sz_t) = \psi(s)$. For instance, standard policy optimization [53] and its modern extensions [45, 46] construct surrogates such that $\nabla J(\theta_t) = \nabla\bar{F}(\theta_t)$ at the current iterate. Define the expected directional alignment quantity $q_t = \mathbb{E}_t[\langle\nabla J(\theta_t), z_t\rangle]$. We start by stating two main classical assumptions that we use in the sequel.

Assumption 5 (Smooth objective)**.** *The true objective $J : \mathbb{R}^d \to \mathbb{R}$ has an $\mathcal{L}_J$-Lipschitz gradient and is bounded above by $J^\star$. Consequently, for all x, y, $J(y) \geq J(x) + \langle\nabla J(x), y - x\rangle - \frac{\mathcal{L}_J}{2}\|y - x\|^2$, and $J(\theta_t) \leq J^\star$ for all iterations t. Additionally, the expected surrogate $\bar{F}$ is also $\mathcal{L}_{\bar{F}}$-smooth.*

Assumption 6 (Aligned update direction)**.** *There exists $\mu > 0$ such that, for all t, $q_t = \mathbb{E}_t[\langle\nabla J(\theta_t), z_t\rangle] \geq \mu\|\nabla J(\theta_t)\|$.*

Assumption 6 is natural since gradient methods estimate an ascent direction for the objective [46]. Although optimizers such as AdamW rescale this direction, the resulting update remains positively oriented with $\nabla J(\theta_t)$ in expectation [3], supporting the assumption of $\mu > 0$.

We begin with a lemma that bounds the step-selection error at iteration t of the objective: $\max_{s\in I_t} J(\theta_t + sz_t) - J(\theta_t + s_t z_t)$. This quantity measures the loss in objective value by using the ZFO step s_t instead of the best step available in the interval I_t for the true objective J. We assume that the value and slope concentration bounds hold jointly for the batch-selected direction z_t, conditional on the history before iteration t, with probability at least $1 - \delta$, with $0 < \delta < 3\mu^2/4$.

Lemma 1 (Surrogate vs. Empirical Difference)**.** *Let Assumptions 1-5 hold, then $\max_{s\in I_t} J(\theta_t + sz_t) - J(\theta_t + s_t z_t) \leq \Delta_t$, holds with probability at least $1 - \delta$, where*

$$\Delta_t \coloneqq (\mathcal{L}_J + \mathcal{L}_{\bar{F}})R_t^2 + 2\sigma(1 + R_t)\sqrt{\frac{2\log(4/\delta)}{n}} + 2WR_t^2 + \phi(s_\phi) - \phi(s_M).$$

Table 1: **Evaluation performance for baseline** first-order (FO) fine-tuning (using AdamW) and zeroth-order (ZO) fine-tuning (using MeZO) against ZFO methods across models and reasoning benchmarks. All runs report the final evaluation and mean results over 3 seeds. Green is the best and red is the second best result.

Model	Dataset	Baseline		ZFO Methods (This Paper)			
		AdamW (FO)	MeZO (ZO)	Taylor_2	Taylor_3	Padé_2	Padé_3
Qwen-2.5-Math-1.5B	GSM8K	78.24 ± 0.59	0.00	81.00 ± 0.11	75.56 ± 0.24	80.95 ± 0.49	80.97 ± 0.73
	MATH	31.77 ± 1.63	22.01 ± 3.93	40.10 ± 1.19	20.18 ± 3.03	37.50 ± 1.79	39.58 ± 0.45
	SVAMP	88.11 ± 0.69	68.67 ± 10.33	90.78 ± 0.96	89.11 ± 0.51	90.89 ± 1.02	90.33 ± 2.19
	AsDiv	91.45 ± 0.36	64.72 ± 3.51	94.09 ± 0.29	88.27 ± 0.35	93.04 ± 0.22	93.46 ± 0.18
	OpenBookQA	26.07 ± 2.53	8.00 ± 10.05	37.80 ± 17.85	39.73 ± 20.99	60.67 ± 0.95	64.07 ± 1.22
Phi-2	SVAMP	55.44 ± 4.91	22.44 ± 3.89	55.67 ± 7.86	39.33 ± 0.88	64.22 ± 0.19	65.11 ± 5.39
	AsDiv	56.51 ± 18.76	10.15 ± 4.88	65.42 ± 2.59	55.85 ± 2.55	65.58 ± 3.74	69.27 ± 1.92
	OpenBookQA	51.20 ± 1.11	20.33 ± 10.90	58.27 ± 13.62	51.27 ± 0.95	49.40 ± 0.20	52.07 ± 0.23
Gemma-2-2B	SVAMP	52.00 ± 1.76	14.44 ± 5.93	59.56 ± 0.51	47.22 ± 2.01	59.00 ± 1.76	58.22 ± 3.69
Llama-3.2-1B	GSM8K	4.85 ± 5.47	1.30 ± 0.98	4.02 ± 3.98	9.00 ± 6.68	5.05 ± 6.94	1.72 ± 0.88
	AsDiv	16.60 ± 28.12	3.33 ± 0.48	13.72 ± 23.17	44.77 ± 1.42	25.23 ± 21.97	52.16 ± 4.75
	OpenBookQA	25.53 ± 0.31	0.00	34.40 ± 1.56	31.53 ± 0.81	34.60 ± 1.04	25.64 ± 0.23

This lemma quantifies the loss in objective value when the ZFO step s_t, selected through the local model, is used instead of the best step in the interval I_t. The error term Δ_t combines smoothness mismatch, finite-sample concentration error, approximation error, and the model-selection gap $\phi(s_\phi) - \phi(s_M)$. E.g., when the local model is the second-order Taylor model T_2, using Theorem 2 we get

$$\Delta_t \le (\mathcal{L}_J + \mathcal{L}_{\bar{F}})R_t^2 + 2\sigma(1+R_t)\sqrt{\frac{2\log(4/\delta)}{n}} + 2WR_t^2 + \frac{MR^3}{3} + \frac{B\epsilon^2 R^2}{12}.$$

Finally, Theorem 3 converts the local quality of the ZFO step-selection rule into a global stationarity guarantee.

Theorem 3 (Stationarity of the ZFO update)**.** *Let Assumptions 1–6 hold. Let* $\mathcal{K} := \mathcal{L}_J + \mathcal{L}_{\bar{F}} + 2W$ *and* $\chi := 3\mu^2 - 4\delta > 0$*, with* $q_t/(2\mathcal{K}) \in I_t$ *almost surely. Then, for every* $T > 0$*, we have*

$$\min_{0\le t<T} \mathbb{E}\|\nabla J(\theta_t)\|^2 \le \frac{8\mathcal{K}(J^\star - J(\theta_0))}{\chi T} + \frac{8\mathcal{K}\tilde{\Delta}_T}{\chi}, \text{ where } \tilde{\Delta}_T = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\Delta_t].$$

The upper bound in Theorem 3 has two terms. The first term, $\frac{8\mathcal{K}(J^\star - J(\theta_0))}{\chi T}$, is the usual optimization term and decays at rate $1/T$. The second term, $\frac{8\mathcal{K}\tilde{\Delta}_T}{\chi}$, is the price paid for imperfect step selection. Thus, ZFO does not necessarily converge exactly to a stationary point unless the average step-selection error vanishes. Instead, it converges to a neighborhood of stationarity controlled by the error $\tilde{\Delta}_T$.

The three results establish the theoretical foundation of ZFO. The CRN concentration theorem establishes the statistical reliability of the finite-difference curvature estimates; the deterministic approximation theorem shows that maximizing the ZFO model yields a near-optimal local step; and the global stationarity theorem shows that controlled step-selection error preserves convergence to stationarity. Together, these results justify the central mechanism of ZFO: a small number of fixed-batch objective evaluations can recover useful local curvature information and improve step selection over a fixed first-order update while retaining a principled convergence interpretation.

4 Experiments

Organization of the Results. We organize our experimental results as follows: In the *main paper*, (*i*) we compare ZFO against first-order fine-tuning with AdamW and zeroth-order fine-tuning with MeZO across several LLMs and datasets in Table 1; (*ii*) we conduct a focused Qwen-2.5-Math-1.5B sub-study over additional reasoning benchmarks in Table 2; (*iii*) we evaluate ZFO on classical optimization problems to assess its benefits beyond LLM fine-tuning in Figure 2; (*iv*) we summarize the ablation studies in the appendix; and (*v*) we report wall-clock and memory overheads in Figures 3 and 4 to quantify the computational cost of ZFO. We defer the remaining experiments and expanded

Table 2: **Qwen-2.5-Math-1.5B study** on ZFO methods in our settings via additional reasoning benchmarks. Results report evaluation performance averaged over 3 seeds for FO and ZFO methods.

Method	ARC-Challenge	StrategyQA	FOLIO	CODAH
AdamW (FO)	49.29 ± 16.08	68.56 ± 0.91	34.48 ± 1.13	57.91 ± 3.24
Taylor_2	58.10 ± 1.33	69.63 ± 1.82	36.60 ± 3.26	62.83 ± 3.14
Taylor_3	54.24 ± 0.94	61.67 ± 1.46	34.48 ± 1.13	55.28 ± 1.66
Padé_2	53.58 ± 2.77	65.55 ± 4.43	36.27 ± 2.73	53.96 ± 17.72
Padé_3	57.05 ± 1.24	70.06 ± 0.61	36.44 ± 4.53	41.85 ± 20.58

diagnostics to *Appendix* C.2, where (*vi*) we provide the complete six-problem classical optimization study in Table 8 and Figure 5, (*vii*) we validate the local ZFO approximants on analytical functions in Figure 6 and the step-selection mechanism on controlled toy RLVR environments in Figure 7, (*viii*) we provide learning-rate sensitivity curves in Figure 8 and study the interaction between α and β in Figure 9, including the complementary $\beta = 1$ control in Table 12, (*ix*) we analyze the frequency of the steps selected by each ZFO method in Table 14, (*x*) we summarize cross-setting improvements in Table 9 and compare against additional *first-order baselines* (Prodigy [37] and AdamW [31]) in Table 10, and (*xi*) we evaluate the sequential `ZFO-Seq` against line-search baselines (Armijo [39], BB [4], Polyak [42], and PLS [33]) in Table 11. Unless otherwise stated, all reported results are averaged over three seeds. The LLM experiments use the implementation described in §C.1.1. Additional experimental configuration details are provided in §C.1.2.

RLVR, LLMs, and Datasets used for Fine-tuning. We report the performance of various high-dimensional models configured for an RLVR setup using the GRPO policy gradient [47] and the AdamW optimizer [31], including Qwen-2.5-MATH-1.5B [54], Phi-2 [24], Gemma-2-2B [52], and Llama-3.2-1B [19] on the datasets GSM8K [12], SVAMP [41], AsDiv [35], OpenBookQA [36], and a (512/256 train/evaluation) subset of MATH [21]; Table 3 summarizes the model and dataset details.

Learning-rate selection. For a fair comparison, we fix the learning rate α to the value that performs the best for the FO baseline, chosen from the set $\{10^{-6}, 5 \times 10^{-6}, 10^{-5}\}$, and consequently select the best bound β from $\{3, 5, 10\}$; the remainder of the configuration details are in §C.1.2.

4.1 ZFO for LLM fine-tuning and classic optimization

(*i*) **Does directional ZO step selection improve FO optimization?** We show the performance of the ZFO framework on LLM fine-tuning in Table 1, where at least one ZFO variant improves on the first-order fine-tuning (FO FT) baseline for every model-dataset pair, indicating that ZFO's benefits are not confined to a single model family or benchmark. We observe the largest gains where the FO baseline is relatively weak or highly variable, including SVAMP for Phi-2 and Gemma-2-2B, as well as AsDiv for Llama and MATH and OpenBookQA for Qwen. When the FO baseline is already strong, such as Qwen on SVAMP, ZFO yields smaller but still positive improvements in mean performance. Although the best-performing ZFO variant changes across tasks and models, Taylor_2 provides a strong representative instance of the framework, improving over FO in 14 of the 16 LLM settings summarized in Table 9. Taylor_2 is a strong representative default in our experiments, although the preferred local model remains task-dependent.

(*ii*) **Qwen Sub-Study.** Table 2 extends the main results by isolating Qwen-2.5-Math-1.5B on four reasoning benchmarks: ARC-Challenge [11], StrategyQA [15], FOLIO [20], and CODAH [10]. At least one ZFO method outperforms the FT baseline on every dataset, with the largest improvements on ARC-Challenge ($\uparrow$8.81 over FO FT) and CODAH ($\uparrow$4.92 over FO FT). Even under the best learning rate configurations, the main results and the Qwen sub-study indicate that FO FT exhibits seed variance under our settings.

(*iii*) **Classic Optimization.** To show that ZFO is not intrinsically tied to RLVR and to examine ZFO's performance beyond LLM fine-tuning, Figure 2 presents three classic optimization problems; see Table 8 for the comprehensive summary of the six classic optimization problems. At least one ZFO variant achieves a substantially lower median objective than its corresponding FO baseline on every displayed problem after 100 updates. Taylor_3–AdamW yields the strongest results on nonlinear least squares and Beale, while Padé_3–AdamW performs best on Rosenbrock. Consistent with our

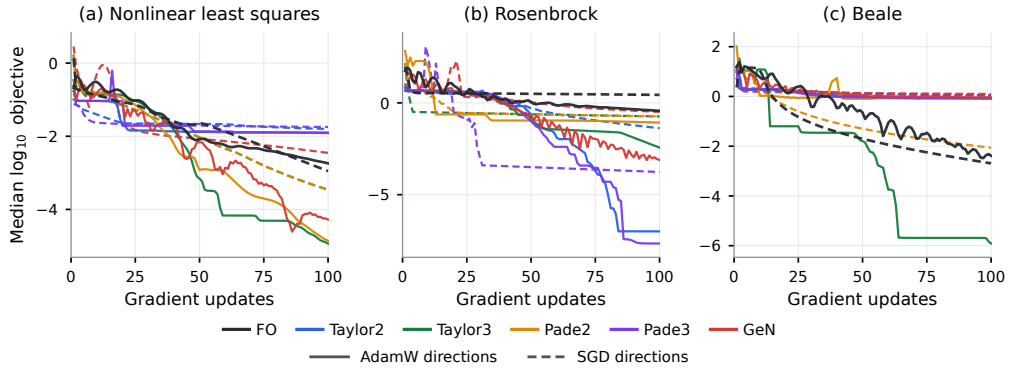

Figure 2: **Results on classic optimization problems.** We plot the median $\log_{10}$ objective over three perturbed initializations. GeN has a similar computational budget as ZFO, as both methods perform two additional function evaluations per iteration. We study the extended classic optimization problems in Table 8 and Figure 5.

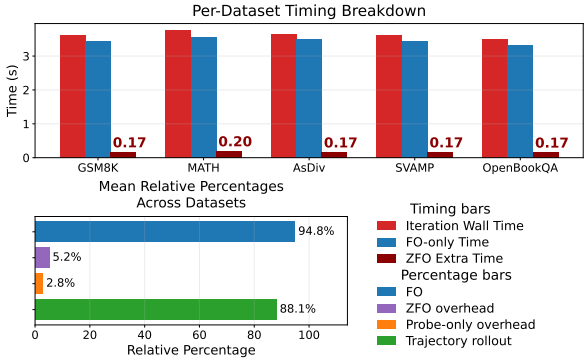

Figure 3: **Wall-clock time comparison.** ZFO step selection adds little wall-clock overhead. Across five benchmarks, trajectory rollout dominates iteration time, while the additional ZFO computation contributes only a small fraction of the time that FO-only updates cost.

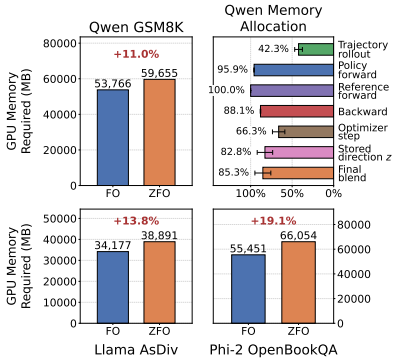

Figure 4: **GPU memory requirement comparisons.** Comparison of FO FT and ZFO FT across three configurations. The top right panel depicts the relative memory allocation for FO and ZFO FT.

LLM experiments, the best approximant varies across problems, indicating that the gains from step selection depend on the local objective.

(*iv*) **Ablation Studies.** We provide additional ablation studies in Appendix C.2 to better understand when and how ZFO improves step selection. Figure 8 compares ZFO against AdamW across fixed learning rates, showing that gains in Table 1 cannot generally be recovered by simply changing the FO learning rate. Figure 9 further studies the interaction between the learning rate α and search bound β. Increasing β is generally beneficial at smaller learning rates but can become unstable at larger learning rates, consistent with the dependence of the search radius on the product $\alpha\beta$. Table 14 shows that the four ZFO instances use the search interval differently, ranging from primarily interior steps for Taylor$_3$ to mostly endpoint decisions for Padé$_2$.

4.2 Computational Cost

The relative computational overhead of ZFO is small. A standard iteration involves four key components: a rollout call for the trajectory τ_t, a reward evaluation $R(\tau_t)$, a forward pass, and a backward pass. In contrast, our ZFO methods leave all four operations untouched and only call two additional forward-pass probes. Since these probes are only forward passes, the overhead is small and comparable to the forward-only cost structure of ZO methods such as MeZO [34]. From this perspective, we present both wall-clock time costs in Figure 3 and memory overhead costs in Figure 4. We train the models on an NVIDIA H100 80GB GPU.

Figure 3 decomposes per-iteration runtime for Qwen-2.5-MATH-1.5B on the benchmarks reported in Table 1. ZFO adds $\sim 5.2\%$ relative wall-clock overhead in this configuration. The corresponding GPU memory costs are examined below.

Figure 4 reports peak GPU memory across three model–dataset pairs, alongside a per-phase breakdown for Qwen-2.5-MATH-1.5B. The overhead ranges from 11.0% (Qwen/GSM8K) to 19.1% (Phi-2/OpenBookQA). We attribute this range to configuration differences; ZFO probes are absorbed with a smaller relative increase, whereas configurations with a small fixed group size and a larger batch see the overhead grow more. The memory allocation panel shows that the dominant cost remains the reference forward pass, while storing the update direction and applying the final blend sit below the existing bottleneck; hence, ZFO does not introduce a new peak; it widens the existing one. We note that the memory-saving trick MeZO relies on does not transfer here, since our update direction is the FO gradient rather than a random direction.

5 Limitations, Discussion, and Conclusion

What can two probes recover? The backward pass already supplies the local slope; two symmetric probes add estimates of curvature and its variation. Theorem 2 translates local-model error into a bound on the value lost relative to the best step in I. A wider interval permits larger updates but increases approximation error. Bounds and Padé fallbacks limit extrapolation and avoid poles, but do not guarantee that every step improves the surrogate. Theorem 3 then connects step-selection accuracy to convergence. Under the stated assumptions, ZFO therefore retains a stationarity guarantee even with approximate steps.

Padé$_2$ Behavior. $\texttt{Padé}_2$ provides a useful diagnostic because, as shown in the toy validation in §C, it behaves the most like FO since its updates select the endpoint of the search region unless a pole is detected, in which case it falls back to $\texttt{Taylor}_2$. This makes $\texttt{Padé}_2$ a conservative reference for distinguishing whether ZFO gains come from expanding the search region alone or intelligent step selection. In Tables 1 and 2, the richer surrogates frequently outperform $\texttt{Padé}_2$ (e.g., Qwen on MATH, ARC-Challenge, and CODAH). Conversely, when $\texttt{Padé}_2$ improves over FO, the expanded search region contributes in beneficial ways that are not obtainable simply by scaling α.

Beyond the Four Instances. We view $\texttt{Taylor}_2$, $\texttt{Taylor}_3$, $\texttt{Padé}_2$, and $\texttt{Padé}_3$ as initial ZFO instances rather than the strongest possible realizations of the framework. Each method is a simple analytic model built from lightweight probes, yet these lightweight probes already yield measurable gains in evaluation performance and training stability across several benchmarks. The main opportunity ZFO provides is not limited to the specific empirical gains reported in this paper, but instead lies in the observation that a meaningful curvature signal can be extracted along the FO direction at low additional cost.

Conclusion. Step-size selection is a classic problem in large-scale optimization. The direction selected by a first-order optimizer is valuable, but the step length remains uncertain. We show that two additional function evaluations along that direction are sufficient to construct useful local models of the one-dimensional objective. This creates a general fixed-query step-selection framework, ZFO. In contrast to ZO methods, ZFO does not estimate a gradient with zeroth-order queries; it estimates how the objective changes along a gradient direction whose orientation has already been supplied by a first-order optimizer.

Existing methods have already explored function-based step selection along a first-order direction, most notably GeN. Our contribution elevates this idea from a single-step rule to a fixed-query directional modeling framework—the same two probes can support multiple local models, including higher-order Taylor and rational Padé approximations, while the analysis separates finite-difference estimation error, local-model approximation error, and global optimization error. Therefore, ZFO is not primarily a new Taylor or Padé step-size formula. It is a fixed-query framework for exploiting zeroth-order information along an already trusted first-order direction, together with a theory that explains the statistical and optimization error introduced by this separation. In that sense, GeN is a special case of the broader ZFO abstraction, rather than an unrelated baseline. In the same spirit, we encourage practitioners to add their favorite functional approximators to the ZFO framework.

Across our experiments, ZFO often improves step selection over the fixed FO update. The quality of the selected step remains dependent on the local model, search radius, and alignment of the underlying first-order direction.

Acknowledgment. During this work, Aritra Dutta was partially supported by the Florida Department of Health Grant AWD00007072.

References

[1] Shihab Ahmed, El Houcine Bergou, Yue Wang, and Aritra Dutta. Stabilizing policy gradient methods via reward profiling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 19560–19568, 2026.

[2] Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-order (non)-convex stochastic optimization via conditional gradient and gradient updates. *Advances in Neural Information Processing Systems*, 31, 2018.

[3] Lukas Balles and Philipp Hennig. Dissecting Adam: The Sign, Magnitude and Variance of Stochastic Gradients. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 404–413. PMLR, 2018.

[4] Jonathan Barzilai and Jonathan M Borwein. Two-point step size gradient methods. *IMA journal of numerical analysis*, 8(1):141–148, 1988.

[5] El Houcine Bergou, Eduard Gorbunov, and Peter Richtárik. Stochastic three points method for unconstrained smooth minimization. *SIAM Journal on Optimization*, 30(4):2726–2749, 2020.

[6] Adel Bibi, El Houcine Bergou, Ozan Sener, Bernard Ghanem, and Peter Richtárik. A stochastic derivative-free optimization method with importance sampling: Theory and learning to control. *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, 2020.

[7] Soumia Boucherouite, Grigory Malinovsky, Peter Richtárik, and El Houcine Bergou. Minibatch stochastic three points method for unconstrained smooth minimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(18), 2024.

[8] Richard P Brent. Algorithms for minimization without derivatives. *Prentice-Hall, Englewood Cliffs, New Jersey*, 1973.

[9] Zhiqi Bu and Shiyun Xu. Gradient descent with generalized newton's method. In *The Thirteenth International Conference on Learning Representations*, 2025.

[10] Michael Chen, Mike D'Arcy, Alisa Liu, Jared Fernandez, and Doug Downey. CODAH: An adversarially-authored question answering dataset for common sense. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 63–69, 2019.

[11] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

[12] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[13] Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021.

[14] Aritra Dutta, El Houcine Bergou, Soumia Boucherouite, Nicklas Werge, Melih Kandemir, and Xin Li. Demystifying the Myths and Legends of Nonconvex Convergence of SGD. *arXiv preprint arXiv:2310.12969*, 2023.

[15] Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021.

[16] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, pages 2341–2368, 2013.

[17] Eduard Gorbunov, Adel Bibi, Ozan Sener, El Houcine Bergou, and Peter Richtarik. A stochastic derivative free optimization method with momentum. In *International Conference on Learning Representations*, 2020.

[18] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch SGD: Training ImageNet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[19] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[20] Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, et al. Folio: Natural language reasoning with first-order logic. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22017–22031, 2024.

[21] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[22] 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.

[23] Kevin G Jamieson, Robert Nowak, and Ben Recht. Query complexity of derivative-free optimization. *Advances in neural information processing systems*, 25, 2012.

[24] Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Caio Cêsar Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, et al. Phi-2: The surprising power of small language models. *Microsoft Research Blog*, 2023.

[25] Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024.

[26] Bumsu Kim, Daniel McKenzie, Hanqin Cai, and Wotao Yin. Curvature-aware derivative-free optimization. *Journal of Scientific Computing*, 103(2):43, 2025.

[27] Yu-Ang Lee, Ching-Yun Ko, Pin-Yu Chen, and Mi-Yen Yeh. Learning rate matters: Vanilla loRA may suffice for LLM fine-tuning. In *ICML'26 workshop on CoLoRAI - The 2nd Workshop on Connecting Low-rank Representations in AI*, 2026.

[28] Zhe Li, Bicheng Ying, Zidong Liu, Chaosheng Dong, and Haibo Yang. Converge faster, talk less: Hessian-informed federated zeroth-order optimization. In *International Conference on Learning Representations*, pages 37066–37099, 2026.

[29] 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, pages 21139–21169, 2025.

[30] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

[31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[32] Dougal Maclaurin, David Duvenaud, and Ryan Adams. Gradient-based hyperparameter optimization through reversible learning. In *International conference on machine learning*, pages 2113–2122. PMLR, 2015.

[33] Maren Mahsereci and Philipp Hennig. Probabilistic line searches for stochastic optimization. *Journal of Machine Learning Research*, 18(119):1–59, 2017.

[34] Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D Lee, Danqi Chen, and Sanjeev Arora. Fine-tuning language models with just forward passes. *Advances in Neural Information Processing Systems*, 36:53038–53075, 2023.

[35] Shen-Yun Miao, Chao-Chun Liang, and Keh-Yih Su. A diverse corpus for evaluating and developing English math word problem solvers. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, pages 975–984, 2020.

[36] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2381–2391, 2018.

[37] Konstantin Mishchenko and Aaron Defazio. Prodigy: an expeditiously adaptive parameter-free learner. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

[38] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

[39] Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer, 2nd edition, 2006.

[40] Courtney Paquette and Katya Scheinberg. A stochastic line search method with expected complexity analysis. *SIAM Journal on Optimization*, 30(1):349–376, 2020.

[41] Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 2080–2094, 2021.

[42] Boris T. Polyak. Minimization of nonsmooth functionals. *USSR Computational Mathematics and Mathematical Physics*, 9(3):14–29, 1969.

[43] Junbin Qiu, Zhengpeng Xie, Xiangda Yan, Yongjie Yang, and Yao Shu. Zeroth-Order Optimization is Secretly Single-Step Policy Optimization. *arXiv preprint arXiv:2506.14460*, 2025.

[44] Ruizhong Qiu and Hanghang Tong. Gradient compressed sensing: A query-efficient gradient estimator for high-dimensional zeroth-order optimization. In *Proceedings of the 41st International Conference on Machine Learning*, pages 41717–41748, 2024.

[45] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR, 2015.

[46] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[47] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[48] Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pages 464–472. IEEE, 2017.

[49] Leslie N Smith. A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820*, 2018.

[50] Jacob Mitchell Springer, Sachin Goyal, Kaiyue Wen, Tanishq Kumar, Xiang Yue, Sadhika Malladi, Graham Neubig, and Aditi Raghunathan. Overtrained language models are harder to fine-tune. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pages 56719–56789, 2025.

[51] Joe Suk and Yaqi Duan. On the optimization dynamics of RLVR: Gradient gap and step size thresholds. *arXiv preprint arXiv:2510.08539*, 2025.

[52] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[53] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

[54] An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-Math Technical Report: Toward Mathematical Expert Model via Self-Improvement, 2024.

[55] Yihua Zhang, Pingzhi Li, Junyuan Hong, Jiaxiang Li, Yimeng Zhang, Wenqing Zheng, Pin-Yu Chen, Jason Lee, Wotao Yin, Mingyi Hong, et al. Revisiting Zeroth-Order Optimization for Memory-Efficient LLM Fine-Tuning: A Benchmark. In *International Conference on Machine Learning*, volume 235, pages 59173–59190, 2024.

[56] Yanjun Zhao, Sizhe Dang, Haishan Ye, Guang Dai, Yi Qian, and Ivor Tsang. Second-Order Fine-Tuning without Pain for LLMs: A Hessian-Informed Zeroth-Order Optimizer. In *The Thirteenth International Conference on Learning Representations*, 2025.

Contents

1 Introduction **1**

2 Methodology **4**

3 Theoretical Analysis **5**

3.1 Local Convergence Analysis 5

3.2 Global Convergence Analysis 6

4 Experiments **7**

4.1 ZFO for LLM fine-tuning and classic optimization 8

4.2 Computational Cost 9

5 Limitations, Discussion, and Conclusion **10**

A Related work **15**

B Addendum to Theoretical Analysis **16**

B.1 ZFO Maximum Error 18

B.2 ZFO Global Stationarity 22

B.3 Finite Difference Lemmas 25

C Addendum to Experimental Results **26**

C.1 Configuration 26

C.1.1 Surrogate Objective Formulation 27

C.1.2 Empirical Configurations 28

C.2 Additional Insights 29

D Notation **36**

Organization. We organize the Appendix as follows: In Section A, we detail the related work. In Section B, we quote the general assumptions and provide detailed proofs of our theoretical results in the main paper. Finally, in Section C, we discuss the configuration details for our experiments in §C.1 and provide additional performance insights in §C.2.

A Related work

We position ZFO within three lines of work that motivate or contrast with our approach. Together, these related works inspire ZFO's central premise that cheap query calls can help model local objective behavior along the gradient direction.

Policy Optimization Stability & Update Regulation. Stability in policy optimization has motivated a range of update-regulation strategies. In policy optimization, methods such as TRPO [45] enforce a KL-constrained trust region, while PPO [46] replaces explicit constraints with a clipped surrogate objective to stabilize updates. More recent approaches, such as GRPO [47], further emphasize that training stability is highly sensitive to the way policy updates are regulated. Related works also consider wrapper-based stabilization strategies for policy-gradient methods that regulate whether an update should be accepted [1]. Our work is motivated by the same stability concern; however,

we approach it from a different perspective by explicitly modeling the objective along the update direction path.

Zeroth-Order Optimization. Our work is also related to ZO optimization for fine-tuning large language models. Methods such as MeZO and CARS use central finite differences along random directions to approximate gradient or curvature behavior [26, 34]; particularly, MeZO uses finite differences along random directions to approximate the gradient, and CARS uses a second-order finite difference to approximate the Hessian. Extensions of MeZO include curvature, sparsity, and query-efficient estimators. Hi-ZOO incorporates a diagonal Hessian estimate at the cost of an extra forward pass [56]. Sparse-MeZO restricts updates to noise-resistant weights to enable larger step sizes [29]. GraCe leverages gradient sparsity to achieve a dependence on dimension in query complexity [44]. HiSo carries Hessian-informed ZO into the federated setting [28]. Complementary benchmarking has mapped the landscape of ZO optimizers across fine-tuning paradigms [55]. Recent theoretical work further establishes that ZO optimization is equivalent to single-step policy optimization [43], suggesting a natural connection between zeroth-order methods and RLVR that ZFO exploits.

Line-Search & Adaptive Step Size. Broadly, our method is similar to line-search and adaptive learning-rate methods, including Polyak step size [42], Barzilai-Borwein [4], and stochastic variants such as probabilistic line-search [33] and stochastic Armijo line-search [40]. These approaches adapt step sizes by repeatedly evaluating candidate points along a univariate objective, often requiring multiple queries. In contrast, our method uses only two symmetric probes to approximately recover higher-order directional information to construct a local model of the objective. This results in a fixed-budget alternative to classical line-search methods.

Generalized Newton Step Selection. Among closely related recent methods, generalized Newton step selection or GeN [9] has already explored function-based step selection along a first-order direction, providing an early example of combining FO directions with low-query step-size adaptation. Our contribution is to elevate this idea from a single-step rule to a fixed-query directional modeling framework: the same two probes can support multiple local models, including higher-order Taylor and rational Padé approximations, while the analysis separates finite-difference estimation error, local-model approximation error, and global optimization error. GeN is a special case of the broader ZFO abstraction, rather than an unrelated baseline.

B Addendum to Theoretical Analysis

This section provides the theoretical foundations needed to characterize when, why, and how ZFO can reliably improve step selection. We first establish concentration guarantees for the finite-difference curvature estimates, then bound how closely the resulting models recover the best step within their local interval, and we finally connect these local guarantees to global stationarity of the ZFO update. Together, these results separate the sources of error introduced by stochastic estimation, local function approximation, and step selection, providing a theoretical justification for the mechanisms used by ZFO.

To justify using finite-difference curvature information for ZFO step selection, we first ask whether the second- and third-order directional derivatives estimated from a shared batch can reliably approximate their population counterparts.

Hoeffding's Inequality. Let X be a random variable such that $a \leq X \leq b$ almost surely. Then Hoeffding's lemma gives

$$\mathbb{E}\left[\exp\left(\lambda(X - \mathbb{E}[X])\right)\right] \leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right), \qquad \text{for all } \lambda \in \mathbb{R}. \tag{4}$$

Thus, every bounded random variable is sub-Gaussian with parameter $\sigma = (b-a)/2$.

Let $X_1, \ldots, X_n$ be independent sub-Gaussian random variables with parameter $\sigma > 0$, i.e.,

$$\mathbb{E}\left[e^{\lambda X_i}\right] \leq e^{\lambda^2\sigma^2/2}, \quad \forall \lambda \in \mathbb{R}.$$

Then the sub-Gaussian concentration inequality gives:

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \leq \sigma\sqrt{\frac{2\log(2/\zeta)}{n}}\right) \geq 1 - \zeta. \tag{5}$$

Theorem 1 (CRN curvature concentration)**.** *Suppose Assumptions 1–4 hold. Then, with probability at least* $1-\delta$, $\left|\widehat{\phi''}(0)-\psi''(0)\right| \le \frac{B\epsilon^2}{12}+\sigma\sqrt{\frac{2\log(4/\delta)}{n}}$, *and* $\left|\widehat{\phi'''}(0)-\psi'''(0)\right| \le \frac{L\epsilon^2}{20}+\sigma\sqrt{\frac{2\log(4/\delta)}{n}}$.

Proof. Fix an index $i \in \{1,\dots,n\}$. By Assumption 1 and the second-order central finite-difference expansion, there exists $\hat{\xi}_i \in (-\epsilon,\epsilon)$ such that

$$\frac{\phi_i(\epsilon)-2\phi_i(0)+\phi_i(-\epsilon)}{\epsilon^2} = \phi_i''(0)+\frac{\epsilon^2}{12}\phi_i^{(4)}(\hat{\xi}_i).$$

Using Assumption 2, the remainder satisfies

$$\left|\frac{\epsilon^2}{12}\phi_i^{(4)}(\hat{\xi}_i)\right| \le \frac{B\epsilon^2}{12}.$$

Averaging the above identity over the shared batch gives

$$\widehat{\phi''}(0) = \frac{1}{n}\sum_{i=1}^n \phi_i''(0)+\frac{1}{n}\sum_{i=1}^n \frac{\epsilon^2}{12}\phi_i^{(4)}(\hat{\xi}_i).$$

Therefore, by the triangle inequality,

$$\left|\widehat{\phi''}(0)-\psi''(0)\right| \le \left|\frac{1}{n}\sum_{i=1}^n (\phi_i''(0)-\psi''(0))\right| + \frac{B\epsilon^2}{12}.$$

By Assumptions 3 and 4, the random variables $\phi_i''(0)-\psi''(0), \quad i=1,\dots,n$, are independent, centered, and sub-Gaussian with parameter σ. Applying Inequality 5 with $\zeta=\delta/2$ yields, with probability at least $1-\delta/2$,

$$\left|\frac{1}{n}\sum_{i=1}^n (\phi_i''(0)-\psi''(0))\right| \le \sigma\sqrt{\frac{2\log(4/\delta)}{n}}.$$

Consequently,

$$\left|\widehat{\phi''}(0)-\psi''(0)\right| \le \frac{B\epsilon^2}{12}+\sigma\sqrt{\frac{2\log(4/\delta)}{n}}.$$

We now repeat the same argument for the third derivative estimate. By the third-order finite-difference expansion, for each i there exists $\tilde{\xi}_i \in (-\epsilon,\epsilon)$ such that

$$\frac{3\left(\phi_i(\epsilon)-\phi_i(-\epsilon)-2\epsilon\phi_i'(0)\right)}{\epsilon^3} = \phi_i'''(0)+\frac{\epsilon^2}{20}\phi_i^{(5)}(\tilde{\xi}_i).$$

Assumption 2 gives

$$\left|\frac{\epsilon^2}{20}\phi_i^{(5)}(\tilde{\xi}_i)\right| \le \frac{L\epsilon^2}{20}.$$

Averaging over the shared batch and applying the triangle inequality gives

$$\left|\widehat{\phi'''}(0)-\psi'''(0)\right| \le \left|\frac{1}{n}\sum_{i=1}^n (\phi_i'''(0)-\psi'''(0))\right| + \frac{L\epsilon^2}{20}.$$

Again, by Assumptions 3 and 4, the centered variables $\phi_i'''(0)-\psi'''(0), \quad i=1,\dots,n$, are independent and sub-Gaussian with parameter σ. Applying Inequality 5 with $\zeta=\delta/2$ yields, with probability at least $1-\delta/2$,

$$\left|\frac{1}{n}\sum_{i=1}^n (\phi_i'''(0)-\psi'''(0))\right| \le \sigma\sqrt{\frac{2\log(4/\delta)}{n}}.$$

Thus,

$$\left|\widehat{\phi'''}(0)-\psi'''(0)\right| \le \frac{L\epsilon^2}{20}+\sigma\sqrt{\frac{2\log(4/\delta)}{n}}.$$

Finally, taking a union bound over the two concentration events shows that both inequalities hold simultaneously with probability at least $1-\delta$. Therefore,

$$\left|\widehat{\phi''}(0)-\psi''(0)\right| \le \frac{B\epsilon^2}{12}+\sigma\sqrt{\frac{2\log(4/\delta)}{n}},$$

and

$$\left|\widehat{\phi'''}(0) - \psi'''(0)\right| \leq \frac{L\epsilon^2}{20} + \sigma\sqrt{\frac{2\log(4/\delta)}{n}}.$$

This completes the proof. □

B.1 ZFO Maximum Error

Theorem 2 (Second-order Taylor approximation bound)**.** *Suppose Assumptions 1 and 2 hold. Then the step selected by maximizing the second-order Taylor model satisfies* $\phi(s_M) \geq \phi(s_\phi) - \frac{MR^3}{3} - \frac{B\epsilon^2R^2}{12}$.

Proof. By Assumption 1 and Taylor's theorem, there exists ξ_s such that

$$\phi(s) = \phi(0) + s\phi'(0) + s^2\frac{\phi''(0)}{2} + s^3\frac{\phi^{(3)}(\xi_s)}{6}.$$

Similarly, for some $\hat{\xi}_s$, the approximation T_2 admits the expansion,

$$T_2(s) = \phi(0) + s\phi'(0) + s^2\frac{\phi''(0)}{2} + \epsilon^2 s^2\frac{\phi^{(4)}(\hat{\xi}_s)}{24}.$$

Subtracting these identities gives

$$\phi(s) - T_2(s) = s^3\frac{\phi^{(3)}(\xi_s)}{6} - \epsilon^2 s^2\frac{\phi^{(4)}(\hat{\xi}_s)}{24}.$$

Therefore, by the triangle inequality and Assumption 2 we have,

$$|\phi(s) - T_2(s)| \leq \left|s^3\frac{\phi^{(3)}(\xi_s)}{6}\right| + \left|\epsilon^2 s^2\frac{\phi^{(4)}(\hat{\xi}_s)}{24}\right| \leq \left|s^3\frac{M}{6}\right| + \left|\epsilon^2 s^2\frac{B}{24}\right| \leq \frac{MR^3}{6} + \frac{\epsilon^2 BR^2}{24}.$$

Consequently,

$$|\phi(s) - T_2(s)| \leq \frac{MR^3}{6} + \frac{\epsilon^2 BR^2}{24} =: \varphi.$$

We have

$$s_\phi \coloneqq \arg\max_{s\in[0,R]} \phi(s).$$

By the definitions of s_ϕ and s_M,

$$\phi(s_\phi) \geq \phi(s_M) \qquad \text{and} \qquad T_2(s_M) \geq T_2(s_\phi).$$

Chaining these inequalities with the uniform bound φ yields

$$\phi(s_M) \geq T_2(s_M) - \varphi \geq T_2(s_\phi) - \varphi \geq \phi(s_\phi) - 2\varphi.$$

Therefore,

$$\phi(s_M) \geq \phi(s_\phi) - \frac{MR^3}{3} - \frac{\epsilon^2 BR^2}{12}.$$

□

Additional Model Error Bounds. To understand how accurately ZFO's local models can choose a step, we next ask how much objective value may be lost when maximizing a finite-difference Taylor or Padé approximation. The proof technique used in Theorem 2 generalizes naturally to higher-order ZFO models. This allows us to characterize the approximation gap of $\texttt{Taylor}_3$ and Padé over the search region $[0, R]$. Corollary 1 establishes a similar bound for $\texttt{Taylor}_3$ as Theorem 2 did for $\texttt{Taylor}_2$. Corollaries 2 and 3 adapt a similar structure through the residual $r(s) \coloneqq D(s)\phi(s) - N(s)$. All three results yield finite-difference and model truncation errors.

Alternative Padé3 Formulation. The third-order Padé approximation is not restricted to the $[1/2]$ formulation considered in our empirical experiments ($P_3(s)$). An alternative $[2/1]$ formulation allocates the additional polynomial degree to the numerator rather than the denominator, yielding

$$P_{[2,1]}(s) = \phi(0) + \frac{as + \rho s^2}{1 + qs}, \qquad q = -\frac{c}{3b}, \qquad \rho = \frac{b}{2} + aq.$$

Both formulations recover the same third-order local Taylor expansion, provided their coefficients are well-defined, but differ in their extrapolation behavior and pole structure. The $[2/1]$ formulation therefore provides another possible realization of Padé3 within the ZFO framework.

Definition 2. *Let $P_k(s)$ be the k-th order Padé approximation of $\phi(s)$ over $[0, R]$, where $P_k(s) = \frac{N(s)}{D(s)}$, $N(s) = a_0 + a_1 s$, with $D(s) = 1 + d_1 s$ for $k = 2$ and $D(s) = 1 + d_1 s + d_2 s^2$ for $k = 3$,*

and $R = s_{\mathrm{FO}}\beta$, such that $\beta > 0$. The coefficients of $N(s)$ and $D(s)$ are chosen so that $P_k(s)$ agrees with the k-th order Taylor model $T_k(s)$ when Taylor expanded. Define the Padé residual $r(s) := D(s)\phi(s) - N(s)$.

When we use the Padé$_2$ and Padé$_3$ models we assume the following:

Assumption 7. *The Padé denominator is bounded away from zero on $[0, R]$, i.e., there exists $\eta > 0$ such that $|D(s)| \geq \eta, \forall s \in [0, R]$, for both Padé$_2$ and Padé$_3$.*

Assumption 8. *The Padé residual has bounded higher-order derivatives, i.e., there exist $Q, P > 0$ such that $|r^{(3)}(s)| \leq Q$ (Padé$_2$), $|r^{(4)}(s)| \leq P$ (Padé$_3$), $\forall s \in [0, R]$.*

Assumption 7 is consistent with our safeguard of `Padé`$_3$ falling back to `Taylor`$_2$ when encountering a pole.

Corollary 1 (Third-order Taylor approximation bound)**.** *Suppose Assumptions 1-2 hold. Then the step selected by maximizing the third-order Taylor model satisfies $\phi(s_M) \geq \phi(s_\phi) - \frac{BR^4}{12} - \frac{B\epsilon^2R^2}{12} - \frac{L\epsilon^2R^3}{60}$.*

Proof. By Assumption 1 and Taylor's theorem, there exists ξ_s such that

$$\phi(s) = \phi(0) + s\phi'(0) + s^2\frac{\phi''(0)}{2} + s^3\frac{\phi'''(0)}{6} + s^4\frac{\phi^{(4)}(\xi_s)}{24}.$$

Similarly, the approximation T_3 admits the expansion, for some $\hat{\xi}_s$ and $\tilde{\xi}_s$,

$$T_3(s) = \phi(0) + s\phi'(0) + s^2\frac{\phi''(0)}{2} + s^3\frac{\phi'''(0)}{6} + \epsilon^2 s^2\frac{\phi^{(4)}(\hat{\xi}_s)}{24} + \epsilon^2 s^3\frac{\phi^{(5)}(\tilde{\xi}_s)}{120}.$$

Subtracting these identities gives

$$\phi(s) - T_3(s) = s^4\frac{\phi^{(4)}(\xi_s)}{24} - \epsilon^2 s^2\frac{\phi^{(4)}(\hat{\xi}_s)}{24} - \epsilon^2 s^3\frac{\phi^{(5)}(\tilde{\xi}_s)}{120}.$$

Therefore, by the triangle inequality and Assumption 2,

$$|\phi(s)-T_3(s)| \leq \left|s^4\frac{\phi^{(4)}(\xi_s)}{24}\right| + \left|\epsilon^2 s^2\frac{\phi^{(4)}(\hat{\xi}_s)}{24}\right| + \left|\epsilon^2 s^3\frac{\phi^{(5)}(\tilde{\xi}_s)}{120}\right| \leq \left|s^4\frac{B}{24}\right| + \left|\epsilon^2 s^2\frac{B}{24}\right| + \left|\epsilon^2 s^3\frac{L}{120}\right|$$

$$\leq \frac{BR^4}{24} + \frac{\epsilon^2 BR^2}{24} + \frac{\epsilon^2 LR^3}{120}.$$

Consequently,

$$|\phi(s) - T_3(s)| \leq \frac{BR^4}{24} + \frac{\epsilon^2 BR^2}{24} + \frac{\epsilon^2 LR^3}{120} =: \varphi.$$

Define

$$s_\phi := \arg\max_{s\in[0,R]} \phi(s).$$

By the definitions of s_ϕ and s_M,

$$\phi(s_\phi) \geq \phi(s_M) \qquad \text{and} \qquad T_3(s_M) \geq T_3(s_\phi).$$

Chaining these inequalities with the uniform bound φ yields

$$\phi(s_M) \geq T_3(s_M) - \varphi \geq T_3(s_\phi) - \varphi \geq \phi(s_\phi) - 2\varphi.$$

Therefore,

$$\phi(s_M) \geq \phi(s_\phi) - \frac{BR^4}{12} - \frac{\epsilon^2 BR^2}{12} - \frac{\epsilon^2 LR^3}{60}.$$

□

Corollary 2 (Padé$_2$ approximation bound)**.** *Let Assumptions 1-2 and 7-8 hold. Then the step selected by maximizing the `Padé`$_2$ model satisfies $\phi(s_M) \geq \phi(s_\phi) - \frac{B\epsilon^2R^2}{12\eta} - \frac{QR^3}{3\eta}$.*

Proof. Since $P_2(s) = N(s)/D(s)$, we have

$$\phi(s) - P_2(s) = \frac{D(s)\phi(s) - N(s)}{D(s)} = \frac{r(s)}{D(s)}.$$

By construction, $P_2(s)$ agrees with the second-order Taylor model $T_2(s)$ when Taylor expanded. Since $\phi(0)$ and $\phi'(0)$ are exact, this gives

$$r(0) = 0, \qquad r'(0) = 0.$$

Notice that

$$r(s) = D(s)(\phi(s) - P_2(s)) = (1 + d_1 s)\left(\frac{\phi''(0) - \widehat{\phi''}(0)}{2} s^2 + \mathcal{O}(s^3)\right).$$

Therefore, the second-order coefficient satisfies

$$\frac{r''(0)}{2} = \frac{\phi''(0) - \widehat{\phi''}(0)}{2},$$

and using the finite-difference error from Lemma 2 yields

$$\left|\frac{r''(0)}{2}\right| \le \frac{B\epsilon^2}{24}.$$

Taylor expanding $r(s)$ gives

$$r(s) = \frac{s^2}{2} r''(0) + \frac{s^3}{6} r^{(3)}(\xi_s)$$

for some $\xi_s \in [0, s]$. Therefore, by Assumption 8,

$$|r(s)| \le \frac{B\epsilon^2 s^2}{24} + \frac{Q s^3}{6}.$$

Since $|D(s)| \ge \eta$ by Assumption 7,

$$|\phi(s) - P_2(s)| \le \frac{1}{\eta}\left[\frac{B\epsilon^2 R^2}{24} + \frac{QR^3}{6}\right] =: \varphi.$$

By the definition of $s_M = \arg\max_{s\in[0,R]} P_2(s)$,

$$P_2(s_M) \ge P_2(s_\phi).$$

Chaining this inequality with the uniform bound φ yields

$$\phi(s_M) \ge P_2(s_M) - \varphi \ge P_2(s_\phi) - \varphi \ge \phi(s_\phi) - 2\varphi.$$

Substituting the value of φ gives

$$\phi(s_M) \ge \phi(s_\phi) - \frac{B\epsilon^2 R^2}{12\eta} - \frac{QR^3}{3\eta}.$$

Hence, the result. □

Corollary 3 (Padé$_3$ approximation bound)**.** *Let Assumptions 1-2 and 7-8 hold. Then the step selected by maximizing the* `Padé`$_3$ *model satisfies* $\phi(s_M) \ge \phi(s_\phi) - \frac{B\epsilon^2 R^2}{12\eta} - \frac{L\epsilon^2 R^3}{60\eta} - \frac{|d_1| B\epsilon^2 R^3}{12\eta} - \frac{PR^4}{12\eta}$.

Proof. Since $P_3(s) = N(s)/D(s)$, we have

$$\phi(s) - P_3(s) = \frac{D(s)\phi(s) - N(s)}{D(s)} = \frac{r(s)}{D(s)}.$$

By construction, $P_3(s)$ agrees with the third-order Taylor model $T_3(s)$ when Taylor expanded. Since $\phi(0)$ and $\phi'(0)$ are exact, this gives

$$r(0) = 0, \qquad r'(0) = 0.$$

Notice that

$$r(s) = D(s)(\phi(s) - p_3(s)) = (1 + d_1 s + d_2 s^2)\left(\frac{\phi''(0) - \hat{\phi}''(0)}{2} s^2 + \frac{\phi'''(0) - \hat{\phi}'''(0)}{6} s^3 + \mathcal{O}(s^4)\right).$$

Therefore, the second-order coefficient satisfies

$$\frac{r''(0)}{2} = \frac{\phi''(0) - \hat{\phi}''(0)}{2},$$

and using the finite-difference error yields

$$\left|\frac{r''(0)}{2}\right| \le \frac{B\epsilon^2}{24}.$$

Similarly, the third-order coefficient satisfies

$$\frac{r'''(0)}{6} = \frac{\phi'''(0) - \hat{\phi}'''(0)}{6} + d_1 \frac{\phi''(0) - \hat{\phi}''(0)}{2}.$$

Consequently,

$$\left|\frac{r'''(0)}{6}\right| \le \frac{L\epsilon^2}{120} + |d_1| \frac{B\epsilon^2}{24}.$$

Taylor expanding $r(s)$ gives

$$r(s) = \frac{s^2}{2}r''(0) + \frac{s^3}{6}r'''(0) + \frac{s^4}{24}r^{(4)}(\xi_s)$$

for some $\xi_s \in [0, s]$. Therefore, by Assumption 8,

$$|r(s)| \leq \frac{B\epsilon^2 s^2}{24} + \left(\frac{L\epsilon^2}{120} + |d_1|\frac{B\epsilon^2}{24}\right)s^3 + \frac{Ps^4}{24}.$$

Since $|D(s)| \geq \eta$ by Assumption 7,

$$|\phi(s) - P_3(s)| \leq \frac{1}{\eta}\left[\frac{B\epsilon^2 R^2}{24} + \left(\frac{L\epsilon^2}{120} + |d_1|\frac{B\epsilon^2}{24}\right)R^3 + \frac{PR^4}{24}\right] =: \varphi.$$

By the definition of $s_M = \arg\max_{s\in[0,R]} P_3(s)$,

$$P_3(s_M) \geq P_3(s_\phi).$$

Chaining this inequality with the uniform bound φ yields

$$\phi(s_M) \geq P_3(s_M) - \varphi \geq P_3(s_\phi) - \varphi \geq \phi(s_\phi) - 2\varphi.$$

Substituting the value of φ gives

$$\phi(s_M) \geq \phi(s_\phi) - \frac{B\epsilon^2 R^2}{12\eta} - \frac{L\epsilon^2 R^3}{60\eta} - \frac{|d_1|B\epsilon^2 R^3}{12\eta} - \frac{PR^4}{12\eta}.$$

□

Together, these three corollaries and Theorem 2 quantify how the additional curvature information used by ZFO models affects the worst-case approximation gap. The $\texttt{Taylor}_3$ bound replaces the $\mathcal{O}(R^3)$ term in Theorem 2 with an $\mathcal{O}(R^4)$ term. Hence the relationship between the errors of $\texttt{Taylor}_2$ and $\texttt{Taylor}_3$ can be most significantly characterized by the power of their search regions R. The $\texttt{Padé}_2$ bound remains closer in order to the $\texttt{Taylor}_2$ bound, but introduces the residual-controlled $\frac{QR^3}{3\eta}$ term. The $\texttt{Padé}_3$ bound exhibits a similar $\mathcal{O}(R^4)$ term as $\texttt{Taylor}_3$ and also uses a residual error term. Overall, the two corollaries reinforce the same conclusion as Theorem 2; when the search region is moderate and the perturbation scale is small, then ZFO maximizers achieve values close to the best step within a local trust region.

ZFO Dominance. As a supplementary result beyond the main-text analysis, we next ask whether the ZFO-selected $\texttt{Taylor}_2$ step can be guaranteed to perform at least as well as the original first-order step under suitable local curvature conditions.

Under Assumption 1 and Assumptions 9-11, we compare the `Taylor2` step with the typical FO step by bounding the curvature approximation error, imposing strong concavity on $[0, R]$, and requiring the Taylor-model maximizer to be an interior point. The theorem gives conditions to guarantee $\phi(s_M) \geq \phi(s_{\mathrm{FO}})$.

Assumption 9. *Let the difference between the second derivative of ϕ and the second-order approximation of ϕ at $s = 0$ be bounded, i.e., there exists $D > 0$ such that $|\phi''(s) - \frac{\phi(\epsilon)-2\phi(0)+\phi(-\epsilon)}{\epsilon^2}| \leq D; \forall s \in [0, R]$.*

Assumption 10. *Let ϕ be ν-strongly concave on $[0, R]$, i.e., $\phi(u) \leq \phi(s) + \phi'(s)(u-s) - \frac{\nu}{2}(u-s)^2$; for all $s, u \in [0, R]$ where $\nu > 0$.*

Assumption 11. *Let the maximum of $T_2(s)$ be an interior point, i.e., $-\frac{\epsilon^2\phi'(0)}{\phi(\epsilon)-2\phi(0)+\phi(-\epsilon)} \in [0, R]$ and $\phi(\epsilon) - 2\phi(0) + \phi(-\epsilon) < 0$.*

Theorem 4 (Second-order Taylor win guarantee)**.** *If Assumption 1 and Assumptions 9-11 hold, and if $s_{\mathrm{FO}} \in [0, R]$, then the following relation holds between $\phi(s_M)$ and $\phi(s_{\mathrm{FO}})$:*

$$\textit{if } |s_{\mathrm{FO}} - s_M| = 0 \textit{ or } |s_{\mathrm{FO}} - s_M| \geq \frac{2Ds_M}{\nu}, \textit{ then } \phi(s_M) \geq \phi(s_{\mathrm{FO}}).$$

Proof. We have

$$T_2(s) = \phi(0) + \phi'(0)s + \frac{1}{2}\widehat{\phi''}(0)s^2,$$

which gives

$$T_2'(s) = \phi'(0) + \widehat{\phi''}(0)s.$$

By Assumption 11, there exists $s_M \in [0, R]$ such that

$$T_2'(s_M) = 0,$$

which gives

$$\phi'(0) + \widehat{\phi''}(0)s_M = 0,$$

and therefore

$$s_M = -\frac{\phi'(0)}{\widehat{\phi''}(0)}.$$

Notice that

$$\phi'(s) - T_2'(s) = \phi'(s) - \phi'(0) - \widehat{\phi''}(0)s = \int_0^s \left(\phi''(u) - \widehat{\phi''}(0)\right) du.$$

Therefore, by Assumption 9,

$$|\phi'(s_M) - T_2'(s_M)| = \left|\int_0^{s_M} \left(\phi''(u) - \widehat{\phi''}(0)\right) du\right| \leq \int_0^{s_M} \left|\phi''(u) - \widehat{\phi''}(0)\right| du \leq \int_0^{s_M} Ddu = Ds_M.$$

Since $T_2'(s_M) = 0$, this gives

$$|\phi'(s_M)| \leq Ds_M. \qquad (\star)$$

Let

$$d = s_{\mathrm{FO}} - s_M.$$

If $|d| = 0$, then $s_{\mathrm{FO}} = s_M$, and hence

$$\phi(s_M) = \phi(s_{\mathrm{FO}}).$$

Now suppose $|d| > 0$. By Assumption 10,

$$\phi(s_{\mathrm{FO}}) \leq \phi(s_M) + \phi'(s_M)d - \frac{\nu}{2}d^2,$$

which gives

$$\phi(s_{\mathrm{FO}}) - \phi(s_M) \leq \phi'(s_M)d - \frac{\nu}{2}d^2.$$

Furthermore, by $(\star)$,

$$\phi'(s_M)d \leq |\phi'(s_M)||d| \leq Ds_M|d|.$$

Therefore,

$$\phi(s_{\mathrm{FO}}) - \phi(s_M) \leq Ds_M|d| - \frac{\nu}{2}d^2 = |d|\left(Ds_M - \frac{\nu}{2}|d|\right).$$

Thus, if

$$|d| \geq \frac{2Ds_M}{\nu},$$

then

$$Ds_M - \frac{\nu}{2}|d| \leq 0,$$

and therefore

$$\phi(s_{\mathrm{FO}}) - \phi(s_M) \leq 0.$$

Hence

$$\phi(s_M) \geq \phi(s_{\mathrm{FO}}).$$

Finally, since $d = s_{\mathrm{FO}} - s_M$, the sufficient condition is

$$|s_{\mathrm{FO}} - s_M| \geq \frac{2Ds_M}{\nu}.$$

If $s_{\mathrm{FO}} = \alpha\phi'(0)$, then

$$|s_{\mathrm{FO}} - s_M| = \left|\alpha\phi'(0) - \left(-\frac{\phi'(0)}{\widehat{\phi''}(0)}\right)\right| = |\phi'(0)|\left|\alpha + \frac{1}{\widehat{\phi''}(0)}\right|.$$

Using $s_M = -\frac{\phi'(0)}{\widehat{\phi''}(0)}$, this gives the equivalent condition

$$\left|\alpha + \frac{1}{\widehat{\phi''}(0)}\right| \geq -\frac{2D}{\nu\widehat{\phi''}(0)}.$$

Therefore, if $|s_{\mathrm{FO}} - s_M| = 0$ or $|s_{\mathrm{FO}} - s_M| \geq \frac{2Ds_M}{\nu}$, then $\phi(s_M) \geq \phi(s_{\mathrm{FO}})$. □

B.2 ZFO Global Stationarity

To connect the local ZFO step-selection guarantees to overall optimization behavior, we finally ask whether repeatedly applying these approximate directional steps still drives the training process toward a stationary point of the population objective.

ZFO is computed on the sampled surrogate $F_t(\theta) = F(\theta; S_t)$, where S_t is generated at the current iterate θ_t and then fixed for nearby perturbations. We define the expected surrogate obtained under the current sampling distribution $\bar{F}_t(\theta) \coloneqq \mathbb{E}_S[F(\theta; S)]$. In our theoretical analysis, the true objective J and the expected surrogate $\bar{F}_t$ have the same gradient $\nabla \bar{F}_t(\theta) = \nabla J(\theta_t)$. In this formulation F_t is used for parameter updates, while its expectation $\bar{F}_t$ develops a reliable objective signal.

Lemma 1 (Surrogate vs Empirical Difference)**.** *Let Assumptions 1-5 hold, then*

$\max_{s \in I_t} J(\theta_t + s z_t) - J(\theta_t + s_t z_t) \le \Delta_t$, *holds with probability at least* $1 - \delta$, *where*

$$\Delta_t \coloneqq (\mathcal{L}_J + \mathcal{L}_{\bar{F}}) R_t^2 + 2\sigma(1 + R_t)\sqrt{\frac{2\log(4/\delta)}{n}} + 2W R_t^2 + \phi(s_\phi) - \phi(s_M).$$

Proof. Let $j_t(s) \coloneqq J(\theta_t + s z_t)$, so that

$$\max_{s \in I_t} J(\theta_t + s z_t) - J(\theta_t + s_M z_t) = j_t(s^*) - j_t(s_M),$$

where s^* maximizes $j_t(s)$ over $s \in I_t$.

Here $\psi_t(s) = \bar{F}_t(\theta_t + s z_t)$, where $\bar{F}_t$ is defined before substituting the batch-selected direction. Since $\nabla J(\theta_t) = \nabla \bar{F}_t(\theta_t)$, we have $j_t'(0) = \psi_t'(0)$. Therefore, for any $s \in I_t$,

$$j_t(s) - j_t(0) = s j_t'(0) + \int_0^s (s - u) j_t''(u) du.$$

Similarly,

$$\psi_t(s) - \psi_t(0) = s \psi_t'(0) + \int_0^s (s - u) \psi_t''(u) du.$$

Subtracting these identities gives

$$[j_t(s) - j_t(0)] - [\psi_t(s) - \psi_t(0)] = \int_0^s (s - u)[j_t''(u) - \psi_t''(u)] du.$$

By Assumption 5, $|j_t''(u)| \le \mathcal{L}_J$ and $|\psi_t''(u)| \le \mathcal{L}_{\bar{F}}$ almost everywhere, and therefore, by the triangle inequality,

$$[j_t(s) - j_t(0)] - [\psi_t(s) - \psi_t(0)] \le (\mathcal{L}_J + \mathcal{L}_{\bar{F}}) \int_0^s (s - u) du = \frac{\mathcal{L}_J + \mathcal{L}_{\bar{F}}}{2} s^2.$$

Consequently,

$$j_t(s^*) - j_t(0) \le \frac{\mathcal{L}_J + \mathcal{L}_{\bar{F}}}{2} (s^*)^2 + \psi_t(s^*) - \psi_t(0),$$

and similarly,

$$j_t(s_M) - j_t(0) \ge \psi_t(s_M) - \psi_t(0) - \frac{\mathcal{L}_J + \mathcal{L}_{\bar{F}}}{2} (s_M)^2.$$

Combining these inequalities yields

$$\max_{s \in I_t} J(\theta_t + s z_t) - J(\theta_t + s_M z_t) \le \psi_t(s^*) - \psi_t(s_M) + (\mathcal{L}_J + \mathcal{L}_{\bar{F}}) R_t^2.$$

We now bound the difference $\psi_t(s^*) - \psi_t(s_M)$. By the triangle inequality,

$$\psi_t(s^*) - \psi_t(s_M) \le |\psi_t(s^*) - \phi_t(s^*)| + \phi_t(s^*) - \phi_t(s_M) + |\phi_t(s_M) - \psi_t(s_M)|.$$

By the definition of s_ϕ,

$$\phi_t(s^*) - \phi_t(s_M) \le \phi_t(s_\phi) - \phi_t(s_M).$$

Furthermore,

$$|\psi_t(s^*) - \phi_t(s^*)| + |\phi_t(s_M) - \psi_t(s_M)| \le 2 \max_{s \in I_t} |\psi_t(s) - \phi_t(s)|.$$

To bound $\max_{s \in I_t} |\psi_t(s) - \phi_t(s)|$, Taylor expanding gives

$$\begin{aligned} |\psi_t(s) - \phi_t(s)| &= \left| (\psi_t - \phi_t)(0) + s(\psi_t' - \phi_t')(0) + \int_0^s (s - u)(\psi_t'' - \phi_t'')(u) du \right| \\ &\le |(\psi_t - \phi_t)(0)| + |(\psi_t' - \phi_t')(0)| R_t + W R_t^2, \end{aligned}$$

where we take $W \ge \mathcal{L}_{\bar{F}}$ without loss of generality. Therefore, by the conditional concentration premise,

$$\max_{s \in I_t} |\psi_t(s) - \phi_t(s)| \le \sigma(1 + R_t)\sqrt{\frac{2\log(4/\delta)}{n}} + W R_t^2.$$

Consequently, with conditional probability at least $1 - \delta$, $j_t(s^*) - j_t(s_M) \le \Delta_t$, where

$$\Delta_t \coloneqq (\mathcal{L}_J + \mathcal{L}_{\bar{F}}) R_t^2 + 2\sigma(1 + R_t)\sqrt{\frac{2\log(4/\delta)}{n}} + 2W R_t^2 + (\phi_t(s_\phi) - \phi_t(s_M)).$$

□

Theorem 3 (Stationarity of the ZFO update)**.** *Let Assumptions 1–6 hold. Let* $\mathcal{K} \coloneqq \mathcal{L}_J + \mathcal{L}_{\bar{F}} + 2W$ *and* $\chi \coloneqq 3\mu^2 - 4\delta > 0$*, with* $q_t/(2\mathcal{K}) \in I_t$ *almost surely. Then, for every* $T > 0$*, we have*

$$\min_{0\le t<T} \mathbb{E}\|\nabla J(\theta_t)\|^2 \le \frac{8\mathcal{K}(J^\star - J(\theta_0))}{\chi T} + \frac{8\mathcal{K}\tilde{\Delta}_T}{\chi}, \text{ where } \tilde{\Delta}_T = \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\Delta_t].$$

Proof. Since $\frac{q_t}{2\mathcal{K}} \in I_t$, the comparison step exists within the relevant search interval. Therefore,

$$\max_{s\in I_t} J(\theta_t + sz_t) \ge J\left(\theta_t + \frac{q_t}{2\mathcal{K}} z_t\right).$$

By Assumption 5,

$$J\left(\theta_t + \frac{q_t}{2\mathcal{K}} z_t\right) \ge J(\theta_t) + \frac{q_t}{2\mathcal{K}}\langle \nabla J(\theta_t), z_t\rangle - \frac{\mathcal{L}_J}{2}\frac{q_t^2}{4\mathcal{K}^2}.$$

Taking expectation over the randomness used to construct z_t, with the history before iteration t held fixed, and using $q_t = \mathbb{E}_t[\langle \nabla J(\theta_t), z_t\rangle]$, gives

$$\mathbb{E}_t\left[J\left(\theta_t + \frac{q_t}{2\mathcal{K}} z_t\right)\right] \ge J(\theta_t) + \frac{q_t^2}{2\mathcal{K}} - \frac{\mathcal{L}_J}{2}\frac{q_t^2}{4\mathcal{K}^2} \ge J(\theta_t) + \frac{3q_t^2}{8\mathcal{K}}.$$

Consequently,

$$\mathbb{E}_t\left[\max_{s\in I_t} J(\theta_t + sz_t)\right] \ge J(\theta_t) + \frac{3q_t^2}{8\mathcal{K}}.$$

By Lemma 1, with conditional probability at least $1-\delta$,

$$J(\theta_{t+1}) = J(\theta_t + s_t z_t) \ge \max_{s\in I_t} J(\theta_t + sz_t) - \Delta_t.$$

On the remaining event, since both steps belong to $I_t = [0, R_t]$, Assumption 5 gives

$$\max_{s\in I_t} J(\theta_t + sz_t) - J(\theta_{t+1}) \le R_t\|\nabla J(\theta_t)\| + \frac{\mathcal{L}_J}{2}R_t^2.$$

By Young's inequality,

$$R_t\|\nabla J(\theta_t)\| + \frac{\mathcal{L}_J}{2}R_t^2 \le \frac{\|\nabla J(\theta_t)\|^2}{2\mathcal{K}} + \frac{\mathcal{K}+\mathcal{L}_J}{2}R_t^2 \le \frac{\|\nabla J(\theta_t)\|^2}{2\mathcal{K}} + \Delta_t,$$

where the final inequality uses $\mathcal{K} \ge \mathcal{L}_J$ and $\Delta_t \ge \mathcal{K}R_t^2$. Since $\|\nabla J(\theta_t)\|^2$ is fixed given the history before iteration t, and the remaining event has conditional probability at most δ, taking $\mathbb{E}_t$ over both cases therefore gives

$$\mathbb{E}_t[J(\theta_{t+1})] \ge J(\theta_t) + \frac{3q_t^2}{8\mathcal{K}} - \mathbb{E}_t[\Delta_t] - \frac{\delta}{2\mathcal{K}}\|\nabla J(\theta_t)\|^2.$$

Assumption 6 gives $q_t^2 \ge \mu^2\|\nabla J(\theta_t)\|^2$, and therefore

$$\frac{\chi}{8\mathcal{K}}\|\nabla J(\theta_t)\|^2 \le \mathbb{E}_t[J(\theta_{t+1})] - J(\theta_t) + \mathbb{E}_t[\Delta_t].$$

Taking full expectation yields

$$\frac{\chi}{8\mathcal{K}}\mathbb{E}\|\nabla J(\theta_t)\|^2 \le \mathbb{E}[J(\theta_{t+1})] - \mathbb{E}[J(\theta_t)] + \widehat{\Delta}_t,$$

where $\widehat{\Delta}_t = \mathbb{E}[\Delta_t]$. Summing from $t = 0$ to $T-1$ and using Assumption 5 yields

$$\frac{\chi}{8\mathcal{K}}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla J(\theta_t)\|^2 \le J^\star - J(\theta_0) + \sum_{t=0}^{T-1}\widehat{\Delta}_t.$$

Finally,

$$\min_{0\le t<T}\mathbb{E}\|\nabla J(\theta_t)\|^2 \le \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla J(\theta_t)\|^2,$$

and therefore, using $\tilde{\Delta}_T = \frac{1}{T}\sum_{t=0}^{T-1}\widehat{\Delta}_t$,

$$\min_{0\le t<T}\mathbb{E}\|\nabla J(\theta_t)\|^2 \le \frac{8\mathcal{K}(J^\star - J(\theta_0))}{\chi T} + \frac{8\mathcal{K}\tilde{\Delta}_T}{\chi}.$$

□

B.3 Finite Difference Lemmas

Finite-Difference Theorems and Proofs. We provide the proofs for two finite-difference approximations used by ZFO in the form of Lemmas.

Lemma 2. *Let $\phi(s)$ be four times differentiable in an ϵ-neighborhood, then:*

$$\widehat{\phi''}(0) \coloneqq \frac{\phi(\epsilon) - 2\phi(0) + \phi(-\epsilon)}{\epsilon^2} = \phi''(0) + \mathcal{O}(\epsilon^2)$$

Proof. The Taylor expansion of $\phi(s)$ gives

$$\phi(s) = \phi(0) + s\phi'(0) + s^2\frac{\phi''(0)}{2} + s^3\frac{\phi'''(0)}{6} + s^4\frac{\phi^{(4)}(\xi_s)}{24},$$

hence

$$\phi(\epsilon) = \phi(0) + \epsilon\phi'(0) + \epsilon^2\frac{\phi''(0)}{2} + \epsilon^3\frac{\phi'''(0)}{6} + \epsilon^4\frac{\phi^{(4)}(\xi_\epsilon)}{24}$$

and

$$\phi(-\epsilon) = \phi(0) - \epsilon\phi'(0) + \epsilon^2\frac{\phi''(0)}{2} - \epsilon^3\frac{\phi'''(0)}{6} + \epsilon^4\frac{\phi^{(4)}(\xi_{-\epsilon})}{24}.$$

Adding the two gives

$$\phi(\epsilon) + \phi(-\epsilon) = 2\phi(0) + \epsilon^2\phi''(0) + \epsilon^4\frac{\phi^{(4)}(\hat{\xi})}{12},$$

subtracting $2\phi(0)$ gives

$$\phi(\epsilon) - 2\phi(0) + \phi(-\epsilon) = \epsilon^2\phi''(0) + \epsilon^4\frac{\phi^{(4)}(\hat{\xi})}{12},$$

and dividing by ϵ^2 gives

$$\frac{\phi(\epsilon) - 2\phi(0) + \phi(-\epsilon)}{\epsilon^2} = \phi''(0) + \epsilon^2\frac{\phi^{(4)}(\hat{\xi})}{12} = \phi''(0) + \mathcal{O}(\epsilon^2).$$

□

The equality above represents a natural second-order central finite difference; we prove the asymmetric finite difference for Lemma 3 below,

Lemma 3. *Let $\phi(s)$ be five times differentiable in an ϵ-neighborhood, then:*

$$\widehat{\phi'''}(0) \coloneqq 3\frac{\phi(\epsilon) - \phi(-\epsilon) - 2\epsilon\phi'(0)}{\epsilon^3} = \phi'''(0) + \mathcal{O}(\epsilon^2).$$

Proof. Consider the new expansion of $\phi(s)$

$$\phi(s) = \phi(0) + s\phi'(0) + s^2\frac{\phi''(0)}{2} + s^3\frac{\phi'''(0)}{6} + s^4\frac{\phi^{(4)}(0)}{24} + s^5\frac{\phi^{(5)}(\xi_s)}{120}.$$

Then taking the difference between $\phi(\epsilon)$ and $\phi(-\epsilon)$ returns

$$\phi(\epsilon) - \phi(-\epsilon) = 2\epsilon\phi'(0) + \epsilon^3\frac{\phi'''(0)}{3} + \epsilon^5\frac{\phi^{(5)}(\hat{\xi})}{60}.$$

Using $\phi'(0)$ gives

Table 3: **Summary of the models and datasets used in the large-model experiments.** Parameter counts denote the total number of model parameters. Dataset sizes are reported as the number of training/evaluation examples used in our experiments. MATH uses the 512/256 train/evaluation subset described in the main text.

Study	Model	Model Parameters	Dataset	Dataset Size (Train / Eval)
Main Study	Qwen-2.5-Math-1.5B	1.54B	GSM8K	7,473 / 1,319
			MATH	512 / 256
			SVAMP	700 / 300
			AsDiv	1,844 / 461
			OpenBookQA	4,957 / 500
	Phi-2	2.70B	SVAMP	700 / 300
			AsDiv	1,844 / 461
			OpenBookQA	4,957 / 500
	Gemma-2-2B	2.61B	SVAMP	700 / 300
	Llama-3.2-1B	1.23B	GSM8K	7,473 / 1,319
			AsDiv	1,844 / 461
			OpenBookQA	4,957 / 500
Qwen Sub-Study	Qwen-2.5-Math-1.5B	1.54B	ARC-Challenge	1,119 / 1,172
			StrategyQA	2,061 / 229
			FOLIO	1,001 / 204
			CODAH	2,498 / 278

$$\phi(\epsilon) - \phi(-\epsilon) - 2\epsilon\phi'(0) = \epsilon^3 \frac{\phi'''(0)}{3} + \epsilon^5 \frac{\phi^{(5)}(\hat{\xi})}{60},$$

and dividing by $\frac{\epsilon^3}{3}$ returns

$$\frac{3(\phi(\epsilon) - \phi(-\epsilon) - 2\epsilon\phi'(0))}{\epsilon^3} = \phi'''(0) + \epsilon^2 \frac{\phi^{(5)}(\hat{\xi})}{20} = \phi'''(0) + \mathcal{O}(\epsilon^2).$$

□

C Addendum to Experimental Results

This addendum collects supplementary diagnostics that complement the main results. We first validate the ZFO step-selection mechanism on two low-dimensional toy RLVR environments, then discuss the local-approximation behavior of each ZFO method on a panel of analytic test functions, and finally report cross-setting win ratios that summarize how consistently each variant improves over the FO baseline.

C.1 Configuration

This section specifies the objective and implementation choices used throughout our experiments. First, we define the GRPO surrogate, optimized by each method, then report the hyperparameter configurations used for the main large-model runs, Qwen sub-study, and toy validation.

Parameter	Setting
Optimizer	AdamW
Policy	`grpo`
Temperature	0.8
Top-p	0.95
KL coefficient $\beta^\star$	0.02
Perturbation scale ϵ	1e-3
Seed	{100, 101, 102}

Table 4: **Fixed experimental settings** used across all large language RLVR-ZFO runs.

Config	Batch	Group	Epochs	Max Len.	Output Len.
a	4	4	10	512	192
b	2	2	8	384	192
c	8	2	8	384	192
d	4	2	5	256	192

Table 5: **Training configuration types used across runs.** Batch represents batch size, Group represents group size, Max Len. and Output Len. represent Max Length and Maximum Output Length.

Table 6: **Hyperparameter Configuration.** Learning rates, bounds, and configuration types used for the main empirical study. FO LR and ZO LR denote the learning rates used for first-order and zeroth-order fine-tuning, respectively. BT2, BT3, BP2, and BP3 represent the selected bounds for Taylor2, Taylor3, Padé2, and Padé3, respectively. Learning rate search is over $\{1\mathrm{e}-6, 5\mathrm{e}{-6}, 1\mathrm{e}{-5}\}$ and bound search is over $\{3, 5, 10\}$. ALL indicates that every lr for MeZO produced the same result.

Model	Dataset	Configuration						
		FO LR	ZO LR	BT2	BT3	BP2	BP3	Config Type
Qwen-2.5-Math-1.5B	GSM8K	5e-6	ALL	10	5	10	10	*a*
	MATH	5e-6	1e-6	5	5	5	5	*c*
	SVAMP	1e-5	5e-7	3	5	3	3	*a*
	AsDiv	5e-6	5e-7	5	5	5	5	*c*
	OpenBookQA	1e-5	1e-7	10	5	5	5	*a*
Phi-2	SVAMP	5e-6	5e-7	3	5	5	3	*a*
	AsDiv	5e-6	1e-7	3	3	5	3	*d*
	OpenBookQA	5e-6	1e-7	3	5	3	3	*d*
Gemma-2-2B	SVAMP	1e-6	5e-9	3	3	3	3	*b*
Llama-3.2-1B	GSM8K	5e-6	1e-7	3	3	3	3	*a*
	AsDiv	5e-6	5e-8	5	5	3	3	*c*
	OpenBookQA	1e-6	ALL	5	5	5	3	*b*

Table 7: **Learning rate, bound, and configuration type used for the Qwen sub-study on additional reasoning benchmarks.** Learning rate search is over {5e-6, 1e-5} and bound search is over $\{3, 5\}$.

Model	Dataset	Configuration					
		Learning Rate	BT2	BT3	BP2	BP3	Config Type
Qwen-2.5-Math-1.5B	ARC-Challenge	1e-5	3	5	3	3	*c*
	StrategyQA	1e-5	5	5	3	5	*c*
	FOLIO	1e-5	3	3	3	3	*c*
	CODAH	1e-5	5	5	5	5	*c*

C.1.1 Surrogate Objective Formulation

GRPO Surrogate. For this paper, the surrogate F_t instantiates the GRPO objective. Let $\mathcal{P}_t$ denote the set of prompts in the current batch, with $|\mathcal{P}_t|$ the batch size and G the number of examples sampled per prompt. For each prompt $i \in \{1, \ldots, |\mathcal{P}_t|\}$ and each sampled example $j \in \{1, \ldots, G\}$, let $\tau_{i,j,t}$ denote the sampled example, $|\tau_{i,j,t}|$ its length, and $(x_{i,j,v}, y_{i,j,v})$ the state-action pair at token position v. The surrogate is

$$F_t(\theta; S_t) = \frac{1}{|\mathcal{P}_t|G} \sum_{i=1}^{|\mathcal{P}_t|} \sum_{j=1}^{G} \widetilde{A}_{i,j,t} \log \pi_\theta(\tau_{i,j,t}) - \kappa \cdot \mathrm{KL}(\pi_\theta \| \pi_{\mathrm{ref}}),$$

where $\kappa > 0$ is the KL coefficient, $\widetilde{A}_{i,j,t}$ is the group-relative advantage given as

$$\widetilde{A}_{i,j,t} = \rho(\tau_{i,j,t}) - \frac{1}{G} \sum_{k=1}^{G} \rho(\tau_{i,k,t}),$$

and ρ is the verifiable reward given by

$$\rho(\tau_{i,j,t}) = \mathbb{1}_{\tau_{i,j,t} \in \mathcal{C}_t},$$

with $\mathcal{C}_t$ denoting the set of sampled examples deemed correct by the verifier at iteration t. The KL penalty is computed at the token level and averaged across the batch:

$$\mathrm{KL}(\pi_\theta \| \pi_{\mathrm{ref}}) = \frac{1}{|\mathcal{P}_t|G} \sum_{i=1}^{|\mathcal{P}_t|} \sum_{j=1}^{G} \frac{1}{|\tau_{i,j,t}|} \sum_{v=1}^{|\tau_{i,j,t}|} \big[\log \pi_\theta(y_{i,j,v} \mid x_{i,j,v}) - \log \pi_{\mathrm{ref}}(y_{i,j,v} \mid x_{i,j,v})\big].$$

Implementation-Level Interpretation of α and β. The implementation uses an equivalent parameterization of the search interval defined in the methodology. Let θ_t and $\widehat{\theta}_t$ denote the parameters

before and after a tentative AdamW update with learning rate α, and define the realized optimizer displacement $d_t = \widehat{\theta}_t - \theta_t$.

For $d_t \neq 0$, searching over $\theta_t + ad_t$ with $a \in [0, \beta]$ is equivalent to searching over $\theta_t + sz_t$, where

$$z_t = \frac{d_t}{\|d_t\|}, \quad s = a\|d_t\|, \quad s \in [0, R_t], \quad R_t = \beta\|d_t\|.$$

These are two coordinate representations of the same parameter-space interval. For a fixed parameter vector, gradient, and optimizer state, the exact-arithmetic AdamW displacement scales linearly with α. Under these conditions, matched-product pairs such as $(\alpha, \beta) = (10^{-6}, 5)$ and $(5 \times 10^{-6}, 1)$ define the same physical search interval.

This proportionality need not hold in the numerical implementation. The tentative update is applied to finite-precision parameters, and d_t is obtained by subtracting the stored parameter vectors. Consequently, rounding can cause a displacement to differ. In particular, multiplying a small, already-rounded update by β need not reproduce an update formed using a larger base learning rate. Matched $\alpha\beta$ products therefore need not yield identical realized search radii or directions. Numerical safeguards can introduce additional departures from scaling equivalence, and subsequent differences in parameters, sampled responses, and optimizer states can amplify these discrepancies over training. The observed matched-product differences thus concern the implemented numerical procedure and do not contradict the exact-arithmetic equivalence.

LLM implementation. GRPO supplies the optimizer direction and first-order coefficient. Symmetric probes evaluate a separate reward-minus-KL score on fixed sampled sequences, with rewards held constant. The probe values and GRPO slope jointly define the local model used for step selection. The theory analyzes an idealized ZFO setting in which the directional derivative and function probes are evaluations of the same sampled surrogate. The RLVR implementation approximates this setting by holding sampled responses and rewards fixed during the probes.

Empirical Candidate Selection. Interior search controls which candidate steps are considered within the same search interval. Using the displacement parameterization above, Taylor2 and Taylor3 compare the endpoints with feasible real stationary points, whereas the accepted Padé2 $[1/1]$ and Padé3 $[1/2]$ models compare the two endpoints. A regular $[1/1]$ model is monotone or constant, so its maximum can be attained at an endpoint. A regular $[1/2]$ model can instead possess interior extrema; endpoint selection for Padé3 is therefore an empirical restriction on its candidate set, although the safeguarded Padé3 procedure can still select an interior step through its Taylor2 fallback when a pole or near-pole is detected.

C.1.2 Empirical Configurations

All large-model runs share the fixed hyperparameters listed in Table 4. We hold these values constant so that observed differences in performance can be attributed to the update method as opposed to incidental variance in sampling, regularization, or optimization settings. Each reported result is averaged over the three seeds $\{100, 101, 102\}$, and the standard deviations reported in Tables 1 and 2 are based on these various runs.

We evaluate the models after each epoch, and all runs use the same early-stopping rule; we terminate the training when evaluation accuracy is unchanged for three consecutive epochs. Otherwise, we continue training until the maximum epoch count specified for the corresponding configuration. Per-run hyperparameters differ across model-dataset pairs and include the learning rate α, the bound hyperparameter β, and the selected configuration from Table 5. We summarize four configuration types in Table 5, which represent a performance-versus-cost trade-off. Config *a* is the default, with the largest group size and longest context length. Configs *b–d* are progressively cheaper alternatives selected when *a* is infeasible based on either memory or time constraints. Config *b* reduces both batch and group sizes for memory-constrained runs, *c* trades group size for a larger batch when throughput is more valuable than rollout diversity, and *d* additionally shortens the context and reduces epochs.

These per-run details are reported in Table 6 for the main results, including the FO learning rate, ZO learning rate, and each ZFO bound β. The FO learning rate is selected from the sweep {1e-6, 5e-6, 1e-5 }, and the corresponding bound for each ZFO method is then chosen from a validated sweep over $\{3, 5, 10\}$. The ZO learning rate is tuned independently from the sweep $\{1e{-}8, 5e{-}8, 1e{-}7, 5e{-}7\}$, since ZO methods generally require a smaller effective step scale than FO updates. The performance

Table 8: **Compact summary of the Classical Optimization Problems used in Figure 5.** Before optimization, each listed initial position θ_0 is perturbed independently per seed as $\theta_{0,s} = \theta_0 + \xi_s$, where $\xi_s \sim \mathcal{N}(0, 0.04I)$ applies independent Gaussian noise (standard deviation 0.2) to every coordinate, drawn from a dedicated, task-specific random stream and shared across all FO, ZFO, and GeN methods evaluated on that task–seed pair.

Problem	Objective	Initial position
Linear least squares	$\min_\theta \frac{1}{2n}\|A\theta - y\|^2$	$\theta_0 = \mathbf{0}_{20}$
Nonlinear least squares	$\min_{a,b,c} \frac{1}{2N}\sum_i \left(a\, e^{bx_i} + c - y_i\right)^2$	$(a_0, b_0, c_0) = (0.4,\ 0.8,\ 1.0)$
Logistic regression	$\min_\theta \frac{1}{n}\sum_i \left[\log(1 + e^{x_i^\top \theta}) - y_i x_i^\top \theta\right] + \frac{\lambda}{2}\|\theta\|^2$	$\theta_0 = \mathbf{0}_{16}$
Low-rank matrix factorization	$\min_{U,V} \frac{1}{2mn}\|UV^\top - M\|_F^2 + \frac{\lambda}{2\|\theta\|}\|\theta\|^2$	$(U_0, V_0) \sim \mathcal{N}(0, 0.15^2)$
Rosenbrock	$\min_{x_1,x_2} 100(x_2 - x_1^2)^2 + (1 - x_1)^2$	$(x_1, x_2)_0 = (-1.2,\ 1.0)$
Beale	$\min_{x_1,x_2} \sum_{k=1}^3 \left(c_k - x_1 + x_1 x_2^k\right)^2$	$(x_1, x_2)_0 = (-1.0,\ 1.0)$

reported is the mean verifiable reward on the evaluation set. The Qwen sub-study from Table 2 follows a similar protocol but fixes config *c* across runs (a memory-conservative choice given the sub-study sweeps over four datasets), restricts the learning rate sweep to {5e-6, 1e-5 }, and limits the bound sweep to $\{3, 5\}$. The Qwen sub-study configurations are reported in Table 7.

The toy models in Figure 7 use a lightweight configuration designed to validate the ZFO step-selection mechanism in a controlled low-dimensional setting. Both environments share a context dimension of 6, a sequence length of 4, and a two-layer policy network of hidden width 48. Each run uses a dataset of 200 contexts trained for 50 steps with batch size 16, group size $G = 4$, perturbation scale $\epsilon = 10^{-2}$, and KL coefficient of 0.1. The bound hyperparameter β is fixed to 1, so the ZFO search interval is restricted to steps no larger than the standard FO update.

Table 3 depicts the models and datasets used throughout our large-model experiments, along with their corresponding parameter counts and dataset sizes. We consider models ranging from 1.23B to 2.7B parameters across a diverse set of mathematical reasoning, commonsense reasoning, and knowledge-based benchmarks. We additionally report the datasets used in our Qwen sub-study to provide a complete summary of the main experimental settings considered.

Our MeZO baseline uses a forward-only, two-sided finite-difference update along an unnormalized Gaussian perturbation direction. At each iteration, sampled responses, group-centered reward advantages, and reference-policy log probabilities are held fixed across the positive and negative perturbations. The evaluated objective combines advantage-weighted completion log probabilities with a nonnegative k_3 KL penalty, averaged over completion tokens within each sequence. Both terms use completion-only masking, and responses are decoded from the end of the padded prompt. Objective evaluations run in evaluation mode, with log-ratio clipping for numerical stability. The perturbation direction is regenerated from the same random seed for both probes and the parameter update, avoiding backpropagation.

Interpreting MeZO Performance. Beyond low average performance, MeZO reaches zero final accuracy in some settings. In two settings, all three seeds decline from nonzero first-epoch accuracy to zero, indicating deterioration during training. Random zeroth-order updates, weak group-relative rewards, and learning rate sensitivity may contribute to this instability. These explanations remain hypotheses rather than isolated causes, and the results concern MeZO's specific adaptation to RLVR.

C.2 Additional Insights

Optimization Beyond Neural Network Training. To evaluate whether ZFO extends beyond neural network optimization, we consider six classical optimization problems: linear least squares, nonlinear least squares, logistic regression, low-rank matrix factorization, Rosenbrock, and Beale. A summary of the problems can be found in Table 8. We evaluate Taylor2, Taylor3, Padé2, and Padé3 with both SGD and AdamW-style directions, alongside their first-order baselines and GeN [9] and present the results in Figure 5. Each method is independently tuned over the same 18 learning rates,

$\{1, 2, 5\} \times 10^{-5,\ldots,0}$, while ZFO methods use a fixed bound $\beta = 3$ and perturbation radius $\epsilon = 10^{-3}$. GeN uses $\gamma = 0.9$ and $\Phi = 1$ for fair comparison against the ZFO methods. Each experiment runs for 100 gradient updates over three seeds.

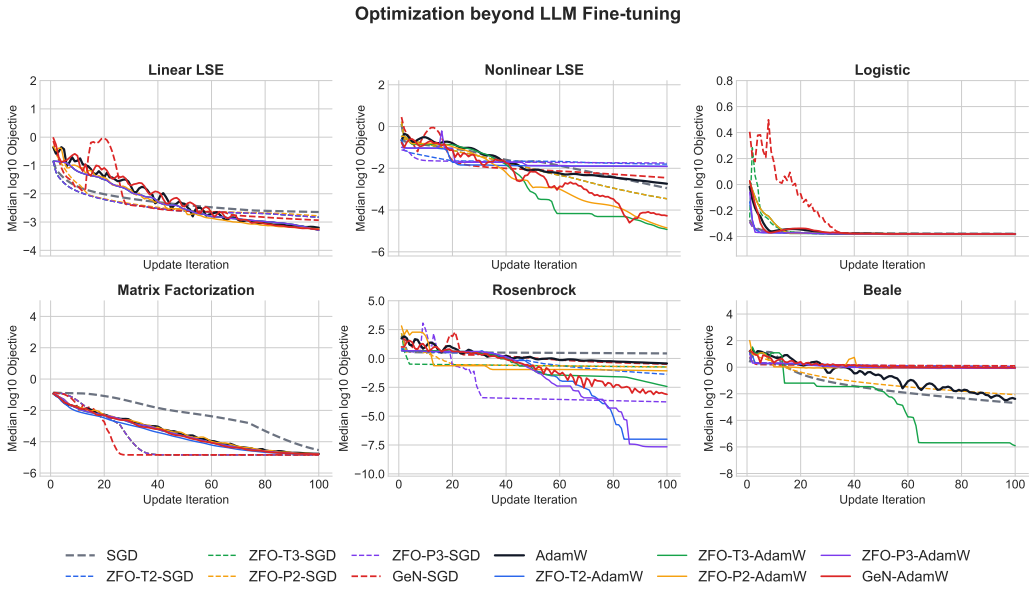

Figure 5: **Median $\log_{10}$ objective over three seeds for six optimization problems.** Solid curves use AdamW directions, and dashed curves use SGD directions. Each method uses its best learning rate from the same 18-point search. Lower values indicate better optimization.

The results show that ZFO's benefits are substantial but depend on both the local model and direction. Taylor_3–AdamW attains the lowest final median objectives on nonlinear least squares and Beale, whereas Padé_3–AdamW performs best on Rosenbrock. Several ZFO and GeN variants also accelerate matrix factorization relative to ordinary SGD. On linear least squares, most adaptive methods improve moderately over their corresponding first-order baselines, whereas logistic regression converges to essentially the same objective across methods. Some combinations also stall or underperform, particularly on the nonconvex objectives. Thus, the experiment supports ZFO's applicability beyond LLMs and its ability to yield order-of-magnitude improvements in favorable geometries.

All Roads Lead to ϕ. Figure 6 examines the four ZFO surrogates, $\texttt{Taylor}_2$, $\texttt{Taylor}_3$, $\texttt{Padé}_2$, and $\texttt{Padé}_3$, on a panel of locally analytic functions. In each panel, $f(x_0)$ and $f'(x_0)$ are computed exactly, while the second and third derivatives are approximated by the same finite differences discussed in this paper. The top row consists of functions that are either sub-linear or polynomial on a tight interval; hence the approximations are nearly indistinguishable from the actual function. The bottom row deliberately widens the window by introducing functions that do not have sub-linear properties. Across both rows, the approximants closely agree with the true objective near the expansion point, while their behavior diverges as the approximation interval widens and the objective becomes more nonlinear.

Toy Model Validation. We first validate the ZFO framework on two low-dimensional toy RLVR environments shown in Figure 7. The first toy model, *Grouped Sequence Bandit*, is a sequence-decision task that requires a learning algorithm to choose a short sequence of tokens to match a hidden target sequence. Specifically, the model samples a complete sequence based on a small input and updates the parameters based on the reward feedback. The second model, *Grouped Path-Matching*, uses a similar idea, but makes the reward structure more dependent on the entire sequence path. Instead of rewarding independent tokens, the policy receives a higher reward when its sampled actions remain aligned with the hidden target path. Both environments use a stochastic policy trained with grouped advantages and KL regularization against a frozen reference, replicating the surrogate structure of $F_t(\theta; \mathcal{S}_t)$ used in our larger experiments. Each method is averaged over 5 seeds with the bound hyperparameter $\beta = 1$; the remainder of the configuration can be found in §C.1.2.

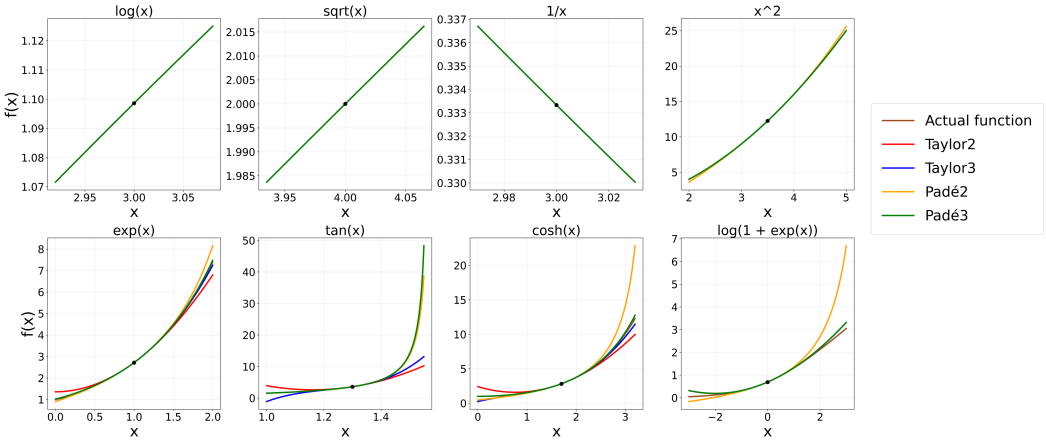

Figure 6: **ZFO approximations on a variety of locally analytic functions.** Each panel plots $\texttt{Taylor}_2$, $\texttt{Taylor}_3$, $\texttt{Padé}_2$, and $\texttt{Padé}_3$ around an x_0. The top row consists of panels where all four interpolants are visually indistinguishable from the actual function. The bottom row consists of panels that expose the qualitative differences between the Taylor and Padé approximants.

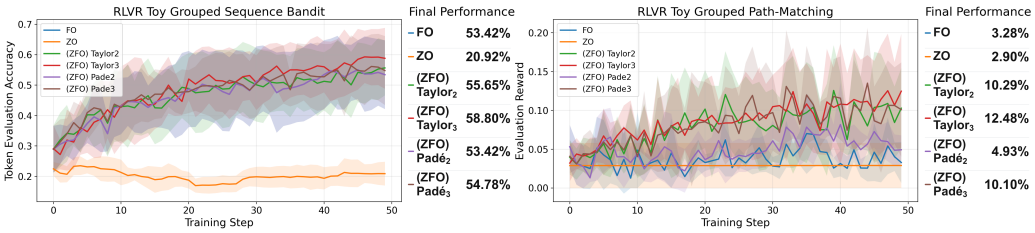

Figure 7: **Toy RLVR validation on the *Grouped Sequence Bandit* and *Grouped Path-Matching* environments.** ZFO methods improve over FO and ZO baselines under the same search scale $\beta = 1$, showing that local curvature models can select more effective steps along update directions.

Figure 7 reports the training curves and final performance across both environments. Across both toy environments, the ZFO methods improve over the FO and ZO baselines, supporting the central claim that additional function evaluations can improve step-size selection along a fixed direction. Importantly, these improvements occur with the same search scale, since all methods use $\beta = 1$. Thus, the ZFO methods are not benefiting from a larger trust region; rather, they are using local curvature behavior to select more effective intermediate steps within the same update interval. This is most visible in the separation between $\texttt{Taylor}_2$, $\texttt{Taylor}_3$, and $\texttt{Padé}_3$ relative to FO, showing that the additional curvature information can meaningfully alter the chosen step even when the learning rate is fixed. The behavior of $\texttt{Padé2}$ further supports this interpretation; among the ZFO methods, $\texttt{Padé2}$ is the most restricted in the sense that it contains only a linear numerator and denominator. Consequently, when $\beta = 1$, $\texttt{Padé2}$ often selects a step close to the standard FO update, making it the ZFO variant most similar to FO. This trend is reflected in the final performance summaries with $\texttt{Taylor}_3$ performing best in both environments followed closely by $\texttt{Taylor}_2$ and $\texttt{Padé}_3$. Overall, these toy results provide validation of the ZFO step-selection mechanism before moving to higher-dimensional models.

Learning Rate Sensitivity Curves. Figure 8 establishes that ZFO's gains arise from intelligent step selection rather than a hidden learning-rate effect. In each panel, we plot ZFO trajectories alongside FO FT runs surrounding the base learning rate to compare against ZFO runs. The three panels highlight complementary behaviors; e.g., Qwen on OpenBookQA shows ZFO trajectories outperforming FO runs with the same nominal search regions, reinforcing the extremely large margin in Table 1. Phi-2 on SVAMP shows ZFO trajectories steadily improving over the FO baselines. Llama on AsDiv shows that the most conservative FO baseline eventually destabilizes during training, whereas ZFO methods remain stable. Together, these curves indicate that ZFO occupies a regime that FO cannot reach by fixed learning rates.

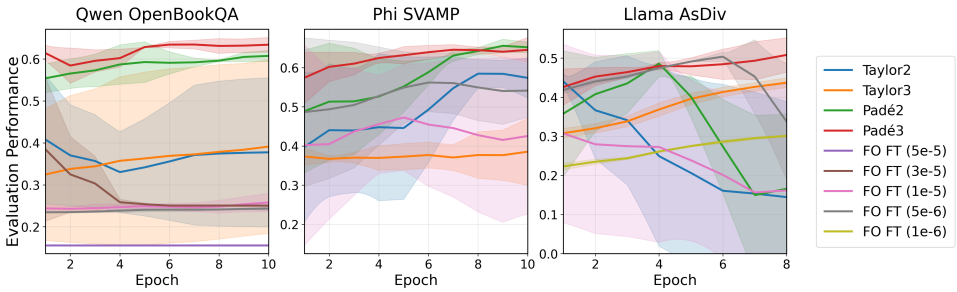

Figure 8: **Stability of ZFO.** Empirical evaluation curves demonstrating ZFO's capacity to stabilize and improve over fine-tuning with AdamW (FO FT). Demonstrations involve AdamW FT (FO FT) curves of multiple learning rates.

Method	Main Results Wins	Qwen Sub-Study Wins	Win Ratio
$\texttt{Taylor}_2$	10/12	4/4	14/16
$\texttt{Taylor}_3$	6/12	1/4	7/16
$\texttt{Padé}_2$	11/12	2/4	13/16
$\texttt{Padé}_3$	11/12	3/4	14/16

Table 9: **Win ratios for each ZFO method across the main results and Qwen sub-study.** A win denotes a setting in which the method achieves a higher mean accuracy than the corresponding FO baseline.

First-Order Variability and Training Stability. High variability across seeds can reflect late training collapse. For ASDiv on Llama-3.2-1B, two FO seeds initially exceed 50% accuracy before falling below 0.4%, while the third finishes at 49.07%. The resulting 16.60 ± 28.12 therefore reflects sharply different training outcomes. The corresponding Padé3 runs finish between 47.33% and 56.83%. Since ZFO can suppress or shorten a proposed update, its role extends beyond searching for larger steps: adaptive step selection may also reduce destructive updates. These observations support a stabilizing role in some configurations.

The additional FO results further illustrate this sensitivity: for OpenBookQA on Qwen, linear warmup followed by cosine decay improves AdamW from 26.07 ± 2.53 to 59.60 ± 0.87, although the schedule does not improve every dataset. This highlights the importance of how step sizes evolve during fine-tuning.

Bound vs. Learning Rate Ablation. Figure 9 reports the full ablation over the learning rate α and search bound β that was summarized in the main paper. We sweep both hyperparameters across all four ZFO methods on Qwen-2.5-Math-1.5B, evaluating after 10 epochs on subsets (512/256 train/eval) of GSM8K and SVAMP. As discussed in the main paper, at smaller learning rates, increasing β produces near-monotonic improvements, while at larger learning rates the gains saturate as the search region R grows too large. Ideally, at a fixed parameter vector, gradient, and optimizer state, matched-product configurations such as $(\alpha = 10^{-6}, \beta = 5)$ and $(\alpha = 5 \times 10^{-6}, \beta = 1)$ define the same search interval in exact arithmetic. Empirically, finite-precision updates and numerical safeguards can break this equivalence, with discrepancies potentially amplified over training; see Section C.1.1.

Examining Figure 9 closely, we observe the high-reward region forms a diagonal band that shifts toward smaller β as α grows. At $\alpha = 10^{-6}$ and $\alpha = 5 \times 10^{-6}$, performance improves steadily as β increases from 1 to 100 across all four methods. At $\alpha = 10^{-5}$, the best results pull back toward small β. Combining the largest bound with $\alpha = 10^{-5}$ collapses the reward to near-zero for almost all methods.

Cross-Setting Improvement Frequency. As an additional appendix diagnostic, we summarize how often each ZFO variant improves over the corresponding FO baseline across the reported evaluation settings in Table 9. Across the 12 main result settings and the 4 Qwen sub-study settings, $\texttt{Taylor}_2$, $\texttt{Taylor}_3$, $\texttt{Padé}_2$, and $\texttt{Padé}_3$ improve over FO in 14/16, 7/16, 13/16, and 14/16 settings, respectively. This supports the view that ZFO is driven by a recurring advantage in adaptive step selection across settings.

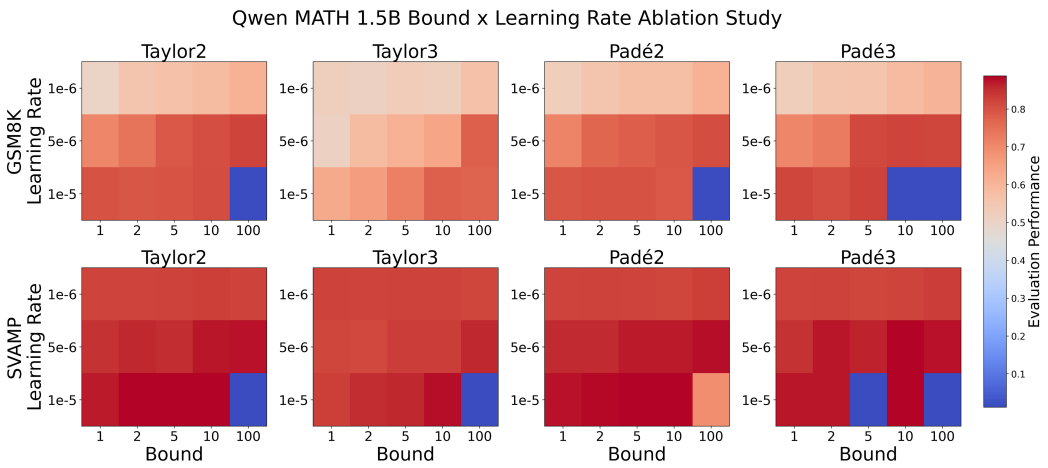

Figure 9: **Ablation study inspecting the relationship between the learning rate α and search hyperparameter β on the reward rate.** We report the performance over a training set of 512 elements and an evaluation set of 256 elements on the Qwen-2.5-MATH-1.5B model and datasets, GSM8K and SVAMP, on the top and bottom row, respectively. From left to right, the panels represent all four ZFO methods ($\texttt{Taylor}_2$, $\texttt{Taylor}_3$, $\texttt{Padé}_2$, $\texttt{Padé}_3$) presented in this paper.

Table 10: **Comparison with additional first-order baselines.** AdamW with linear warmup and cosine decay and Prodigy are compared with the worst- and best-performing ZFO variants for each Qwen-2.5-Math-1.5B dataset. The ZFO extrema are selected by mean final evaluation performance across Taylor_2, Taylor_3, Padé_2, and Padé_3; the selected variant is shown in parentheses. The best result in each row is bold.

Model	Dataset	AdamW + Cosine	Prodigy	Worst ZFO	Best ZFO
Qwen-2.5-Math-1.5B	GSM8K	72.88 ± 0.54	64.32 ± 0.90	75.56 ± 0.24 (Taylor_3)	**81.00** $\pm$ **0.11** (Taylor_2)
	SVAMP	86.00 ± 0.33	83.22 ± 0.19	89.11 ± 0.51 (Taylor_3)	**90.89** $\pm$ **1.02** (Padé_2)
	AsDiv	81.98 ± 0.52	80.82 ± 0.30	88.27 ± 0.35 (Taylor_3)	**94.09** $\pm$ **0.29** (Taylor_2)
	OpenBookQA	59.60 ± 0.87	38.73 ± 0.31	37.80 ± 17.85 (Taylor_2)	**64.07** $\pm$ **1.22** (Padé_3)
	ARC-Challenge	55.15 ± 1.11	30.80 ± 0.15	53.58 ± 2.77 (Padé_2)	**58.10** $\pm$ **1.33** (Taylor_2)
	StrategyQA	66.28 ± 0.51	60.80 ± 0.95	61.67 ± 1.46 (Taylor_3)	**70.06** $\pm$ **0.61** (Padé_3)
	FOLIO	34.48 ± 1.13	34.31 ± 0.49	34.48 ± 1.13 (Taylor_3)	**36.60** $\pm$ **3.26** (Taylor_2)
	CODAH	48.32 ± 14.52	30.58 ± 1.57	41.85 ± 20.58 (Padé_3)	**62.83** $\pm$ **3.14** (Taylor_2)

Additional first-order baselines. To compare ZFO with different learning-rate adaptation strategies, we additionally evaluate AdamW with a warmup-cosine schedule [30] and the parameter-free Prodigy optimizer [37] in Table 10. We preserve the Qwen-2.5-Math-1.5B model, dataset-specific training configurations, GRPO objective, evaluation protocol, and seeds used in the corresponding main and sub-study experiments. For AdamW with cosine decay, the learning rate selected for the constant-rate FO baseline is used as the peak rate; it gets reached by linear warmup over the first 10% of optimizer steps and then decayed to zero according to a cosine schedule. Prodigy instead uses its recommended configuration with adjustment factor 1.0, $d_{\text{coef}} = 1.0$, decoupled weight decay 0.01, and no external learning-rate schedule. To contextualize these baselines within the range of behavior exhibited by ZFO, the table also reports the lowest- and highest-mean ZFO variants for each dataset.

Table 10 shows that the strongest ZFO variant outperforms both AdamW with cosine decay and Prodigy across all eight datasets on Qwen. AdamW with cosine decay remains the stronger of the two additional FO baselines, while Prodigy consistently trails both the scheduled AdamW baseline and the best ZFO method. However, the worst-performing ZFO variant falls below AdamW with cosine decay on several datasets, indicating that the advantage does not arise uniformly across every local model. Overall, these results suggest that ZFO can improve over many step-adaptive baselines, while also signaling that the choice of local approximation remains important.

ZFO Sequential Line Search. The four ZFO methods in the main paper select a step using a local model constructed from a fixed budget of two probes. `ZFO-Seq` extends this design to a sequential budget of $k > 2$ fixed-batch evaluations. The first two probes construct the initial model

Table 11: **Comparison with line-search algorithms.** Extended line-search results comparing the basic ZFO-Seq method against Armijo, Barzilai-Borwein, Polyak, and a simplified probabilistic line search (PLS) on the Qwen Model. The best result is in **bold**, and the second best result is underlined.

Dataset	ZFO-Seq	Armijo	BB	Polyak	PLS
CODAH	**55.47**	32.81	52.73	30.47	31.25
ARC-Challenge	55.86	**57.81**	54.69	35.16	35.16
SVAMP	83.59	**85.55**	**85.55**	82.81	76.17

$M \in \{T_2, T_3, P_2, P_3\}$ as in Algorithm 1, and the remaining $k-2$ probes are spent refining the model's proposal.

At each refinement step, the current candidate maximizer is evaluated on the shared minibatch S_t, and a new candidate is proposed by fitting a parabola through the best evaluated point in I_t and its two nearest evaluated neighbors, aligning with safeguarded successive parabolic interpolation [8]. The returned step is the best evaluated point rather than a model maximizer. Every candidate must survive an actual evaluation of ϕ_t before it can be selected. Since $(0, \phi_t(0)) \in E_t$ by construction, `ZFO-Seq` guarantees $\phi_t(s_t) \geq \phi_t(0)$ at every iteration, a monotone non-decrease of the empirical surrogate. This replaces the model-selection gap $\phi(s_\phi) - \phi(s_M)$ with the gap attained by the best of k evaluated points, which shrinks under refinement. Algorithm 2 summarizes the procedure.

Table 11 reports the evaluation accuracy for a variety of line-search methods on the Qwen-2.5-Math-1.5B model (seed 100, configuration C, learning rate 1e-5, search bound $\beta = 3$). Each ZFO-Seq model used P_3 as its default initialization.

Algorithm 2 ZFO-Seq

Require: θ_0, base optimizer, learning rate α, $\epsilon > 0$, $\beta > 0$, budget $k \geq 3$, initial model $M \in \{T_2, T_3, P_2, P_3\}$

for $t = 0, 1, 2, \ldots$ **do**
  Algorithm 1 Inner Loop
  for $j = 1, \ldots, k-2$ **do**
    Evaluate $\phi_t(s^{(j)})$ on S_t and add $(s^{(j)}, \phi_t(s^{(j)}))$ to E_t
    if $j < k-2$ **then**
      Fit a parabola $\mathcal{M}$ through the best point in E_t and its two nearest neighbors
      Propose $s^{(j+1)} \in \arg\max_{s \in I_t} \mathcal{M}(s)$
    end if
  end for
  Select $s_t \in \arg\max_{s:\,(s,\phi_t(s)) \in E_t} \phi_t(s)$
  Update $\theta_{t+1} = \theta_t + s_t z_t$
end for

Table 11 compares `ZFO-Seq` against several established line-search methods. `ZFO-Seq` achieves the strongest result on CODAH, while remaining within approximately two points of the best-performing method on both ARC-Challenge and SVAMP. Armijo and Barzilai-Borwein remain slightly stronger on these latter settings, whereas Polyak and PLS generally trail `ZFO-Seq`. Thus, `ZFO-Seq` is competitive with classical line-search and adaptive step-size methods across the three evaluated settings.

Unit-Bound Ablation. To determine whether ZFO benefits require a search interval larger than the standard first-order update, we additionally evaluate the four ZFO models with the bound fixed to $\beta = 1$ on Qwen-2.5-Math-1.5B. Under this setting, ZFO may shorten or reshape the AdamW update but cannot select a step beyond the endpoint corresponding to the nominal FO displacement. ARC-Challenge and StrategyQA use three-epoch versions of configuration *a*, while ASDiv uses a three-epoch version of configuration *c*; we denote these shortened configurations by *a'* and *c'*, respectively. Their batch size, group size, context length, and maximum output length otherwise remain unchanged. All methods use the fixed RLVR settings in Table 4, including perturbation scale $\epsilon = 10^{-3}$, and results are averaged over seeds $\{100, 101, 102\}$. Table 12 reports the resulting final evaluation accuracy.

Table 12: **Qwen-2.5-Math-1.5B sub-study ablation with $\beta = 1$.** Results are reported as mean $\pm$ standard deviation over three seeds. All datasets are trained with 3 epochs (ARC-Challenge, StrategyQA; config a'), (ASDiv; config c'). Under this 3-epoch configuration, reducing the ARC FO learning rate from 10^{-5} to 5×10^{-6} improves accuracy from 31.71 ± 8.11 to 55.12 ± 0.60. The ASDiv FO result uses the third-epoch checkpoint of the matched FO runs.

Dataset	**Base LR**	**AdamW (FO)**	**Taylor$_2$**	**Taylor$_3$**	**Padé$_2$**	**Padé$_3$**
ARC-Challenge	1×10^{-5}	31.71 ± 8.11	$\mathbf{56.77 \pm 0.61}$	56.51 ± 0.36	55.49 ± 0.98	49.86 ± 0.27
StrategyQA	1×10^{-5}	59.24 ± 10.14	70.69 ± 5.10	$\mathbf{74.87 \pm 0.37}$	67.25 ± 13.64	48.18 ± 0.44
ASDiv	5×10^{-6}	$\mathbf{82.08 \pm 0.08}$	81.82 ± 0.26	78.05 ± 1.41	81.87 ± 0.26	81.92 ± 0.46

Table 12 shows that ZFO can modify training behavior even when it is not permitted to move beyond the nominal FO update. On ARC-Challenge, all four ZFO variants outperform AdamW at the shared learning rate of 10^{-5}; however, reducing the separately tuned FO learning rate to 5×10^{-6} raises its accuracy to 55.12, so the ARC result is best interpreted as ZFO recovering a competitive step scale. On StrategyQA, $\texttt{Taylor}_3$ performs best at 74.87, followed by $\texttt{Taylor}_2$ and $\texttt{Padé}_2$, whereas $\texttt{Padé}_3$ falls below FO. On ASDiv, FO remains narrowly best, although $\texttt{Taylor}_2$, $\texttt{Padé}_2$, and $\texttt{Padé}_3$ remain within 0.26 points. Thus, $\beta = 1$ is sufficient for gains in some settings, but we argue that the ZFO methods' largest gains arise from the ability to safely navigate larger learning rates.

Does ZFO require tuning the search bound? To examine performance without selecting β separately for each method, we fix $\beta = 3$, permitting steps up to three times the nominal FO displacement. We retain the corresponding main-experiment learning rates and training configurations. All results use seeds $\{100, 101, 102\}$.

Table 13: Final evaluation accuracy (%, mean $\pm$ standard deviation) with $\beta = 3$ for all ZFO variants. These selected settings illustrate different local-model behaviors. Bold denotes the highest mean in each row. Matching runs are reused from the main experiments; the Phi-2/OpenBookQA Taylor$_2$ entry uses a separate three-seed repeat of the same recorded configuration.

Model / Dataset	FO	Taylor$_2$	Taylor$_3$	Padé$_2$	Padé$_3$
Phi-2 / OpenBookQA	51.20 ± 1.11	$\mathbf{58.87} \pm 12.96$	50.67 ± 0.81	49.40 ± 0.20	52.07 ± 0.23
Gemma / SVAMP	52.00 ± 1.76	$\mathbf{59.56} \pm 0.51$	47.22 ± 2.01	59.00 ± 1.76	58.22 ± 3.69
Llama / GSM8K	4.85 ± 5.47	4.02 ± 3.98	$\mathbf{9.00} \pm 6.68$	5.05 ± 6.94	1.72 ± 0.88
Phi-2 / AsDiv	56.51 ± 18.76	65.42 ± 2.59	55.85 ± 2.55	62.05 ± 7.38	$\mathbf{69.27} \pm 1.92$
Qwen / SVAMP	88.11 ± 0.69	90.78 ± 0.96	86.33 ± 0.33	$\mathbf{90.89} \pm 1.02$	90.33 ± 2.19

Table 13 shows that improvements over FO remain possible with a fixed search bound. Each local model achieves the highest mean in at least one selected setting: Taylor$_2$ on Phi-2/OpenBookQA and Gemma/SVAMP, Taylor$_3$ on Llama/GSM8K, Padé$_2$ on Qwen/SVAMP, and Padé$_3$ on Phi-2/AsDiv. As the smallest expanded bound in our sweep, $\beta = 3$ provides a conservative starting point relative to $\beta = 5$ or 10, allowing exploration beyond the nominal FO step while limiting the extent of extrapolation. Its suitability still depends on the base learning rate, since the search radius scales with $\alpha\beta$. Thus, $\beta = 3$ offers a simple initial choice whose effective search scale should be interpreted alongside the task-specific learning-rate range.

Location of ZFO Steps. We next examine where the ZFO methods select steps within the search interval $I_t = [0, R]$. This diagnostic distinguishes between selecting no update ($s_t = 0$), the largest permitted update ($s_t = R$), and an interior step. Because numerical solutions may lie arbitrarily close to an endpoint, we report both the exact interior frequency, $0 < s_t < R$, and a 1%-trimmed interior frequency, $0.01 < s_t/R < 0.99$. We also report the median normalized step s_t/R. Statistics are computed within each run. Table 14 summarizes the $\beta = 1$ Family-B cohort, consisting of 48 runs across four configurations, and the fixed-$\beta = 3$ cohort, consisting of 108 runs across nine model-dataset pairs.

Table 14 reveals distinct step-selection behaviors across the four ZFO models. $\texttt{Taylor}_3$ makes the most frequent use of the interval interior under both bounds. Its exact interior frequency increases from 65.83% at $\beta = 1$ to 80.77% at $\beta = 3$, while its 1%-trimmed interior frequency rises more substantially from 23.88% to 70.28%. The difference between the two interior measures at $\beta = 1$, together with a median normalized step of 0.00156, indicates that many of its nominally interior

Table 14: **Distribution of ZFO-selected steps over the search interval** $[0, R]$**.** Frequencies are macro-averaged across runs. Exact and 1%-trimmed interior selections are reported separately, and median steps are normalized by R. For $\texttt{Padé}_3$, fallback frequency denotes activation of the quadratic safeguard. The largest interior frequency for each value of β is bold.

β	Method	Selected-step frequency (%)				Median s_t/R	Fallback frequency (%)
		$s_t = 0$	$0 < s_t < R$	$0.01 < s_t/R < 0.99$	$s_t = R$		
1	$\texttt{Taylor}_2$	1.37	40.97	3.76	57.66	1.00000	0.61
	$\texttt{Taylor}_3$	0.86	**65.83**	**23.88**	33.31	0.00156	0.67
	$\texttt{Padé}_2$	43.41	0.00	0.00	56.59	1.00000	0.76
	$\texttt{Padé}_3$	76.27	3.45	2.55	20.27	0.00000	12.49
3	$\texttt{Taylor}_2$	5.04	30.37	15.87	64.59	1.00000	< 0.01
	$\texttt{Taylor}_3$	4.10	**80.77**	**70.28**	15.13	0.02287	0.01
	$\texttt{Padé}_2$	27.65	0.00	0.00	72.35	1.00000	< 0.01
	$\texttt{Padé}_3$	23.41	12.23	6.93	64.36	1.00000	54.95

selections lie close to the zero endpoint. At $\beta = 3$, however, most selections remain interior even after trimming, showing that $\texttt{Taylor}_3$ frequently chooses steps that are meaningfully separated from both boundaries.

The other models rely more heavily on endpoint selections. $\texttt{Taylor}_2$ most often selects the full step $s_t = R$, doing so in 57.66% of updates at $\beta = 1$ and 64.59% at $\beta = 3$. $\texttt{Padé}_2$ never selects an interior point in either cohort and therefore acts empirically as a binary rule between $s_t = 0$ and $s_t = R$. $\texttt{Padé}_3$ also remains largely endpoint-driven. The median of $\texttt{Padé}_3$ shifts from zero at $\beta = 1$ to the upper endpoint at $\beta = 3$. Its fallback frequency simultaneously increases from 12.49% to 54.95%, indicating that the quadratic safeguard becomes substantially more important over the larger search interval. Overall, the results show that the cubic Taylor model uses higher-order information primarily to select intermediate step lengths, whereas the remaining models more often use their local approximations to choose between suppressing and taking the full permitted update.

D Notation

Table 15: Table of notations.

Notation	Definition
$\|\cdot\|$	Euclidean norm.
$\pi_\theta(y \mid x)$	Conditional language model with parameters θ, prompt x, and response y.
θ_t	Model parameters at iteration t, updated as $\theta_{t+1} = \theta_t + s_t z_t$.
S_t	Sampled information held fixed during iteration t, defining the current surrogate and shared across its evaluations.
$F_t(\theta) = F(\theta; S_t)$	Fixed-batch surrogate objective optimized during iteration t.
$\bar{F}_t(\theta)$	Expected surrogate under the sampling distribution at iteration t.
$J(\theta), J^\star$	Global objective used in the stationarity analysis and its upper bound, $J(\theta) \le J^\star$.
g_t	Update direction produced by the base first-order optimizer.
z_t	Normalized search direction, $z_t = g_t/\|g_t\|$ for $g_t \neq 0$.
α	Base learning rate of the first-order optimizer.
β	Search-radius multiplier relative to the standard first-order displacement.
s_{FO}	Length of the standard first-order step, $s_{\mathrm{FO}} = \alpha\|g_t\|$.
R_t, I_t	Search radius $R_t = \beta\alpha\|g_t\|$ and feasible step interval $I_t = [0, R_t]$ in the theoretical formulation.
$\phi_t(s)$	Directional fixed-batch surrogate, $\phi_t(s) = F_t(\theta_t + sz_t)$. The iteration index is omitted when unambiguous.
ϵ	Physical probe radius for the symmetric evaluations $\phi_t(\epsilon)$ and $\phi_t(-\epsilon)$.

Continued on next page

Table 15: Table of notations continued.

Notation	Definition
a_t	Directional slope supplied by the backward pass, $a_t = \phi_t'(0) = \langle \nabla F_t(\theta_t), z_t \rangle$.
b_t, c_t	Finite-difference estimates of the second and third directional derivatives, $b_t = \widehat{\phi_t''}(0)$ and $c_t = \widehat{\phi_t'''}(0)$.
$T_2(s), T_3(s)$	Quadratic and cubic Taylor models constructed from $\phi_t(0)$ and the available directional derivative estimates.
$P_2(s), P_3(s)$	Padé $[1/1]$ and $[1/2]$ local models used by Padé$_2$ and Padé$_3$, respectively.
$M(s)$	Chosen local model, $M \in \{T_2, T_3, P_2, P_3\}$. Distinct from the scalar derivative bound M below.
s_M	Model-maximizing step, $s_M \in \arg\max_{s \in I_t} M(s)$.
s_t	Step applied at iteration t, selected using the local model and any applicable safeguard.
s_ϕ	Maximizer of the empirical directional surrogate, $s_\phi \in \arg\max_{s \in I_t} \phi_t(s)$.
τ_i, n	Shared sampled example τ_i and number of examples n in the CRN analysis.
$\phi_i(s)$	Per-sample directional objective, $\phi_i(s) = F(\theta + sz; \tau_i)$, with $\phi(s) = \frac{1}{n}\sum_{i=1}^n \phi_i(s)$.
$\psi(s)$	Population directional objective, $\psi(s) = \mathbb{E}_\tau[\phi_i(s)]$; at iteration t, $\psi(s) = \bar{F}_t(\theta_t + sz_t)$.
W, M, B, L	Uniform bounds on $\lvert\phi_i''(s)\rvert$, $\lvert\phi_i'''(s)\rvert$, $\lvert\phi_i^{(4)}(s)\rvert$, and $\lvert\phi_i^{(5)}(s)\rvert$, respectively, for all i and $s \in [-\epsilon, R]$.
σ	Sub-Gaussian parameter controlling deviations of the per-sample directional derivatives from their population counterparts in Assumption 4.
δ	Failure-probability parameter in the concentration and step-selection bounds.
$N(s), D(s)$	Numerator and denominator of a Padé model, $P_k(s) = N(s)/D(s)$.
$r(s)$	Padé residual, $r(s) = D(s)\phi(s) - N(s)$.
η	Positive lower bound on the denominator magnitude, $\lvert D(s)\rvert \geq \eta > 0$ for $s \in [0, R]$.
Q, P	Uniform residual derivative bounds: $\lvert r^{(3)}(s)\rvert \leq Q$ for Padé$_2$ and $\lvert r^{(4)}(s)\rvert \leq P$ for Padé$_3$ on $[0, R]$.
$\mathcal{L}_J, \mathcal{L}_{\bar{F}}$	Lipschitz constants of the gradients of J and the expected surrogate $\bar{F}$, respectively.
$\mathbb{E}_t$	Conditional expectation given the history before iteration t.
q_t	Expected directional alignment, $q_t = \mathbb{E}_t[\langle \nabla J(\theta_t), z_t \rangle]$.
μ	Positive alignment constant satisfying $q_t \geq \mu \lVert \nabla J(\theta_t)\rVert$.
Δ_t	Error bound from Lemma 1 on the loss in J relative to the best step in I_t, holding with the stated conditional probability.
$\widetilde{\Delta}_T$	Average expected step-selection error over T iterations, $\widetilde{\Delta}_T = \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}[\Delta_t]$.
K	Combined smoothness constant, $K := \mathcal{L}_J + \mathcal{L}_{\bar{F}} + 2W$.
χ	Positive coefficient in Theorem 3, $\chi := 3\mu^2 - 4\delta > 0$.

Impact Statement. This work improves step-size selection for first-order optimization, making fine-tuning more stable and less sensitive to learning-rate tuning, which can reduce the compute spent on hyperparameter search. As a general-purpose optimization method, ZFO does not introduce application-specific risks beyond those already associated with training and fine-tuning machine learning models.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction state the ZFO framework, its two-probe step-selection mechanism, the theoretical guarantees, and the empirical scope evaluated in the paper.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We discuss limitations and future directions in the discussion and conclusion.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: See §B for complete proofs. The assumptions are stated in our theoretical section.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: We report the configurations and relevant concepts like the surrogate and optimizer used.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: Our code is publicly available at `https://github.com/nizswan/Zeroth-First-Order-Framework`, and all datasets used are publicly available benchmarks.

Guidelines:

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

6. **Experimental setting/details**

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

Answer: [Yes]

Justification: We provide information on the datasets and the training details that are necessary in §C.1.2.

Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: For experiments, we averaged our results over three seeds and reported both the mean and standard deviation.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We provide both the GPU that we trained on and the computational demand of the methods.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research complies with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: We provide an impact statement at the end of our appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: We do not foresee any high risk for misuse of this work.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: We properly cite all assets used in the paper and respect the license and terms of use of said assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: The paper's released code can be found at `https://github.com/nizswan/Zeroth-First-Order-Framework`.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: We used LLMs to better understand related concepts, find similar works, polish the camera ready paper, and assist in code debugging.

    Guidelines:

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