Revisiting Zeroth-Order Optimization: Minimum-Variance Two-Point Estimators and Directionally Aligned Perturbations

Shaocong Ma

August, 2026

Computer Science, University of Maryland, College Park

UNIVERSITY OF MARYLAND

Background: Zeroth-Order Optimization in LLM Fine-tuning

- A single GPU cannot handle backpropagation for entire large models.

Table 1: VRAM Requirements and GPU Configuration

Model Size	First-Order (Full FT)	Est. GPU Setup
OPT-1.3B	$\approx$ 27 GB	1 $\times$ A100
OPT-6.7B	$\approx$ 156 GB	2 $\times$ A100
OPT-13B	$\approx$ 356 GB	4 $\times$ A100
OPT-30B	$\approx$ 633 GB	8 $\times$ A100

Source: Malladi, Sadhika, et al. "Fine-tuning language models with just forward passes." NeurIPS 2023.

- Full-parameter fine-tuning requires storing the **computation graph** or **gradients** for backpropagation.
- **Solution:** Use a gradient-free approach.

Zeroth-Order Optimization (ZOO)

Optimization problem:

$$\min_{x \in \mathbb{R}^d} f(x).$$

Gradient Descent:

$$x' \leftarrow x - \eta \nabla f(x)$$

Notation:

- $f(x)$: The loss function.
- η: The learning rate.

Zeroth-Order Optimization (ZOO)

Optimization problem:

$$\min_{x \in \mathbb{R}^d} f(x).$$

Gradient Descent:

$$x' \leftarrow x - \eta \nabla f(x)$$

Notation:

- $f(x)$: The loss function.
- η: The learning rate.

Memory-Consuming: Deep Neural Network $\frac{\partial f}{\partial x} = \frac{\partial f}{\partial L_N} \frac{\partial L_N}{\partial L_{N-1}} \cdots \frac{\partial L_1}{\partial x}$.

Maintain all intermediate states for backpropagation.

Zeroth-Order Optimization (ZOO)

Core Formula (Two-Point Estimator):

$$\nabla f(x) \approx \hat{\nabla} f(x) = \frac{f(x + \mu v) - f(x)}{\mu} v$$

$$x' \leftarrow x - \eta \hat{\nabla} f(x)$$

Notation:

- $f(x)$: The loss function.
- v: A random perturbation vector (e.g., drawn from a Gaussian distribution $\mathcal{N}(0, I_d)$).
- μ: The perturbation stepsize.

Zeroth-Order Optimization (ZOO)

Core Formula (Two-Point Estimator):

$$\nabla f(x) \approx \hat{\nabla} f(x) = \frac{f(x+\mu v) - f(x)}{\mu} v$$

$$x' \leftarrow x - \eta \hat{\nabla} f(x)$$

Notation:

- $f(x)$: The loss function.
- v: A random perturbation vector (e.g., drawn from a Gaussian distribution $\mathcal{N}(0, I_d)$).
- μ: The perturbation stepsize.

A high-level explanation:

- $\frac{f(x+\mu v)-f(x)}{\mu} > 0 \implies$ Loss increases $\implies$ Move to the opposite direction of v;
- $\frac{f(x+\mu v)-f(x)}{\mu} < 0 \implies$ Loss decreases $\implies$ Move to the direction of v.

Advantages and Challenges of ZOO

Core Advantage:

- Requires only the **Forward Pass**.
- Memory footprint is comparable to **Inference** only.

Advantages and Challenges of ZOO

Core Advantage:

- Requires only the **Forward Pass**.
- Memory footprint is comparable to **Inference** only.

Key Challenges:

- **High Variance:** Gradient estimates are volatile/noisy.

Minimum Variance: Directionally Aligned Perturbation (DAP)

Recap: Zero-Order Gradient Estimator

$$\nabla f(x) \approx \hat{\nabla} f(x) := \frac{f(x + \mu v) - f(x)}{\mu} v$$

where v is typically sampled from a Gaussian distribution.

Q: Is the standard choice of distribution actually optimal?

Minimum Variance: Directionally Aligned Perturbation (DAP)

Recap: Zero-Order Gradient Estimator

$$\nabla f(x) \approx \hat{\nabla} f(x) := \frac{f(x+\mu v) - f(x)}{\mu} v$$

where v is typically sampled from a Gaussian distribution.

Goal: Minimize Estimation Error

Find the optimal distribution V for the perturbation vector v:

$$\begin{aligned} \min_{V} \quad & \mathbb{E}_{v\sim V} \left\| \frac{f(x+\mu v) - f(x)}{\mu} v - \nabla f(x) \right\|^2, \\ \text{s.t.} \quad & \mathbb{E}_{v\sim V}[vv^\top] = I_d. \end{aligned}$$

Minimum Variance: Directionally Aligned Perturbation (DAP)

Recap: Zero-Order Gradient Estimator

$$\nabla f(x) \approx \hat{\nabla} f(x) := \frac{f(x+\mu v) - f(x)}{\mu} v$$

where v is typically sampled from a Gaussian distribution.

Goal: Minimize Estimation Error

Find the optimal distribution V for the perturbation vector v:

$$\begin{aligned} \min_{V} \quad & \mathbb{E}_{v\sim V}\left\| \frac{f(x+\mu v) - f(x)}{\mu} v - \nabla f(x) \right\|^2, \\ \text{s.t.} \quad & \mathbb{E}_{v\sim V}[vv^\top] = I_d. \end{aligned}$$

Optimization Challenges:

- **Functional space:** Optimization is taken over all probability distributions.
- **Constraints with an empty interior:** The empty interior precludes the use of Interior Point Methods.

Minimize Estimation Error (Part I)

Find the optimal distribution V for the perturbation vector v:

$$\begin{aligned} \min_{V} \quad & \mathbb{E}_{v\sim V}\left\| \frac{f(x+\mu v)-f(x)}{\mu} v - \nabla f(x) \right\|^2, \\ \text{s.t.} \quad & \mathbb{E}_{v\sim V}[vv^\top] = I_d. \end{aligned}$$

Taylor Expansion:

$$\begin{aligned} f(x+\mu v) - f(x) &\approx \mu \nabla f(x)^\top v + \mu^2 \cdot \text{Bias} \\ \frac{f(x+\mu v)-f(x)}{\mu} v &\approx v v^\top \nabla f(x) + \mu \cdot \text{Bias} \\ \frac{f(x+\mu v)-f(x)}{\mu} v - \nabla f(x) &\approx (vv^\top - I_d)\nabla f(x) + \underbrace{\mu \cdot \text{Bias}}_{\text{Ignored}} \end{aligned}$$

Definition of Bias: $\mu \cdot \text{Bias} := \mathbb{E}[\hat{\nabla} f(x)] - \nabla f(x)$

Minimize Estimation Error (Part II)

Find the optimal distribution V for the perturbation vector v:

$$\begin{aligned} \min_{V} \quad & \mathbb{E}_{v\sim V}\left\| \frac{f(x+\mu v)-f(x)}{\mu} v - \nabla f(x) \right\|^2, \\ \text{s.t.} \quad & \mathbb{E}_{v\sim V}[vv^\top] = I_d. \end{aligned}$$

Plug in the objective function:

$$\mathbb{E}_{v\sim V}\left\| \frac{f(x+\mu v)-f(x)}{\mu} v - \nabla f(x) \right\|^2 = \mathbb{E}_{v\sim V} \nabla f(x)^\top (vv^\top) \nabla f(x) - \|\nabla f(x)\|^2$$

Minimize Estimation Error (Part II)

Find the optimal distribution V for the perturbation vector v:

$$\begin{aligned} \min_{V} \quad & \mathbb{E}_{v\sim V}\left\| \frac{f(x+\mu v)-f(x)}{\mu} v - \nabla f(x) \right\|^2, \\ \text{s.t.} \quad & \mathbb{E}_{v\sim V}[vv^\top] = I_d. \end{aligned}$$

Plug in the objective function:

$$\mathbb{E}_{v\sim V}\left\| \frac{f(x+\mu v)-f(x)}{\mu} v - \nabla f(x) \right\|^2 = \mathbb{E}_{v\sim V} \nabla f(x)^\top (vv^\top) \nabla f(x) - \|\nabla f(x)\|^2$$

Simplified Objective:

$$\begin{aligned} \min_{V} \quad & \mathbb{E}_{v\sim V} a^\top (vv^\top)^2 a \\ \text{s.t.} \quad & \mathbb{E}_{v\sim V}[vv^\top] = I_d. \end{aligned}$$

Minimize Estimation Error (Part II)

We analytically solve this functional optimization problem.

Derive the lower bound + Tell when the lower bound is achieved

Theorem

Let v be a random vector following the distribution V with $\mathbb{E}_{v\sim V} vv^\top = I_d$ and $a \in \mathbb{R}$ be a fixed vector. Then

$$d\|a\|^2 \le \mathbb{E}_{v\sim V} a^\top (vv^\top)^2 a \le d\|a\|^2 + \frac{\|a\|^2}{2}\left(\rho_V + \sqrt{\rho_V^2 + 4(d-1)\rho_V}\right)$$

where $\rho_V := \mathbb{E}\|v\|^4 - d^2$.

Ma, S. & Huang, H. Revisiting Zeroth-Order Optimization: Minimum-Variance Two-Point Estimators and Directionally Aligned Perturbations. ICLR, 2025. **Spotlight**

Minimize Estimation Error (Part II)

We analytically solve this functional optimization problem.

Derive the lower bound + Tell when the lower bound is achieved

Theorem

Let v be a random vector following the distribution V with $\mathbb{E}_{v\sim V} vv^\top = I_d$ and $a \in \mathbb{R}$ be a fixed vector. Then

$$d\|a\|^2 \le \mathbb{E}_{v\sim V} a^\top (vv^\top)^2 a \le d\|a\|^2 + \frac{\|a\|^2}{2}\left(\rho_V + \sqrt{\rho_V^2 + 4(d-1)\rho_V}\right)$$

where $\rho_V := \mathbb{E}\|v\|^4 - d^2$.

Ma, S. & Huang, H. Revisiting Zeroth-Order Optimization: Minimum-Variance Two-Point Estimators and Directionally Aligned Perturbations. ICLR, 2025. **Spotlight**

What kind of distribution actually achieves this lower bound?

DAPs: Directionally Aligned Perturbation

We analytically solve this functional optimization problem.

(Equality Condition)

- Constant Magnitude Perturbations:
  - $\mathbb{E}_{v\sim V}[vv^\top] = I_d$.
  - $\|v\|$ is fixed.
- Directionally Aligned Perturbations (DAPs):
  - $\mathbb{E}_{v\sim V}[vv^\top] = I_d$.
  - $\nabla f(x)^\top v$ is fixed.

$\Longrightarrow$ Both estimators achieve the minimum variance.

$\Longrightarrow$ DAPs have some nice properties.

Traditional Methods Cannot Identify the Important Directions

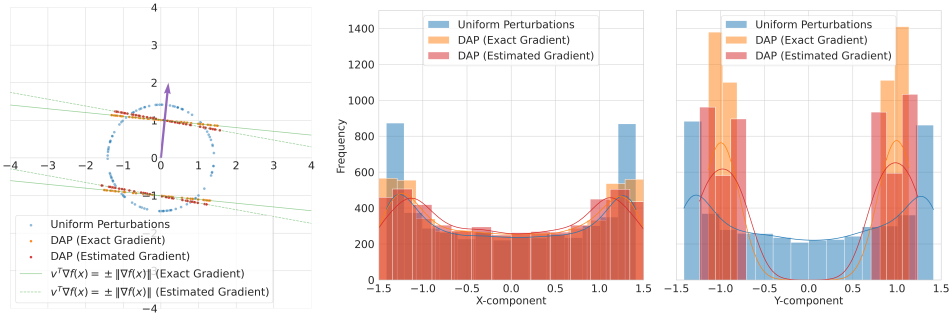

Figure 1: Illustration of the *directional alignment property* of DAP in $d = 2$ with estimating the gradient of $f(x) = x_1^2 + x_2^2$ at $x = \begin{bmatrix} 0.1 & 1 \end{bmatrix}^\top$. Traditional estimator is symmetric, but we need a non-symmetric estimator.

Traditional Methods Cannot Identify the Important Directions

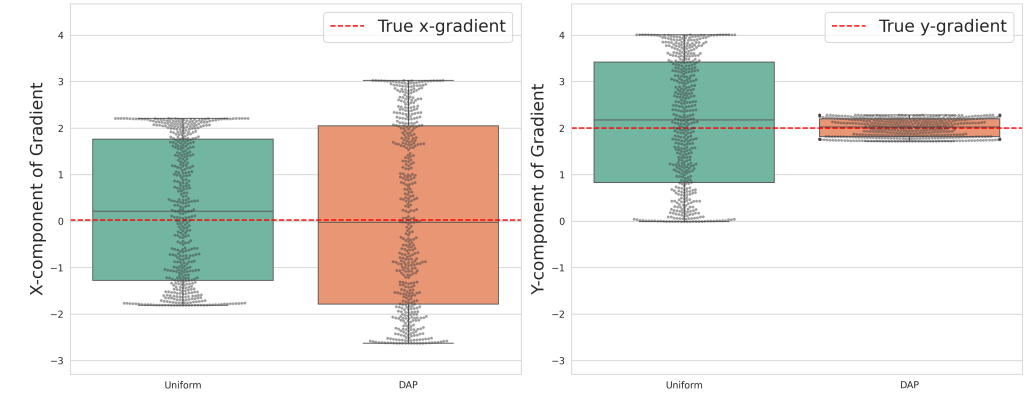

Figure 2: Comparison of gradient estimation performance with estimating the gradient of $f(x) = x_1^2 + x_2^2$ at $x = \begin{bmatrix} 0.1 & 1 \end{bmatrix}^\top$ between uniform random perturbations and DAPs. The non-symmetric estimator is more accurate in the direction with larger gradient.

Applications in LLM Fine-Tuning

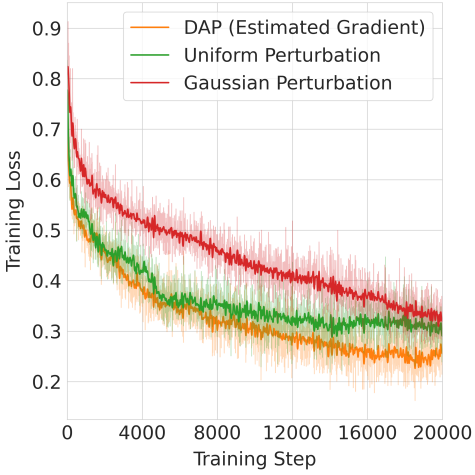

Figure 3: Comparison of training loss curves among different random perturbations on Large Language Model Fine Tuning.

Applications in Scientific Optimization

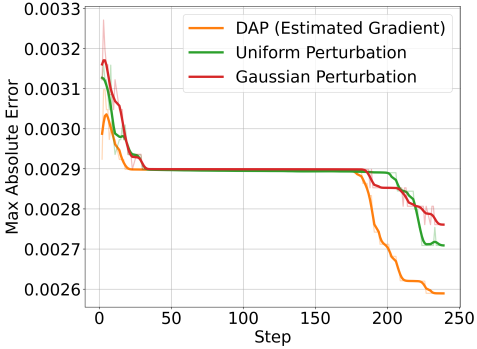

Figure 4: Comparison of training loss curves among different random perturbations on Mesh Optimization for the Physical Numerical Solver.

Summary

- Derived the optimal perturbation distributions for achieving **minimum-variance** zeroth-order gradient estimation:

$$\nabla f(x) \approx \hat{\nabla} f(x) := \frac{f(x + \mu v) - f(x)}{\mu} v.$$

Summary

- Derived the optimal perturbation distributions for achieving **minimum-variance** zeroth-order gradient estimation:

$$\nabla f(x) \approx \hat{\nabla} f(x) := \frac{f(x + \mu v) - f(x)}{\mu} v.$$

- **Main result:** The minimum variance is attained by perturbations satisfying

$$\mathbb{E}[vv^\top] = I_d,$$

together with either

$$\|v\| = \text{constant} \qquad \text{or} \qquad \nabla f(x)^\top v = \text{constant}.$$

The second condition gives the proposed Directionally Aligned Perturbations (DAPs).

Local Geometry-Aware ZOO

- **Geometry identification:** Can we identify important local directions from few function queries?
- **Adaptive perturbations:** Can perturbation distribution V adapt to the local geometry of f?
- **Intrinsic dimension:** Can ZOO exploit low-dimensional local structure in high dimensions?

Thanks!