GRADIENT MATCHING MEETS MUON: EFFICIENT ONLINE MINIBATCH SELECTION FOR LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Training large language models with larger batch sizes is often prohibitive due to memory constraints. Recent works in both language and vision models have shown that training on significantly smaller, carefully selected subsets can achieve performance closer to that of full-batch training, while being efficient. A substantial line of work utilises gradient matching for such minibatch sampling, where approximated last-layer gradients provide a useful proxy for aligning training samples with validation gradients, with importance scores tailored to standard SGD or Adam updates. However, recent spectral optimizers such as Muon, which have demonstrated gains over Adam, require a fundamentally different criterion: Muon orthogonalises the gradient momentum matrix before the parameter update, inducing an anisotropic update. We show that the Euclidean inner-product score used by SGD/Adam-based gradient matching is systematically misaligned with Muon's update direction, while an orthogonalised, momentum-aware score provides a tighter approximation to sample utility. We further study how Muon's orthogonalisation affects coreset quality, particularly sample diversity. Building on these observations, we introduce **Mu**on **S**pectral **M**atching for **A**daptive **S**election (μ-SMAS), an efficient greedy method that linearises the polar-factor orthogonalisation operator ($\mathcal{P}_{\circledcirc}$) via its Fréchet derivative and applies a spectral correction to sample importance scores. Experiments on both large scale finetuning and pretraining regimes show the efficacy of gradient matching via Muon optimizer w.r.t Adam counterparts as evident across multiple subset selection baselines.

1 INTRODUCTION

Despite their effectiveness, pretraining and full-parameter fine-tuning of Large Language Models (LLMs) remain prohibitively resource intensive. Training requires substantial GPU memory for model weights, gradients, activations, and optimizer states. For instance, full-parameter training of Llama 2 7B (Touvron et al., 2023) requires approximately 112 GB just for model states, before accounting for activations and temporary buffers. Since activation memory grows with batch size and sequence length, large-batch training often requires model-state sharding, activation checkpointing, or CPU offloading. This memory bottleneck limits data parallelism and batch sizes, which can improve hardware utilization and reduce gradient variance, although their benefits eventually saturate beyond a problem-dependent critical batch size (McCandlish et al., 2018; Zhang et al., 2025). This has led to a growing line of work on designing more effective minibatch selection strategies for language model training (Nguyen et al., 2025a; Chanda et al., 2026a; Agrawal et al., 2026). A prominent class of these methods formulates minibatch selection as a *gradient matching* problem, aiming to select samples whose gradients best capture the training objective. We formalize this perspective next.

Gradient divergences as a metric for sample selection. Gradient-matching based minibatch subset selection methods such as (Mirzasoleiman et al., 2020; Killamsetty et al., 2021a; Wang et al., 2024; Nguyen et al., 2025a) aims at selecting a subset $\mathcal{S}_t \subseteq \mathcal{B}_t$ at each training step t with $|\mathcal{S}_t| \leq \kappa \ll |\mathcal{B}_t|$ whose (preconditioned) aggregate gradient best aligns with the validation sample gradients:

$$\min_{\mathcal{S}\subseteq\mathcal{B}_t} \mathbb{D}\Big(\sum_{\mathbf{z}\in\mathcal{S}} \mathbf{g}_{\boldsymbol{\theta}_t}(\mathbf{z}) \;||\; \mathbf{G}^{\mathsf{val}}_{\boldsymbol{\theta}_t}\Big), \quad \mathbf{g}_{\boldsymbol{\theta}_t}(\mathbf{z}) := \frac{\boldsymbol{m}_t(\mathbf{z})}{\sqrt{\boldsymbol{v}_t(\mathbf{z})}+\varepsilon}, \quad \begin{matrix} \boldsymbol{m}_{\boldsymbol{\theta}_t}(\mathbf{z}) = \beta_1\, \boldsymbol{m}_{\boldsymbol{\theta}_{t-1}}(\mathbf{z}) + (1-\beta_1)\mathbf{g}_{\boldsymbol{\theta}_t}(\mathbf{z}) \\ \boldsymbol{v}_{\boldsymbol{\theta}_t}(\mathbf{z}) = \beta_2\, \boldsymbol{v}_{\boldsymbol{\theta}_{t-1}}(\mathbf{z}) + (1-\beta_2)\mathbf{g}_{\boldsymbol{\theta}_t}(\mathbf{z})^2 \end{matrix}$$

where $\mathbf{g}_{\boldsymbol{\theta}_t}(\mathbf{z})$ is the raw per-example gradient, and Adam (Kingma & Ba, 2014) maintains exponentially-smoothed moments $\boldsymbol{m}_{\boldsymbol{\theta}_t}(\mathbf{z})$ (first moment) and $\boldsymbol{v}_{\boldsymbol{\theta}_t}(\mathbf{z})$ (second moment). $\mathbb{D}(\cdot \;||\; \cdot)$

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

denotes any typical divergence measure between sample gradients and validation gradient $\mathbf{G}_{\boldsymbol{\theta}_t}^{\mathsf{val}}$ (ℓ_2 norm error (Killamsetty et al., 2021a), ℓ_1 norm (Nguyen et al., 2025b) (Mirzasoleiman et al., 2020), Wasserstein (Chanda et al., 2026a)).

Muon optimizer. Recently, alternative optimizers e.g. Muon (Jordan et al., 2024; Bernstein & Newhouse, 2024) have demonstrated faster convergence than AdamW (Loshchilov & Hutter) on NanoGPT speedrun benchmarks (Jordan et al., 2024). Muon is particularly motivated by spectral update principles that promote stable feature learning (Yang et al., 2023; Pethick et al., 2025). Rather than stepping directly along the momentum-accumulated gradient, Muon first orthogonalizes the gradient matrix and then performs descent in this transformed direction.

Limitations of classical gradient matching under Muon orthogonalization. Existing minibatch subset selection methods (Wang et al., 2024; Killamsetty et al., 2021a;b) derive their marginal-gain approximations assuming a standard first-order update, such as SGD (Kiefer & Wolfowitz, 1952; Hardt et al., 2016) or Adam-preconditioned gradient descent. Under these updates, the selection objective reduces to pairwise inner products between per-example gradients. The same quantities underlie influence estimation and data attribution (Koh & Liang, 2017; Pruthi et al., 2020a; Xia et al., 2024). Spectral optimizers such as Muon and Shampoo (Gupta et al., 2018) break this assumption. They step along a geometry-dependent transformation of the gradient rather than the gradient itself, so the interactions that determine a subset's utility are no longer Euclidean (Davis & Drusvyatskiy, 2025; Chen et al., 2026). Subsets chosen under SGD or Adam assumptions can therefore be suboptimal when training uses a spectral optimizer. For background on the geometric viewpoint, we refer to Edelman et al. (1998); Absil et al. (2008); Roy et al. (2018). We try to answer the following question.

How does minibatch coreset quality change when subset selection is performed in Muon's spectral geometry rather than Adam's diagonally preconditioned Euclidean geometry?

Our contributions. We identify a fundamental mismatch between Euclidean gradient-based coresets and the orthogonalized geometry induced by spectral optimizers such as Muon. We introduce the first geometry-aware coreset framework that explicitly incorporates this optimizer-induced geometry into sample selection.

2 Related Work

Adaptive and spectral gradient methods. Coordinatewise adaptive methods such as AdaGrad (Duchi et al., 2011) and Adam (Kingma & Ba, 2014) precondition each parameter independently by an empirical estimate of its second moment. Muon (Jordan et al., 2024) instead treats each weight matrix as a single geometric object: Bernstein & Newhouse (2024) show that its orthogonalized update implements steepest descent under the spectral-norm ball, a view later placed within a general linear-minimization-oracle (LMO) / Frank–Wolfe framework for norm-constrained updates (Pethick et al., 2025) and extended to a broader family of spectral (Lion-$\mathcal{K}$) duals (Chen et al., 2026). Subsequent work has shown that Shampoo (Gupta et al., 2018) and SOAP (Vyas et al., 2025) reduce to Muon-like updates under simplifying assumptions (Shah et al., 2025), and that orthogonalization improves conditioning by shifting the effective learning-rate bottleneck from the largest singular value of the momentum to its average scale (Nguyen et al., 2026). A parallel line of convergence analysis situates Muon relative to normalized SGD (Li & Hong, 2025) and non-Euclidean trust-region methods (Kovalev, 2025), and Davis & Drusvyatskiy (2025) characterize when such spectral updates provably help over Euclidean ones, tying the benefit to anisotropic (low-stable-rank / spiked) gradient spectra of the kind documented empirically in transformer training (Sonthalia et al., 2026; Gur-Ari et al., 2018).

Gradient-matching coreset and data selection. Online coreset selection methods such as GLISTER (Killamsetty et al., 2021b), GRAD-MATCH (Killamsetty et al., 2021a), and GREATS (Wang et al., 2024) score candidate examples by the alignment of their (Adam-preconditioned) gradient with a validation gradient, and CoLM (Nguyen et al., 2025a) clusters in the same Adam-normalized space. These methods, together with influence-function and data-attribution approaches (Koh & Liang, 2017; Pruthi et al., 2020a; Xia et al., 2024), implicitly assume that the optimizer consumes the selected gradient through a Euclidean (or diagonally-preconditioned) update rule.To our knowledge, existing coreset methods do not account for the anisotropic, matrix-valued geometry induced by spectral optimizers such as Muon, which our framework explicitly addresses.

3 Background and Notation

Standard Notation. The n-dimensional non-negative orthant is $\mathbb{R}^n_+$. Sets are denoted by calligraphic letters (e.g., $\mathcal{S}$). Vectors are lower-case boldface (e.g., $\mathbf{w}$); matrices are upper-case boldface (e.g., $\mathbf{G}$). The standard (Euclidean) inner product of vectors $\mathbf{x}, \mathbf{y}$ is $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y}$. For a vector $\mathbf{w} \geq 0$, $\mathrm{supp}(\mathbf{w}) = \{j : \mathbf{w}_j > 0\}$. The standard (Euclidean) inner product for matrices $\mathbf{A}, \mathbf{B}$ is the Frobenius inner product $\langle \mathbf{A}, \mathbf{B} \rangle_F = \mathrm{Tr}(\mathbf{A}^\top \mathbf{B})$, where $\mathrm{Tr}(\cdot)$ denotes the trace. The Frobenius norm of $\mathbf{A}$ is $\|\mathbf{A}\|_F = \sqrt{\mathrm{Tr}(\mathbf{A}^\top \mathbf{A})}$. For a square matrix $\mathbf{A}$, $\mathrm{symm}(\mathbf{A}) = (\mathbf{A} + \mathbf{A}^\top)/2$ denotes its symmetric part. The probability simplex is $\Delta_n = \{\mathbf{x} \in \mathbb{R}^n_+ : \sum_i x_i = 1\}$. The operator (spectral) norm is $\|\mathbf{A}\|_{\mathrm{op}} = \sup_{\|\mathbf{x}\|=1} \|\mathbf{A}\mathbf{x}\|$.

Polar map and Fréchet derivatives. For $\mathbf{X} \in \mathbb{R}^{d \times r}$ with polar decomposition $\mathbf{X} = \mathbf{Q}\mathbf{P}$, we write $\mathcal{P}_{\circledcirc}(\mathbf{X}) := \mathbf{Q}$, which equals $\mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1/2}$ when $\mathbf{X}$ has full column rank. Muon approximates $\mathcal{P}_{\circledcirc}$ by Newton–Schulz iterations. For a map F on matrices, $\mathscr{D}F(\mathbf{X})[\mathbf{H}]$ denotes its Fréchet derivative at $\mathbf{X}$ in direction $\mathbf{H}$, i.e. the linear map satisfying $F(\mathbf{X}+\mathbf{H}) - F(\mathbf{X}) - \mathscr{D}F(\mathbf{X})[\mathbf{H}] = o(\|\mathbf{H}\|)$. When it exists, it coincides with $\frac{d}{dt} F(\mathbf{X} + t\mathbf{H})\big|_{t=0}$. For maps given by a formula, we write, e.g., $\mathscr{D}(\mathbf{T}^s)(\mathbf{X})[\mathbf{H}]$ for the derivative of $\mathbf{T} \mapsto \mathbf{T}^s$ at $\mathbf{X}$.

Training Preliminaries. We denote the training model parameters as $\boldsymbol{\theta} \in \mathbb{R}^{d \times r}$. Further, let $\mathcal{D}^{\mathrm{tr}}$ denote the training dataset and $\ell : \mathcal{Z} \times \boldsymbol{\theta} \to \mathbb{R}$ the per-example loss, where $\mathcal{Z}$ denotes the space of training instances. We assume access to validation samples $\mathcal{D}^{\mathrm{val}}$, which is a standard practice as shown in previous works (Agrawal et al., 2026; Wang et al., 2026; Chanda et al., 2026a; Wang et al., 2024). We denote the gradient of a training sample w.r.t model parameter at training step t $\boldsymbol{\theta}_t$ as $\boldsymbol{g}_{\boldsymbol{\theta}_t} = \nabla_{\boldsymbol{\theta}_t} \ell(\boldsymbol{\theta}_t; \boldsymbol{z})$

Submodularity ratio. Let $f : 2^{\mathcal{V}} \to \mathbb{R}$ be a set function over ground set $\mathcal{V}$. For any subset $\mathcal{S} \subseteq \mathcal{V}$ and element $e \in \mathcal{V} \setminus \mathcal{S}$, the marginal gain is $\Delta f(e \mid \mathcal{S}) = f(\mathcal{S} \cup \{e\}) - f(\mathcal{S})$. The function f is *submodular* if $\Delta f(e \mid \mathcal{A}) \geq \Delta f(e \mid \mathcal{B})$ for all $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V}$ and $e \notin \mathcal{B}$. The *submodularity ratio* (Das & Kempe, 2018) captures how close a non-submodular function is to being submodular:

$$\gamma_{\mathcal{L},\kappa}(f) = \min_{\mathcal{S} \subseteq \mathcal{L},\ \mathcal{A}:\, |\mathcal{A}| \leq \kappa,\ \mathcal{A} \cap \mathcal{S} = \emptyset} \frac{\sum_{u \in \mathcal{A}} \Delta f(u \mid \mathcal{S})}{f(\mathcal{S} \cup \mathcal{A}) - f(\mathcal{S})},$$

with $0/0 := 1$. If $\gamma > 0$, the greedy algorithm for $\max_{|\mathcal{S}| \leq \kappa} f(\mathcal{S})$ achieves a $(1 - e^{-\gamma})$ approximation to the optimum (Nemhauser et al., 1978; Das & Kempe, 2018).

RSC and RSM. A differentiable function $l : \mathbb{R}^d \to \mathbb{R}$ is *restricted strong concave* (RSC) with parameter c_Ω and *restricted smooth* (RSM) with parameter C_Ω on a domain $\Omega \subset \mathbb{R}^d \times \mathbb{R}^d$ if $\forall\, (\mathbf{x}, \mathbf{y}) \in \Omega$,

$$l(\mathbf{x}) - l(\mathbf{y}) + \langle \nabla l(\mathbf{x}),\, \mathbf{y} - \mathbf{x} \rangle \in [\tfrac{c_\Omega}{2} \|\mathbf{y} - \mathbf{x}\|_2^2,\, \tfrac{C_\Omega}{2} \|\mathbf{y} - \mathbf{x}\|_2^2].$$

RSC and RSM have been studied in the context of weakly submodular functions (Elenberg et al., 2018; Gurumoorthy et al., 2019). Let $\Omega_\kappa := \{(\mathbf{x}, \mathbf{y}) : \|\mathbf{x}\|_0 \leq \kappa,\ \|\mathbf{y}\|_0 \leq \kappa,\ \mathbf{x}, \mathbf{y} \geq 0\}$, and let c_κ and C_κ denote the RSC and RSM parameters of l on Ω_κ. Since $\Omega_k \subseteq \Omega_\kappa$ whenever $k \leq \kappa$, it is easily verified that $c_k \geq c_\kappa$ and $C_k \leq C_\kappa$. Finally, define $\widetilde{\Omega} := \{(\mathbf{x}, \mathbf{y}) : \|\mathbf{x} - \mathbf{y}\|_0 \leq 1\}$, with associated smoothness parameter $\widetilde{C}_1$.

Updates using Muon Optimizer. For each Muon-optimized (Kosson et al., 2024; Bernstein & Newhouse, 2024) weight matrix, the optimizer maintains a Nesterov-style momentum buffer $\boldsymbol{m}_{\boldsymbol{\theta}_t}$. At iteration t, given a mini-batch $\mathcal{B}_t$ and its batch gradient $\overline{\boldsymbol{g}}_{\boldsymbol{\theta}_t} := \frac{1}{|\mathcal{B}_t|} \sum_{\mathbf{z} \in \mathcal{B}_t} \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}_t; \mathbf{z})$, the Muon dynamics take the form

$$\boldsymbol{m}_{\boldsymbol{\theta}_t} = \beta\, \boldsymbol{m}_{\boldsymbol{\theta}_{t-1}} + \overline{\boldsymbol{g}}_{\boldsymbol{\theta}_t}, \tag{1}$$

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t\, \mathcal{P}_{\circledcirc}\big(\boldsymbol{m}_{\boldsymbol{\theta}_t}\big), \tag{2}$$

where $\beta \in (0, 1)$ is the momentum parameter, and $\mathcal{P}_{\circledcirc}(\bullet)$ denotes the polar operator which returns the orthogonal factor in the polar decomposition (equivalently, the closest semi-orthogonal matrix in Frobenius norm) (Amsel et al., 2025) and (Higham, 2008, Chapter 8).

Definition 1 (Newton–Schulz polynomial (Kim & Oh, 2026)). *For $\tau \in \mathbb{N}$, let p_τ be the degree-τ Taylor polynomial of $\lambda^{-1/2}$ about $\lambda = 1$, i.e., $p_\tau^{(s)}(1) = \frac{\mathrm{d}^s}{\mathrm{d}\lambda^s} \lambda^{-1/2}\big|_{\lambda=1}$ for $s = 1, \ldots, \tau$. It has the closed form*

$$p_\tau(\lambda) = \sum_{s=0}^{\tau} c_s\, (1 - \lambda)^s, \qquad c_s := \frac{(2s)!}{4^s\, (s!)^2} > 0. \tag{3}$$

For $\mathbf{A} \in \mathbb{R}^{d\times d}$, define $p_\tau(\mathbf{A})$ by the same polynomial, $p_\tau(\mathbf{A}) := \sum_{s=1}^{\tau} c_s\,(\mathbf{I}_d - \mathbf{A})^s$.

Given $\mathbf{X}_0 \in \mathbb{R}^{d\times r}$, define the degree-$\tau$ Newton–Schulz iterates by $\mathbf{X}_{i+1} := p_\tau(\mathbf{X}_i\mathbf{X}_i^\top)\mathbf{X}_i$ for $i = 0, \ldots, q-1$, where q denotes the Newton-Schulz steps. Since p_τ is a polynomial, equivalently $\mathbf{X}_{i+1} = \mathbf{X}_i\, p_\tau(\mathbf{X}_i^\top\mathbf{X}_i)$. The resulting q-step orthogonalization map (parameterized by the τ degree polynomial $p_\tau(\bullet)$ as defined above) starting from $\mathbf{X}_0$ is given by $\mathcal{P}^{(q)}_{\circledcirc,\tau}(\mathbf{X}_0) := \mathbf{X}_q$. For appropriately scaled $\mathbf{X}_0$, iterating pushes each nonzero singular value toward 1. If $\mathbf{X}_0$ has full column rank, then $\mathbf{X}_q$ converges to the orthogonal polar factor as $q \to \infty$; if $\mathbf{X}_0$ is rank deficient, zero singular values remain zero and the iterates converge to the canonical partial polar factor on the nonzero subspace. In practice, a small fixed q (e.g., $q = 3$) and $\tau = 5$ is commonly used (Kim & Oh, 2026).

4 Problem Formulation

At each step $t \in [T]$, the model receives a mini-batch $\mathcal{B}_t \subseteq \mathcal{D}^{\text{tr}}$ and we assume access to a small validation set $\mathcal{D}^{\text{val}}$ from which a validation mini-batch $\mathcal{B}_t^{\text{val}} \subseteq \mathcal{D}^{\text{val}}$ is drawn. We denote the per-example gradient by $\boldsymbol{g}(\boldsymbol{\theta}_t; \mathbf{z}) := \nabla_{\boldsymbol{\theta}}\ell(\mathbf{z}, \boldsymbol{\theta}_t)$ and the validation gradient by $\boldsymbol{g}_t^{\text{val}} := \boldsymbol{g}(\boldsymbol{\theta}_t; \mathbf{z}_{\text{val}})$.

MiniBatch Coreset Selection. Motivated by recent works (Wang et al., 2024; Koh & Liang, 2017) our goal is to select, at each step t, a subset $\mathcal{S}_t \subseteq \mathcal{B}_t$ with $|\mathcal{S}_t| \le \kappa \ll |\mathcal{B}_t|$ that maximizes a training step dependent utility function $\mathcal{U}_t(\cdot) : \mathcal{S} \mapsto \mathbb{R}$ while adhering to the cardinality constraints κ. Formally, we define the selection problem as:

$$\max_{\mathcal{S}\subseteq\mathcal{B}_t\,|\mathcal{S}|\le\kappa} \mathcal{U}_t(\mathcal{S}) \tag{4}$$

where κ is the per-step selection budget. We hereby discuss the design of $\mathcal{U}_t(\cdot)$.

Design choices of the utility function $\mathcal{U}_t(\cdot)$. Recent works such as (Nguyen et al., 2025a; Agrawal et al., 2026) define the utility of a training subset by measuring the alignment between training and validation gradients, typically using L_1 or L_2 distance. In contrast, (Chanda et al., 2026a) formulates subset selection through optimal transport, using the Wasserstein distance (Villani, 2021) to capture a probabilistic correspondence between training and validation gradients. Another line of work (Wang et al., 2026; 2024; Chanda et al., 2026b) instead uses validation-loss dynamics as a proxy for the utility of candidate subsets, thereby evaluating their impact through the evolution of validation performance.

Validation-Guided Gradient Matching Previous works such as (Killamsetty et al., 2021a; Xia et al., 2024) uses validation gradients as a proxy for selecting training samples by matching training samples' gradients w.r.t validation samples' gradients. The utility function $\mathcal{U}_t(\mathcal{S})$ is designed to measure the incremental contribution of a candidate subset by trading off two factors: consistency with the validation signal (capturing relevance) and diversity within the subset (reducing redundancy and mitigating gradient interference) (Wang et al., 2024; Nguyen et al., 2025a; Chanda et al., 2026a).

In specifying $\mathcal{U}_t(\mathcal{S})$, we employ a validation-guided formulation as shown in (Wang et al., 2024). More specifically, we want to maximize the expected reduction in validation loss after a single gradient step derived from $\mathcal{S}_t$. For a validation sample $\mathbf{z}_{\text{val}} \in \mathcal{B}_t^{\text{val}}$, the utility of a subset $\mathcal{S} \subseteq \mathcal{B}_t$ is:

$$\mathcal{U}_t(\mathcal{S};\, \mathbf{z}_{\text{val}}) := \ell(\mathbf{z}_{\text{val}}; \boldsymbol{\theta}_t) - \ell\Big(\mathbf{z}_{\text{val}};\, \tilde{\boldsymbol{\theta}}_{t+1}(\mathcal{S})\Big) \tag{5}$$

where $\tilde{\boldsymbol{\theta}}_{t+1}(\mathcal{S}) = \boldsymbol{\theta}_t - \eta_t\left(\sum_{z\in\mathcal{S}} \boldsymbol{g}(\boldsymbol{\theta}_t; z)\right)$ denotes the one-step gradient update over the samples in the current set $\mathcal{S}$.

Given the partially constructed minibatch at time step t, $\mathcal{S} := \{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_{i-1}\}$, we compute the marginal gain of the utility function as follows:

$$\begin{aligned}
\boldsymbol{\Delta}\mathcal{U}_t(\mathbf{z}_i \mid \mathcal{S}) &= \mathcal{U}_t(\mathcal{S}\cup\{\mathbf{z}_i\}) - \mathcal{U}_t(\mathcal{S}) \\
&= \ell(\mathbf{z}_{\text{val}}; \tilde{\boldsymbol{\theta}}_{t+1}(\mathcal{S})) - \ell(\mathbf{z}_{\text{val}}; \tilde{\boldsymbol{\theta}}_{t+1}(\mathcal{S}\cup\{\mathbf{z}_i\})) \quad \triangleright \text{ Equation (5)} \\
&\approx \eta_t \langle \boldsymbol{g}_{\boldsymbol{\theta}_t}(\mathbf{z}_i),\, \boldsymbol{g}_{\tilde{\boldsymbol{\theta}}_{t+1}(\mathcal{S})}(\mathbf{z}_{\text{val}})\rangle_F \qquad (6) \\
&\approx \underbrace{\eta_t \langle \boldsymbol{g}_{\boldsymbol{\theta}_t}(\mathbf{z}_i),\, \boldsymbol{g}_{\boldsymbol{\theta}_t}(\mathbf{z}_{\text{val}})\rangle_F}_{\bullet} - \underbrace{\tfrac{\eta_t^2}{2} \langle \boldsymbol{g}_{\boldsymbol{\theta}_t}(\mathbf{z}_i),\, \boldsymbol{\mathcal{H}}_{\mathbf{z}_{\text{val}}}(\boldsymbol{\theta}_t)\big(\textstyle\sum_{\mathbf{z}\in\mathcal{S}} \boldsymbol{g}_{\boldsymbol{\theta}_t}(\mathbf{z})\big)\rangle_F}_{\bullet} \qquad (7)
\end{aligned}$$

Equation (6) and Equation (7) follow from first-order Taylor expansions in η_t of $\ell(\cdot)$ and its gradient, together with the one-step relation $\tilde{\boldsymbol{\theta}}_{t+1}(\mathcal{S}\cup\{\mathbf{z}_i\}) = \tilde{\boldsymbol{\theta}}_{t+1}(\mathcal{S}) - \eta_t\, \boldsymbol{g}_{\boldsymbol{\theta}_t}(\mathbf{z}_i)$.

Remark 1. The first term in Equation (7) (●) quantifies the contribution of $\mathbf{z}_i$ relative to the validation example $\mathbf{z}_{\text{val}}$, effectively measuring how strongly the gradient induced by $\mathbf{z}_i$ reduces the validation loss. This interpretation is closely aligned with influence function–based perspectives (Koh & Liang, 2017; Pruthi et al., 2020b). The second term (●) captures the Hessian-weighted similarity between the current instance and the remaining samples in the batch $\mathcal{B}_t \setminus i$, thereby discouraging redundancy and encouraging diversity in the selection of $\mathcal{S}$.

5 Proposed Approach

We begin from Muon's update (Equations (1) and (2)). For a selected subset $\mathcal{S}$, the resulting momentum and the corresponding pre-projection next iterate can be written as:

$$\boldsymbol{m}_{\boldsymbol{\theta}_t}(\mathcal{S}) = \beta\, \boldsymbol{m}_{\boldsymbol{\theta}_{t-1}} + \sum_{\mathbf{z}\in\mathcal{S}} \boldsymbol{g}_{\boldsymbol{\theta}_t}(\mathbf{z}) \qquad (8) \qquad\qquad \tilde{\boldsymbol{\theta}}_{t+1}(\mathcal{S}) = \boldsymbol{\theta}_t - \eta_t\, \mathcal{P}_{\circledcirc}\big(\boldsymbol{m}_{\boldsymbol{\theta}_t}(\mathcal{S})\big). \qquad (9)$$

Now for a single candidate point $\mathbf{z}_i$, the momentum changes by its per-example contribution $\boldsymbol{m}_{\boldsymbol{\theta}_t}(\mathcal{S}) + \boldsymbol{g}_{\boldsymbol{\theta}_t}(\mathbf{z}_i)$, which in turn perturbs the next iterate: $\tilde{\boldsymbol{\theta}}_{t+1}(\mathcal{S}\cup\{\mathbf{z}_i\}) = \boldsymbol{\theta}_t - \eta_t\, \mathcal{P}_{\circledcirc}\big(\boldsymbol{m}_{\boldsymbol{\theta}_t}(\mathcal{S}) + \boldsymbol{g}_{\boldsymbol{\theta}_t}(\mathbf{z}_i)\big)$.

Fréchet Differential of Polar. To obtain a tractable marginal gain, we linearise the polar operator $\mathcal{P}_{\circledcirc}(\cdot)$ around $\boldsymbol{m}_{\boldsymbol{\theta}_t}(\mathcal{S})$ in the direction $\boldsymbol{g}_{\boldsymbol{\theta}_t}(\mathbf{z}_i)$, using the Fréchet derivative $\mathscr{D}\mathcal{P}_{\circledcirc}(\boldsymbol{m}_{\boldsymbol{\theta}_t}(\mathcal{S}))$ (Gawlik & Leok, 2016, Eq. (1)). For notational convenience, define $\boldsymbol{g}^{\langle\mu\rangle}_{\boldsymbol{\theta}_t}(\mathbf{z}_i) \coloneqq \mathscr{D}\mathcal{P}_{\circledcirc}(\boldsymbol{m}_{\boldsymbol{\theta}_t}(\mathcal{S}))[\boldsymbol{g}_{\boldsymbol{\theta}_t}(\mathbf{z}_i)]$ (see Lemma 5.2).

$$\mathcal{P}_{\circledcirc}\big(\boldsymbol{m}_{\boldsymbol{\theta}_t}(\mathcal{S}) + \boldsymbol{g}_{\boldsymbol{\theta}_t}(\mathbf{z}_i)\big) = \mathcal{P}_{\circledcirc}\big(\boldsymbol{m}_{\boldsymbol{\theta}_t}(\mathcal{S})\big) + \boldsymbol{g}^{\langle\mu\rangle}_{\boldsymbol{\theta}_t}(\mathbf{z}_i) + o(\|\boldsymbol{g}_{\boldsymbol{\theta}_t}(\mathbf{z}_i)\|). \qquad (10)$$

Substituting Equation (10) into the next iterate update and comparing with Equation (9) yields the approximate parameter perturbation $\tilde{\boldsymbol{\theta}}_{t+1}(\mathcal{S}\cup\{\mathbf{z}_i\}) \approx \tilde{\boldsymbol{\theta}}_{t+1}(\mathcal{S}) - \eta_t\, \boldsymbol{g}^{\langle\mu\rangle}_{\boldsymbol{\theta}_t}(\mathbf{z}_i)$

Marginal Utility Gain. The resulting (closed-form) approximation is

$$\boldsymbol{\Delta}\mathcal{U}^{\langle\mu\rangle}_t(\mathbf{z}_i \mid \mathcal{S}) \;\approx\; \eta_t\, \langle \boldsymbol{g}^{\langle\mu\rangle}_{\boldsymbol{\theta}_t}(\mathbf{z}_i),\, \boldsymbol{g}_{\boldsymbol{\theta}_t}(\mathbf{z}_{\text{val}})\rangle_F - \tfrac{\eta_t^2}{2}\, \langle \boldsymbol{g}^{\langle\mu\rangle}_{\boldsymbol{\theta}_t}(\mathbf{z}_i),\, \boldsymbol{\mathcal{H}}_{\mathbf{z}_{\text{val}}}(\boldsymbol{\theta}_t)\, \mathcal{P}_{\circledcirc}\big(\boldsymbol{m}_{\boldsymbol{\theta}_t}(\mathcal{S})\big)\rangle_F\,. \qquad (11)$$

The derivation is deferred to Appendix Section C.1; here we only note how the remaining validation gradient is expressed in terms of quantities at $\boldsymbol{\theta}_t$. Specifically, $\tilde{\boldsymbol{\theta}}_{t+1}(\mathcal{S}) - \boldsymbol{\theta}_t = -\eta_t\, \mathcal{P}_{\circledcirc}\big(\boldsymbol{m}_{\boldsymbol{\theta}_t}(\mathcal{S})\big)$, and we use a first-order Taylor expansion of $\boldsymbol{g}_{\tilde{\boldsymbol{\theta}}_{t+1}(\mathcal{S})}(\mathbf{z}_{\text{val}})$ around $\boldsymbol{\theta}_t$ (see Appendix Section C.1).

Weakly submodular reformulation. Following Agrawal et al. (2026), the first-order approximation in Equation (7) can be cast as maximizing a (weakly) submodular set function obtained by optimizing a concave quadratic utility over nonnegative weights supported on the chosen subset. We define the Muon utility set function $f_{\langle\mu\rangle}(\mathcal{S})$ by optimizing a concave quadratic utility over nonnegative weights supported on $\mathcal{S}$:

$$\max_{\mathcal{S}\subseteq\mathcal{B}_t,\ |\mathcal{S}|\le\kappa} f_{\langle\mu\rangle}(\mathcal{S}) \;\Big(:= \max_{\mathbf{w}\in\mathbb{R}_+^{|\mathcal{B}_t|},\ \text{supp}(\mathbf{w})\subseteq\mathcal{S}} \mathcal{U}^{\langle\mu\rangle}_t(\mathbf{w}, \mathcal{B}_t; \mathbf{z}_{\text{val}})\Big). \qquad (12)$$

where

$$\mathcal{U}^{\langle\mu\rangle}_t(\mathbf{w}, \mathcal{B}_t; \mathbf{z}_{\text{val}}) = \eta_t \langle \mathbf{w}, \boldsymbol{\mu}^{\langle\mu\rangle}_{\boldsymbol{\theta}_t}\rangle - \tfrac{\eta_t^2}{2}\, \langle \mathbf{w}, \mathbf{K}^{\langle\mu\rangle}_{\boldsymbol{\theta}_t}\mathbf{w}\rangle, \qquad (13)$$

with $\boldsymbol{\mu}^{\langle\mu\rangle}_{\boldsymbol{\theta}_t}[i] = \langle \boldsymbol{g}^{\langle\mu\rangle}_{\boldsymbol{\theta}_t}(\mathbf{z}_i), \boldsymbol{g}_{\boldsymbol{\theta}_t}(\mathbf{z}_{\text{val}})\rangle_F + \frac{\eta_t}{2}\big[\mathbf{K}^{\langle\mu\rangle}_{\boldsymbol{\theta}_t}\big]_{ii}$ and $\big[\mathbf{K}^{\langle\mu\rangle}_{\boldsymbol{\theta}_t}\big]_{ij} = \langle \boldsymbol{g}^{\langle\mu\rangle}_{\boldsymbol{\theta}_t}(\mathbf{z}_i), \boldsymbol{g}^{\langle\mu\rangle}_{\boldsymbol{\theta}_t}(\mathbf{z}_j)\rangle_F$.

The resulting Muon optimizer based set function $f_{\langle\mu\rangle}$ is stated alongside its Adam counterpart $f_{\langle\varepsilon\rangle}$, which also compares their induced geometries.

5.1 Algorithm

Here, we showcase the main algorithm for selecting minibatch coresets under Muon optimizer via solving the inner maximization objective of Equation (13).

Orthogonal Matching Pursuit Algorithm. Since the objective Equation (13) is concave quadratic in terms of $\mathbf{w}$ and since theoretically it can be proven that the induced set function $f_{\langle\mu\rangle}(\cdot)$ as defined in Equation (12) is weakly-submodular (as shown in Theorem 5.1), we can employ an orthogonal matching pursuit algorithm to solve for the same.

The orthogonal matching pursuit algorithm starts by maintaining a support set $\mathfrak{R}_i$ corresponding to each iteration i where $i = [1, \kappa]$. At each iteration step, the gradient residual is computed w.r.t $\mathbf{w}$ i.e. $\nabla_{\mathbf{w}} \mathcal{U}_t(\mathbf{w}_{\mathfrak{R}_i})$, and the sample corresponding to the highest gradient residual $\mathbf{x}^\star$ is selected at each step and added to the i-th iterate support set $\mathfrak{R}_i \leftarrow \mathfrak{R}_{i-1} \cup \{\mathbf{x}^\star\}$. This process is continued till κ iterations by repeatedly adding the candidate with the largest residual, re-solving the nonnegative weight subproblem via accelerated projected gradient ascent (APGA), and updating the gradient of the objective with respect to the weights.

Algorithm 1: Orthogonal Matching Pursuit

Require: Objective $\mathcal{U}_t(\cdot)$, ground set $\mathcal{B}_t$, sparsity budget κ
Ensure: Selected support $\mathfrak{R}_\kappa$
Initialise $\mathfrak{R}_0 \leftarrow \emptyset$, $\mathbf{w}_{\mathfrak{R}_0} \leftarrow \mathbf{0}$
$\mathbf{g} \leftarrow \nabla_{\mathbf{w}} \mathcal{U}_t(\mathbf{w}_{\mathfrak{R}_0}) = \eta_t \boldsymbol{\mu}_{\boldsymbol{\theta}_t}^{\langle \mu \rangle} - \eta_t^2 \mathbf{K}_{\boldsymbol{\theta}_t}^{\langle \mu \rangle} \mathbf{w}$
for $i \leftarrow 1$ **to** κ **do**
  $\mathbf{x}^\star \leftarrow \arg\max_{\mathbf{x} \in \mathcal{B}_t \setminus \mathfrak{R}_{i-1}} \mathbf{g}_{\mathbf{x}}$
  $\mathfrak{R}_i \leftarrow \mathfrak{R}_{i-1} \cup \{\mathbf{x}^\star\}$
  $\mathbf{w}_{\mathfrak{R}_i} \leftarrow \arg\max_{\mathbf{w}:\, \mathrm{supp}(\mathbf{w}) \subseteq \mathfrak{R}_i} \mathcal{U}_t(\mathbf{w})$
  $\mathbf{g} \leftarrow \nabla_{\mathbf{w}} \mathcal{U}_t(\mathbf{w}_{\mathfrak{R}_i})$
return $\mathfrak{R}_\kappa$

5.2 Theoretical Results

We provide here some of the important theoretical findings.

Proposition 5.1. *The finite-step Newton–Schulz map $\mathcal{P}^{(q)}_{\circledcirc,\tau}(\bullet)$ defined in Section 3 is a matrix polynomial, hence Fréchet differentiable at every $\mathbf{X}_0 \in \mathbb{R}^{d\times r}$ (including rank-deficient $\mathbf{X}_0$).*

Proof sketch. The finite-step Newton–Schulz map is a matrix polynomial (finite composition of polynomial operations), hence Fréchet differentiable everywhere; see Appendix Section C.3 for details. □

Lemma 5.1 (Fréchet derivative of the finite Newton–Schulz map)**.** *For a perturbation $\boldsymbol{\rho} \in \mathbb{R}^{d\times r}$, define $\mathbf{Y}_i := \mathscr{D}\mathbf{X}_i(\mathbf{X}_0)[\boldsymbol{\rho}]$, with $\mathbf{Y}_0 := \boldsymbol{\rho}$. Let $\mathbf{A}_k := \mathbf{X}_k\mathbf{X}_k^\top$ and $\mathbf{B}_k := \mathbf{Y}_k\mathbf{X}_k^\top + \mathbf{X}_k\mathbf{Y}_k^\top$. Then*

$$\mathbf{Y}_{k+1} = \mathscr{D}p_\tau(\mathbf{A}_k)[\mathbf{B}_k]\mathbf{X}_k + p_\tau(\mathbf{A}_k)\mathbf{Y}_k, \tag{14}$$

where $\mathscr{D}p_\tau(\mathbf{A})[\mathbf{B}] = -\sum_{s=1}^{\tau} c_s \sum_{j=0}^{s-1} (\mathbf{I}-\mathbf{A})^j \mathbf{B} (\mathbf{I}-\mathbf{A})^{s-1-j}$; consequently $\mathscr{D}\mathcal{P}^{(q)}_{\circledcirc,\tau}(\mathbf{X}_0)[\boldsymbol{\rho}] = \mathbf{Y}_q$.

Proof sketch. Differentiate the recursion $\mathbf{A}_k = \mathbf{X}_k\mathbf{X}_k^\top$ and apply product/chain rules; full details are in Appendix Section C.3. □

Lemma 5.2 (Fréchet derivative of the rectangular polar factor)**.** *Let $\boldsymbol{m}_{\boldsymbol{\theta}} \in \mathbb{R}^{d\times r}$ ($d \geq r$) have full column rank and polar decomposition $\boldsymbol{m}_{\boldsymbol{\theta}} = \mathbf{QP}$, where $\mathbf{Q} = \mathcal{P}_{\circledcirc}(\boldsymbol{m}_{\boldsymbol{\theta}}) \in \mathrm{St}(d,r)$ and $\mathbf{P} = (\boldsymbol{m}_{\boldsymbol{\theta}}^\top \boldsymbol{m}_{\boldsymbol{\theta}})^{1/2} \succ 0$. For any perturbation $\boldsymbol{\rho} \in \mathbb{R}^{d\times r}$, $\mathscr{D}\mathcal{P}_{\circledcirc}(\boldsymbol{m}_{\boldsymbol{\theta}})[\boldsymbol{\rho}] = (\mathbf{I} - \mathbf{QQ}^\top)\boldsymbol{\rho}\mathbf{P}^{-1} + \mathbf{Q}\boldsymbol{\Omega}$, where $\boldsymbol{\Omega}^\top = -\boldsymbol{\Omega}$ is the unique solution of $\boldsymbol{\Omega}\mathbf{P} + \mathbf{P}\boldsymbol{\Omega} = \mathbf{Q}^\top\boldsymbol{\rho} - \boldsymbol{\rho}^\top\mathbf{Q} = 2\,\mathrm{skew}(\mathbf{Q}^\top\boldsymbol{\rho})$. If $\mathbf{P} = \mathbf{V}\,\mathrm{diag}(\sigma_1, \ldots, \sigma_r)\mathbf{V}^\top$, then $(\mathbf{V}^\top\boldsymbol{\Omega}\mathbf{V})_{ab} = \dfrac{2\big[\mathbf{V}^\top\,\mathrm{skew}(\mathbf{Q}^\top\boldsymbol{\rho})\mathbf{V}\big]_{ab}}{\sigma_a + \sigma_b}$. Moreover, $\|\mathscr{D}\mathcal{P}_{\circledcirc}(\boldsymbol{m}_{\boldsymbol{\theta}})[\boldsymbol{\rho}]\|_F \leq \dfrac{\|\boldsymbol{\rho}\|_F}{\sigma_{\min}(\boldsymbol{m}_{\boldsymbol{\theta}})}$.*

Proof sketch. The proof is deferred to Appendix D. □

We now establish the key theoretical properties of $f_{\langle\mu\rangle}(\cdot)$.

Lemma 5.3 (Monotonicity)**.** *Let $f_{\langle\mu\rangle}(\cdot)$ be defined in Equation (12) with $\mathbf{K}_{\boldsymbol{\theta}_t}^{\langle\mu\rangle} \succ 0$. Then for any two sets $\mathcal{A} \subseteq \mathcal{B}$, we have $f_{\langle\mu\rangle}(\mathcal{A}) \leq f_{\langle\mu\rangle}(\mathcal{B})$.*

Theorem 5.1 (Weak submodularity)**.** *The set function $f_{\langle\mu\rangle}(\bullet)$ defined in Equation (12) is γ-weakly submodular, with submodularity ratio $\gamma \geq \frac{c_{2\kappa}}{C_1} > 0$, where $c_{2\kappa}$ and C_1 are the RSC and RSM constants from Lemma C.3 at sparsity levels 2κ and 1, respectively.*

Corollary 5.1 (Approximation guarantee)**.** *Let $\widehat{\mathcal{S}}_t$ be the κ-sparse subset returned by Algorithm 1 and $\mathcal{S}^*$ the optimal κ-sparse subset. Then $f_{\langle\mu\rangle}(\widehat{\mathcal{S}}_t) \geq f_{\langle\mu\rangle}(\mathcal{S}^*)\big(1 - e^{-\gamma}\big) = f_{\langle\mu\rangle}(\mathcal{S}^*)\big(1 - e^{-c_{2\kappa}/C_1}\big)$.*

Proof. See Appendix Section C.3.

6 EXPERIMENTS

We evaluate our hypothesis across standard fine-tuning and, more extensively, large-scale pretraining. We first isolate the effect of optimizer-induced geometry by studying how orthogonalization of the momentum matrix and its Fréchet differential through the polar operator affect minibatch selection and convergence under fixed optimizer dynamics. We then evaluate μ-SMAS in large-scale pretraining across three model architectures and different optimizers, demonstrating that polar-orthogonalization-aware gradient matching provides consistent gains over strong state-of-the-art baselines.

6.1 VALIDATION OF ORTHOGONALIZATION FOR BETTER SUBSETS

Experiment Details. We use Qwen2.5-0.5B with LoRA fine-tuning on the molecule-oriented split of MOL-INSTRUCTIONS, treating its four sub-tasks as source domains. We select $\kappa = 16$ examples per minibatch and train for 1500 steps using AdamW, with all runs sharing the same data order and validation stream. Full model, data, LoRA, and optimization details are provided in Section E.3. For selection, we score the LoRA factors $\mathbf{A}, \mathbf{B}$ in blocks 21–23, yielding 36 parameter blocks. Per-example gradients are recovered exactly from a single backward pass using the ghost construction of Wang et al. (2024) and normalized to unit ℓ_2 norm. At each step, we draw a fresh validation batch of $|\mathcal{B}_t^{\text{val}}| = 8$ examples and maintain one selection-only momentum buffer per block with $\beta = 0.95$ as defined in Equation (21).

Budget. We fix $\kappa = 16$ and vary $|\mathcal{B}_t| \in 128, 64, 43, 32$, yielding selection ratios $\rho \in 0.125, 0.250, 0.372, 0.500$. We use $|\mathcal{B}_t| = 43$ instead of the non-integer $\rho = 0.375$. Fixing κ keeps the selected Gram matrices at 16×16, ensuring comparable set-level statistics across ratios; thus, differences reflect selection pressure rather than subset size.

Selection-only geometry. To isolate selection geometry from optimizer dynamics, all methods use the same AdamW update, while Muon's geometry is used only for scoring. We score the LoRA factors $\mathbf{A}, \mathbf{B}$ in blocks 21–23 across six projections, yielding 36 parameter blocks. For each $\mathbf{z}_i \in \mathcal{B}_t$, block-wise gradients $\boldsymbol{g}_{\boldsymbol{\theta}_t}^{(k)}(\mathbf{z}_i) \in \mathbb{R}^{d_k \times r}$ are recovered using the ghost inner product construction of Wang et al. (2024), with $r = 16$, and each block maintains an independent selection-only momentum buffer

$$\mathbf{m}_{\boldsymbol{\theta}_t}^{(k)} = \beta \mathbf{m}_{\boldsymbol{\theta}_{t-1}}^{(k)} + \sum_{i \in \mathcal{B}_t} \hat{\mathbf{g}}_i^{(k)}, \quad \beta = 0.95, \quad k = 1, \ldots, 36. \tag{21}$$

Its polar factor $\mathbf{Q}^{(k)} = \mathbf{U}^{(k)}\mathbf{V}^{(k)\top}$ is used to map pool and anchor gradients through the Fr'echet derivative of the q-step Newton–Schulz map used by Muon. The mapped blocks are concatenated and ℓ_2-normalized to form $\hat{\mathbf{g}}_i$, from which we construct the Gram matrix $[\mathbf{K}_{\boldsymbol{\theta}_t}]_{ij} = \langle \hat{\mathbf{g}}_i, \hat{\mathbf{g}}_j \rangle$ and anchor alignment $\mu_i = \frac{1}{V} \sum_{v=1}^{V} \langle \hat{\mathbf{g}}_i, \hat{\mathbf{g}}_v \rangle$. The same GREATS greedy rule (Wang et al., 2024) then selects $\hat{S}_t$ with $|\hat{S}_t| = \kappa$. The selection buffers and orthogonalization map affect only selection; the selected minibatch is trained with the unchanged, unit-norm-clipped AdamW update.

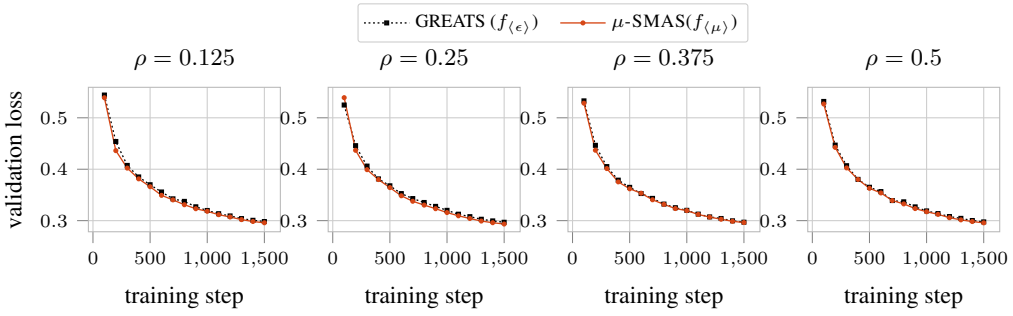


Figure 1: Validation loss for GREATS (dotted black) and μ-SMAS (solid, colored) at $\kappa = 16$ across four data ratios (ρ). μ-SMAS consistently achieves improved convergence and a more favorable validation loss trajectory compared to the GREATS baseline.

Geometry of the selected batch. Our primary goal here is to investigate the quality of coresets obtained via applying the orthogonalization operation during the subset selection process and vice-versa. However, it is difficult to determine this metric since for both cases, the model parameters

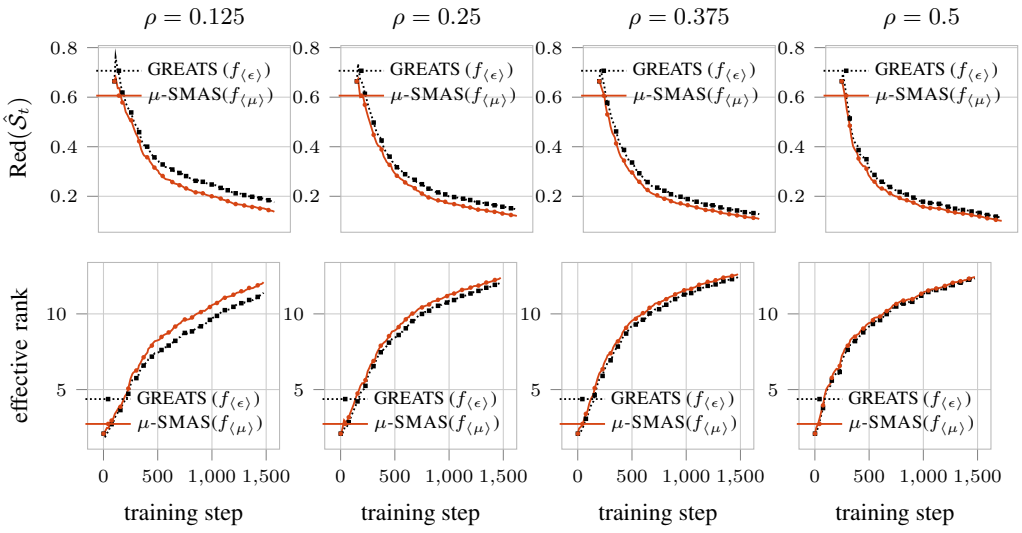


Figure 2: **Redundancy** (top) and **effective rank** (bottom) evaluated on the Gram matrix at each training step. μ-SMAS (solid) consistently outperforms GREATS (dotted) across all ratios ρ by selecting subsets that exhibit notably lower redundancy and higher effective rank, ensuring richer batch diversity.

would be different as in . We thereofre consider a paired design, whereby, at each training step, we compute both the Gram matrices and apply the selection rules to the same candidate batch and thereby evaluate the two selected sets on the same Euclidean Gram matrix. Thus, we can attribute the differences between the two sets solely due to the selection criteria alone.

Redundancy of the selected batch. For a selected subset $\hat{\mathcal{S}}_t \subseteq \mathcal{B}_t$ with $|\hat{\mathcal{S}}_t| = \kappa$, let $[\mathbf{K}_{\boldsymbol{\theta}_t}]_{\hat{\mathcal{S}}_t}$ be the Gram matrix of normalised gradients, with entries $\langle \hat{\mathbf{g}}_i, \hat{\mathbf{g}}_j \rangle_F$. We define redundancy as its mean entry, $\mathrm{Red}(\hat{\mathcal{S}}_t) = \frac{1}{\kappa^2}\mathbf{1}^\top [\mathbf{K}_{\boldsymbol{\theta}_t}]_{\hat{\mathcal{S}}_t}\mathbf{1} = \left\| \frac{1}{\kappa}\sum_{i\in\hat{\mathcal{S}}_t} \hat{\mathbf{g}}_i \right\|_F^2$, the squared norm of the mean normalised gradient. As the diagonal is $\mathbf{1}$, $\mathrm{Red}(\hat{\mathcal{S}}_t) = \frac{1}{\kappa} + \frac{\kappa-1}{\kappa}\,\bar{c}_{\hat{\mathcal{S}}_t}$, where $\bar{c}_{\hat{\mathcal{S}}_t}$ is the mean off-diagonal cosine similarity. Thus Red is 1 for identical gradients, $1/\kappa$ for orthogonal ones, and below $1/\kappa$ for opposing ones.

Effective rank of the selected batch. We measure the effective dimensionality of a selected batch using the Gram-spectrum participation ratio, $\mathrm{ERank}(\mathcal{S}) := (\sum_a \lambda_a)^2 / \sum_a \lambda_a^2$, where λ_a are the eigenvalues of $[\mathbf{K}_{\boldsymbol{\theta}_t}]_{\hat{\mathcal{S}}_t}$. This captures how broadly the batch spans the gradient space and complements redundancy by measuring dimensional diversity. We evaluate it using the same paired design on the shared Euclidean Gram matrix across all selection ratios.

6.2 Gradient Matching via Muon in Language Model Pretraining

Models. We showcase our hypothesis on three family of pretraining models. Firstly, we use a dense decoder model Qwen2.5-0.5B model (Qwen Team, 2024). Secondly, we showcase our findings on GPT2-medium (355M) model, following nanoGPT recipe [1]. Apart from the dense decoder models, we also empirically test our hypothesis with a Mixtral style mixture of experts (MoE) model with total 1.14B parameters with 0.38B active parameters.

Baselines. The pretraining arms compare six rules under both optimisers: Random (uniform κ-subset); GradNorm (Katharopoulos & Fleuret, 2018) (top-κ by per-example gradient norm); FacLoc (greedy facility location on the gradient similarity matrix) (similar to (Nguyen et al., 2025a)); TracIn (Pruthi et al., 2020a) (influence on the held-out anchors); Greats (Wang et al., 2024) (validation-guided greedy matching in the Euclidean/Adam geometry); and μ-SMAS (ours).

Pretraining results. Across three architectures and two optimisers, μ-SMAS attains the best final validation loss in five of six settings and is never worse than Random. Gains are largest on the heterogeneous MoE corpus, especially on the rare arXiv/PubMed domains, consistent with the method's geometry-aware, diversity-promoting design.

[1] https://github.com/kellerjordan/modded-nanogpt

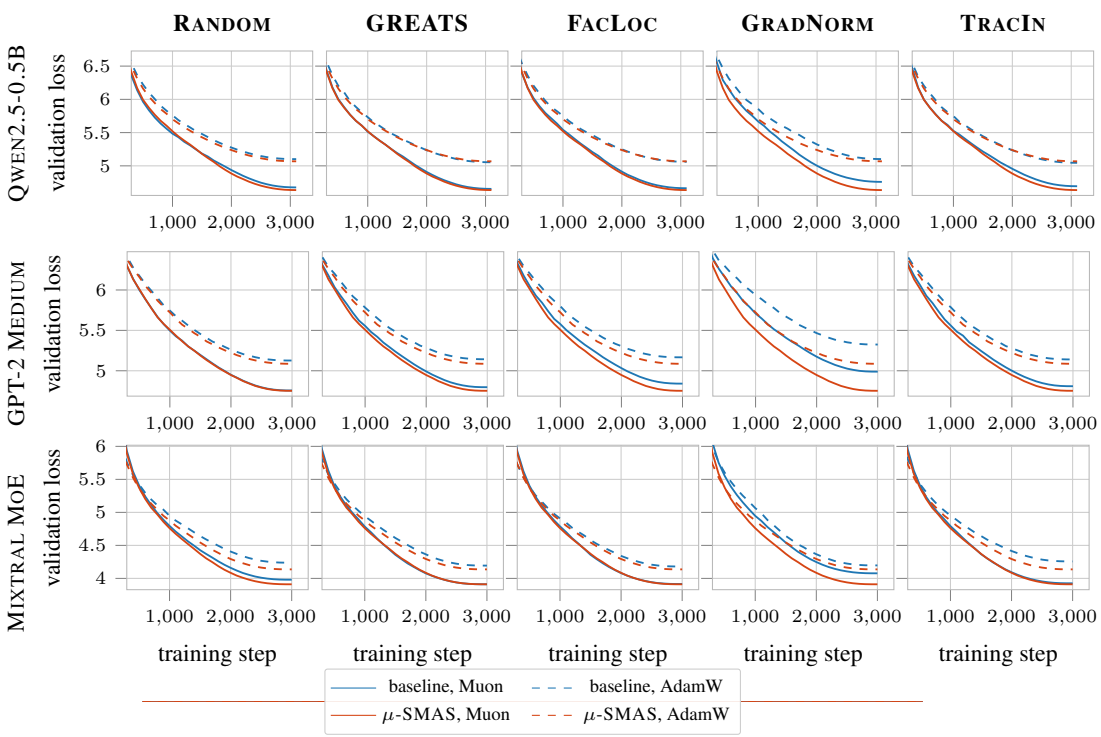


Figure 3: Validation loss pplx curves for each baseline against μ-SMAS.

Domain breakdown. The per-domain view highlights where geometry-aware selection matters most: μ-SMAS consistently improves the rare domains (arXiv and PubMed) under both optimizers, while matching the strongest baselines on the high-share web/code slices. The resulting worst-minus-best *spread* stays competitive, indicating improved tail performance without sacrificing head domains.

	rule (share)	web 30%	code 25%	wiki 15%	math 12%	arXiv 10%	PubMed 8%	spread
Muon	Random	**5.036**	2.106	4.989	4.253	3.440	4.054	**2.930**
	GREATS	5.095	1.998	5.060	4.087	3.327	3.898	3.097
	FacLoc	5.098	2.052	5.070	**4.067**	3.326	3.875	3.046
	GradNorm	5.505	**1.976**	5.500	4.003	3.462	4.015	3.529
	TracIn	5.105	1.995	5.088	4.087	3.344	3.934	3.110
	μ-SMAS	5.099	2.108	**5.033**	4.138	**3.284**	**3.801**	2.991
AdamW	Random	**5.234**	2.345	5.218	4.539	3.676	4.413	**2.888**
	GREATS	5.259	2.356	5.233	4.464	3.574	4.275	2.904
	FacLoc	5.259	2.355	5.222	4.441	3.551	4.252	2.904
	GradNorm	5.418	**2.166**	5.339	**4.191**	3.710	4.356	3.252
	TracIn	5.295	2.403	5.268	4.558	3.646	4.372	2.891
	μ-SMAS	5.255	2.317	**5.202**	4.419	**3.495**	**4.131**	2.937

Table 1: Per-domain validation loss on the six-domain corpus, Mixtral MoE, final checkpoint (single seed). Domain shares of the training corpus are given in the header. Best per column in bold; μ-SMAS shaded. "spread" is the worst-minus-best domain loss.

7 CONCLUSION

Gradient-matching coreset methods typically score examples using Euclidean or Adam-preconditioned inner products with a validation gradient, implicitly assuming that the optimizer applies the selected gradient directly.

We proposed μ-SMAS, a Muon-consistent selector derived from the polar-factor map, and proved monotonicity and γ-weak submodularity (hence a greedy $(1 - e^{-\gamma})$ guarantee). These results suggest that *optimizer-induced geometry* should be treated as a first-class constraint in coreset design.

Spectral optimizers such as Muon violate this assumption: their updates are determined by the polar factor of the momentum, which transforms gradient components according to the momentum's spectral structure. Consequently, Euclidean gradient matching can be misaligned with the actual update direction, with the mismatch increasing as the momentum becomes more anisotropic.

We introduced μ-SMAS to account for this geometry by linearizing the polar map around the current momentum through its Fr'echet derivative. The resulting finite-step Newton–Schulz map is a matrix polynomial and therefore differentiable everywhere, including rank-deficient momenta. Formulating selection as a concave quadratic over sparse nonnegative weights yields a monotone γ-weakly

submodular objective. Empirically, μ-SMAS reduces batch redundancy and improves convergence across Qwen2.5-0.5B, GPT-2 Medium, and a Mixtral-style MoE under Muon and AdamW, achieving the lowest final validation loss in five of six settings and never underperforming random selection.

ETHICS STATEMENT

This work studies online minibatch selection for training language models and does not involve human subjects, user studies, or the collection of personal data. All experiments use publicly available datasets, used under their respective licenses: OpenWebText-100K, a six-domain corpus drawn from FineWeb-Edu, CodeParrot, Wikipedia, OpenWebMath, and the arXiv and PubMed summarization datasets, and the molecule-oriented split of MOL-INSTRUCTIONS. Web-scale corpora can contain biased, toxic, or personally identifying text. Our method does not filter for such content, and the models we train inherit whatever the underlying corpora contain.

A broader consideration specific to validation-guided selection is that it steers training toward the distribution of the held-out validation or anchor set. If that set under-represents certain domains, languages, or populations, selection can amplify the imbalance relative to uniform sampling. For this reason, our mixture-of-experts experiments use domain-balanced anchor sets. Practitioners adopting μ-SMAS should likewise audit the composition of the validation data they select against. The molecular fine-tuning experiments use MOL-INSTRUCTIONS, whose tasks include molecule design and reaction prediction. We use this data only to measure validation loss under different selection rules; we do not evaluate or release models for molecule-generation capability.

AI USE STATEMENT

We used generative AI tools to assist with manuscript editing, condensation, consistency checks, and code development. All citations and reference metadata were independently verified by the authors. The research hypothesis, theoretical development, experimental design, results, analysis, and interpretation were conducted by the authors. We reviewed all AI-assisted content and take full responsibility for the final manuscript and its claims.

REFERENCES

P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

Prayas Agrawal, Prateek Chanda, Ishita Khatri, Ganesh Ramakrishnan, Bamdev Mishra, and Pratik Jawanpuria. Minibatch selection for language models via partition matroid constrained gradient matching. *ICML*, 2026.

Ahmad Ajalloeian and Sebastian U. Stich. On the convergence of sgd with biased gradients, 2021. URL https://arxiv.org/abs/2008.00051.

Noah Amsel, David Persson, Christopher Musco, and Robert M Gower. The polar express: Optimal matrix sign methods and their application to the muon algorithm. *arXiv preprint arXiv:2505.16932*, 2025.

J. Bernstein and L. Newhouse. Old optimizer, new norm. *arXiv preprint arXiv:2409.20325*, 2024.

Prateek Chanda, Prayas Agrawal, Karthik S Gurumoorthy, Ganesh Ramakrishnan, Bamdev Mishra, and Pratik Jawanpuria. Uniprot: Uniform prototype selection via partial optimal transport with submodular guarantees. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026a.

Prateek Chanda, Prayas Agrawal, Saral Sureka, Lokesh Reddy Polu, Atharv Kshirsagar, and Ganesh Ramakrishnan. Bandit guided submodular curriculum for adaptive subset selection. *Advances in Neural Information Processing Systems*, 38:115035–115076, 2026b.

Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=Blz4hjxLwU.

CodeParrot. CodeParrot Dataset Cleaned (valid). Hugging Face dataset repository. URL https://huggingface.co/datasets/codeparrot/codeparrot-clean-valid.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. A discourse-aware attention model for abstractive summarization of long documents, 2018. URL https://arxiv.org/abs/1804.05685.

A. Das and D. Kempe. Approximate submodularity and its applications: Subset selection, sparse approximation and dictionary selection. *Journal of Machine Learning Research*, 19(3):1–34, 2018.

Damek Davis and Dmitriy Drusvyatskiy. When do spectral gradient updates help in deep learning? *arXiv preprint arXiv:2512.04299*, 2025.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

Alan Edelman, Tomás A Arias, and Steven T Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

E.R. Elenberg, R. Khanna, A.G. Dimakis, and S. Negahban. Restricted strong convexity implies weak submodularity. *Annals of Statistics*, 46(6B):3539–3568, 2018.

Evan S Gawlik and Melvin Leok. Computing the fréchet derivative of the polar decomposition. *arXiv preprint arXiv:1608.04491*, 2016.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pp. 1842–1850. PMLR, 2018.

Guy Gur-Ari, Daniel A Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018.

K.S. Gurumoorthy, A. Dhurandhar, G. Cecchi, and C. Aggarwal. Efficient data representation by selecting prototypes with importance weights. In *IEEE International Conference on Data Mining*, pp. 260–269, 2019.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pp. 1225–1234. PMLR, 2016.

Nicholas J Higham. *Functions of matrices: theory and computation*. SIAM, 2008.

Keller Jordan, Yuchen Jin, Vlado Boza, You Jiacheng, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. *URL https://kellerjordan. github. io/posts/muon*, 6(3):4, 2024.

A. Katharopoulos and F. Fleuret. Not all samples are created equal: Deep learning with importance sampling. In *International Conference on Machine Learning*, pp. 2525–2534, 2018.

Jack Kiefer and Jacob Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, pp. 462–466, 1952.

K. Killamsetty, S. Durga, G. Ramakrishnan, A. De, and R. Iyer. GRAD-MATCH: Gradient matching based data subset selection for efficient neural network training. In *International Conference on Machine Learning*, 2021a.

K. Killamsetty, D. Sivasubramanian, G. Ramakrishnan, and R. Iyer. GLISTER: Generalization based data subset selection for efficient and robust learning. In *AAAI Conference on Artificial Intelligence*, 2021b.

Gyu Yeol Kim and Min-hwan Oh. Convergence of muon with newton-schulz. *arXiv preprint arXiv:2601.19156*, 2026.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

P.W. Koh and P. Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, 2017.

A. Kosson, M. Messmer, and M. Jaggi. Rotational equilibrium: How weight decay balances learning across neural networks. In *International Conference on Learning Representations*, 2024.

Dmitry Kovalev. Understanding gradient orthogonalization for deep learning via non-euclidean trust-region optimization. *arXiv preprint arXiv:2503.12645*, 2025.

Jiaxiang Li and Mingyi Hong. A note on the convergence of muon. *arXiv preprint arXiv:2502.02900*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

P. Manupriya, J. Saketha Nath, and P. Jawanpuria. MMD-regularized unbalanced optimal transport. *Transactions on Machine Learning Research*, 2024.

Sam McCandlish, Jared Kaplan, Dario Amodei, and OpenAI Dota Team. An empirical model of large-batch training. *arXiv preprint arXiv:1812.06162*, 2018.

B. Mirzasoleiman, J. Bilmes, and J. Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, 2020.

G.L. Nemhauser, L.A. Wolsey, and M.L. Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical Programming*, 14:265–294, 1978.

D. Nguyen, W. Yang, R. Anand, Y. Yang, and B. Mirzasoleiman. Mini-batch coresets for memory-efficient language model training on data mixtures. In *International Conference on Learning Representations*, 2025a.

Dang Nguyen, Wenhan Yang, Rathul Anand, Yu Yang, and Baharan Mirzasoleiman. Mini-batch coresets for memory-efficient language model training on data mixtures. In *International Conference on Learning Representations*, volume 2025, pp. 82476–82496, 2025b.

Tien-Phat Nguyen, Truong Nguyen, Minh-Phuc Truong, Tuc Nguyen, James Bailey, and Trung Le. Spectral flattening is all muon needs: How orthogonalization controls learning rate and convergence. *arXiv preprint arXiv:2605.13079*, 2026.

Keiran Paster, Marco Dos Santos, Zhangir Azerbayev, and Jimmy Ba. Openwebmath: An open dataset of high-quality mathematical web text, 2023. URL https://arxiv.org/abs/2310.06786.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale, 2024. URL https://arxiv.org/abs/2406.17557.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained lmos. In *International Conference on Machine Learning*, pp. 49069–49104. PMLR, 2025.

G. Pruthi, F. Liu, S. Kale, and M. Sundararajan. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33:19920–19930, 2020a.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33: 19920–19930, 2020b.

Qwen Team. Qwen2.5: A party of foundation models. https://qwenlm.github.io/blog/qwen2.5/, 2024.

Soumava Kumar Roy, Zakaria Mhammedi, and Mehrtash Harandi. Geometry aware constrained optimization techniques for deep learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4460–4469, 2018.

Ishaan Shah, Anthony M Polloreno, Karl Stratos, Philip Monk, Adarsh Chaluvaraju, Andrew Hojel, Andrew Ma, Anil Thomas, Ashish Tanwer, Darsh J Shah, et al. Practical efficiency of muon for pretraining. *arXiv preprint arXiv:2505.02222*, 2025.

Valeria Simoncini. Computational methods for linear matrix equations. *siam REVIEW*, 58(3): 377–441, 2016.

Rishi Sonthalia, Michael Murray, and Guido Montúfar. Low rank gradients and where to find them. *Advances in Neural Information Processing Systems*, 38:152341–152385, 2026.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Cédric Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. Soap: Improving and stabilizing shampoo using adam for language modeling. In *International Conference on Learning Representations*, volume 2025, pp. 93423–93444, 2025.

J.T. Wang, T. Wu, D. Song, P. Mittal, and R. Jia. GREATS: Online selection of high-quality data for LLM training in every iteration. *Advances in Neural Information Processing Systems*, 37: 131197–131223, 2024.

Shaobo Wang, Xuan Ouyang, Tianyi Xu, Yuzheng Hu, Jialin Liu, Guo Chen, Tianyu Zhang, Junhao Zheng, Kexin Yang, Xingzhang Ren, et al. Opus: towards efficient and principled data selection in large language model pre-training in every iteration. *ICML*, 2026.

Wikimedia Foundation. Wikipedia: English Snapshot (20231101.en). Hugging Face dataset, 2023. URL https://huggingface.co/datasets/wikimedia/wikipedia. Configuration: 20231101.en; snapshot: November 1, 2023.

M. Xia, S. Malladi, S. Gururangan, S. Arora, and D. Chen. LESS: Selecting influential data for targeted instruction tuning. In *International Conference on Machine Learning*, 2024.

Greg Yang, James B Simon, and Jeremy Bernstein. A spectral condition for feature learning. *arXiv preprint arXiv:2310.17813*, 2023.

Hanlin Zhang, Depen Morwani, Nikhil Vyas, Jingfeng Wu, Difan Zou, Udaya Ghai, Dean Foster, and Sham Kakade. How does critical batch size scale in pre-training? In *International Conference on Learning Representations*, volume 2025, pp. 66756–66782, 2025.

Supplementary Material: Gradient Matching Meets Muon: Efficient Online Minibatch Selection for Language Models

Contents

Appendices 14
A Algorithm Implementation Details 15
A.1 APGA . . . 15
B Convergence Results 16
B.1 Bias for Muon Updates . . . 16
B.2 Convergence Proof . . . 18
C Additional Notations Usage in Appendix Proofs 22
C.1 Derivation of the Muon marginal utility gain . . . 22
C.2 Auxiliary lemmas for $\mathbf{K}_{\theta_t}^{\langle\mu\rangle}$ and $\mathcal{U}_t$ . . . 22
C.3 Proofs of monotonicity, weak submodularity, and approximation . . . 23
C.4 Applying Muon to LoRA . . . 25
D Proofs of Theoretical Results 26
E Additional Experiment Details 27
E.1 Pretraining Details . . . 27
E.2 Learning-Rate Calibration . . . 29
E.3 Details on Finetuning Experiments . . . 29
E.4 Additional Pretraining Details . . . 38
E.5 LoRA Finetuing . . . 38
E.6 Pretraining: Setup . . . 43
E.7 Learning-Rate Calibration . . . 43
E.8 Qwen2.5-0.5B . . . 44
E.9 GPT-2 medium (355M) . . . 46
E.10 Mixtral MoE (1.14B total, 0.38B active) . . . 46
E.11 Summary . . . 48
E.12 Ablation: Why the Selected Subsets Are Better . . . 48
F Code 51

Supplementary Material: Gradient Matching Meets Muon: Efficient Online Minibatch Selection for Language Models

A Algorithm Implementation Details

Here we provide some high level details regarding our algorithm implementation particularly Accelerated projected gradient descent used in our main OMP algorithm.

A.1 APGA

Algorithm 2 describes the accelerated projected gradient ascent (APGA) solver (Manupriya et al., 2024) we use for this inner problem. It is a standard Nesterov-style accelerated method with a projection step onto the nonnegative orthant, using a step size $1/\mathcal{L}$ where $\mathcal{L}$ is any Lipschitz constant of $\nabla \mathcal{U}_t$. In our implementation, we warm-start APGA from the previous iterate whenever possible, and terminate when the relative improvement in objective becomes small (or after a fixed iteration budget), which keeps the overall overhead negligible.

Algorithm 2: APGA: Accelerated Projected Gradient Ascent

Require: Objective $\mathcal{U}_t(\mathbf{w})$; Lipschitz constant $\mathcal{L}$; initial point $\mathbf{w}_0 \in \mathbb{R}_+^{|\mathcal{B}_t|}$
Ensure: Approximate maximizer $\mathbf{w}$
Initialize $y \leftarrow \mathbf{w}_0$, $\mathbf{w} \leftarrow \mathbf{w}_0$, $t \leftarrow 1$
while *not converged* **do**
  $\mathbf{w}^+ \leftarrow \textsc{Proj}_{\geq 0}\big(y + \frac{1}{\mathcal{L}} \nabla_t(y)\big)$
  $t^+ \leftarrow \frac{1+\sqrt{1+4t^2}}{2}$
  $y \leftarrow \mathbf{w}^+ + \frac{t-1}{t^+}(\mathbf{w}^+ - \mathbf{w})$
  $(\mathbf{w}, t) \leftarrow (\mathbf{w}^+, t^+)$
return $\mathbf{w}$

Proposition A.1 (Polar-metric interpretation)**.** *Let $\mathbf{A}_t^*$ denote the adjoint of $\mathbf{A}_t$ with respect to the Frobenius inner product i.e. ($\langle \mathbf{A}_t[\boldsymbol{\rho}_i], \mathbf{A}_t[\boldsymbol{\rho}_j] \rangle_F = \langle \boldsymbol{\rho}_i, \mathbf{A}_t^* \mathbf{A}_t[\boldsymbol{\rho}_j] \rangle_F$). Then $\mathbf{K}_{ij}^{\mu} = \langle \mathbf{g}_i, \mathbf{A}_t^* \mathbf{A}_t[\mathbf{g}_j] \rangle_F$ and $\mu_i^{\mu} = \langle \mathbf{g}_i, \mathbf{A}_t^*[\mathbf{g}_{\text{val}}] \rangle_F + \frac{\eta_t}{2} \langle \mathbf{g}_i, \mathbf{A}_t^* \mathbf{A}_t[\mathbf{g}_i] \rangle_F$. Consequently, μ-SMAS is equivalent to applying* GREATS *with validation direction $\mathbf{A}_t^*[\mathbf{g}_{\text{val}}]$ and gradient metric $\mathbf{A}_t^* \mathbf{A}_t$, rather than the Euclidean metric.*

Proof. By the definition of the adjoint,

$$\langle \mathbf{A}_t[\mathbf{g}_i], \mathbf{A}_t[\mathbf{g}_j] \rangle_F = \langle \mathbf{g}_i, \mathbf{A}_t^* \mathbf{A}_t[\mathbf{g}_j] \rangle_F.$$

Likewise,

$$\langle \mathbf{A}_t[\mathbf{g}_i], \mathbf{g}_{\text{val}} \rangle_F = \langle \mathbf{g}_i, \mathbf{A}_t^*[\mathbf{g}_{\text{val}}] \rangle_F.$$

Substitution into the definitions of $\mathbf{K}^{\mu}$ and $\boldsymbol{\mu}^{\mu}$ proves the result. □

Proposition A.2 (Geometry-dependent GreedyOMP rule)**.** *Let $\mathbf{w}_{\mathfrak{R}_i}^{\langle \varepsilon \rangle \dagger}$ and $\mathbf{w}_{\mathfrak{R}_i}^{\langle \mu \rangle \dagger}$ be the weight vectors obtained by optimizing the respective objectives over the current support $\mathfrak{R}_i$ at the ith iteration of* **??** *(for $i = 1, \ldots, \kappa$). Define the GreedyOMP residuals*

$$r_{\mathbf{x}}^{\langle \varepsilon \rangle}(\mathfrak{R}_i) = \boldsymbol{\mu}_{\boldsymbol{\theta}_t}^{\langle \mu \rangle}[\mathbf{x}] - \eta_t \sum_{j \in \mathfrak{R}_i} \mathbf{w}_{\mathfrak{R}_i}^{\langle \varepsilon \rangle \dagger}[j] \left[\mathbf{K}_{\boldsymbol{\theta}_t}^{\langle \mu \rangle}\right]_{\mathbf{x}j},$$

$$r_{\mathbf{x}}^{\langle \mu \rangle}(\mathfrak{R}_i) = \boldsymbol{\mu}_{\boldsymbol{\theta}_t}^{\langle \mu \rangle}[\mathbf{x}] - \eta_t \sum_{j \in \mathfrak{R}_i} \mathbf{w}_{\mathfrak{R}_i}^{\langle \mu \rangle \dagger}[j] \left[\mathbf{K}_{\boldsymbol{\theta}_t}^{\langle \mu \rangle}\right]_{\mathbf{x}j}.$$

Consequently,

$$\mathbf{x}_{\langle \varepsilon \rangle}^{\star} \in \operatorname*{arg\,max}_{\mathbf{x} \in \mathcal{B}_t \setminus \mathfrak{R}_i} r_{\mathbf{x}}^{\langle \varepsilon \rangle}(\mathfrak{R}_i), \qquad \mathbf{x}_{\langle \mu \rangle}^{\star} \in \operatorname*{arg\,max}_{\mathbf{x} \in \mathcal{B}_t \setminus \mathfrak{R}_i} r_{\mathbf{x}}^{\langle \mu \rangle}(\mathfrak{R}_i).$$

Hence the two methods select different samples whenever their transformed relevance–redundancy residuals induce different orderings.

Proof. For either geometry $a \in \{\varepsilon, \mu\}$,

$$\mathcal{U}_t^{\langle a \rangle}(\mathbf{w}) = \langle \mathbf{w}, \boldsymbol{\mu}^{\langle a \rangle} \rangle - \frac{\eta_t}{2} \langle \mathbf{w}, \mathbf{K}^{\langle a \rangle} \mathbf{w} \rangle.$$

Therefore,

$$\nabla_{\mathbf{w}} \mathcal{U}_t^a(\mathbf{w}) = \boldsymbol{\mu}^a - \eta_t \mathbf{K}^a \mathbf{w}.$$

Evaluating the i-th coordinate at the optimized vector supported on $\mathcal{S}$ gives the stated residual. GreedyOMP chooses the unselected coordinate with the largest positive residual. □

Making $\langle \varepsilon \rangle$ concrete. Throughout, freeze the Adam second-moment accumulator at its current value $\mathbf{v}_t^{\text{ref}}$, exactly as $\boldsymbol{m}_{\boldsymbol{\theta}_t}^{\text{ref}}$ is frozen for Muon, and define the diagonal (coordinatewise) operator

$$\mathbf{B}_t[\mathbf{X}] \;:=\; \mathbf{X} \oslash \left(\sqrt{\mathbf{v}_t^{\text{ref}}} + \varepsilon\right) \qquad \text{(Hadamard division)},$$

so that $\mathbf{g}_i^{\langle \varepsilon \rangle} := \mathbf{B}_t[\mathbf{g}_i]$ recovers the Adam-preconditioned gradient of Equation (4). Since $\mathbf{B}_t$ acts entrywise by a real diagonal matrix, it is self-adjoint under $\langle \cdot, \cdot \rangle_F$, i.e. $\mathbf{B}_t^* = \mathbf{B}_t$. Define, in exact parallel with $\boldsymbol{\mu}^\mu, \mathbf{K}^\mu$ of Equation (13),

$$K_{ij}^{\langle \varepsilon \rangle} := \langle \mathbf{g}_i^{\langle \varepsilon \rangle}, \mathbf{g}_j^{\langle \varepsilon \rangle} \rangle_F, \qquad \mu_i^{\langle \varepsilon \rangle} := \langle \mathbf{g}_i^{\langle \varepsilon \rangle}, \mathbf{g}_{\mathsf{val}} \rangle_F + \tfrac{\eta_t}{2} K_{ii}^{\langle \varepsilon \rangle}.$$

By Proposition E.1 applied with $\mathbf{A}_t \to \mathbf{B}_t$ (and $\mathbf{B}_t^* \mathbf{B}_t = \mathbf{D}_t := \mathbf{B}_t^2$, since $\mathbf{B}_t$ is already self-adjoint), $\langle \varepsilon \rangle$-GREEDYOMP is exactly GREATS run in the metric $\mathbf{D}_t$ and validation direction $\mathbf{B}_t[\mathbf{g}_{\mathsf{val}}]$ — the coordinatewise counterpart of what μ-SMAS does with $\mathbf{A}_t^* \mathbf{A}_t$.

Recall the per-iteration subproblem is a *constrained* concave program, not free optimization: for support $\mathfrak{R}_i$ and $a \in \{\langle \varepsilon \rangle, \langle \mu \rangle\}$,

$$\mathbf{w}_{\mathfrak{R}_i}^{a\dagger} \in \underset{\substack{\mathbf{w} \geq 0 \\ \operatorname{supp}(\mathbf{w}) \subseteq \mathfrak{R}_i}}{\arg\max} \; \mathcal{U}_t^a(\mathbf{w}) = \langle \mathbf{w}, \boldsymbol{\mu}^a \rangle - \tfrac{\eta_t}{2} \langle \mathbf{w}, \mathbf{K}^a \mathbf{w} \rangle. \tag{$\star$}$$

Since $\mathbf{K}^a$ is a Gram matrix, $\mathbf{K}^a \succeq 0$, so $\mathcal{U}_t^a$ is concave on $\mathbb{R}_{\geq 0}^{|\mathfrak{R}_i|}$ and KKT is both necessary and sufficient for $(\star)$.

Lemma A.1 (KKT characterization on the nonnegative orthant)**.** *Fix $a \in \{\langle \varepsilon \rangle, \langle \mu \rangle\}$ and support $\mathfrak{R}_i$, and write $r_j^a(\mathfrak{R}_i) := \mu_j^a - \eta_t \big(\mathbf{K}^a \mathbf{w}_{\mathfrak{R}_i}^{a\dagger}\big)_j$ for $j \in \mathfrak{R}_i$ (extended by $\mathbf{w}_{\mathfrak{R}_i}^{a\dagger}[j] := 0$ for $j \notin \mathfrak{R}_i$). A vector $\mathbf{w}_{\mathfrak{R}_i}^{a\dagger} \geq 0$ supported on $\mathfrak{R}_i$ solves $(\star)$ if and only if there exists $\boldsymbol{\lambda}^a \geq 0$ supported on $\mathfrak{R}_i$ with*

$$r_j^a(\mathfrak{R}_i) = -\lambda_j^a \;\leq\; 0, \qquad \lambda_j^a \cdot w_{\mathfrak{R}_i}^{a\dagger}[j] = 0, \qquad \forall\, j \in \mathfrak{R}_i.$$

In particular $r_j^a(\mathfrak{R}_i) = 0$ on every strictly active *coordinate ($w_{\mathfrak{R}_i}^{a\dagger}[j] > 0$), but $r_j^a(\mathfrak{R}_i)$ may be strictly negative on a coordinate that is in $\mathfrak{R}_i$ yet zeroed out by the constraint.*

Proof. The Lagrangian of $(\star)$ is $L(\mathbf{w}, \boldsymbol{\lambda}) = \mathcal{U}_t^a(\mathbf{w}) + \boldsymbol{\lambda}^{a\top} \mathbf{w}$, $\boldsymbol{\lambda} \geq 0$ dual to $\mathbf{w} \geq 0$. Stationarity in $\mathbf{w}$ gives $\nabla_{\mathbf{w}} \mathcal{U}_t^a(\mathbf{w}_{\mathfrak{R}_i}^{a\dagger}) + \boldsymbol{\lambda}^a = 0$, i.e. $\lambda_j^a = \eta_t (\mathbf{K}^a \mathbf{w}_{\mathfrak{R}_i}^{a\dagger})_j - \mu_j^a = -r_j^a(\mathfrak{R}_i)$ for $j \in \mathfrak{R}_i$. Dual feasibility $\lambda_j^a \geq 0$ gives $r_j^a(\mathfrak{R}_i) \leq 0$; complementary slackness forces $\lambda_j^a = 0$ (hence $r_j^a = 0$) whenever $w_{\mathfrak{R}_i}^{a\dagger}[j] > 0$. Concavity of $\mathcal{U}_t^a$ (from $\mathbf{K}^a \succeq 0$) and convexity of the feasible set make KKT sufficient as well as necessary for $(\star)$. □

B CONVERGENCE RESULTS

B.1 BIAS FOR MUON UPDATES

We first derive a bound on the bias induced by any subset-selection mechanism in the Muon algorithm.

Consider the gradient computed at sample $\mathbf{z}_i \sim \mathcal{D}$ and model parameter $\boldsymbol{\theta}_t$ as $\mathbf{g}_{\boldsymbol{\theta}_t}^i := \mathbf{g}_{\boldsymbol{\theta}_t}(\mathbf{z}_i) = \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}_t; z_i)$. Let us define the estimation error for a single data sample $\mathbf{z}_i$ as $\Delta \mathbf{g}_i := \mathbf{g}_{\boldsymbol{\theta}_t}^i - \mathbb{E}_{z \sim \mathcal{D}}[\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}_t; z)]$. We simplify the notation as $\Delta \mathbf{g}_i = \mathbf{g}_i - \nabla \ell(\boldsymbol{\theta}_t)$, where $\nabla \ell(\boldsymbol{\theta}_t) = \mathbb{E}_{z \sim \mathcal{D}}[\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}_t; z)]$. Define filtration $\mathcal{F}_{t-1}$ that accounts for all the randomness up to $\boldsymbol{\theta}_t$ and $\boldsymbol{m}_{\boldsymbol{\theta}_{t-1}}$. Thus, under the assumption that $z_1, \ldots, z_n$ are drawn i.i.d. from $\mathcal{D}$, we have $\mathbb{E}[\Delta \mathbf{g}_i \mid \mathcal{F}_{t-1}] = 0$.

For the estimation error, we define covariance operator for $\mathbf{Y} \in \mathbb{R}^{d \times r}$, $C_t[\mathbf{Y}] := \mathbb{E}[\langle \Delta \mathbf{g}_i, \mathbf{Y} \rangle_F \Delta \mathbf{g}_i \mid \mathcal{F}_{t-1}]$. $C_t \succeq 0$ because $\langle \mathbf{Y}, C_t[\mathbf{Y}] \rangle = \mathbb{E}[\langle \Delta \mathbf{g}_i, \mathbf{Y} \rangle^2] \geq 0$. Note that C_t is self-adjoint w.r.t. $\langle \cdot, \cdot \rangle_F$ as $\langle \mathbf{X}, C_t[\mathbf{Y}] \rangle_F = \mathbb{E}[\langle \Delta \mathbf{g}_i, \mathbf{X} \rangle_F, \langle \Delta \mathbf{g}_i, \mathbf{Y} \rangle_F] = \langle C_t[\mathbf{X}], \mathbf{Y} \rangle_F$.

Assumption 1. We assume that the operator norm for the covariance operator C_t is bounded, i.e., $\|C_t\|_{\text{op}} := \lambda_{max}(C_t) \leq \sigma^2$.

μ-SMAS computes $\hat{\mathbf{g}}_{\boldsymbol{\theta}_t} = \sum_{z\in\hat{\mathcal{S}}_t} g_{\boldsymbol{\theta}_t}(z) = \sum_{z\in\hat{\mathcal{S}}_t} \nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}_t; z)$. In this section, for analysis on bounds on the bias term, we consider the averaged gradient $\tilde{\mathbf{g}}_{\boldsymbol{\theta}_t} = \frac{1}{\kappa}\sum_{i\in\hat{\mathcal{S}}_t} \mathbf{g}^i_{\boldsymbol{\theta}_t}$. This produces a scaled momentum $\boldsymbol{m}_{\boldsymbol{\theta}_t}$ by a factor $\frac{1}{\kappa}$. Since, $\mathcal{P}_{\circledcirc}(\cdot)$ is scale-invariant, thus $\mathcal{P}_{\circledcirc}(\frac{1}{\kappa}\boldsymbol{m}_{\boldsymbol{\theta}_t}) = \mathcal{P}_{\circledcirc}(\boldsymbol{m}_{\boldsymbol{\theta}_t})$, and the Muon update remains consistent.

Biased Gradients. We follow biased gradient framework of Ajalloeian & Stich (2021) and define:

$$\tilde{\mathbf{g}}_{\boldsymbol{\theta}_t} = \nabla\ell(\boldsymbol{\theta}_t) + b(\boldsymbol{\theta}_t) + n(\boldsymbol{\theta}_t, \mathbf{z}), \tag{22}$$

where $b(\boldsymbol{\theta}_t)$ is the bias defined as $b(\boldsymbol{\theta}_t) = \mathbb{E}[\tilde{\mathbf{g}}_{\boldsymbol{\theta}_t} - \nabla\ell(\boldsymbol{\theta}_t) \mid \mathcal{F}_{t-1}]$ at parameter $\boldsymbol{\theta}_t$ and $n(\boldsymbol{\theta}_t, \mathbf{z})$ is the noise such that $\mathbb{E}_{z\sim\mathcal{D}}[n(\boldsymbol{\theta}_t, \mathbf{z})] = 0$ and $\mathbb{E}_{z\sim\mathcal{D}}\|n(\boldsymbol{\theta}_t, \mathbf{z})\|^2_{\text{F}} \leq \sigma^2_\kappa$.

We now first derive the exact form of the bias $b(\boldsymbol{\theta}_t)$.

Lemma B.1. *Consider indicators $s_i := \mathbb{1}_{i\in\hat{\mathcal{S}}_t}$, for $i = 1\ldots, n$ then $\sum_{i=1}^n s_i = \kappa$. For $\mathbb{E}[\Delta\mathbf{g}_i \mid \mathcal{F}_{t-1}] = 0$, we have $b(\boldsymbol{\theta}_t) = \frac{1}{\kappa}\sum_{i=1}^n \text{Cov}(s_i, \Delta\mathbf{g}_i \mid \mathcal{F}_{t-1})$.*

Proof. Since $\Delta\mathbf{g}_i = \mathbf{g}^i_{\boldsymbol{\theta}_t} - \nabla\ell(\boldsymbol{\theta}_t)$, thus $\frac{1}{\kappa}\sum_{i\in\hat{\mathcal{S}}_t}\Delta\mathbf{g}_i = \frac{1}{\kappa}\sum_{i\in\hat{\mathcal{S}}_t}(\mathbf{g}^i_{\boldsymbol{\theta}_t} - \nabla\ell(\boldsymbol{\theta}_t)) = \tilde{\mathbf{g}}_{\boldsymbol{\theta}_t} - \nabla\ell(\boldsymbol{\theta}_t)$.

Thus, to compute $b(\boldsymbol{\theta}_t)$, we need to compute $\tilde{\mathbf{g}}_{\boldsymbol{\theta}_t} - \nabla\ell(\boldsymbol{\theta}_t)$.

$$\tilde{\mathbf{g}}_{\boldsymbol{\theta}_t} - \nabla\ell(\boldsymbol{\theta}_t) = \frac{1}{\kappa}\sum_{i\in\hat{\mathcal{S}}_t}(\Delta\mathbf{g}_i + \nabla\ell(\boldsymbol{\theta}_t)) - \nabla\ell(\boldsymbol{\theta}_t) = \frac{1}{\kappa}\sum_{i\in\hat{\mathcal{S}}_t}\Delta\mathbf{g}_i = \frac{1}{\kappa}\sum_{i=1}^n s_i\Delta\mathbf{g}_i.$$

Taking expectation conditioned on $\mathcal{F}_{t-1}$ on both sides, we get

$$b(\boldsymbol{\theta}_t) = \mathbb{E}[\tilde{\mathbf{g}}_{\boldsymbol{\theta}_t} - \nabla\ell(\boldsymbol{\theta}_t) \mid \mathcal{F}_{t-1}] = \frac{1}{\kappa}\sum_{i=1}^n \mathbb{E}[s_i\Delta\mathbf{g}_i \mid \mathcal{F}_{t-1}].$$

s_i and $\mathbf{g}^i_{\boldsymbol{\theta}_t}$ are coupled due to choosing s_i based on $\mathbf{g}^i_{\boldsymbol{\theta}_t}$, thus $\text{Cov}(s_i, \Delta\mathbf{g}_i \mid \mathcal{F}_{t-1}) = \mathbb{E}[s_i\Delta\mathbf{g}_i \mid \mathcal{F}_{t-1}] - \mathbb{E}[s_i \mid \mathcal{F}_{t-1}]\mathbb{E}[\Delta\mathbf{g}_i \mid \mathcal{F}_{t-1}]$. Since $\mathbb{E}[\Delta\mathbf{g}_i \mid \mathcal{F}_{t-1}] = 0$, thus $\mathbb{E}[s_i\Delta\mathbf{g}_i \mid \mathcal{F}_{t-1}] = \text{Cov}(s_i, \Delta\mathbf{g}_i \mid \mathcal{F}_{t-1})$, where covariance is computed entry wise.

□

Lemma B.2. *Under the definition of biased gradient in Equation (22) and Assumption 1, the bound on $\|b(\boldsymbol{\theta}_t)\|_{\text{F}}$ and $\|b(\boldsymbol{\theta}_t)\|_*$ is obtained as*

$$\|b(\boldsymbol{\theta}_t)\|_{\text{F}} \leq \sigma\sqrt{\frac{(1-\frac{\kappa}{n})}{\frac{\kappa}{n}}}, \quad \textit{and} \quad \|b(\boldsymbol{\theta}_t)\|_{\text{op}} \leq \sqrt{r}\sqrt{\frac{(1-\frac{\kappa}{n})}{\frac{\kappa}{n}}}.$$

Proof. Using the definition of dual norm, we write $\|b(\boldsymbol{\theta}_t)\|_{\text{F}} = \sup_{\|\mathbf{Y}\|_{\text{F}}\leq 1}\langle b(\boldsymbol{\theta}_t), \mathbf{Y}\rangle_F$. For some fixed $\mathbf{Y}$ such that $\|\mathbf{Y}\|_{\text{F}} \leq 1$

$$\langle b(\boldsymbol{\theta}_t), \mathbf{Y}\rangle_F = \frac{1}{\kappa}\sum_{i=1}^n \mathbb{E}[s_i\langle\Delta\mathbf{g}_i, \mathbf{Y}\rangle_F \mid \mathcal{F}_{t-1}] = \frac{1}{\kappa}\sum_{i=1}^n \text{Cov}(s_i, \langle\Delta\mathbf{g}_i, \mathbf{Y}\rangle_F \mid \mathcal{F}_{t-1}),$$

since $\mathbb{E}[\langle\Delta\mathbf{g}_i, \mathbf{Y}\rangle_F \mid \mathcal{F}_{t-1}] = 0$. Using Cauchy-Schwarz, $\text{Cov}(X, Z) \leq \sqrt{\text{Var}\, X}\sqrt{\text{Var}\, Z}$. Thus,

$$\langle b(\boldsymbol{\theta}_t), \mathbf{Y}\rangle_F \leq \frac{1}{\kappa}\sum_{i=1}^n \sqrt{\text{Var}(s_i \mid \mathcal{F}_{t-1})}\sqrt{\text{Var}(\langle\Delta\mathbf{g}_i, \mathbf{Y}\rangle_F \mid \mathcal{F}_{t-1})}.$$

Let $p_i := \Pr(i \in \hat{\mathcal{S}}_t \mid \mathcal{F}_{t-1})$, so $s_i \sim \text{Bernoulli}(p_i)$, $\text{Var}(s_i \mid \mathcal{F}_{t-1}$, and $\sum_{i=1}^n p_i = \mathbb{E}[\sum_{i=1}^n s_i \mid \mathcal{F}_{t-1}] = \kappa$.

$$\langle b(\boldsymbol{\theta}_t), \mathbf{Y}\rangle_F \leq \frac{1}{\kappa}\sum_{i=1}^n \sqrt{p_i(1-p_i)}\sqrt{\text{Var}(\langle\Delta\mathbf{g}_i, \mathbf{Y}\rangle_F \mid \mathcal{F}_{t-1})}.$$

Also, $\mathrm{Var}(\langle \Delta \mathbf{g}_i, \mathbf{Y}\rangle_F \mid \mathcal{F}_{t-1}) = \langle \mathbf{Y}, C_t[\mathbf{Y}]\rangle_F \le \|\mathbf{Y}\|_{\mathrm{F}} \|C_t[\mathbf{Y}]\|_{\mathrm{F}} \le \|C_t\|_{\mathrm{op}} \|\mathbf{Y}\|_{\mathrm{F}}^2 \le \|C_t\|_{\mathrm{op}}$. By Cauchy-Schwarz over the n summands and then $\sum_i p_i^2 \ge (\sum_i p_i)^2/n = \kappa^2/n$,

$$\sum_{i=1}^{n} \sqrt{p_i(1-p_i)} \le \sqrt{n \sum_{i=1}^{n} p_i(1-p_i)} = \sqrt{n\Big(\kappa - \sum_i p_i^2\Big)} \le \sqrt{n\kappa - \kappa^2} = \kappa\sqrt{\frac{n}{\kappa} - 1}$$
$$= \kappa\sqrt{\frac{1-\frac{\kappa}{n}}{\frac{\kappa}{n}}}.$$

Thus,

$$\|b(\boldsymbol{\theta}_t)\|_{\mathrm{F}} \le \sqrt{\|\mathbf{C}_t\|_{\mathrm{op}}}\sqrt{\tfrac{(1-\frac{\kappa}{n})}{\frac{\kappa}{n}}}$$
$$\le \sigma\sqrt{\tfrac{(1-\frac{\kappa}{n})}{\frac{\kappa}{n}}}.$$

Finally, apply $\|\cdot\|_* \le \sqrt{r}\|\cdot\|_{\mathrm{F}}$. □

B.2 Convergence Proof

The objective ℓ is nonconvex with $\ell^* := \inf_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) > -\infty$. We denote by $r := \min\{m, n\}$ the maximal possible rank of the matrix. The convergence analysis follows from the convergence of Muon with the Newton-Schultz Algorithm Kim & Oh (2026). We derive the convergence of Muon under the additional bias incurred by the coreset selection mechanism of μ-SMAS. Consider $\hat{\mathbf{g}}_{\boldsymbol{\theta}_t} := \sum_{\mathbf{z}\in\hat{\mathcal{S}}} \boldsymbol{g}_{\boldsymbol{\theta}_t}(\mathbf{z})$, the Muon algorithm is written as

$$\boldsymbol{m}_{\boldsymbol{\theta}_t} = \beta\, \boldsymbol{m}_{\boldsymbol{\theta}_{t-1}} + \hat{\mathbf{g}}, \quad \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t\, \mathcal{P}_{\circledcirc}\big(\boldsymbol{m}_{\boldsymbol{\theta}_t}\big). \tag{23}$$

Assumption 2 (L-smoothness)**.** A differentiable function $\ell : \mathbb{R}^{m\times n} \to \mathbb{R}$ is L-smooth, i.e. for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{m\times n}$, $\|\nabla\ell(\mathbf{x}) - \nabla\ell(\mathbf{y})\|_* \le L\|\|\mathbf{x} - \mathbf{y}\|\|_{\mathrm{op}}$.

In our proof, we use smoothness with respect to the operator norm (and the nuclear norm as its dual).

Assumption 3 (Biased Gradient Oracle)**.** We assume $\hat{\mathbf{g}}_{\boldsymbol{\theta}} := \sum_{\mathbf{z}\in\hat{\mathcal{S}}} \boldsymbol{g}_{\boldsymbol{\theta}}(\mathbf{z})$ is biased estimator of $\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta})$, i.e. map $\mathbf{g} : \mathbb{R}^{m\times n} \times \mathcal{D} \to \mathbb{R}$ is s.t.

$$\hat{\mathbf{g}}_{\boldsymbol{\theta}} = \nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}) + b(\boldsymbol{\theta}) + n(\boldsymbol{\theta}, \mathbf{z})$$

where bias is $b : \mathbb{R}^{m\times n} \to \mathbb{R}$ and noise $n : \mathbb{R}^{m\times n} \times \mathcal{D} \to \mathbb{R}$ has zero-mean noise, i.e $\mathbb{E}_{\mathbf{z}\sim\mathcal{D}}[n(\boldsymbol{\theta}, \mathbf{z})] = 0, \forall \boldsymbol{\theta} \in \mathbb{R}^{m\times n}$.

Based on the above discussion, we assume that the bias and noise terms of the gradient computed on the selected subset are bounded.

Assumption 4 (ζ-bounded bias)**.** There exists a constant $\zeta \ge 0$ s.t. $\forall \boldsymbol{\theta} \in \mathbb{R}^{m\times n}$ $\|b(\boldsymbol{\theta})\|_* \le \zeta$.

Assumption 5 (σ_κ^2-bounded noise)**.** There exist constants $\sigma_\kappa^2 \ge 0$ s.t. $\forall \boldsymbol{\theta} \in \mathbb{R}^{m\times n}$, $\mathbb{E}\|n(\boldsymbol{\theta}, z)\|_{\mathrm{F}}^2 = \mathbb{E}\|\hat{\mathbf{g}}_{\boldsymbol{\theta}} - (\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}) + b(\boldsymbol{\theta}))\|_{\mathrm{F}}^2 \le \sigma_\kappa^2$.

Theorem B.1 (Convergence of Muon under biased gradient)**.** *Suppose Assumptions 2, 3, 5 and 4 hold for running Muon.Choosing* $\eta = \sqrt{\frac{(1-\beta)D}{TL}}$ *and* $\beta = 1 - \min\left\{\frac{\sqrt{LD}}{\sigma_\kappa\sqrt{rT}}, 1\right\}$ *then Muon under biased gradient assumption follows*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}[\nabla\ell(\boldsymbol{\theta}_{t-1})] = \mathcal{O}\left(\sqrt{\frac{LD}{T}} + \frac{\sigma r}{\sqrt{T}} + \left(\frac{r\sigma_\kappa^2 LD}{T}\right)^{1/4}\right) + \mathcal{O}(\zeta).$$

where $\mathcal{O}(\zeta)$ *is the irreducible bias floor.*

Proof. First, we introduce the *scaled* EMA momentum: $N_t := (1-\beta)\boldsymbol{m}_{\boldsymbol{\theta}_t}$. Then, we get $N_t = \beta N_{t-1} + (1-\beta)\hat{\mathbf{g}}_{\boldsymbol{\theta}_t}$. Note that the polar factor is scale-invariant, and let $\mathcal{P}$ be the polar factor of $\boldsymbol{m}_{\boldsymbol{\theta}_t}$, i.e., $\mathcal{P}_{\circledcirc}(\boldsymbol{m}_{\boldsymbol{\theta}_t})$. We define $\boldsymbol{P}_{\boldsymbol{\theta}_t} := \mathcal{P}_{\circledcirc}(N_t) = \mathcal{P}\circledcirc(\boldsymbol{m}_{\boldsymbol{\theta}_t})$.

By Assumption 2, we start from *descent lemma* (Lemma B.3). Since $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \eta\boldsymbol{P}_{\boldsymbol{\theta}_t}$, we have

$$\ell(\boldsymbol{\theta}_t) \le \ell(\boldsymbol{\theta}_{t-1}) + \langle\nabla\ell(\boldsymbol{\theta}_{t-1}), \boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}\rangle_F + \frac{L}{2}\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}\|_{\mathrm{F}}$$
$$= \ell(\boldsymbol{\theta}_{t-1}) - \eta\langle\nabla\ell(\boldsymbol{\theta}_{t-1}), \boldsymbol{P}_{\boldsymbol{\theta}_t}\rangle_F + \frac{L}{2}\eta^2\|\boldsymbol{P}_{\boldsymbol{\theta}_t}\|_{\mathrm{op}}^2.$$

Let $(\boldsymbol{m}_{\boldsymbol{\theta}_t}) = (\mathbf{U}_t, \mathbf{\Sigma}_t, \mathbf{V}_t)$. Then $\boldsymbol{P}_{\boldsymbol{\theta}_t} = \mathrm{polar}(\boldsymbol{m}_{\boldsymbol{\theta}_t}) = \mathbf{U}_t\mathbf{V}_t^\top$. Since $\boldsymbol{P}_{\boldsymbol{\theta}_t} = \mathrm{polar}(\boldsymbol{m}_{\boldsymbol{\theta}_t})$ is a partial isometry, $\|\boldsymbol{P}_{\boldsymbol{\theta}_t}\|_{\mathrm{op}} \le 1$ (and $\|\boldsymbol{P}_{\boldsymbol{\theta}_t}\|_{\mathrm{op}} = 1$ if $(\boldsymbol{m}_{\boldsymbol{\theta}_t}) > 0$). Then, we have

$$\begin{aligned}
\ell(\boldsymbol{\theta}_t) &\le \ell(\boldsymbol{\theta}_{t-1}) - \eta\langle \nabla\ell(\boldsymbol{\theta}_{t-1}), \boldsymbol{P}_{\boldsymbol{\theta}_t}\rangle_F + \frac{L}{2}\eta^2 \\
&= \ell(\boldsymbol{\theta}_{t-1}) - \eta\langle N_t, \boldsymbol{P}_{\boldsymbol{\theta}_t}\rangle_F + \eta\langle N_t - \nabla\ell(\boldsymbol{\theta}_{t-1}), \boldsymbol{P}_{\boldsymbol{\theta}_t}\rangle_F + \frac{L\eta^2}{2} \\
&= \ell(\boldsymbol{\theta}_{t-1}) - \eta\|N_t\|_* + \eta\langle N_t - \nabla\ell(\boldsymbol{\theta}_{t-1}), \boldsymbol{P}_{\boldsymbol{\theta}_t}\rangle_F + \frac{L\eta^2}{2},
\end{aligned}$$

where the last equality is due to $\langle N_t, \boldsymbol{P}_{\boldsymbol{\theta}_t}\rangle_F = \|N_t\|_*$. By Hölder's inequality $|\langle A, B\rangle_F| \le \|A\|_*\|B\|_{\mathrm{op}}$ and the triangle inequality, we have

$$\begin{aligned}
\ell(\boldsymbol{\theta}_t) &\le \ell(\boldsymbol{\theta}_{t-1}) - \eta\|N_t\|_* + \eta\|N_t - \nabla\ell(\boldsymbol{\theta}_{t-1})\|_*\|\boldsymbol{P}_{\boldsymbol{\theta}_t}\|_{\mathrm{op}} + \frac{L\eta^2}{2} \\
&\le \ell(\boldsymbol{\theta}_{t-1}) - \eta\|\nabla\ell(\boldsymbol{\theta}_{t-1})\|_* + \eta\|\nabla\ell(\boldsymbol{\theta}_{t-1}) - N_t\|_* + \eta\|N_t - \nabla\ell(\boldsymbol{\theta}_{t-1})\|_* + \frac{L\eta^2}{2} \\
&= \ell(\boldsymbol{\theta}_{t-1}) - \eta\|\nabla\ell(\boldsymbol{\theta}_{t-1})\|_* + 2\eta\|\nabla\ell(\boldsymbol{\theta}_{t-1}) - N_t\|_* + \frac{L\eta^2}{2}.
\end{aligned}$$

Rearranging and taking expectation gives

$$\mathbb{E}[\|\nabla\ell(\boldsymbol{\theta}_{t-1})\|_*] \le \frac{\mathbb{E}[\ell(\boldsymbol{\theta}_{t-1}) - \ell(\boldsymbol{\theta}_t)]}{\eta} + 2\mathbb{E}[\|\nabla\ell(\boldsymbol{\theta}_{t-1}) - N_t\|_*] + \frac{L\eta}{2} \tag{24}$$

Bounding $\mathbb{E}[\|\nabla\ell(\boldsymbol{\theta}_{t-1}) - N_t\|_*]$.
In order to bound $\mathbb{E}[\|\nabla\ell(\boldsymbol{\theta}_{t-1}) - N_t\|_*]$, we introduce the true scaled momentum $\bar{N}_t$ defined by the true (full-batch) gradient $\nabla\ell(\boldsymbol{\theta}_t)$ instead of $\hat{\mathbf{g}}_{\boldsymbol{\theta}_t}$ for each step t:

- $\bar{N}_t = \beta\bar{N}_{t-1} + (1-\beta)\nabla\ell(\boldsymbol{\theta}_{t-1})$ for $t > 0$ and $\bar{N}_0 = 0$.
- $N_t = \beta N_{t-1} + (1-\beta)\hat{\mathbf{g}}_{\boldsymbol{\theta}_t}$ for $t > 0$ and $N_0 = 0$.

Substituting $\hat{\mathbf{g}}_{\boldsymbol{\theta}_t} = \nabla\ell(\boldsymbol{\theta}_{t-1}) + b(\boldsymbol{\theta}_{t-1}) + n(\boldsymbol{\theta}_t)$ in N_t, we have $N_t = \beta N_{t-1} + (1-\beta)[\nabla\ell(\boldsymbol{\theta}_{t-1}) + b(\boldsymbol{\theta}_{t-1}) + n(\boldsymbol{\theta}_t)]$. When we unroll the EMA recursion of both N_t and $\bar{N}_t$, we have

$$\begin{aligned}
N_t &= \beta N_{t-1} + (1-\beta)\hat{\mathbf{g}}_{\boldsymbol{\theta}_t} = \beta^t N_0 + (1-\beta)\sum_{i=1}^{t}\beta^{t-i}[\nabla\ell(\boldsymbol{\theta}_{i-1}) + b(\boldsymbol{\theta}_{i-1}) + n(\boldsymbol{\theta}_i)] \\
&= (1-\beta)\sum_{i=1}^{t}\beta^{t-i}[\nabla\ell(\boldsymbol{\theta}_{i-1}) + b(\boldsymbol{\theta}_{i-1}) + n(\boldsymbol{\theta}_i)], \text{ and} \\
\bar{N}_t &= \beta\bar{N}_{t-1} + (1-\beta)\nabla\ell(\boldsymbol{\theta}_{t-1}) = \beta^t\bar{N}_0 + (1-\beta)\sum_{i=1}^{t}\beta^{t-i}\nabla\ell(\boldsymbol{\theta}_{i-1}) \\
&= (1-\beta)\sum_{i=1}^{t}\beta^{t-i}\nabla\ell(\boldsymbol{\theta}_{i-1}).
\end{aligned}$$

$$\begin{aligned}
\|\nabla\ell(\boldsymbol{\theta}_{t-1}) - N_t\|_* &= \|\nabla\ell(\boldsymbol{\theta}_{t-1}) - (1-\beta)\sum_{i=1}^{t}\beta^{t-i}[\nabla\ell(\boldsymbol{\theta}_{i-1}) + b(\boldsymbol{\theta}_{i-1}) + n(\boldsymbol{\theta}_i)]\|_* \\
&\le \|\nabla\ell(\boldsymbol{\theta}_{t-1}) - (1-\beta)\sum_{i=1}^{t}\beta^{t-i}[\nabla\ell(\boldsymbol{\theta}_{i-1})]\|_* \\
&\quad + (1-\beta)\|\sum_{i=1}^{t}\beta^{t-i}b(\boldsymbol{\theta}_{i-1})\|_* + (1-\beta)\|\sum_{i=1}^{t}\beta^{t-i}n(\boldsymbol{\theta}_i)]\|_* \\
&= \|\nabla\ell(\boldsymbol{\theta}_{t-1}) - \bar{N}_t\|_* + (1-\beta)\|\sum_{i=1}^{t}\beta^{t-i}b(\boldsymbol{\theta}_{i-1})\|_* + (1-\beta)\|\sum_{i=1}^{t}\beta^{t-i}n(\boldsymbol{\theta}_i)]\|_*.
\end{aligned}$$

Then we can decompose $\mathbb{E}[\|\nabla\ell(\boldsymbol{\theta}_{t-1}) - N_t\|_*]$ as

$$\mathbb{E}[\|\nabla\ell(\boldsymbol{\theta}_{t-1}) - N_t\|_*] \leq \underbrace{\mathbb{E}[\|\nabla\ell(\boldsymbol{\theta}_{t-1}) - \bar{N}_t\|_*]}_{\text{Term (A)}} + \underbrace{(1-\beta)\|\mathbb{E}[\sum_{i=1}^{t}\beta^{t-i}b(\boldsymbol{\theta}_{i-1})\|_*]}_{\text{Term (B)}}$$
$$+ \underbrace{(1-\beta)\|\mathbb{E}[\sum_{i=1}^{t}\beta^{t-i}n(\boldsymbol{\theta}_i)\|_*]}_{\text{Term (C)}}. \tag{25}$$

Bounding the Term (A).
Let $D_t = \|\nabla\ell(\boldsymbol{\theta}_{t-1}) - \bar{N}_t\|_*$. From the recursion, $\nabla\ell(\boldsymbol{\theta}_{t-1}) - \bar{N}_t = \beta(\nabla\ell(\boldsymbol{\theta}_{t-1}) - \bar{N}_{t-1})$. Hence, we have

$$\begin{aligned} D_t &= \beta\|\nabla\ell(\boldsymbol{\theta}_{t-1}) - \bar{N}_{t-1}\|_* \leq \beta\|\nabla\ell(\boldsymbol{\theta}_{t-2}) - \bar{N}_{t-1}\| + \beta\|\nabla\ell(\boldsymbol{\theta}_{t-1}) - \nabla\ell(\boldsymbol{\theta}_{t-2})\|_* \\ &= \beta D_{t-1} + \beta\|\nabla\ell(\boldsymbol{\theta}_{t-1}) - \nabla\ell(\boldsymbol{\theta}_{t-2})\|_*. \end{aligned}$$

Applying Assumption 2 and using the fact that $\|\boldsymbol{P}_{\boldsymbol{\theta}_{t-1}}\|_{\text{op}} \leq 1$, we have

$$\|\nabla\ell(\boldsymbol{\theta}_{t-1}) - \nabla\ell(\boldsymbol{\theta}_{t-2})\|_* \leq L\|\boldsymbol{\theta}_{t-1} - \boldsymbol{\theta}_{t-2}\|_{\text{op}} = L\eta\|\boldsymbol{P}_{\boldsymbol{\theta}_{t-1}}\|_{\text{op}} \leq L\eta.$$

Hence, we have the following recursion,

$$D_t \leq \beta D_{t-1} + \beta L\eta.$$

Since $\bar{N}_0 = 0$, we have

$$D_1 = \|\nabla\ell(\boldsymbol{\theta}_0) - \bar{N}_1\|_* = \|\nabla\ell(\boldsymbol{\theta}_0) - (\beta\bar{N}_0 + (1-\beta)\nabla\ell(\boldsymbol{\theta}_0))\|_* = \beta\|\nabla\ell(\boldsymbol{\theta}_0)\|_*.$$

By unrolling the recursion and taking the expectation, we get

$$\mathbb{E}[D_t] \leq \beta^t\mathbb{E}[\|\nabla\ell(\boldsymbol{\theta}_0)\|_*] + \sum_{i=1}^{t}\beta^i L\eta \leq \beta^t\|\nabla\ell(\boldsymbol{\theta}_0)\|_* + \frac{\beta L\eta}{1-\beta}. \tag{26}$$

Bounding Term (B).
Using linearity of expectation, triangle inequality, Jensen's inequality, and Assumption 4, we expand and bound the term step-by-step:

$$\begin{aligned} (1-\beta)\left\|\mathbb{E}\left[\sum_{i=1}^{t}\beta^{t-i}b(\boldsymbol{\theta}_{i-1})\right]\right\|_* &= (1-\beta)\left\|\sum_{i=1}^{t}\beta^{t-i}\mathbb{E}\left[b(\boldsymbol{\theta}_{i-1})\right]\right\|_* \\ &\leq (1-\beta)\sum_{i=1}^{t}\beta^{t-i}\left\|\mathbb{E}\left[b(\boldsymbol{\theta}_{i-1})\right]\right\|_* \\ &\leq (1-\beta)\sum_{i=1}^{t}\beta^{t-i}\mathbb{E}\left[\|b(\boldsymbol{\theta}_{i-1})\|_*\right] \\ &\leq (1-\beta)\zeta\sum_{i=1}^{t}\beta^{t-i} \\ &= \zeta(1-\beta^t) \leq \zeta. \end{aligned}$$

The last inequality is due to $\beta \in [0, 1)$.

Bounding the Term (C).
Applying Jensen's inequality and the linearity of expectation gives

$$\begin{aligned} (1-\beta)\mathbb{E}\left[\left\|\sum_{i=1}^{t}\beta^{t-i}n(\boldsymbol{\theta}_i)\right\|_F\right] &\leq \sqrt{(1-\beta)^2\mathbb{E}\left[\left\|\sum_{i=1}^{t}\beta^{t-i}n(\boldsymbol{\theta}_i)\right\|_F^2\right]} \\ &= (1-\beta)\sqrt{\mathbb{E}\left[\sum_{i=1}^{t}\beta^{2(t-i)}\left\|n(\boldsymbol{\theta}_i)\right\|_F^2\right]} \\ &= (1-\beta)\sqrt{\sum_{i=1}^{t}\beta^{2(t-i)}\mathbb{E}[\|n(\boldsymbol{\theta}_i)\|_F^2]}. \end{aligned} \tag{27}$$

By Assumption 5,

$$(1-\beta)\mathbb{E}\left[\left\|\sum_{i=1}^{t}\beta^{t-i}n(\boldsymbol{\theta}_i)\right\|_F\right] \le (1-\beta)\sqrt{\frac{1-\beta^{2t}}{1-\beta^2}\sigma_\kappa^2} \le \sqrt{\frac{1-\beta}{1+\beta}}\sigma_\kappa.$$

Using the fact that $\|\mathbf{x}\|_* \le \sqrt{r}\|\mathbf{x}\|_F$ for $\mathbf{x}\in\mathbb{R}^{m\times n}$ with $r=\min\{m,n\}$, we have

$$(1-\beta)\|\mathbb{E}[\sum_{i=1}^{t}\beta^{t-i}n(\boldsymbol{\theta}_i)\|_* \le \sqrt{\frac{1-\beta}{1+\beta}}\sigma_\kappa.$$

Plugging into Equation 25, we obtain

$$\mathbb{E}[\|\nabla\ell(\boldsymbol{\theta}_{t-1}) - N_t\|_*] \le \frac{\beta L\eta}{1-\beta} + \beta^t\|\nabla\ell(\boldsymbol{\theta}_0)\|_* + \zeta + \sqrt{\frac{1-\beta}{1+\beta}}\sqrt{r}\sigma_\kappa. \tag{28}$$

Finally, plugging Equation 28 into Equation 24, we get

$$\mathbb{E}[\|\nabla\ell(\boldsymbol{\theta}_{t-1})\|_*] \le \frac{\mathbb{E}[\ell(\boldsymbol{\theta}_{t-1}) - \ell(\boldsymbol{\theta}_t)]}{\eta} + 2\left(\frac{\beta L\eta}{1-\beta} + \beta^t\|\nabla\ell(\boldsymbol{\theta}_0)\|_* + \zeta + \sqrt{\frac{1-\beta}{1+\beta}}\sqrt{r}\sigma_\kappa\right) + \frac{L\eta}{2}.$$

Averaging over $t=1,\dots T$, we obtain

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla\ell(\boldsymbol{\theta}_{t-1})\|_*] \le \frac{D}{T\eta} + 2\left(\frac{\beta L\eta}{1-\beta} + \zeta + \sqrt{\frac{1-\beta}{1+\beta}}\sqrt{r}\sigma_\kappa\right) + \frac{2}{T}\sum_{t=1}^{T}\beta^t\|\nabla\ell(\boldsymbol{\theta}_0)\|_* + \frac{L\eta}{2}.$$

Using proposition $\|\mathbf{x}\|_* \le \sqrt{r}\|\mathbf{x}\|_F$ and Gradient-gap inequality, $\|\nabla\ell(\boldsymbol{\theta}_0)\|_* \le \sqrt{r}\,\|\nabla\ell(\boldsymbol{\theta}_0)\|_F \le \sqrt{2rLD}$, where $D=\ell(\boldsymbol{\theta}_0)-\ell^*$. Then, we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla\ell(\boldsymbol{\theta}_{t-1})\|_*] \le \frac{D}{T\eta} + 2\left(\frac{\beta L\eta}{1-\beta} + \zeta + \sqrt{\frac{1-\beta}{1+\beta}}\sqrt{r}\sigma_\kappa\right) + \frac{2\beta\sqrt{2rLD}}{T(1-\beta)} + \frac{L\eta}{2},$$

where $D=\ell(\boldsymbol{\theta}_0)-\ell^*$. By setting $\eta=\sqrt{\frac{(1-\beta)D}{TL}}$, we obtain

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla\ell(\boldsymbol{\theta}_{t-1})\|_*] &\le \frac{D}{T\eta} + \frac{L\eta}{2} + \frac{2\beta L\eta}{1-\beta} + 2\zeta + 2\sigma_\kappa\sqrt{r}\sqrt{\frac{1-\beta}{1+\beta}} + \frac{2\beta\sqrt{2rLD}}{T(1-\beta)}\\ &= \sqrt{\frac{LD}{T}}\left(\frac{1+2\beta}{\sqrt{1-\beta}} + \frac{\sqrt{1-\beta}}{2}\right) + 2\zeta + 2\sigma_\kappa\sqrt{r}\sqrt{\frac{1-\beta}{1+\beta}} + \frac{2\beta\sqrt{2rLD}}{T(1-\beta)}.\end{aligned}$$

Setting $\beta = 1-\min\left\{\frac{\sqrt{LD}}{\sigma_\kappa\sqrt{rT}}, 1\right\}$, we get

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla\ell(\boldsymbol{\theta}_{t-1})\|_*] &\le \frac{\sqrt{1-\beta}}{2}\sqrt{\frac{LD}{T}} + \frac{1+2\beta}{\sqrt{1-\beta}}\sqrt{\frac{LD}{T}} + 2\sigma_\kappa\sqrt{r}\sqrt{\frac{1-\beta}{1+\beta}}\\ &\quad + \frac{2\beta\sqrt{2rLD}}{T(1-\beta)} + 2\zeta\\ &= \frac{\sqrt{1-\beta}}{2}\sqrt{\frac{LD}{T}} + \left(1+2\beta+\frac{2}{\sqrt{1+\beta}}\right)\left(\frac{r\sigma_\kappa^2 LD}{T}\right)^{1/4}\\ &\quad + 2\sqrt{2}\beta\frac{\sigma_\kappa r}{\sqrt{T}} + 2\zeta\\ &= \mathcal{O}\left(\sqrt{\frac{LD}{T}} + \frac{\sigma r}{\sqrt{T}} + \left(\frac{r\sigma_\kappa^2 LD}{T}\right)^{1/4}\right) + \mathcal{O}\left(\zeta\right).\end{aligned}$$

□

Lemma B.3 (Descent Lemma). *Let $\ell:\mathbb{R}^{m\times n}\to\mathbb{R}$ be L-smooth. Then, for all $\mathbf{x},\mathbf{y}\in\mathbb{R}^{m\times n}$,*

$$\ell(\mathbf{y}) \le \ell(\mathbf{x}) + \langle\nabla\ell(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle_F + \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|_{\text{op}}^2.$$

Proof. Define an auxiliary scalar function $h:[0,1]\to\mathbb{R}$ by considering the value of ℓ along the line segment from $\mathbf{x}$ to $\mathbf{y}$:

$$h(s) := \ell(\mathbf{x}+s(\mathbf{y}-\mathbf{x}))$$

By the fundamental theorem of calculus,

$$\ell(\mathbf{y}) - \ell(\mathbf{x}) = h(1) - h(0) = \int_0^1 h'(s)\, ds$$

Using the chain rule, $h'(s) = \langle \nabla \ell(\mathbf{x} + s(\mathbf{y} - \mathbf{x})), \mathbf{y} - \mathbf{x} \rangle_F$. Substituting this back into the integral equation gives:

$$\ell(\mathbf{y}) - \ell(\mathbf{x}) = \int_0^1 \langle \nabla \ell(\mathbf{x} + s(\mathbf{y} - \mathbf{x})), \mathbf{y} - \mathbf{x} \rangle_F ds$$

Add and subtract $\langle \nabla \ell(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle_F = \int_0^1 \langle \nabla \ell(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle_F ds$, we have:

$$\ell(\mathbf{y}) - \ell(\mathbf{x}) = \langle \nabla \ell(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle_F + \int_0^1 \langle \nabla \ell(\mathbf{x} + s(\mathbf{y} - \mathbf{x})) - \nabla \ell(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle_F ds$$

Using Hölder's inequality which states that for any matrices $A, B \in \mathbb{R}^{m \times n}$, $|\langle A, B \rangle_F| \leq \|A\|_* \|B\|_{\text{op}}$. Applying this to the integrand:

$$\langle \nabla \ell(\mathbf{x} + s(\mathbf{y} - \mathbf{x})) - \nabla \ell(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle_F \leq \|\nabla \ell(\mathbf{x} + s(\mathbf{y} - \mathbf{x})) - \nabla \ell(\mathbf{x})\|_* \|\mathbf{y} - \mathbf{x}\|_{\text{op}}$$

By the L-smoothness of ℓ:

$$\|\nabla \ell(\mathbf{x} + s(\mathbf{y} - \mathbf{x})) - \nabla \ell(\mathbf{x})\|_* \leq L \|s(\mathbf{y} - \mathbf{x})\|_{\text{op}} = Ls \|\mathbf{y} - \mathbf{x}\|_{\text{op}}$$

Combining these inequalities, we obtain a bound:

$$\ell(\mathbf{y}) - \ell(\mathbf{x}) \leq \langle \nabla \ell(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle_F + \int_0^1 Ls \|\mathbf{y} - \mathbf{x}\|_{\text{op}}^2 ds = \langle \nabla \ell(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle_F + \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|_{\text{op}}^2$$

□

C Additional Notations Usage in Appendix Proofs

We denote $\mathbf{A}_t = \mathscr{DP}_{\circledcirc}(\boldsymbol{m}_{\boldsymbol{\theta}_t})$ as the linear polar operator, for further notation usage in Appendix.

C.1 Derivation of the Muon marginal utility gain

We provide the derivation omitted in the main text.

$$\begin{aligned}
\boldsymbol{\Delta}\mathcal{U}_t^{\langle\mu\rangle}(\mathbf{z}_i \mid \mathcal{S}) &= \mathcal{U}_t^{\langle\mu\rangle}(\mathcal{S} \cup \{\mathbf{z}_i\}) - \mathcal{U}_t^{\langle\mu\rangle}(\mathcal{S}) \\
&= \ell(\mathbf{z}_{\text{val}}; \tilde{\boldsymbol{\theta}}_{t+1}(\mathcal{S})) - \ell(\mathbf{z}_{\text{val}}; \tilde{\boldsymbol{\theta}}_{t+1}(\mathcal{S} \cup \{\mathbf{z}_i\})) \quad \triangleright \text{ Equation (5)} \\
&\approx \eta_t \langle \mathscr{DP}_{\circledcirc}(\boldsymbol{m}_{\boldsymbol{\theta}_t}(\mathcal{S}))[\boldsymbol{g}_{\boldsymbol{\theta}_t}(\mathbf{z}_i)],\, \boldsymbol{g}_{\tilde{\boldsymbol{\theta}}_{t+1}(\mathcal{S})}(\mathbf{z}_{\mathsf{val}}) \rangle_F \qquad (29) \\
&\approx \eta_t \langle \mathscr{DP}_{\circledcirc}(\boldsymbol{m}_{\boldsymbol{\theta}_t}(\mathcal{S}))[\boldsymbol{g}_{\boldsymbol{\theta}_t}(\mathbf{z}_i)],\, \boldsymbol{g}_{\boldsymbol{\theta}_t}(\mathbf{z}_{\mathsf{val}}) \rangle_F \\
&\quad - \eta_t^2 \langle \mathscr{DP}_{\circledcirc}(\boldsymbol{m}_{\boldsymbol{\theta}_t}(\mathcal{S}))[\boldsymbol{g}_{\boldsymbol{\theta}_t}(\mathbf{z}_i)],\, \boldsymbol{\mathcal{H}}_{\mathbf{z}_{\mathsf{val}}}(\boldsymbol{\theta}_t)\, \mathcal{P}_{\circledcirc}\big(\boldsymbol{m}_{\boldsymbol{\theta}_t}(\mathcal{S})\big) \rangle_F . \qquad (30)
\end{aligned}$$

C.2 Auxiliary lemmas for $\mathbf{K}_{\boldsymbol{\theta}_t}^{\langle\mu\rangle}$ and $\mathcal{U}_t$

Lemma C.1 (Positive semidefiniteness of $\mathbf{K}_{\boldsymbol{\theta}_t}^{\langle\mu\rangle}$)**.** *Let $\mathbf{K}_{\boldsymbol{\theta}_t}^{\langle\mu\rangle}[i,j] = \langle \boldsymbol{g}_{\boldsymbol{\theta}_t}^{\langle\mu\rangle}(\mathbf{z}_i),\, \boldsymbol{g}_{\boldsymbol{\theta}_t}^{\langle\mu\rangle}(\mathbf{z}_j) \rangle$ be the Gram matrix of compressed Muon-aligned features. Then $\mathbf{K}_{\boldsymbol{\theta}_t}^{\langle\mu\rangle}$ is symmetric positive semidefinite. Moreover, $\mathbf{K}_{\boldsymbol{\theta}_t}^{\langle\mu\rangle}$ is positive definite if and only if the compressed feature vectors $\{\boldsymbol{g}_{\boldsymbol{\theta}_t}^{\langle\mu\rangle}(\mathbf{z}_i)\}_{i \in \mathcal{B}_t}$ are linearly independent.*

Lemma C.2 ($\mathcal{L}$-smoothness)**.** *The function $\mathcal{U}_t^{\langle\mu\rangle}(\bullet) : \mathbb{R}_{\geq 0}^{|\mathcal{B}_t|} \to \mathbb{R}$ defined in Equation (13) is $\mathcal{L}$-smooth with Lipschitz constant $\mathcal{L} = \eta_t \|\mathbf{K}_{\boldsymbol{\theta}_t}^{\langle\mu\rangle}\|_{\text{op}}$.*

Proof. We have $\nabla_{\mathbf{w}} \mathcal{U}_t^{\langle\mu\rangle}(\mathbf{w}) = \boldsymbol{\mu}_{\boldsymbol{\theta}_t}^{\langle\mu\rangle} - \eta_t \mathbf{K}_{\boldsymbol{\theta}_t}^{\langle\mu\rangle} \mathbf{w}$. Hence, for any $\mathbf{w}_1, \mathbf{w}_2$,

$$\begin{aligned}
\|\nabla \mathcal{U}_t^{\langle\mu\rangle}(\mathbf{w}_1) - \nabla \mathcal{U}_t^{\langle\mu\rangle}(\mathbf{w}_2)\| &= \eta_t \, \|\mathbf{K}_{\boldsymbol{\theta}_t}^{\langle\mu\rangle}(\mathbf{w}_1 - \mathbf{w}_2)\| \\
&\leq \eta_t \, \|\mathbf{K}_{\boldsymbol{\theta}_t}^{\langle\mu\rangle}\|_{\text{op}} \, \|\mathbf{w}_1 - \mathbf{w}_2\|.
\end{aligned}$$

□

Lemma C.3 (Finite RSC/RSM). *Given $\mathbf{K}_{\boldsymbol{\theta}_t}^{\langle\mu\rangle} \in \mathbb{R}^{|\mathcal{B}_t|\times|\mathcal{B}_t|}$ is symmetric positive definite, then $\mathcal{U}_t^{\langle\mu\rangle}(\mathbf{w}) = \eta_t \mathbf{w}^\top \boldsymbol{\mu}_{\boldsymbol{\theta}_t}^{\langle\mu\rangle} - \frac{\eta_t^2}{2}\mathbf{w}^\top \mathbf{K}_{\boldsymbol{\theta}_t}^{\langle\mu\rangle}\mathbf{w}$ is RSC and RSM on κ-sparse non-negative vectors, with:*

$$c_\kappa = \eta_t^2 \cdot \lambda_{\min}^{(\kappa)}(\mathbf{K}_{\boldsymbol{\theta}_t}^{\langle\mu\rangle}), \qquad C_\kappa = \eta_t^2 \cdot \lambda_{\max}^{(\kappa)}(\mathbf{K}_{\boldsymbol{\theta}_t}^{\langle\mu\rangle})$$

where $\lambda_{\min}^{(\kappa)}$ and $\lambda_{\max}^{(\kappa)}$ denote the κ-sparse extremal eigenvalues of $\mathbf{K}_{\boldsymbol{\theta}_t}^{\langle\mu\rangle}$.

Proof. Fix any $\mathbf{w}_1, \mathbf{w}_2$ with $\|\mathbf{w}_1 - \mathbf{w}_2\|_0 \le \kappa$. Then

$$\mathcal{U}_t^{\langle\mu\rangle}(\mathbf{w}_1) - \mathcal{U}_t^{\langle\mu\rangle}(\mathbf{w}_2) - \langle \nabla\mathcal{U}_t^{\langle\mu\rangle}(\mathbf{w}_2), \mathbf{w}_1 - \mathbf{w}_2\rangle = -\frac{\eta_t^2}{2}(\mathbf{w}_1 - \mathbf{w}_2)^\top \mathbf{K}_{\boldsymbol{\theta}_t}^{\langle\mu\rangle}(\mathbf{w}_1 - \mathbf{w}_2).$$

Let $\Omega_\kappa := \{(\mathbf{x}, \mathbf{y}) : \|\mathbf{x}\|_0 \le \kappa,\ \|\mathbf{y}\|_0 \le \kappa,\ \mathbf{x}, \mathbf{y} \ge 0\}$, and let c_κ and C_κ denote the RSC/RSM parameters of $\mathcal{U}_t^{\langle\mu\rangle}$ on Ω_κ (cf. Section 3). Since $(\mathbf{w}_1, \mathbf{w}_2) \in \Omega_\kappa$ and $\mathbf{w}_1 - \mathbf{w}_2$ is κ-sparse, we have

$$\begin{aligned}\frac{\eta_t^2}{2}\lambda_{\min}^{(\kappa)}(\mathbf{K}_{\boldsymbol{\theta}_t}^{\langle\mu\rangle})\ \|\mathbf{w}_1 - \mathbf{w}_2\|^2 &\le \frac{\eta_t^2}{2}(\mathbf{w}_1 - \mathbf{w}_2)^\top \mathbf{K}_{\boldsymbol{\theta}_t}^{\langle\mu\rangle}(\mathbf{w}_1 - \mathbf{w}_2) \\ &\le \frac{\eta_t^2}{2}\lambda_{\max}^{(\kappa)}(\mathbf{K}_{\boldsymbol{\theta}_t}^{\langle\mu\rangle})\ \|\mathbf{w}_1 - \mathbf{w}_2\|^2.\end{aligned}$$

Therefore $c_\kappa = \eta_t^2 \lambda_{\min}^{(\kappa)}(\mathbf{K}_{\boldsymbol{\theta}_t}^{\langle\mu\rangle})$ and $C_\kappa = \eta_t^2 \lambda_{\max}^{(\kappa)}(\mathbf{K}_{\boldsymbol{\theta}_t}^{\langle\mu\rangle})$.

□

C.3 Proofs of monotonicity, weak submodularity, and approximation

Lemma 5.3 (Monotonicity). *Let $f_{\langle\mu\rangle}(\cdot)$ be defined in Equation (12) with $\mathbf{K}_{\boldsymbol{\theta}_t}^{\langle\mu\rangle} \succ 0$. Then for any two sets $\mathcal{A} \subseteq \mathcal{B}$, we have $f_{\langle\mu\rangle}(\mathcal{A}) \le f_{\langle\mu\rangle}(\mathcal{B})$.*

Proof. Let $\mathcal{F}(\mathcal{A}) = \{\mathbf{w} \ge 0 \mid \mathrm{supp}(\mathbf{w}) \subseteq \mathcal{A}\}$ denote the feasible set for $f_{\langle\mu\rangle}(\mathcal{A})$, and let $\mathbf{w}^*(\mathcal{A}) \in \arg\max_{\mathbf{w}\in\mathcal{F}(\mathcal{A})} \mathcal{U}_t^{\langle\mu\rangle}(\mathbf{w})$ be a maximizer of the objective over this set, so that $f_{\langle\mu\rangle}(\mathcal{A}) = \mathcal{U}_t^{\langle\mu\rangle}(\mathbf{w}^*(\mathcal{A}))$ and $\mathbf{w}^*(\mathcal{A}) \in \mathcal{F}(\mathcal{A})$ by construction.

Since $\mathcal{A} \subseteq \mathcal{B}$, every $\mathbf{w}$ with $\mathrm{supp}(\mathbf{w}) \subseteq \mathcal{A}$ also satisfies $\mathrm{supp}(\mathbf{w}) \subseteq \mathcal{B}$, so $\mathcal{F}(\mathcal{A}) \subseteq \mathcal{F}(\mathcal{B})$; in particular, $\mathbf{w}^*(\mathcal{A}) \in \mathcal{F}(\mathcal{B})$.

Since $\mathcal{U}_t^{\langle\mu\rangle}(\mathbf{w})$ is concave in $\mathbf{w}$ (as $\mathbf{K}_{\boldsymbol{\theta}_t}^{\langle\mu\rangle} \succ 0$), its maximum over $\mathcal{F}(\mathcal{B})$ is global, so $\mathcal{U}_t^{\langle\mu\rangle}(\mathbf{w}^*(\mathcal{B})) \ge \mathcal{U}_t^{\langle\mu\rangle}(\mathbf{w}^\dagger)$ for every $\mathbf{w}^\dagger \in \mathcal{F}(\mathcal{B})$. Taking $\mathbf{w}^\dagger := \mathbf{w}^*(\mathcal{A}) \in \mathcal{F}(\mathcal{B})$ gives

$$f_{\langle\mu\rangle}(\mathcal{B}) = \mathcal{U}_t^{\langle\mu\rangle}(\mathbf{w}^*(\mathcal{B})) \ \ge\ \mathcal{U}_t^{\langle\mu\rangle}(\mathbf{w}^*(\mathcal{A})) = f_{\langle\mu\rangle}(\mathcal{A}).$$

□

Theorem 5.1 (Weak submodularity). *The set function $f_{\langle\mu\rangle}(\bullet)$ defined in Equation (12) is γ-weakly submodular, with submodularity ratio $\gamma \ge \frac{c_{2\kappa}}{C_1} > 0$, where $c_{2\kappa}$ and C_1 are the RSC and RSM constants from Lemma C.3 at sparsity levels 2κ and 1, respectively.*

Proof of Theorem 5.1. Let $\mathcal{S}, \mathcal{A} \subseteq \mathcal{B}_t$ be disjoint with $|\mathcal{S}| \le \kappa$, $|\mathcal{A}| \le \kappa$ (so $\bar\kappa := |\mathcal{S}| + |\mathcal{A}| \le 2\kappa$), and let $\boldsymbol{\zeta}^{(\mathcal{S})}$, $\boldsymbol{\zeta}^{(\mathcal{S}\cup\mathcal{A})}$ denote the maximizers of $\mathcal{U}_t^{\langle\mu\rangle}(\mathbf{w})$ over $\mathrm{supp}(\mathbf{w}) \subseteq \mathcal{S}$ and $\mathrm{supp}(\mathbf{w}) \subseteq \mathcal{S}\cup\mathcal{A}$ respectively, so $f_{\langle\mu\rangle}(\mathcal{S}) = \mathcal{U}_t^{\langle\mu\rangle}(\boldsymbol{\zeta}^{(\mathcal{S})})$. Write $[\cdot]_+ := \max(\cdot, 0)$.

Upper bound on the denominator. Since $\boldsymbol{\zeta}^{(\mathcal{S})}$ is feasible but generally suboptimal for the larger support $\mathcal{S}\cup\mathcal{A}$, restricted strong concavity of $\mathcal{U}_t^{\langle\mu\rangle}$ with constant $c_{\bar\kappa}$ gives the standard optimality-gap bound

$$f_{\langle\mu\rangle}(\mathcal{S}\cup\mathcal{A}) - f_{\langle\mu\rangle}(\mathcal{S}) \ \le\ \frac{1}{2c_{\bar\kappa}}\left\|\left[\nabla_{\mathcal{A}}\mathcal{U}_t^{\langle\mu\rangle}(\boldsymbol{\zeta}^{(\mathcal{S})})\right]_+\right\|^2, \tag{31}$$

where $\nabla_{\mathcal{A}}\mathcal{U}_t^{\langle\mu\rangle}(\boldsymbol{\zeta}^{(\mathcal{S})})$ is the gradient restricted to coordinates in $\mathcal{A}$ (coordinates in $\mathcal{S}$ contribute nothing, since $\boldsymbol{\zeta}^{(\mathcal{S})}$ is already stationary there up to the non-negativity constraint).

Lower bound on the numerator. Fix $j \in \mathcal{A}$ and set $g_j := \left[\nabla_j \mathcal{U}_t^{\langle\mu\rangle}(\boldsymbol{\zeta}^{(\mathcal{S})})\right]_+ \geq 0$. Define $\mathbf{y}^{(j)} = \boldsymbol{\zeta}^{(\mathcal{S})} + \alpha \mathbf{1}^{(j)}$ with step size $\alpha = g_j / C_1 \geq 0$, so $\mathbf{y}^{(j)}$ is feasible for support $\mathcal{S} \cup \{j\}$. Restricted smoothness of $\mathcal{U}_t^{\langle\mu\rangle}$ with constant C_1 gives the standard sufficient-ascent guarantee

$$\mathcal{U}_t^{\langle\mu\rangle}(\mathbf{y}^{(j)}) - \mathcal{U}_t^{\langle\mu\rangle}(\boldsymbol{\zeta}^{(\mathcal{S})}) \;\geq\; \frac{1}{2C_1} g_j^2,$$

and since $\boldsymbol{\zeta}^{(\mathcal{S}\cup\{j\})}$ maximizes over a feasible set containing $\mathbf{y}^{(j)}$,

$$f_{\langle\mu\rangle}(\mathcal{S} \cup \{j\}) - f_{\langle\mu\rangle}(\mathcal{S}) \;\geq\; \mathcal{U}_t^{\langle\mu\rangle}(\mathbf{y}^{(j)}) - \mathcal{U}_t^{\langle\mu\rangle}(\boldsymbol{\zeta}^{(\mathcal{S})}) \;\geq\; \frac{1}{2C_1} g_j^2. \tag{32}$$

Summing equation 32 over $j \in \mathcal{A}$ (using $\sum_{j\in\mathcal{A}} g_j^2 = \|[\nabla_{\mathcal{A}} \mathcal{U}_t^{\langle\mu\rangle}(\boldsymbol{\zeta}^{(\mathcal{S})})]_+\|^2$) and comparing with equation 31,

$$\frac{\sum_{j\in\mathcal{A}}\left[f_{\langle\mu\rangle}(\mathcal{S} \cup \{j\}) - f_{\langle\mu\rangle}(\mathcal{S})\right]}{f_{\langle\mu\rangle}(\mathcal{S} \cup \mathcal{A}) - f_{\langle\mu\rangle}(\mathcal{S})} \;\geq\; \frac{c_{\bar{\kappa}}}{C_1} \;\geq\; \frac{c_{2\kappa}}{C_1},$$

using that RSC constants are non-increasing in sparsity level, so $\bar{\kappa} \leq 2\kappa$ gives $c_{\bar{\kappa}} \geq c_{2\kappa}$. Taking the minimum over all such $\mathcal{S}, \mathcal{A}$ (with $|\mathcal{S}|, |\mathcal{A}| \leq \kappa$) gives $\gamma \geq c_{2\kappa}/C_1 > 0$. □

Proof of Corollary 5.1. This follows directly from Theorems 5.3 and 5.1, together with the standard approximation result for greedy maximisation of monotone γ-weakly submodular functions under a cardinality constraint (Nemhauser et al., 1978; Das & Kempe, 2018; Elenberg et al., 2018). □

Proposition 5.1. *The finite-step Newton–Schulz map $\mathcal{P}^{(q)}_{\circledcirc,\tau}(\bullet)$ defined in Section 3 is a matrix polynomial, hence Fréchet differentiable at every $\mathbf{X}_0 \in \mathbb{R}^{d\times r}$ (including rank-deficient $\mathbf{X}_0$).*

Proof. We prove by induction on q that $\mathcal{P}^{(q)}_{\circledcirc,\tau}$ is a matrix polynomial map. The base case for $q = 0$: $\mathcal{P}^{(q)}_{\circledcirc,\tau}(\mathbf{Z}) = \mathbf{Z}$ is trivially polynomial. For the inductive step, if $\mathcal{P}^{(q)}_{\circledcirc,\tau}$ is a matrix polynomial in $\mathbf{Z}$, then $\mathcal{P}^{(q)}_{\circledcirc,\tau}(\mathbf{Z})\mathcal{P}^{(q)}_{\circledcirc,\tau}(\mathbf{Z})^\top$ is too (a product of polynomials), $p_\tau\big(\mathcal{P}^{(q)}_{\circledcirc,\tau}(\mathbf{Z})\mathcal{P}^{(q)}_{\circledcirc,\tau}(\mathbf{Z})^\top\big) = \sum_{s=1}^{\tau} c_s \big(\mathbf{I} - \mathcal{P}^{(q)}_{\circledcirc,\tau}(\mathbf{Z})\mathcal{P}^{(q)}_{\circledcirc,\tau}(\mathbf{Z})^\top\big)^s$ is too (a finite sum of finite powers of a polynomial), and hence $\mathcal{P}^{(q+1)}_{\circledcirc,\tau}(\mathbf{Z}) = p_\tau(\cdots)\mathcal{P}^{(q)}_{\circledcirc,\tau}(\mathbf{Z})$ is too (from Definition 1). Thus $\mathcal{P}^{(q)}_{\circledcirc,\tau}$ is a matrix polynomial map of finite degree, and every entry of a matrix polynomial is a polynomial function $\mathbb{R}^{d\times r} \to \mathbb{R}$ of the entries of its argument, hence C^∞ at every point $\mathbf{Z} \in \mathbb{R}^{d\times r}$; in particular at $\mathbf{Z} = \mathbf{X}_0$, including rank-deficient $\mathbf{X}_0$. □

Lemma 5.1 (Fréchet derivative of the finite Newton–Schulz map)**.** *For a perturbation $\boldsymbol{\rho} \in \mathbb{R}^{d\times r}$, define $\mathbf{Y}_i := \mathscr{D}\mathbf{X}_i(\mathbf{X}_0)[\boldsymbol{\rho}]$, with $\mathbf{Y}_0 := \boldsymbol{\rho}$. Let $\mathbf{A}_k := \mathbf{X}_k\mathbf{X}_k^\top$ and $\mathbf{B}_k := \mathbf{Y}_k\mathbf{X}_k^\top + \mathbf{X}_k\mathbf{Y}_k^\top$. Then*

$$\mathbf{Y}_{k+1} = \mathscr{D}p_\tau(\mathbf{A}_k)[\mathbf{B}_k]\mathbf{X}_k + p_\tau(\mathbf{A}_k)\mathbf{Y}_k, \tag{14}$$

where $\mathscr{D}p_\tau(\mathbf{A})[\mathbf{B}] = -\sum_{s=1}^{\tau} c_s \sum_{j=0}^{s-1} (\mathbf{I}-\mathbf{A})^j \mathbf{B} (\mathbf{I}-\mathbf{A})^{s-1-j}$; consequently $\mathscr{D}\mathcal{P}^{(q)}_{\circledcirc,\tau}(\mathbf{X}_0)[\boldsymbol{\rho}] = \mathbf{Y}_q$.

Proof. Step 2: Fréchet derivative of a matrix power. We claim that for $\mathbf{T} \mapsto \mathbf{T}^s$,

$$\mathscr{D}(\mathbf{T}^s)[\mathbf{H}] = \sum_{j=0}^{s-1} \mathbf{T}^j \mathbf{H} \mathbf{T}^{s-1-j}.$$

We prove this by induction on s. The base case $s = 1$ is immediate: $\mathscr{D}(\mathbf{T})[\mathbf{H}] = \mathbf{H} = \mathbf{T}^0\mathbf{H}\mathbf{T}^0$. For the inductive step, write $\mathbf{T}^s = \mathbf{T} \cdot \mathbf{T}^{s-1}$ and apply the product rule:

$$\mathscr{D}(\mathbf{T}^s)[\mathbf{H}] = \mathbf{H}\,\mathbf{T}^{s-1} + \mathbf{T} \cdot \mathscr{D}(\mathbf{T}^{s-1})[\mathbf{H}] = \mathbf{H}\mathbf{T}^{s-1} + \mathbf{T}\sum_{j=0}^{s-2} \mathbf{T}^j \mathbf{H} \mathbf{T}^{s-2-j} = \sum_{j=0}^{s-1} \mathbf{T}^j \mathbf{H} \mathbf{T}^{s-1-j}.$$

Applying this with $\mathbf{T} = \mathbf{I} - \mathbf{A}$ and using $\mathscr{D}(\mathbf{I} - \mathbf{A})[\mathbf{B}] = -\mathbf{B}$, the chain rule gives

$$\mathscr{D}\big[(\mathbf{I} - \mathbf{A})^s\big](\mathbf{A})[\mathbf{B}] = -\sum_{j=0}^{s-1} (\mathbf{I} - \mathbf{A})^j \mathbf{B} (\mathbf{I} - \mathbf{A})^{s-1-j}.$$

Summing over $s = 0, \dots, \tau$ with weights c_s (the $s = 0$ term is an empty sum, hence zero) gives equation 15:

$$\mathscr{D}p_\tau(\mathbf{A})[\mathbf{B}] = -\sum_{s=1}^{\kappa} c_s \sum_{j=0}^{s-1} (\mathbf{I}-\mathbf{A})^j \mathbf{B} (\mathbf{I}-\mathbf{A})^{s-1-j}.$$

Step 3: Fréchet derivative of $\mathbf{Z} \mapsto \mathbf{Z}\mathbf{Z}^\top$. For any $\mathbf{H}$ and scalar t,

$$(\mathbf{Z}+t\mathbf{H})(\mathbf{Z}+t\mathbf{H})^\top = \mathbf{Z}\mathbf{Z}^\top + t\big(\mathbf{H}\mathbf{Z}^\top + \mathbf{Z}\mathbf{H}^\top\big) + t^2\mathbf{H}\mathbf{H}^\top,$$

and reading off the coefficient of t gives $\mathscr{D}(\mathbf{Z}\mathbf{Z}^\top)(\mathbf{Z})[\mathbf{H}] = \mathbf{H}\mathbf{Z}^\top + \mathbf{Z}\mathbf{H}^\top$.

Step 4: assembling the recursion. We prove $\mathscr{D}p_\tau(\mathbf{X}_0)[\mathbf{H}] = \mathbf{Y}_k$ by induction on q. The base case $q = 0$ holds since $\mathcal{P}^{(0)}_{\circledcirc,\tau}(\mathbf{Z}) = \mathbf{Z}$ is the identity map, so $\mathscr{D}p_0(\mathbf{X}_0)[\mathbf{H}] = \mathbf{H} = \mathbf{Y}_0$. For the inductive step, assume $\mathscr{D}\mathcal{P}^{(q)}_{\circledcirc,\tau}(\mathbf{X}_0)[\mathbf{H}] = \mathbf{Y}_k$. Since $\mathcal{P}^{(q+1)}_{\circledcirc,\tau}(\mathbf{Z}) = p_\tau\big(\mathcal{P}^{(q)}_{\circledcirc,\tau}(\mathbf{Z})\mathcal{P}^{(q)}_{\circledcirc,\tau}(\mathbf{Z})^\top\big) \cdot \mathcal{P}^{(q)}_{\circledcirc,\tau}(\mathbf{Z})$ is a product of two functions of $\mathbf{Z}$, the product rule evaluated at $\mathbf{Z} = \mathbf{X}_0$ gives

$$\mathbf{Y}_{k+1} = \mathscr{D}\Big[p_\tau\big(\mathcal{P}^{(q)}_{\circledcirc,\tau}(\mathbf{Z})\mathcal{P}^{(q)}_{\circledcirc,\tau}(\mathbf{Z})^\top\big)\Big](\mathbf{X}_0)[\mathbf{H}] \cdot \mathbf{X}_k + p_\tau(\mathbf{A}_k) \cdot \mathscr{D}\mathcal{P}^{(q)}_{\circledcirc,\tau}(\mathbf{X}_0)[\mathbf{H}].$$

By the inductive hypothesis, the second factor of the second term is $\mathbf{Y}_k$. For the first term, chain through $\mathbf{Z} \mapsto \mathcal{P}^{(q)}_{\circledcirc,\tau}(\mathbf{Z})\mathcal{P}^{(q)}_{\circledcirc,\tau}(\mathbf{Z})^\top$: by Step 3 and the chain rule, its derivative at $\mathbf{X}_0$ in direction $\mathbf{H}$ is

$$\mathscr{D}\mathcal{P}^{(q)}_{\circledcirc,\tau}(\mathbf{X}_0)[\mathbf{H}]\,\mathbf{X}_k^\top + \mathbf{X}_k\,\mathscr{D}\mathcal{P}^{(q)}_{\circledcirc,\tau}(\mathbf{X}_0)[\mathbf{H}]^\top = \mathbf{Y}_k\mathbf{X}_k^\top + \mathbf{X}_k\mathbf{Y}_k^\top = \mathbf{B}_k;$$

then by Step 2 (chain rule through p_τ),

$$\mathscr{D}\Big[p_\tau\big(\mathcal{P}^{(q)}_{\circledcirc,\tau}(\mathbf{Z})\mathcal{P}^{(q)}_{\circledcirc,\tau}(\mathbf{Z})^\top\big)\Big](\mathbf{X}_0)[\mathbf{H}] = \mathscr{D}p_\tau(\mathbf{A}_k)[\mathbf{B}_k].$$

Substituting both pieces gives exactly equation 14:

$$\mathbf{Y}_{k+1} = \mathscr{D}p_\tau(\mathbf{A}_k)[\mathbf{B}_k]\,\mathbf{X}_k + p_\tau(\mathbf{A}_k)\,\mathbf{Y}_k.$$

By induction this holds for every $\kappa = 0, \dots, K$; taking $\kappa = K$ by definition gives $\mathscr{D}\mathcal{P}^{(q)}_{\circledcirc,\tau}(\mathbf{X}_0)[\mathbf{H}] = \mathbf{Y}_K$, which is equation 16.

□

C.4 Applying Muon to LoRA

In LoRA, a frozen weight $\boldsymbol{\theta}_0 \in \mathbb{R}^{n\times m}$ is adapted via a low-rank update $\boldsymbol{\theta}_t = \boldsymbol{\theta}_0 + \alpha\,\mathbf{B}_t\mathbf{A}_t$ with $\mathbf{B}_t \in \mathbb{R}^{n\times r}$, $\mathbf{A}_t \in \mathbb{R}^{r\times m}$ and small r. Each trainable factor ($\mathbf{A}_t$ and $\mathbf{B}_t$) is a 2D parameter, so Muon can be applied per factor:

$$\mathbf{M}_t^{(\mathbf{A})} \leftarrow \beta\,\mathbf{M}_{t-1}^{(\mathbf{A})} + \nabla_{\mathbf{A}}\mathcal{L}, \qquad \mathbf{A}_{t+1} \leftarrow \mathbf{A}_t - \eta_{\mathbf{A}}\,\mathrm{NS}\big(\mathbf{M}_t^{(\mathbf{A})}\big), \tag{33}$$

$$\mathbf{M}_t^{(\mathbf{B})} \leftarrow \beta\,\mathbf{M}_{t-1}^{(\mathbf{B})} + \nabla_{\mathbf{B}}\mathcal{L}, \qquad \mathbf{B}_{t+1} \leftarrow \mathbf{B}_t - \eta_{\mathbf{B}}\,\mathrm{NS}\big(\mathbf{M}_t^{(\mathbf{B})}\big). \tag{34}$$

Note that acting on the two factors separately makes Muon non–reparameterization-invariant: its per-factor orthogonalization depends on arbitrary scalings or rotations, skewing the weight-space step and often letting one factor dominate.

Extension to LoRA When fine-tuning with LoRA, only the low-rank factors are updated; we therefore rewrite the approximations above in terms of the trainable parameters $\boldsymbol{\theta}_t := (\mathbf{A}_t, \mathbf{B}_t)$ (with base model weights frozen). We denote the per-example gradient as $\boldsymbol{g}_{\boldsymbol{\theta}_t}(\mathbf{z}) := \big(\boldsymbol{g}_{\mathbf{A}_t}(\mathbf{z}),\, \boldsymbol{g}_{\mathbf{B}_t}(\mathbf{z})\big)$ and equip this product space with the sum-of-Frobenius inner product. Under this convention, the first-order (GREATS) term in (11) decomposes additively over the two LoRA factors:

$$\langle \boldsymbol{g}_{\boldsymbol{\theta}_t}(\mathbf{z}_i),\, \boldsymbol{g}_{\boldsymbol{\theta}_t}(\mathbf{z}_{\mathsf{val}})\rangle_F = \langle \boldsymbol{g}_{\mathbf{A}_t}(\mathbf{z}_i),\, \boldsymbol{g}_{\mathbf{A}_t}(\mathbf{z}_{\mathsf{val}})\rangle_F + \langle \boldsymbol{g}_{\mathbf{B}_t}(\mathbf{z}_i),\, \boldsymbol{g}_{\mathbf{B}_t}(\mathbf{z}_{\mathsf{val}})\rangle_F.$$

The corresponding one-step LoRA update (used inside $\tilde{\boldsymbol{\theta}}_{t+1}(\mathcal{S})$) is simply factorwise gradient descent,

$$\tilde{\mathbf{A}}_{t+1}(\mathcal{S}) = \mathbf{A}_t - \eta_{\mathbf{A}} \sum_{\mathbf{z}\in\mathcal{S}} \boldsymbol{g}_{\mathbf{A}_t}(\mathbf{z}), \qquad \tilde{\mathbf{B}}_{t+1}(\mathcal{S}) = \mathbf{B}_t - \eta_{\mathbf{B}} \sum_{\mathbf{z}\in\mathcal{S}} \boldsymbol{g}_{\mathbf{B}_t}(\mathbf{z}).$$

Finally, the second-order correction in (11) involves the Hessian with respect to LoRA parameters, which we denote (following our convention elsewhere) by $\boldsymbol{\mathcal{H}}_{\mathbf{z}_{\mathsf{val}}}(\boldsymbol{\theta}_t) = \nabla^2_{\boldsymbol{\theta}}\ell(\mathbf{z}_{\mathsf{val}}; \boldsymbol{\theta}_t)$. Writing this Hessian in block form over $(\mathbf{A}_t, \mathbf{B}_t)$, we define the blocks by the corresponding mixed partials, e.g.,

$$\boldsymbol{\mathcal{H}}_{\mathbf{A}_t\mathbf{A}_t} := \nabla^2_{\mathbf{A}_t\mathbf{A}_t}\ell(\mathbf{z}_{\mathsf{val}}; \boldsymbol{\theta}_t), \quad \boldsymbol{\mathcal{H}}_{\mathbf{A}_t\mathbf{B}_t} := \nabla^2_{\mathbf{A}_t\mathbf{B}_t}\ell(\mathbf{z}_{\mathsf{val}}; \boldsymbol{\theta}_t),$$

with $\boldsymbol{\mathcal{H}}_{\mathbf{B}_t\mathbf{A}_t}$ and $\boldsymbol{\mathcal{H}}_{\mathbf{B}_t\mathbf{B}_t}$ defined analogously. Then,

$$\boldsymbol{\mathcal{H}}_{\mathbf{z}_{\mathsf{val}}}(\boldsymbol{\theta}_t) = \begin{pmatrix} \boldsymbol{\mathcal{H}}_{\mathbf{A}_t\mathbf{A}_t} & \boldsymbol{\mathcal{H}}_{\mathbf{A}_t\mathbf{B}_t} \\ \boldsymbol{\mathcal{H}}_{\mathbf{B}_t\mathbf{A}_t} & \boldsymbol{\mathcal{H}}_{\mathbf{B}_t\mathbf{B}_t} \end{pmatrix},$$

the Hessian–vector inner product expands as

$$\langle g_{\boldsymbol{\theta}_t}(\mathbf{z}_i),\, \mathcal{H}_{\mathbf{z}_{\text{val}}}(\boldsymbol{\theta}_t)\Big(\sum_{\mathbf{z}\in\mathcal{S}} g_{\boldsymbol{\theta}_t}(\mathbf{z})\Big)\rangle_F = \langle g_{\mathbf{A}_t}(\mathbf{z}_i),\, \mathcal{H}_{\mathbf{A}_t\mathbf{A}_t}\Big(\sum_{\mathbf{z}\in\mathcal{S}} g_{\mathbf{A}_t}(\mathbf{z})\Big) + \mathcal{H}_{\mathbf{A}_t\mathbf{B}_t}\Big(\sum_{\mathbf{z}\in\mathcal{S}} g_{\mathbf{B}_t}(\mathbf{z})\Big)\rangle_F$$
$$+ \langle g_{\mathbf{B}_t}(\mathbf{z}_i),\, \mathcal{H}_{\mathbf{B}_t\mathbf{A}_t}\Big(\sum_{\mathbf{z}\in\mathcal{S}} g_{\mathbf{A}_t}(\mathbf{z})\Big) + \mathcal{H}_{\mathbf{B}_t\mathbf{B}_t}\Big(\sum_{\mathbf{z}\in\mathcal{S}} g_{\mathbf{B}_t}(\mathbf{z})\Big)\rangle_F.$$

In practice, one often drops the cross-block terms ($\mathcal{H}_{\mathbf{A}_t\mathbf{B}_t}, \mathcal{H}_{\mathbf{B}_t\mathbf{A}_t}$), yielding a sum of independent $\mathbf{A}_t$- and $\mathbf{B}_t$-curvature corrections.

D Proofs of Theoretical Results

Lemma 5.2 (Fréchet derivative of the rectangular polar factor)**.** *Let $\boldsymbol{m}_{\boldsymbol{\theta}} \in \mathbb{R}^{d\times r}$ ($d \geq r$) have full column rank and polar decomposition $\boldsymbol{m}_{\boldsymbol{\theta}} = \mathbf{QP}$, where $\mathbf{Q} = \mathcal{P}_{\circledcirc}(\boldsymbol{m}_{\boldsymbol{\theta}}) \in \mathrm{St}(d,r)$ and $\mathbf{P} = (\boldsymbol{m}_{\boldsymbol{\theta}}^\top \boldsymbol{m}_{\boldsymbol{\theta}})^{1/2} \succ 0$. For any perturbation $\boldsymbol{\rho} \in \mathbb{R}^{d\times r}$, $\mathscr{D}\mathcal{P}_{\circledcirc}(\boldsymbol{m}_{\boldsymbol{\theta}})[\boldsymbol{\rho}] = (\mathbf{I} - \mathbf{QQ}^\top)\boldsymbol{\rho}\mathbf{P}^{-1} + \mathbf{Q}\boldsymbol{\Omega}$, where $\boldsymbol{\Omega}^\top = -\boldsymbol{\Omega}$ is the unique solution of $\boldsymbol{\Omega}\mathbf{P} + \mathbf{P}\boldsymbol{\Omega} = \mathbf{Q}^\top\boldsymbol{\rho} - \boldsymbol{\rho}^\top\mathbf{Q} = 2\,\mathrm{skew}(\mathbf{Q}^\top\boldsymbol{\rho})$. If $\mathbf{P} = \mathbf{V}\,\mathrm{diag}(\sigma_1,\ldots,\sigma_r)\mathbf{V}^\top$, then $(\mathbf{V}^\top\boldsymbol{\Omega}\mathbf{V})_{ab} = \dfrac{2\big[\mathbf{V}^\top\,\mathrm{skew}(\mathbf{Q}^\top\boldsymbol{\rho})\mathbf{V}\big]_{ab}}{\sigma_a + \sigma_b}$. Moreover, $\|\mathscr{D}\mathcal{P}_{\circledcirc}(\boldsymbol{m}_{\boldsymbol{\theta}})[\boldsymbol{\rho}]\|_F \leq \dfrac{\|\boldsymbol{\rho}\|_F}{\sigma_{\min}(\boldsymbol{m}_{\boldsymbol{\theta}})}$.*

Proof. Let $\mathbf{Z} := \mathscr{D}\mathcal{P}_{\circledcirc}(\boldsymbol{m}_{\boldsymbol{\theta}})[\mathbf{H}]$ denote the directional (Fréchet derivative) of the matrix polar-factor map at $\boldsymbol{m}_{\boldsymbol{\theta}}$, applied to the perturbation direction $\mathbf{H}$. By definition in Eq. 1 of Gawlik & Leok (2016), we know that $\mathcal{P}_{\circledcirc}(\boldsymbol{m}_{\boldsymbol{\theta}} + \epsilon\mathbf{H}) = \mathcal{P}_{\circledcirc}(\boldsymbol{m}_{\boldsymbol{\theta}}) + \epsilon\mathbf{Z} + \mathcal{O}(\epsilon\|\mathbf{H}\|)$.

Further, let $\mathbf{S} = \mathscr{D}\mathbf{P}[\mathbf{H}]$ be the derivative of the positive polar factor. We note that for the same ϵ, we have $\mathbf{Q}(\varepsilon) := \mathcal{P}_{\circledcirc}(\boldsymbol{m}_{\boldsymbol{\theta}} + \varepsilon\mathbf{H}) = \mathbf{Q} + \varepsilon\mathbf{Z} + \mathcal{O}(\epsilon\|\mathbf{H}\|)$. Similarly $\mathbf{P}(\varepsilon) = \mathbf{P} + \varepsilon\mathbf{S} + \mathcal{O}(\varepsilon\|\mathbf{H}\|)$.

For a sufficiently small ε, $(\mathbf{Q} + \varepsilon\mathbf{H})(\mathbf{P} + \varepsilon\mathbf{H}) = \mathbf{QP} + \varepsilon(\mathbf{QH} + \mathbf{HP}) = \boldsymbol{m}_{\boldsymbol{\theta}} + \epsilon$.

$\boldsymbol{m}_{\boldsymbol{\theta}}(\varepsilon) = \mathbf{Q}(\varepsilon)\mathbf{P}(\varepsilon)$

Differentiating $\boldsymbol{m}_{\boldsymbol{\theta}}(\varepsilon)$ w.r.t ε we get

$$\frac{\partial \boldsymbol{m}_{\boldsymbol{\theta}}(\varepsilon)}{\partial\varepsilon} = \frac{\partial\mathbf{Q}(\varepsilon)}{\partial\varepsilon}\,\mathbf{P}(\varepsilon) + \mathbf{Q}(\varepsilon)\,\frac{\partial\mathbf{P}(\varepsilon)}{\partial\varepsilon}, \tag{35}$$
$$\mathbf{H} = \mathbf{Z}\,\mathbf{P}(\varepsilon) + \mathbf{Q}(\varepsilon)\,\mathbf{S} + \text{const.} \tag{36}$$

Thus for any general ε,

Differentiating $\boldsymbol{m}_{\boldsymbol{\theta}} = \mathbf{QP}$ gives

$$\mathbf{H} = \mathbf{ZP} + \mathbf{QS}, \qquad \mathbf{S}^\top = \mathbf{S}. \tag{37}$$

We know that $\mathbf{Q}(\varepsilon)^\top\mathbf{Q}(\varepsilon) = \mathbf{I}$. Differentiating gives

$$0 = \frac{\partial}{\partial\varepsilon}\big(\mathbf{Q}(\varepsilon)^\top\mathbf{Q}(\varepsilon)\big) = \Big(\frac{\partial\mathbf{Q}(\varepsilon)}{\partial\varepsilon}\Big)^\top\mathbf{Q}(\varepsilon) + \mathbf{Q}(\varepsilon)^\top\frac{\partial\mathbf{Q}(\varepsilon)}{\partial\varepsilon},$$
$$0 = \mathbf{Z}^\top\mathbf{Q}(\varepsilon) + \mathbf{Q}(\varepsilon)^\top\mathbf{Z}.$$

Hence, $\mathbf{Q}(\varepsilon)^\top\mathbf{Z}$ is skew-symmetric.

Differentiating $\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}$ shows that $\boldsymbol{\Omega} := \mathbf{Q}^\top\mathbf{Z}$ is skew-symmetric. Left-multiply (37) by $\mathbf{Q}^\top$ and subtract its transpose to obtain (18).

$$\begin{aligned}
\mathbf{Q}^\top\mathbf{H} - \mathbf{H}^\top\mathbf{Q} &= \mathbf{Q}^\top(\mathbf{Z}\,\mathbf{P} + \mathbf{Q}\,\mathbf{S}) - (\mathbf{Z}\,\mathbf{P} + \mathbf{Q}\,\mathbf{S})^\top\mathbf{Q} && \text{(substitute } \mathbf{H} = \mathbf{ZP} + \mathbf{QS}\text{)}\\
&= \boldsymbol{\Omega}\,\mathbf{P} + \mathbf{S} - (\mathbf{P}^\top\mathbf{Z}^\top + \mathbf{S}^\top\mathbf{Q}^\top)\mathbf{Q} && \text{(use } \mathbf{Q}^\top\mathbf{Z} = \boldsymbol{\Omega} \text{ and } \mathbf{Q}^\top\mathbf{Q} = \mathbf{I}\text{)}\\
&= \boldsymbol{\Omega}\,\mathbf{P} + \mathbf{S} - \mathbf{P}\,\mathbf{Z}^\top\mathbf{Q} - \mathbf{S}^\top && \text{(transpose + regroup)}\\
&= \boldsymbol{\Omega}\,\mathbf{P} + \mathbf{S} + \mathbf{P}\,\boldsymbol{\Omega} - \mathbf{S} && \text{(since } \mathbf{Z}^\top\mathbf{Q} = -\mathbf{Q}^\top\mathbf{Z} = -\boldsymbol{\Omega} \text{ and } \mathbf{S}^\top = \mathbf{S}\text{)}\\
&= \boldsymbol{\Omega}\,\mathbf{P} + \mathbf{P}\,\boldsymbol{\Omega}.
\end{aligned}$$

which proves $\boldsymbol{\Omega}\,\mathbf{P} + \mathbf{P}\,\boldsymbol{\Omega} = 2\,\mathrm{skew}(\mathbf{Q}^\top\mathbf{H})$

Since $\mathbf{P} \succ 0$, the Sylvester operator $\mathbf{\Omega} \mapsto \mathbf{\Omega P} + \mathbf{P\Omega}$ is invertible on skew-symmetric matrices, so the solution is unique.

Applying $\mathbf{I} - \mathbf{QQ}^\top$ to (37) gives

$$(\mathbf{I} - \mathbf{QQ}^\top)\mathbf{Z} = (\mathbf{I} - \mathbf{QQ}^\top)\mathbf{HP}^{-1} - \cancelto{= 0 \,\because\, \mathbf{Q}^\top\mathbf{Q}=\mathbf{I}}{(\mathbf{I} - \mathbf{QQ}^\top)\mathbf{QS}}.$$

Together with $\mathbf{QQ}^\top\mathbf{Z} = \mathbf{Q\Omega}$, this proves (17); diagonalizing $\mathbf{P}$ gives (19).

For the diagonalizable one check (Simoncini, 2016)

For (20), the two terms in equation 17 are Frobenius-orthogonal. In the eigenbasis of $\mathbf{P}$, $|\Omega_{ab}| \le |\operatorname{skew}(\mathbf{Q}^\top\mathbf{H})_{ab}|/\sigma_{\min}(\mathbf{P})$. Hence

$$\|\mathbf{Z}\|_F^2 \le \frac{\|(\mathbf{I} - \mathbf{QQ}^\top)\mathbf{H}\|_F^2 + \|\operatorname{skew}(\mathbf{Q}^\top\mathbf{H})\|_F^2}{\sigma_{\min}(\mathbf{P})^2} \le \frac{\|\mathbf{H}\|_F^2}{\sigma_{\min}(\boldsymbol{m}_{\boldsymbol{\theta}})^2}.$$

□

Adam. Let $\mathbf{v}_t \in \mathbb{R}^{d\times r}$ be the (matrix-shaped) second-moment estimate. Adam's update is

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \underbrace{\mathbf{B}_t \odot \boldsymbol{g}_{\boldsymbol{\theta}_t}}_{=:\, \boldsymbol{g}^{\langle\varepsilon\rangle}_{\boldsymbol{\theta}_t}}, \qquad \mathbf{B}_t := 1\big/\big(\sqrt{\mathbf{v}_t} + \varepsilon\big) \quad \text{(entrywise)}. \tag{38}$$

The **Adam effective-gradient map** is the diagonal (coordinatewise) operator $\mathbf{B}_t$, self-adjoint under $\langle\cdot,\,\cdot\rangle_F$.

Muon (corrected). Muon's *actual parameter step* uses the polar factor itself, not a tangent-space projection:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t\, \mathbf{Q}_t, \qquad \mathbf{Q}_t := \mathcal{P}_{\circledcirc}(\boldsymbol{m}_{\boldsymbol{\theta}_t}) \in \mathrm{St}(d, r). \tag{39}$$

The object that governs *marginal-gain sensitivity* to a new candidate $\mathbf{z}_i$ — i.e. how the selected set changes the next iterate, cf. Equation (10) — is the **Fréchet derivative of the polar map** at the reference momentum (Lemma 5.2), not a manifold tangent projector:

$$\mathbf{A}_t := \mathscr{D}\mathcal{P}_{\circledcirc}(\boldsymbol{m}_{\boldsymbol{\theta}_t}), \qquad \boldsymbol{g}^{\langle\mu\rangle}_{\boldsymbol{\theta}_t}(\mathbf{z}_i) = \mathbf{A}_t[\boldsymbol{g}_{\boldsymbol{\theta}_t}(\mathbf{z}_i)] = (\mathbf{I} - \mathbf{Q}_t\mathbf{Q}_t^\top)\boldsymbol{g}_{\boldsymbol{\theta}_t}(\mathbf{z}_i)\mathbf{P}_t^{-1} + \mathbf{Q}_t\mathbf{\Omega}_i, \tag{40}$$

with $\mathbf{\Omega}_i$ the Sylvester solution of Equation (18) and $\mathbf{P}_t = (\boldsymbol{m}_{\boldsymbol{\theta}_t}^\top \boldsymbol{m}_{\boldsymbol{\theta}_t})^{1/2}$. Crucially, $\mathbf{A}_t$ carries the $\mathbf{P}_t^{-1}$ (momentum-*magnitude*-dependent) term that a bare tangent-space projector lacks; this term is exactly what makes $\mathbf{A}_t$ a spectral *reweighting*, not a projection.

The data-selection kernels are, as elsewhere in the paper,

$$K^{\langle\varepsilon\rangle}(\mathbf{z}_i, \mathbf{z}_j) := \langle \mathbf{B}_t[\boldsymbol{g}_i],\, \boldsymbol{g}_j\rangle_F, \qquad K^{\langle\mu\rangle}(\mathbf{z}_i, \mathbf{z}_j) := \langle \mathbf{A}_t[\boldsymbol{g}_i],\, \mathbf{A}_t[\boldsymbol{g}_j]\rangle_F.$$

E Additional Experiment Details

E.1 Pretraining Details

Schedule and budget. All runs use a cosine schedule with 100 linear warm-up steps, weight decay 0.1, gradient clipping at 1.0, bf16 autocast for the forward pass with fp32 master weights and optimiser state, and activation checkpointing (required to fit the 64-row candidate pool). The Qwen arm runs one epoch (3159 steps); the GPT-2 and MoE arms run a fixed 3000 steps. At 16 trained rows of 512 tokens per step this is ≈25M trained tokens drawn from a ≈100M-token candidate stream — a deliberately short horizon: these runs are designed to separate *selection rules*, not to produce converged language models.

Corpora and tokenisation. The two dense arms train on OpenWebText-100K[2], tokenised once per tokenizer and concatenated into a single stream that is then cut into fixed-length chunks; the last 5% of chunks are held out for evaluation. **Every arm uses a sequence length of 512 tokens.** This is worth stating explicitly because the GPT-2 configuration file requests 1024: the dataset constructor is invoked without a `max_length` argument and therefore falls back to its 512-token default, so the GPT-2 runs are in fact 512-token runs. We report the effective value rather than the configured one.

[2] https://huggingface.co/datasets/Elriggs/openwebtext-100k

The MoE arm cannot use a single-domain corpus: a router has nothing to specialise on when every token comes from the same distribution, and the eight experts collapse towards interchangeable copies. We therefore build a **six-domain, deliberately imbalanced corpus** of 110M tokens (215,092 chunks), streamed and tokenised from six real sources — it is not synthetic text:

domain	source	share
web	`HuggingFaceFW/fineweb-edu (sample-10BT)`	30%
code	`codeparrot/codeparrot-clean-valid`	25%
wiki	`wikimedia/wikipedia (20231101.en)`	15%
math	`open-web-math/open-web-math`	12%
arXiv	`ccdv/arxiv-summarization`	10%
PubMed	`ccdv/pubmed-summarization`	8%

Two of the six (code, math) are not prose: web, wiki, arXiv and PubMed differ in register but share token statistics, whereas source code and LaTeX do not, and it is that contrast which makes routing observable. Chunks never straddle a domain boundary, and each chunk carries its domain label. The imbalance is deliberate: it is simultaneously what the router must learn to exploit and what a selection rule can correct for.

Held-out splits for the MoE arm. Rather than hold out a fraction of a shuffled pool — which inherits the corpus imbalance — the MoE arm carves two *domain-balanced* hold-outs, each containing the same number of chunks from every domain: an evaluation set ($6 \times 200 = 1200$ chunks) and, disjoint from it, an anchor set ($6 \times 100 = 600$ chunks). The remaining 213,292 chunks form the training split and retain the natural, imbalanced mixture. Consequently the reported validation loss is the unweighted mean over the six domains — a model that improves on web at the expense of code cannot hide behind web's 30% share — and the anchor gradient means "do well on all six domains equally".

Optimisers. We compare AdamW (Loshchilov & Hutter) with Muon (Jordan et al., 2024) (momentum $\beta = 0.95$, Nesterov, $K = 10$ Newton–Schulz iterations of the degree-2 Taylor map, `max1` scaling, float32 iteration). Departing from common practice, the token embeddings and the LM head remain on the *Muon* branch rather than being routed to an auxiliary AdamW optimiser, because the selection step needs their Stiefel-orthogonal momentum buffer; AdamW handles only the 1-D parameters (biases, layer norms), at a fixed 3×10^{-4}.

The MoE arm requires one addition. In current implementations Mixtral's experts are not `Linear` modules but batched three-dimensional tensors, one slab per block (`gate_up_proj` of shape $(E, 2I, H)$ and `down_proj` of shape (E, H, I)), holding 1.01B of the model's 1.14B parameters. The standard Muon implementation asserts two-dimensional parameters and would reject them. We orthogonalise each expert's own (out, in) matrix separately and scale its learning rate by that per-expert shape, so an expert receives the update a dense layer of the same size would. Flattening the slab to $(E,\ \text{out} \cdot \text{in})$ instead — the alternative found in reference implementations — would make the E flattened experts mutually orthogonal *as vectors*, which is a different and far less natural object; we report the per-expert variant throughout.

Selection protocol. Every step loads a candidate pool of $|\mathcal{B}_t| = 64$ sequences and trains on the $\kappa = 16$ the selector keeps (selection ratio 0.25). Because κ is fixed, *every* rule — including `full`, which simply loads 16 — takes identically sized gradient steps, so a difference between rules is attributable to *which* rows were chosen and never to how many. Anchors for the gradient-matching rules are drawn afresh at each step from the held-out split: 16 anchors for the dense arms. The MoE arm uses 18, so that the six domains divide it evenly, and draws them with a domain-stratified sampler that places exactly three chunks of each domain in *every* anchor batch. The selector's Taylor weighting η is refreshed from the live learning rate at each step, so warm-up and decay reach the selection objective and not just the update.

Scored layer. Selection scores the last in-block linear layer. For the dense arms this is an MLP projection. For the MoE this resolves to the final block's attention output projection, which is the desirable outcome: attention is *dense*, so every token traverses it and a per-example gradient there is a clean per-example quantity. Had the experts been ordinary linear layers, an expert would only see the $\sim 2/8$ of tokens routed to it and its per-example gradient would track routing frequency rather

Table 2: Configuration of the three pretraining arms. Sequence length is the *effective* value (see "Corpora and tokenisation").

	Qwen2.5-0.5B	GPT-2 medium	Mixtral MoE
parameters (total)	0.49B	0.355B	1.14B
parameters (active)	0.49B	0.355B	0.38B
blocks / hidden	24 / 896	24 / 1024	10 / 1024
heads (Q / KV)	14 / 2	16 / 16	16 / 4
experts (top-k)	—	—	8 (top-2)
tokenizer (vocab)	Qwen (151,936)	GPT-2 (50,304)	GPT-2 (50,257)
corpus	OpenWebText-100K	OpenWebText-100K	6-domain, 110M tok
sequence length	512	512	512
held-out split	last 5%	last 5%	balanced, 200/domain
anchors per step	16	16	18 (3/domain)
candidate pool / κ	64 / 16	64 / 16	64 / 16
steps	3159 (1 epoch)	3000	3000
η_{Muon}	1×10^{-3}	1×10^{-3}	4×10^{-4}
η_{AdamW}	5×10^{-4}	2×10^{-4}	2×10^{-4}

than utility. The corresponding limitation is that selection never observes the experts, i.e. 90% of the MoE's parameters.

Evaluation. Validation loss is the token-weighted mean cross-entropy on the held-out split, evaluated every 100 steps. For the MoE arm the reported loss is the *language-modelling* term only: with `output_router_logits` enabled the optimised loss is $\mathcal{L}_{\text{LM}} + \lambda \mathcal{L}_{\text{aux}}$, and the auxiliary term is a load-balancing regulariser — a property of the router, not of the data — so folding it in would compare selection rules partly on how evenly their batches happened to route. It is logged separately, alongside per-domain losses, expert load statistics and routing entropy.

Controls and reproducibility. Within an arm, all rules share the model initialisation, the data order, the anchor stream, the schedule and the seed (42); the selection rule is the only thing that varies. Each configuration is run **once**: the pretraining numbers below are single-seed point estimates, and we therefore report them without error bars and refrain from significance claims. Differences of a few thousandths of a nat should be read as ties. (The multi-seed protocol with paired bootstrap tests applies to the LoRA fine-tuning studies, where run cost permits it.)

E.2 Learning-Rate Calibration

A comparison between optimisers is only meaningful at each optimiser's own operating point, so Muon and AdamW were swept independently before any selection experiment was run. Both sweeps use the `full` rule, so the learning rate is calibrated on the optimiser alone and no selection effect can contaminate it.

The sweeps select $\eta_{\text{Muon}} = 1 \times 10^{-3}$ for both dense models, with $\eta_{\text{AdamW}} = 5 \times 10^{-4}$ for Qwen and 2×10^{-4} for GPT-2. For the MoE arm we use $\eta_{\text{Muon}} = 4 \times 10^{-4}$ and $\eta_{\text{AdamW}} = 2 \times 10^{-4}$. Every selection rule within an arm is then run at its optimiser's calibrated rate, so a gap between two rules cannot be an artefact of a learning rate that happens to suit one of them.

E.3 Details on Finetuning Experiments

Model. We fine-tune Qwen2.5-0.5B (24 blocks, hidden size 896, intermediate size 4864, 14 query and 2 key/value heads; `bfloat16`, SDPA attention) with LoRA adapters ($r = 16$, $\alpha = 32$, dropout 0.05) on all six projections (`q`, `k`, `v`, `gate`, `up`, `down`) of every block, giving 8.11M trainable parameters (1.62% of 502M).

Data. We use the molecule-oriented split of Mol-Instructions, whose $D = 4$ sub-tasks we treat as source domains: description-guided molecule design, forward reaction prediction, reagent prediction, and retrosynthesis. From each sub-task we hold out 1000 test examples, then concatenate and shuffle the remaining training data (seed 42), tagging each example with its domain. Sequences use the Alpaca template, truncated to 512 tokens, with the loss computed over both prompt and

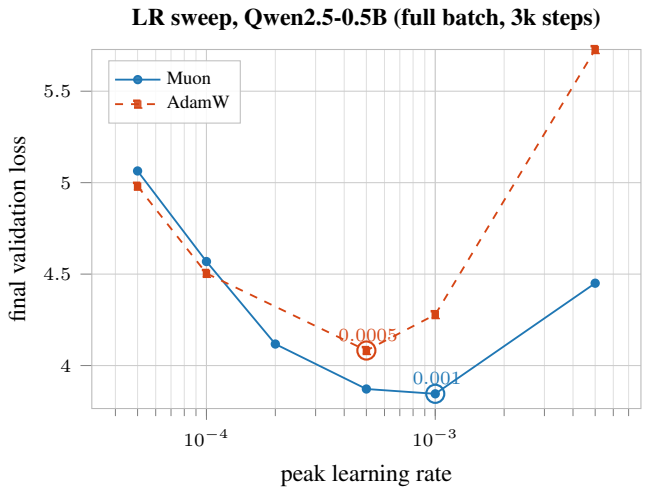


Figure 4: **Learning-rate sweep for Qwen2.5-0.5B, Muon *vs* AdamW**, trained with the full batch for one epoch (6317 steps) at batch 32×512, cosine schedule with 100 warm-up steps, seed 42. Circled points mark each optimiser's optimum. Muon peaks at $\eta = 0.001$ (final validation loss 3.8455), AdamW at $\eta = 0.0005$ (4.0822) — a factor of 2 apart, which is why a single shared learning rate would misrepresent one of the two.

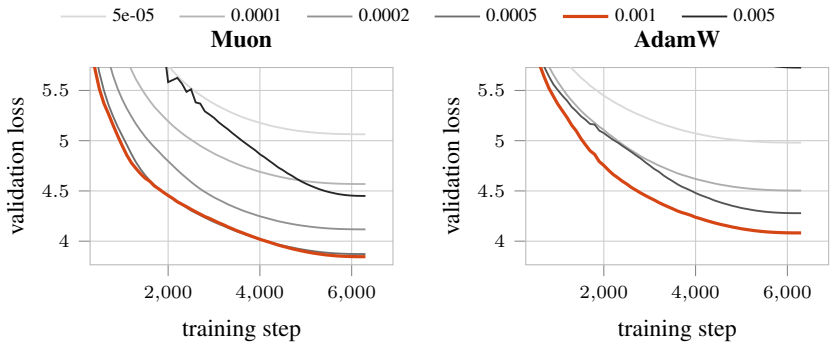


Figure 5: The same Qwen2.5-0.5B sweep as validation curves, evaluated every 100 steps; the selected learning rate is drawn in colour and the remaining grid points in grey. The loss is single-peaked in $\log \eta$ for both optimisers, so neither optimum is an artefact of the grid's endpoints.

answer, following the dataset's reference recipe. A held-out 2% of the training data serves as the validation set, evaluated every 100 steps.

Optimizer. We train with AdamW ($\eta = 2 \times 10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$, weight decay 0.01, gradient clipping at 1.0), using a 3% linear warmup followed by linear decay to zero, for 1500 steps on the $\kappa = 16$ selected examples. All runs share the same data order and validation stream.

Proposition E.1 (Polar-metric interpretation)**.** *Let $\mathbf{A}_t^*$ denote the adjoint of $\mathbf{A}_t$ with respect to the Frobenius inner product i.e. $(\langle \mathbf{A}_t[\boldsymbol{\rho}_i], \mathbf{A}_t[\boldsymbol{\rho}_j] \rangle_F = \langle \boldsymbol{\rho}_i, \mathbf{A}_t^* \mathbf{A}_t[\boldsymbol{\rho}_j] \rangle_F)$. Then $\mathbf{K}_{ij}^{\mu} = \langle \mathbf{g}_i, \mathbf{A}_t^* \mathbf{A}_t[\mathbf{g}_j] \rangle_F$ and $\mu_i^{\mu} = \langle \mathbf{g}_i, \mathbf{A}_t^*[\mathbf{g}_{\mathrm{val}}] \rangle_F + \frac{\eta_t}{2} \langle \mathbf{g}_i, \mathbf{A}_t^* \mathbf{A}_t[\mathbf{g}_i] \rangle_F$. Consequently, μ-SMAS is equivalent to applying* GREATS *with validation direction $\mathbf{A}_t^*[\mathbf{g}_{\mathrm{val}}]$ and gradient metric $\mathbf{A}_t^* \mathbf{A}_t$, rather than the Euclidean metric.*

Proof. By the definition of the adjoint,

$$\langle \mathbf{A}_t[\mathbf{g}_i], \mathbf{A}_t[\mathbf{g}_j] \rangle_F = \langle \mathbf{g}_i, \mathbf{A}_t^* \mathbf{A}_t[\mathbf{g}_j] \rangle_F.$$

Likewise,

$$\langle \mathbf{A}_t[\mathbf{g}_i], \mathbf{g}_{\mathrm{val}} \rangle_F = \langle \mathbf{g}_i, \mathbf{A}_t^*[\mathbf{g}_{\mathrm{val}}] \rangle_F.$$

Substitution into the definitions of $\mathbf{K}^{\mu}$ and $\boldsymbol{\mu}^{\mu}$ proves the result. □

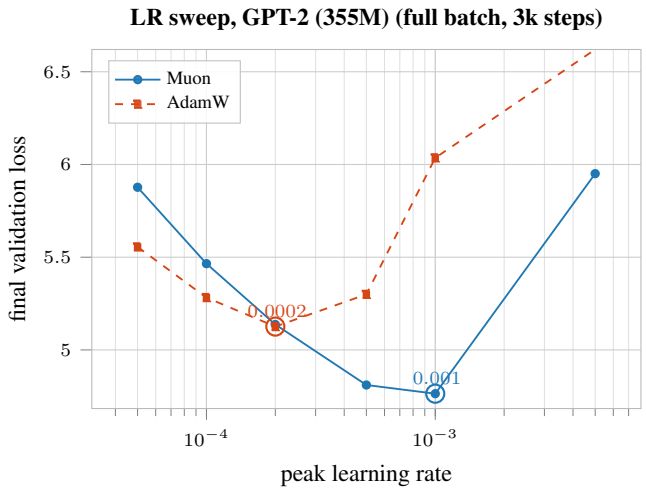


Figure 6: **Learning-rate sweep for GPT-2 medium (355M), Muon *vs* AdamW**, trained with the full batch for 3000 steps at batch 16×512, cosine schedule with 100 warm-up steps, seed 42. Circled points mark each optimiser's optimum. Muon peaks at $\eta = 0.001$ (final validation loss 4.764), AdamW at $\eta = 0.0002$ (5.1259) — a factor of 5 apart, which is why a single shared learning rate would misrepresent one of the two.

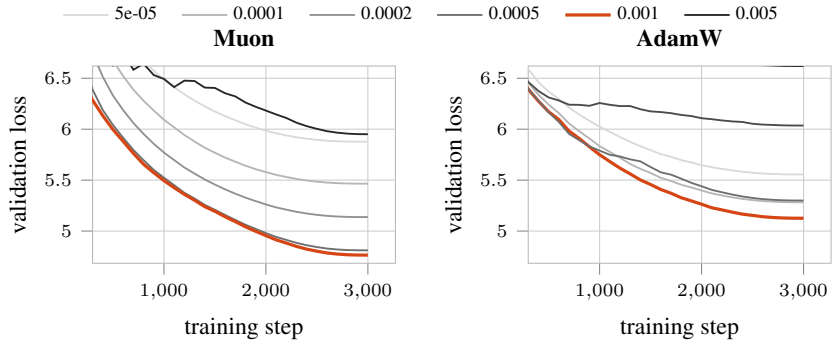


Figure 7: The same GPT-2 medium (355M) sweep as validation curves, evaluated every 100 steps; the selected learning rate is drawn in colour and the remaining grid points in grey. The loss is single-peaked in $\log \eta$ for both optimisers, so neither optimum is an artefact of the grid's endpoints.

Proposition E.2 (Geometry-dependent GreedyOMP rule)**.** *Let $\mathbf{w}_{\mathfrak{R}_i}^{\langle\varepsilon\rangle\dagger}$ and $\mathbf{w}_{\mathfrak{R}_i}^{\langle\mu\rangle\dagger}$ be the weight vectors obtained by optimizing the respective objectives over the current support $\mathfrak{R}_i$ at the ith iteration of* **??** *(for $i = 1, \ldots, \kappa$). Define the GreedyOMP residuals*

$$r_{\mathbf{x}}^{\langle\varepsilon\rangle}(\mathfrak{R}_i) = \boldsymbol{\mu}_{\boldsymbol{\theta}_t}^{\langle\mu\rangle}[\mathbf{x}] - \eta_t \sum_{j \in \mathfrak{R}_i} \mathbf{w}_{\mathfrak{R}_i}^{\langle\varepsilon\rangle\dagger}[j] \left[\mathbf{K}_{\boldsymbol{\theta}_t}^{\langle\mu\rangle}\right]_{\mathbf{x}j},$$

$$r_{\mathbf{x}}^{\langle\mu\rangle}(\mathfrak{R}_i) = \boldsymbol{\mu}_{\boldsymbol{\theta}_t}^{\langle\mu\rangle}[\mathbf{x}] - \eta_t \sum_{j \in \mathfrak{R}_i} \mathbf{w}_{\mathfrak{R}_i}^{\langle\mu\rangle\dagger}[j] \left[\mathbf{K}_{\boldsymbol{\theta}_t}^{\langle\mu\rangle}\right]_{\mathbf{x}j}.$$

Consequently,

$$\mathbf{x}_{\langle\varepsilon\rangle}^{\star} \in \arg\max_{\mathbf{x} \in \mathcal{B}_t \setminus \mathfrak{R}_i} r_{\mathbf{x}}^{\langle\varepsilon\rangle}(\mathfrak{R}_i), \qquad \mathbf{x}_{\langle\mu\rangle}^{\star} \in \arg\max_{\mathbf{x} \in \mathcal{B}_t \setminus \mathfrak{R}_i} r_{\mathbf{x}}^{\langle\mu\rangle}(\mathfrak{R}_i).$$

Hence the two methods select different samples whenever their transformed relevance–redundancy residuals induce different orderings.

Proof. For either geometry $a \in \{\varepsilon, \mu\}$,

$$\mathcal{U}_t^{\langle a\rangle}(\mathbf{w}) = \langle \mathbf{w}, \boldsymbol{\mu}^{\langle a\rangle} \rangle - \frac{\eta_t}{2} \langle \mathbf{w}, \mathbf{K}^{\langle a\rangle} \mathbf{w} \rangle.$$

Therefore,

$$\nabla_{\mathbf{w}} \mathcal{U}_t^a(\mathbf{w}) = \boldsymbol{\mu}^a - \eta_t \mathbf{K}^a \mathbf{w}.$$

Evaluating the i-th coordinate at the optimized vector supported on $\mathcal{S}$ gives the stated residual. GreedyOMP chooses the unselected coordinate with the largest positive residual. □

Throughout, freeze the Adam second-moment accumulator at its current value $\mathbf{v}_t^{\mathrm{ref}}$, exactly as $\boldsymbol{m}_{\boldsymbol{\theta}_t}^{\mathrm{ref}}$ is frozen for Muon, and define the diagonal (coordinatewise) operator

$$\mathbf{B}_t[\mathbf{X}] \;:=\; \mathbf{X} \oslash \left(\sqrt{\mathbf{v}_t^{\mathrm{ref}}} + \varepsilon\right) \qquad \text{(Hadamard division)},$$

so that $\mathbf{g}_i^{\langle\varepsilon\rangle} := \mathbf{B}_t[\mathbf{g}_i]$ recovers the Adam-preconditioned gradient of Equation (4). Since $\mathbf{B}_t$ acts entrywise by a real diagonal matrix, it is self-adjoint under $\langle\cdot,\cdot\rangle_F$, i.e. $\mathbf{B}_t^* = \mathbf{B}_t$. Define, in exact parallel with $\boldsymbol{\mu}^\mu, \mathbf{K}^\mu$ of Equation (13),

$$K_{ij}^{\langle\varepsilon\rangle} := \langle \mathbf{g}_i^{\langle\varepsilon\rangle}, \mathbf{g}_j^{\langle\varepsilon\rangle}\rangle_F, \qquad \mu_i^{\langle\varepsilon\rangle} := \langle \mathbf{g}_i^{\langle\varepsilon\rangle}, \mathbf{g}_{\mathsf{val}}\rangle_F + \tfrac{\eta_t}{2} K_{ii}^{\langle\varepsilon\rangle}.$$

By Proposition E.1 applied with $\mathbf{A}_t \to \mathbf{B}_t$ (and $\mathbf{B}_t^*\mathbf{B}_t = \mathbf{D}_t := \mathbf{B}_t^2$, since $\mathbf{B}_t$ is already self-adjoint), $\langle\varepsilon\rangle$-GREEDYOMP is exactly GREATS run in the metric $\mathbf{D}_t$ and validation direction $\mathbf{B}_t[\mathbf{g}_{\mathsf{val}}]$ — the coordinatewise counterpart of what μ-SMAS does with $\mathbf{A}_t^*\mathbf{A}_t$.

Recall the per-iteration subproblem is a *constrained* concave program, not free optimization: for support $\mathfrak{R}_i$ and $a \in \{\langle\varepsilon\rangle, \langle\mu\rangle\}$,

$$\mathbf{w}_{\mathfrak{R}_i}^{a\dagger} \in \underset{\substack{\mathbf{w}\geq 0 \\ \mathrm{supp}(\mathbf{w})\subseteq\mathfrak{R}_i}}{\arg\max} \; \mathcal{U}_t^a(\mathbf{w}) = \langle \mathbf{w}, \boldsymbol{\mu}^a\rangle - \tfrac{\eta_t}{2}\langle \mathbf{w}, \mathbf{K}^a\mathbf{w}\rangle. \tag{$\star$}$$

Since $\mathbf{K}^a$ is a Gram matrix, $\mathbf{K}^a \succeq 0$, so $\mathcal{U}_t^a$ is concave on $\mathbb{R}_{\geq 0}^{|\mathfrak{R}_i|}$ and KKT is both necessary and sufficient for ($\star$).

Lemma E.1 (KKT characterization on the nonnegative orthant)**.** *Fix $a \in \{\langle\varepsilon\rangle, \langle\mu\rangle\}$ and support $\mathfrak{R}_i$, and write $r_j^a(\mathfrak{R}_i) := \mu_j^a - \eta_t\big(\mathbf{K}^a \mathbf{w}_{\mathfrak{R}_i}^{a\dagger}\big)_j$ for $j \in \mathfrak{R}_i$ (extended by $\mathbf{w}_{\mathfrak{R}_i}^{a\dagger}[j] := 0$ for $j \notin \mathfrak{R}_i$). A vector $\mathbf{w}_{\mathfrak{R}_i}^{a\dagger} \geq 0$ supported on $\mathfrak{R}_i$ solves ($\star$) if and only if there exists $\boldsymbol{\lambda}^a \geq 0$ supported on $\mathfrak{R}_i$ with*

$$r_j^a(\mathfrak{R}_i) = -\lambda_j^a \;\leq\; 0, \qquad \lambda_j^a \cdot w_{\mathfrak{R}_i}^{a\dagger}[j] = 0, \qquad \forall\, j \in \mathfrak{R}_i.$$

In particular $r_j^a(\mathfrak{R}_i) = 0$ on every strictly active *coordinate ($w_{\mathfrak{R}_i}^{a\dagger}[j] > 0$), but $r_j^a(\mathfrak{R}_i)$ may be strictly negative on a coordinate that is in $\mathfrak{R}_i$ yet zeroed out by the constraint.*

Proof. The Lagrangian of ($\star$) is $L(\mathbf{w}, \boldsymbol{\lambda}) = \mathcal{U}_t^a(\mathbf{w}) + {\boldsymbol{\lambda}^a}^\top \mathbf{w}$, $\boldsymbol{\lambda} \geq 0$ dual to $\mathbf{w} \geq 0$. Stationarity in $\mathbf{w}$ gives $\nabla_{\mathbf{w}}\mathcal{U}_t^a(\mathbf{w}_{\mathfrak{R}_i}^{a\dagger}) + \boldsymbol{\lambda}^a = 0$, i.e. $\lambda_j^a = \eta_t(\mathbf{K}^a\mathbf{w}_{\mathfrak{R}_i}^{a\dagger})_j - \mu_j^a = -r_j^a(\mathfrak{R}_i)$ for $j \in \mathfrak{R}_i$. Dual feasibility $\lambda_j^a \geq 0$ gives $r_j^a(\mathfrak{R}_i) \leq 0$; complementary slackness forces $\lambda_j^a = 0$ (hence $r_j^a = 0$) whenever $w_{\mathfrak{R}_i}^{a\dagger}[j] > 0$. Concavity of $\mathcal{U}_t^a$ (from $\mathbf{K}^a \succeq 0$) and convexity of the feasible set make KKT sufficient as well as necessary for ($\star$). □

Lemma E.2 (Adam and Muon are generically not simultaneously diagonalizable)**.** *Let $\boldsymbol{m}_{\boldsymbol{\theta}_t} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ be a compact SVD with $\sigma_1, \ldots, \sigma_r > 0$ pairwise distinct, and extend $\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ (the columns of $\mathbf{U}$) to an orthonormal basis $\{\mathbf{u}_1, \ldots, \mathbf{u}_d\}$ of $\mathbb{R}^d$; $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$ (columns of $\mathbf{V}$) already form an orthonormal basis of $\mathbb{R}^r$. Assume the Adam accumulator is non-degenerate: the dr values $(B_t)_{pq}^2$, $p \in [d], q \in [r]$, are pairwise distinct.*

(a) $\mathbf{D}_t = \mathbf{B}_t^2$ *has* simple *spectrum in the elementary basis $\{\mathbf{e}_p\mathbf{e}_q^\top\}_{p\in[d],q\in[r]}$: each $\mathbf{e}_p\mathbf{e}_q^\top$ is an eigenvector with eigenvalue $(B_t)_{pq}^2$, and by non-degeneracy this is the* unique *eigenbasis of $\mathbf{D}_t$ (up to per-vector sign/scale).*

(b) $\mathbf{A}_t^*\mathbf{A}_t$ *decomposes the elementary basis $\{\mathbf{u}_a\mathbf{v}_b^\top\}_{a\in[d],b\in[r]}$ into invariant blocks:*

- *for $a > r$: $\mathbf{u}_a\mathbf{v}_b^\top$ is* itself *an exact eigenvector of both $\mathbf{A}_t$ and $\mathbf{A}_t^*$ (hence of $\mathbf{A}_t^*\mathbf{A}_t$), with eigenvalue σ_b^{-2} (Theorem E.1);*
- *for $a \neq b$, both $\leq r$: $\mathbf{A}_t^*\mathbf{A}_t$ acts within the 2-dimensional span $\{\mathbf{u}_a\mathbf{v}_b^\top, \mathbf{u}_b\mathbf{v}_a^\top\}$, via the Sylvester weights of Equation (19);*
- *for $a = b \leq r$: $\mathbf{u}_a\mathbf{v}_a^\top \in \ker\big(\mathscr{D}\mathcal{P}_{\circledcirc}(\boldsymbol{m}_{\boldsymbol{\theta}_t})\big)$ — it is of the symmetric form $\mathbf{QC}$ with $\mathbf{C} = \mathbf{v}_a\mathbf{v}_a^\top$ symmetric — so this is a 1-dimensional zero-eigenspace.*

In every case $\mathbf{A}_t^*\mathbf{A}_t$'s eigenbasis is built entirely from $\{\mathbf{u}_a\mathbf{v}_b^\top\}$, block-diagonal with blocks of size ≤ 2; the $a > r$ blocks, indexed by b, carry pairwise distinct eigenvalues σ_b^{-2} since the σ_b are pairwise distinct.

(c) *Suppose $r < d$ (always true in practice: LoRA rank or row-count r is $\ll$ hidden/embedding dimension d) and $\mathbf{V}$ is not a signed permutation matrix, i.e. some standard basis vector $\mathbf{e}_q \in \mathbb{R}^r$ is not proportional to any single column $\mathbf{v}_b$ — a mild non-degeneracy satisfied by every generic momentum matrix. Then $\mathbf{D}_t$ and $\mathbf{A}_t^*\mathbf{A}_t$ do not commute. Consequently there is no scalar c under which $K_{ij}^{\langle\varepsilon\rangle} = c\,K_{ij}^{\mu}$ for all i, j: unlike Corollary E.5 (where* GREATS *is the isotropic special case of μ-*SMAS within the same eigenbasis*), $\langle\varepsilon\rangle$ and $\langle\mu\rangle$ cannot be related this way even when $\sigma_1 = \cdots = \sigma_r$, because their eigenbases already differ.*

Proof. **(a)** is immediate from the entrywise action of the Hadamard-diagonal operator $\mathbf{B}_t$, together with the standard fact that a self-adjoint operator with pairwise distinct eigenvalues has a unique eigenbasis.

(b), case $a > r$. By Theorem E.1, $\mathbf{A}_t[\mathbf{u}_a\mathbf{v}_b^\top] = \sigma_b^{-1}\mathbf{u}_a\mathbf{v}_b^\top$ exactly. To see this basis vector is also an eigenvector of the *adjoint* $\mathbf{A}_t^*$: for any $\mathbf{u}_c\mathbf{v}_d^\top$ with $c > r$, orthonormality of $\{\mathbf{u}_a\mathbf{v}_b^\top\}$ gives $\langle \mathbf{A}_t[\mathbf{u}_c\mathbf{v}_d^\top], \mathbf{u}_a\mathbf{v}_b^\top\rangle_F = \sigma_d^{-1}\delta_{ca}\delta_{db}$; for $c \le r$, $\mathbf{A}_t[\mathbf{u}_c\mathbf{v}_d^\top]$ lies in $\operatorname{span}\{\mathbf{u}_c\mathbf{v}_d^\top, \mathbf{u}_d\mathbf{v}_c^\top\}$ (both first-indices $\le r$), which is orthogonal to $\mathbf{u}_a\mathbf{v}_b^\top$ ($a > r$), so this inner product is 0. Matching coefficients in the defining relation $\langle \mathbf{A}_t[\mathbf{u}_c\mathbf{v}_d^\top], \mathbf{u}_a\mathbf{v}_b^\top\rangle_F = \langle \mathbf{u}_c\mathbf{v}_d^\top, \mathbf{A}_t^*[\mathbf{u}_a\mathbf{v}_b^\top]\rangle_F$ for every (c, d) forces $\mathbf{A}_t^*[\mathbf{u}_a\mathbf{v}_b^\top] = \sigma_b^{-1}\mathbf{u}_a\mathbf{v}_b^\top$, so $\mathbf{A}_t^*\mathbf{A}_t[\mathbf{u}_a\mathbf{v}_b^\top] = \sigma_b^{-2}\mathbf{u}_a\mathbf{v}_b^\top$ exactly, as claimed.

(b), case $a \le r$. The projection term in Equation (17) vanishes since $\mathbf{u}_a \in \operatorname{col}(\mathbf{Q})$, so the derivative of $\rho = \mathbf{u}_a\mathbf{v}_b^\top$ comes entirely from the Sylvester correction $\mathbf{Q}\boldsymbol{\Omega}$; here $\mathbf{Q}^\top\rho = \mathbf{v}_a\mathbf{v}_b^\top$, so $\operatorname{skew}(\mathbf{Q}^\top\rho) = \frac{1}{2}(\mathbf{v}_a\mathbf{v}_b^\top - \mathbf{v}_b\mathbf{v}_a^\top)$ is supported entirely on coordinates (a, b) and (b, a) of $\boldsymbol{\Omega}$ in the $\mathbf{V}$-basis, and Equation (18) then forces $\boldsymbol{\Omega}$'s solution to be supported there too — proving the derivative of $\mathbf{u}_a\mathbf{v}_b^\top$ is a combination of only $\mathbf{u}_a\mathbf{v}_b^\top$ and $\mathbf{u}_b\mathbf{v}_a^\top$. When $a = b$, $\operatorname{skew}(\mathbf{Q}^\top\rho) = 0$ identically, so the unique solution of equation 18 with $\mathbf{P} \succ 0$ is $\boldsymbol{\Omega} = 0$, giving the stated kernel membership.

(c). Suppose toward contradiction that $\mathbf{D}_t$ and $\mathbf{A}_t^*\mathbf{A}_t$ commute. Both are self-adjoint, hence normal, so they are simultaneously orthogonally diagonalizable. Since $\mathbf{D}_t$ has simple spectrum by (a), its eigenbasis is rigid: *any* operator commuting with $\mathbf{D}_t$ must itself be diagonal in $\{\mathbf{e}_p\mathbf{e}_q^\top\}$ or equivalently, every $\mathbf{e}_p\mathbf{e}_q^\top$ must be an eigenvector of $\mathbf{A}_t^*\mathbf{A}_t$.

We exhibit one that is not. By hypothesis, fix q with $\mathbf{e}_q = \sum_{b=1}^r d_b\mathbf{v}_b$ having at least two nonzero coefficients $d_{b_1}, d_{b_2} \neq 0$. Since $r < d$, fix any $a_0 > r$ and, since $\mathbf{u}_{a_0} \neq \mathbf{0}$, fix p with $c_{a_0} := \langle \mathbf{e}_p, \mathbf{u}_{a_0}\rangle \neq 0$ in the expansion $\mathbf{e}_p = \sum_{a=1}^d c_a\mathbf{u}_a$. Then

$$\mathbf{e}_p\mathbf{e}_q^\top = \sum_{a=1}^{d}\sum_{b=1}^{r} c_a d_b\,\mathbf{u}_a\mathbf{v}_b^\top$$

has, among its terms, $c_{a_0}d_{b_1}\mathbf{u}_{a_0}\mathbf{v}_{b_1}^\top$ and $c_{a_0}d_{b_2}\mathbf{u}_{a_0}\mathbf{v}_{b_2}^\top$, both nonzero and both lying in $a_0 > r$ blocks (by (b)) with eigenvalues $\sigma_{b_1}^{-2} \neq \sigma_{b_2}^{-2}$ respectively. A vector with nonzero projection onto two eigenspaces of distinct eigenvalues cannot itself be an eigenvector, so $\mathbf{e}_p\mathbf{e}_q^\top$ is not an eigenvector of $\mathbf{A}_t^*\mathbf{A}_t$ — contradicting the rigidity requirement above. Hence $\mathbf{D}_t$ and $\mathbf{A}_t^*\mathbf{A}_t$ do not commute.

Finally, non-commuting self-adjoint operators cannot be proportional (proportionality $\mathbf{D}_t = c\,\mathbf{A}_t^*\mathbf{A}_t$ trivially implies commutativity), so no scalar c can give $K_{ij}^{\langle\varepsilon\rangle} = \langle \mathbf{g}_i, \mathbf{D}_t\mathbf{g}_j\rangle_F = c\,\langle \mathbf{g}_i, \mathbf{A}_t^*\mathbf{A}_t\mathbf{g}_j\rangle_F = c\,K_{ij}^{\mu}$ for all i, j simultaneously, since this would require the operator identity $\mathbf{D}_t = c\,\mathbf{A}_t^*\mathbf{A}_t$. □

Corollary E.1 (Quantitative disagreement at the first greedy step)**.** *At $i = 1$ ($\mathfrak{R}_1 = \emptyset$), ($\star$) has no active support and $r_{\mathbf{x}}^a(\emptyset) = \mu_{\mathbf{x}}^a$ exactly for $a \in \{\langle\varepsilon\rangle, \langle\mu\rangle\}$. Define the mismatch operator $\mathbf{M}_t := \mathbf{D}_t - \mathbf{A}_t^*\mathbf{A}_t$ (self-adjoint; by Lemma E.2(c), generically nonzero — not claimed sign-definite). Then for every candidate $\mathbf{x}$,*

$$\left|\mu_{\mathbf{x}}^{\langle\varepsilon\rangle} - \mu_{\mathbf{x}}^{\langle\mu\rangle}\right| \;\le\; \|\mathbf{g}_{\mathbf{x}}\|_F\,\left\|(\mathbf{B}_t - \mathbf{A}_t^*)[\mathbf{g}_{\mathsf{val}}]\right\|_F + \tfrac{\eta_t}{2}\,\|\mathbf{M}_t\|_{\mathrm{op}}\,\|\mathbf{g}_{\mathbf{x}}\|_F^2 \;=:\; \epsilon_{\mathbf{x}}.$$

Let $\mathbf{x}_{\langle\varepsilon\rangle}^\star = \arg\max_{\mathbf{x}} \mu_{\mathbf{x}}^{\langle\varepsilon\rangle}$ with margin $\Delta_1^{\langle\varepsilon\rangle} := \mu_{\mathbf{x}_{\langle\varepsilon\rangle}^\star}^{\langle\varepsilon\rangle} - \max_{\mathbf{x}\neq\mathbf{x}_{\langle\varepsilon\rangle}^\star} \mu_{\mathbf{x}}^{\langle\varepsilon\rangle}$. If $\Delta_1^{\langle\varepsilon\rangle} > 2\max_{\mathbf{x}\in\mathcal{B}_t}\epsilon_{\mathbf{x}}$, then $\mathbf{x}_{\langle\mu\rangle}^\star = \mathbf{x}_{\langle\varepsilon\rangle}^\star$ necessarily. Conversely, $\mathbf{x}_{\langle\varepsilon\rangle}^\star \neq \mathbf{x}_{\langle\mu\rangle}^\star$ is only possible when $\Delta_1^{\langle\varepsilon\rangle} \le 2\max_{\mathbf{x}}\epsilon_{\mathbf{x}}$ —

disagreement requires the Adam score gap to be dominated by $\|\mathbf{M}_t\|_{\mathrm{op}}$, which Lemma E.2(c) shows is generically nonzero (and hence a real, non-vacuous threshold to cross).

Proof. Since $\mathbf{B}_t$ is self-adjoint, $K_{ij}^{\langle\varepsilon\rangle} = \langle \mathbf{B}_t[\mathbf{g}_i],\, \mathbf{B}_t[\mathbf{g}_j]\rangle_F = \langle \mathbf{g}_i,\, \mathbf{B}_t^2[\mathbf{g}_j]\rangle_F = \langle \mathbf{g}_i,\, \mathbf{D}_t[\mathbf{g}_j]\rangle_F$, and likewise $\langle \mathbf{B}_t[\mathbf{g}_i],\, \mathbf{g}_{\mathsf{val}}\rangle_F = \langle \mathbf{g}_i,\, \mathbf{B}_t[\mathbf{g}_{\mathsf{val}}]\rangle_F$, giving

$$\mu_{\mathbf{x}}^{\langle\varepsilon\rangle} = \langle \mathbf{g}_{\mathbf{x}},\, \mathbf{B}_t[\mathbf{g}_{\mathsf{val}}]\rangle_F + \tfrac{\eta_t}{2}\langle \mathbf{g}_{\mathbf{x}},\, \mathbf{D}_t[\mathbf{g}_{\mathbf{x}}]\rangle_F.$$

Proposition E.1 gives the analogous expression for $\mu_{\mathbf{x}}^{\langle\mu\rangle}$ in terms of $\mathbf{A}_t^*$. Subtracting,

$$\mu_{\mathbf{x}}^{\langle\varepsilon\rangle} - \mu_{\mathbf{x}}^{\langle\mu\rangle} = \langle \mathbf{g}_{\mathbf{x}},\, (\mathbf{B}_t - \mathbf{A}_t^*)[\mathbf{g}_{\mathsf{val}}]\rangle_F + \tfrac{\eta_t}{2}\langle \mathbf{g}_{\mathbf{x}},\, \mathbf{M}_t[\mathbf{g}_{\mathbf{x}}]\rangle_F.$$

Cauchy–Schwarz on the first term and the operator-norm bound $|\langle \mathbf{g}_{\mathbf{x}},\, \mathbf{M}_t[\mathbf{g}_{\mathbf{x}}]\rangle_F| \le \|\mathbf{M}_t\|_{\mathrm{op}} \|\mathbf{g}_{\mathbf{x}}\|_F^2$ on the second give $\epsilon_{\mathbf{x}}$. For the margin claim: if $\Delta_1^{\langle\varepsilon\rangle} > 2\max_{\mathbf{x}} \epsilon_{\mathbf{x}}$, then for any $\mathbf{x}' \neq \mathbf{x}_{\langle\varepsilon\rangle}^\star$,

$$\mu_{\mathbf{x}_{\langle\varepsilon\rangle}^\star}^{\langle\mu\rangle} - \mu_{\mathbf{x}'}^{\langle\mu\rangle} \ge \big(\mu_{\mathbf{x}_{\langle\varepsilon\rangle}^\star}^{\langle\varepsilon\rangle} - \epsilon_{\mathbf{x}_{\langle\varepsilon\rangle}^\star}\big) - \big(\mu_{\mathbf{x}'}^{\langle\varepsilon\rangle} + \epsilon_{\mathbf{x}'}\big) \ge \Delta_1^{\langle\varepsilon\rangle} - 2\max_{\mathbf{x}} \epsilon_{\mathbf{x}} > 0,$$

so $\mathbf{x}_{\langle\varepsilon\rangle}^\star$ strictly beats every alternative under $\langle\mu\rangle$ as well, i.e. $\mathbf{x}_{\langle\mu\rangle}^\star = \mathbf{x}_{\langle\varepsilon\rangle}^\star$. □

Proposition E.3. *Let $\mathbf{w}_{\mathfrak{R}_i}^{\langle\varepsilon\rangle\dagger}, \mathbf{w}_{\mathfrak{R}_i}^{\langle\mu\rangle\dagger}$ solve $(\star)$ for $\mathcal{U}_t^{\langle\varepsilon\rangle}, \mathcal{U}_t^{\langle\mu\rangle}$ respectively at the i-th iteration of Algorithm 1, and let $r_{\mathbf{x}}^{\langle\varepsilon\rangle}(\mathfrak{R}_i), r_{\mathbf{x}}^{\langle\mu\rangle}(\mathfrak{R}_i)$ be as in Lemma E.1, defined for all $\mathbf{x} \in \mathcal{B}_t$. Then:*

(i) *$r_j^a(\mathfrak{R}_i) \le 0$ for every $j \in \mathfrak{R}_i$ and every $a \in \{\langle\varepsilon\rangle, \langle\mu\rangle\}$, with equality exactly on the strictly-active coordinates.*

(ii) *If the greedy step is non-degenerate — some $\mathbf{x} \in \mathcal{B}_t \setminus \mathfrak{R}_i$ has $r_{\mathbf{x}}^a(\mathfrak{R}_i) > 0$ — then the constrained-domain argmax and the full-domain argmax coincide:*

$$\mathbf{x}_a^\star \in \arg\max_{\mathbf{x}\in\mathcal{B}_t\setminus\mathfrak{R}_i} r_{\mathbf{x}}^a(\mathfrak{R}_i) = \arg\max_{\mathbf{x}\in\mathcal{B}_t} r_{\mathbf{x}}^a(\mathfrak{R}_i), \qquad a \in \{\langle\varepsilon\rangle, \langle\mu\rangle\}.$$

(iii) *Whenever $\mathbf{D}_t$ and $\mathbf{A}_t^*\mathbf{A}_t$ do not commute (Lemma E.2), $\mathbf{x}_{\langle\varepsilon\rangle}^\star \neq \mathbf{x}_{\langle\mu\rangle}^\star$ for an open, non-empty set of candidate configurations.*

Proof. (i) is Lemma E.1. For (ii): by (i), every $j \in \mathfrak{R}_i$ has $r_j^a(\mathfrak{R}_i) \le 0$, while by the non-degeneracy hypothesis the true maximizer $\mathbf{x}^\star$ over $\mathcal{B}_t \setminus \mathfrak{R}_i$ satisfies $r_{\mathbf{x}^\star}^a(\mathfrak{R}_i) > 0 \ge r_j^a(\mathfrak{R}_i)$ for every $j \in \mathfrak{R}_i$; hence widening the search domain to include $\mathfrak{R}_i$ cannot change the maximizer. (iii) is Corollary E.1 below, instantiated at $\mathfrak{R}_1 = \emptyset$, where $(\star)$ has no active support to reason about and $r_{\mathbf{x}}^a(\emptyset) = \mu_{\mathbf{x}}^a$ exactly, so the constraint plays no role at this step. □

Theorem E.1 (Inverse-spectral balancing of sample importance)**.** *Let $\boldsymbol{m}_{\boldsymbol{\theta}_t} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ be a compact SVD with $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$, $\sigma_k > 0$, and write $\mathbf{Q} := \mathcal{P}_\circledcirc(\boldsymbol{m}_{\boldsymbol{\theta}_t}) = \mathbf{U}\mathbf{V}^\top$. Fix $k_i \in [r]$ and consider a sample gradient $\mathbf{g}_i = \mathbf{x}_i\mathbf{v}_{k_i}^\top$ with $\mathbf{U}^\top\mathbf{x}_i = 0$. Then*

$$\mathscr{D}\mathcal{P}_\circledcirc(\boldsymbol{m}_{\boldsymbol{\theta}_t})[\mathbf{g}_i] = \sigma_{k_i}^{-1}\mathbf{g}_i, \tag{41}$$

$$a_i^\mu := \big\langle \mathscr{D}\mathcal{P}_\circledcirc(\boldsymbol{m}_{\boldsymbol{\theta}_t})[\mathbf{g}_i], \mathbf{g}_{\mathrm{val}}\big\rangle_F = \sigma_{k_i}^{-1}\langle \mathbf{g}_i, \mathbf{g}_{\mathrm{val}}\rangle_F = \frac{a_i^{\langle\varepsilon\rangle}}{\sigma_{k_i}}, \tag{42}$$

where $a_i^{\langle\varepsilon\rangle} := \langle \mathbf{g}_i, \mathbf{g}_{\mathrm{val}}\rangle_F$ is the vanilla GREATS *relevance score. In particular, if $a_i^{\langle\varepsilon\rangle} = a_j^{\langle\varepsilon\rangle} > 0$ and $\sigma_{k_i} < \sigma_{k_j}$, then $a_i^\mu > a_j^\mu$, while vanilla* GREATS *is indifferent.*

Proof. The derivative of the rectangular polar factor is

$$\mathscr{D}\mathcal{P}_\circledcirc(\boldsymbol{m}_{\boldsymbol{\theta}_t})[\mathbf{H}] = (\mathbf{I} - \mathbf{Q}\mathbf{Q}^\top)\mathbf{H}\mathbf{P}^{-1} + \mathbf{Q}\boldsymbol{\Omega},$$

where

$$\mathbf{P} = (\boldsymbol{m}_{\boldsymbol{\theta}_t}^\top \boldsymbol{m}_{\boldsymbol{\theta}_t})^{1/2} = \mathbf{V}\boldsymbol{\Sigma}\mathbf{V}^\top.$$

Since $\mathbf{U}^\top\mathbf{x}_i = 0$, we have

$$\mathbf{Q}^\top\mathbf{g}_i = \mathbf{V}\mathbf{U}^\top\mathbf{x}_i\mathbf{v}_{k_i}^\top = 0.$$

Therefore, the Sylvester equation defining $\boldsymbol{\Omega}$ has zero right-hand side, so $\boldsymbol{\Omega} = 0$. Moreover,

$$(\mathbf{I} - \mathbf{Q}\mathbf{Q}^\top)\mathbf{g}_i = \mathbf{g}_i.$$

Since

$$\mathbf{P}^{-1}\mathbf{v}_{k_i} = \frac{1}{\sigma_{k_i}}\mathbf{v}_{k_i},$$

we obtain

$$\mathscr{DP}_{\circledcirc}(\boldsymbol{m}_{\boldsymbol{\theta}_t})[\mathbf{g}_i] = \mathbf{g}_i\mathbf{P}^{-1} = \frac{1}{\sigma_{k_i}}\mathbf{g}_i.$$

Taking its Frobenius inner product with the validation gradient proves the score identity. □

Remark 2 (The self-term rescales at the same rate, squared)**.** Equation (41) gives the full transformed gradient, not merely its inner product with $\mathbf{g}_{\text{val}}$, so the same identity immediately gives the self-interaction term

$$K^{\mu}_{ii} = \|\mathscr{DP}_{\circledcirc}(\boldsymbol{m}_{\boldsymbol{\theta}_t})[\mathbf{g}_i]\|_F^2 = \|\mathbf{g}_i\|_F^2/\sigma_{k_i}^2 = K^{\langle\varepsilon\rangle}_{ii}/\sigma_{k_i}^2 :$$

the redundancy/self term rescales by $\sigma_{k_i}^{-2}$, the *square* of the linear relevance term's $\sigma_{k_i}^{-1}$ rescaling in Equation (42). Corollary E.3 extends this to the multi-mode case used throughout the rest of the paper, and to the full Gram matrix K_{ij} rather than only its diagonal.

We now show that the same failure mode is the *generic* one under a mild, literature-motivated structural assumption on how momentum and gradient variance co-evolve during training, thereby connecting the polar-Jacobian machinery of Section 5.2 directly to the question of *why* Muon's anisotropic update yields a better selection criterion than Adam's, rather than merely a different one.

Conceptually it's also the more interesting case anyway: content that reinforces an already-active direction barely needs smart selection, since momentum is already carrying it forward for free (that's the whole content of Corollary coverage-stability later) — the hard, decision-relevant question is which genuinely new direction is worth spending selection budget on, and that's exactly the regime this condition isolates.

Proposition E.4 (Modewise bias correction of relevance scores)**.** *Given $\boldsymbol{m}_{\boldsymbol{\theta}_t} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ with $\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1, \dots, \sigma_r)$, $\sigma_k > 0$, and $\mathbf{A}_t := \mathscr{DP}_{\circledcirc}(\boldsymbol{m}_{\boldsymbol{\theta}_t})$. Suppose a candidate gradient decomposes as*

$$\mathbf{g}_i = \sum_{k=1}^{r} \mathbf{x}_{i,k}\mathbf{v}_k^\top, \qquad \mathbf{U}^\top\mathbf{x}_{i,k} = 0 \ \ \forall k,$$

and write $\mathbf{y}_k := \mathbf{g}_{\text{val}}\mathbf{v}_k$ and $s_{i,k} := \mathbf{x}_{i,k}^\top\mathbf{y}_k$. Then

$$a_i^{\langle\varepsilon\rangle} := \langle\mathbf{g}_i, \mathbf{g}_{\text{val}}\rangle_F = \sum_{k=1}^{r} s_{i,k}, \qquad a_i^{\mu} := \langle\mathbf{A}_t[\mathbf{g}_i], \mathbf{g}_{\text{val}}\rangle_F = \sum_{k=1}^{r} \frac{s_{i,k}}{\sigma_k}.$$

Proof. By Theorem E.1, $\mathbf{A}_t[\mathbf{x}_{i,k}\mathbf{v}_k^\top] = \sigma_k^{-1}\mathbf{x}_{i,k}\mathbf{v}_k^\top$ for each k; sum over k using linearity of the Fréchet derivative $\mathbf{A}_t$, then take the Frobenius inner product with $\mathbf{g}_{\text{val}}$. □

GREATS scores a candidate by the *unweighted* sum of its per-mode relevances $s_{i,k}$, while μ-SMAS scores it by the *harmonic*, σ_k^{-1}-weighted sum , down-weighting modes the momentum has already accumulated substantial mass in, and up-weighting under-exploited modes, regardless of which sample happens to activate them.

Proposition E.4 states only the linear (relevance) term a_i; the following corollary supplies the quadratic (self/redundancy) term K_{ii} that the full GREEDYOMP score $\mu_i = a_i + \frac{\eta_t}{2}K_{ii}$ actually requires, and in fact the entire Gram matrix K_{ij}.

Corollary E.2 (Full Gram-matrix decomposition: linear and quadratic terms reweight at different rates)**.** *In the setting of Proposition E.4, for any two candidates $\mathbf{z}_i, \mathbf{z}_j \in \mathcal{B}_t$ write $\chi_{ij,k} := \mathbf{x}_{i,k}^\top\mathbf{x}_{j,k}$ for their overlap in mode k (so $\chi_{ii,k} = \|\mathbf{x}_{i,k}\|_2^2$). Then the Euclidean and Muon Gram matrices decompose modewise as*

$$K_{ij}^{\langle\varepsilon\rangle} = \langle\mathbf{g}_i, \mathbf{g}_j\rangle_F = \sum_{k=1}^{r}\chi_{ij,k}, \qquad K_{ij}^{\langle\mu\rangle} = \langle\mathbf{A}_t[\mathbf{g}_i], \mathbf{A}_t[\mathbf{g}_j]\rangle_F = \sum_{k=1}^{r}\frac{\chi_{ij,k}}{\sigma_k^2}.$$

Combined with Proposition E.4, the GREEDYOMP *relevance vector $\mu_i^a = a_i^a + \frac{\eta_t}{2}K_{ii}^a$, $a \in \{\langle\varepsilon\rangle, \langle\mu\rangle\}$ (Equation (13)), decomposes as*

$$\mu_i^{\langle\varepsilon\rangle} = \sum_{k=1}^{r}\Big(s_{i,k} + \tfrac{\eta_t}{2}\,\chi_{ii,k}\Big), \qquad \mu_i^{\langle\mu\rangle} = \sum_{k=1}^{r}\Big(\frac{s_{i,k}}{\sigma_k} + \frac{\eta_t}{2}\,\frac{\chi_{ii,k}}{\sigma_k^2}\Big).$$

Thus the linear (relevance) term is reweighted by σ_k^{-1}, whereas every quadratic term (the self-interaction K_{ii} in μ_i and the pairwise overlaps K_{ij}, $i \neq j$, in the penalty $\frac{\eta_t^2}{2}\langle\mathbf{w}, \mathbf{K}\mathbf{w}\rangle$) is reweighted

by σ_k^{-2}. This is not a single uniform rescaling, because $\mathbf{A}_t$ enters a bilinear form twice but a linear form only once.

Relative to any other mode l, a dominant mode k ($\sigma_k > \sigma_l$) therefore has its relevance discounted by the factor σ_l/σ_k and its quadratic contributions by the stronger factor $(\sigma_l/\sigma_k)^2$. This comparison is invariant to the overall scale of $\boldsymbol{m}_{\boldsymbol{\theta}_t}$. It has two consequences.

(i) *The self-term $\frac{\eta_t}{2}K_{ii}$ rewards gradient magnitude independently of validation relevance. Under μ-SMAS, the part of this reward that comes from the dominant mode is suppressed more strongly than that mode's relevance. This reinforces the relevance debiasing of Proposition E.4.*

(ii) *Two candidates whose overlap $\chi_{ij,k}$ is concentrated in the dominant mode are penalised as less redundant under $\mathbf{K}^{\langle\mu\rangle}$ than under $\mathbf{K}^{\langle\varepsilon\rangle}$.*

Indeed, $K_{ij}^{\langle\mu\rangle} = \sum_k \tau_k^2 \chi_{ij,k}$ with $\tau_k = 1/\sigma_k$ is exactly the realised feature inner product whose second moment computes in expectation. The present identity holds deterministically, for every realisation, and requires no population model.

Proof. Write $\mathbf{g}_i = \sum_k \mathbf{x}_{i,k}\mathbf{v}_k^\top$. By bilinearity of the Frobenius inner product,

$$\langle \mathbf{g}_i, \mathbf{g}_j \rangle_F = \sum_{k,l=1}^{r} (\mathbf{x}_{i,k}^\top \mathbf{x}_{j,l})(\mathbf{v}_k^\top \mathbf{v}_l) = \sum_{k=1}^{r} \chi_{ij,k},$$

since $\mathbf{v}_k^\top \mathbf{v}_l = \delta_{kl}$ by orthonormality of the right singular vectors. By linearity of $\mathbf{A}_t$ and Theorem E.1 (applicable because $\mathbf{U}^\top \mathbf{x}_{i,k} = 0$), we have $\mathbf{A}_t[\mathbf{g}_i] = \sum_k \sigma_k^{-1}\mathbf{x}_{i,k}\mathbf{v}_k^\top$. The same computation then gives $K_{ij}^{\langle\mu\rangle} = \sum_k \sigma_k^{-2}\chi_{ij,k}$. Setting $j = i$ and substituting K_{ii}^a together with the decomposition of a_i^a from Proposition E.4 into $\mu_i^a = a_i^a + \frac{\eta_t}{2}K_{ii}^a$ yields the stated forms. For the relative-weight statement, the ratio of the mode-k to mode-l weights is σ_l/σ_k for the linear term and $(\sigma_l/\sigma_k)^2$ for the quadratic terms, and both ratios are unchanged under $\boldsymbol{m}_{\boldsymbol{\theta}_t} \mapsto \alpha\, \boldsymbol{m}_{\boldsymbol{\theta}_t}$. □

Corollary E.3 (Full Gram-matrix decomposition: linear and quadratic terms reweight at different rates)**.** *In the setting of Proposition E.4, write $t_{ij,k} := \mathbf{x}_{i,k}^\top \mathbf{x}_{j,k}$ for any two candidates $\mathbf{z}_i, \mathbf{z}_j$ (so $t_{ii,k} = \|\mathbf{x}_{i,k}\|^2$). The Gram matrix entries decompose by the same modewise argument, using orthogonality of $\mathbf{v}_1, \ldots, \mathbf{v}_r$ in place of linearity:*

$$K_{ij}^{\langle\varepsilon\rangle} = \langle \mathbf{g}_i, \mathbf{g}_j \rangle_F = \sum_{k=1}^{r} t_{ij,k}, \qquad K_{ij}^{\mu} = \langle \mathbf{A}_t[\mathbf{g}_i], \mathbf{A}_t[\mathbf{g}_j] \rangle_F = \sum_{k=1}^{r} \frac{t_{ij,k}}{\sigma_k^2}.$$

Combined with Proposition E.4, the full GREEDYOMP *score $\mu_i = a_i + \frac{\eta_t}{2}K_{ii}$ decomposes as*

$$\mu_i^{\langle\varepsilon\rangle} = \sum_{k=1}^{r} \Big(s_{i,k} + \tfrac{\eta_t}{2}\|\mathbf{x}_{i,k}\|^2 \Big), \qquad \mu_i^{\mu} = \sum_{k=1}^{r} \Big(\frac{s_{i,k}}{\sigma_k} + \frac{\eta_t}{2}\frac{\|\mathbf{x}_{i,k}\|^2}{\sigma_k^2} \Big).$$

The linear (relevance) term reweights by σ_k^{-1} and the quadratic (redundancy) term by σ_k^{-2} not a single uniform rescaling, since $\mathbf{A}_t$ enters a bilinear form twice but a linear form only once. An over-exploited mode (σ_k large) therefore has its contribution to redundancy suppressed more aggressively *than its contribution to relevance, so μ-SMAS's diversity-encouraging term discounts an already-dominant mode even more strongly than its relevance term does reinforcing, not merely complementing, Proposition E.4. This is also the exact, population-model-free counterpart, specialized to $\tau_k = 1/\sigma_k$, of the general τ_k^2-weighted second moment computed in expectation.*

Proof. By linearity and orthogonality of $\mathbf{v}_1, \ldots, \mathbf{v}_r$: $\langle \mathbf{g}_i, \mathbf{g}_j \rangle_F = \big\langle \sum_k \mathbf{x}_{i,k}\mathbf{v}_k^\top, \sum_l \mathbf{x}_{j,l}\mathbf{v}_l^\top \big\rangle_F = \sum_{k,l}(\mathbf{x}_{i,k}^\top \mathbf{x}_{j,l})(\mathbf{v}_k^\top \mathbf{v}_l) = \sum_k t_{ij,k}$, since $\mathbf{v}_k^\top \mathbf{v}_l = \delta_{kl}$. Applying the identical computation to $\mathbf{A}_t[\mathbf{g}_i] = \sum_k \sigma_k^{-1}\mathbf{x}_{i,k}\mathbf{v}_k^\top$ (established in the proof of Proposition E.4) and $\mathbf{A}_t[\mathbf{g}_j]$ gives K_{ij}^{μ}. Setting $i = j$ recovers K_{ii}; substituting into $\mu_i = a_i + \frac{\eta_t}{2}K_{ii}$ together with Proposition E.4's decomposition of a_i gives the stated forms of $\mu_i^{\langle\varepsilon\rangle}, \mu_i^{\mu}$. □

Assumption 6 (Momentum–variance proportionality)**.** For each mode $k \in [r]$, the population covariance of candidate-gradient blocks satisfies $\mathrm{Cov}(\mathbf{x}_{i,k}) = \lambda_k \mathbf{I}_d$, and the momentum spectrum has equilibrated to the gradient variance in each mode, i.e. $\sigma_k^2 = c^2\lambda_k$ for a constant $c > 0$ common to all k. This is the structural (matrix/spectral) analogue of the stationarity condition underlying

Adam's own diagonal second-moment estimate $\boldsymbol{v}_t$ (Section 3), applied instead to the singular-value spectrum of the momentum *matrix* rather than to individual coordinates. It is consistent with reported low-stable-rank / spiked structure in transformer gradients and momenta (Davis & Drusvyatskiy, 2025).

Corollary E.4 (Anisotropy-robust submodularity guarantee)**.** *Applying Proposition 5.2 to* $\boldsymbol{\Sigma}^{\mu}$ *and to* $\boldsymbol{\Sigma}^{\langle\varepsilon\rangle}$ *respectively (in place of its generic* $\boldsymbol{\Sigma}$*), with probability* $1-\delta$*,*

$$\gamma_{\mu} \;\geq\; 1 - O(\varepsilon), \qquad \gamma_{\langle\varepsilon\rangle} \;\geq\; \mathrm{cond}(\boldsymbol{\Sigma}^{\langle\varepsilon\rangle})^{-1} - O(\varepsilon).$$

Consequently μ-SMAS*'s greedy guarantee* $(1-e^{-\gamma_{\mu}})$ *attains (near-)the best possible generic submodular ratio* $(1-1/e)$ independently of the ambient gradient spectrum*, whereas the guaranteed ratio available to* GREATS *degrades directly with* $\mathrm{cond}(\boldsymbol{\Sigma}^{\langle\varepsilon\rangle})$*, the anisotropy of the raw gradient distribution. This is a comparison of guarantees, not of realized values: it does not by itself show* $\gamma_{\langle\varepsilon\rangle}$ *is small, only that no comparable anisotropy-independent guarantee is available for it. Under empirically documented spiked / low-effective-rank gradient spectra in transformer training (Sonthalia et al., 2026; Gur-Ari et al., 2018; Davis & Drusvyatskiy, 2025),* $\mathrm{cond}(\boldsymbol{\Sigma}) \gg 1$ *is the typical regime, making that worst case the practically relevant one rather than a pathology.*

Remark 3 (Why Adam cannot do this)**.** Adam's preconditioner $\boldsymbol{v}_t$ is diagonal in the *parameter* basis: it corrects coordinatewise variance but cannot see correlated, structural anisotropy across a weight matrix's row/column space. The polar Jacobian $\mathbf{A}_t$, by contrast, whitens with respect to the *singular* basis of the momentum matrix itself (Lemma 5.2). For matrix-shaped gradients where anisotropy is predominantly low-rank/structural rather than per-coordinate. Adam's correction is the wrong shape to remove the bias identified in Proposition E.4, while Muon's orthogonalization removes it by construction. This picture is corroborated by an independent line of optimization-theoretic work showing that Muon's orthogonalization shifts the effective learning-rate bottleneck from the largest singular value of the update to an *average* singular scale (Nguyen et al., 2026).

The results so far establish *that* GREATS and μ-SMAS diverge and *why* the divergence favors μ-SMAS under realistic anisotropy. We close the theoretical development with three results that together make the case that Muon-aware selection is not an incremental variant of GREATS, but strictly subsumes it, compounds its advantage over the course of training, and has a directly interpretable resource cost when ignored.

Corollary E.5 (GREATS as the isotropic degenerate case of μ-SMAS)**.** *Under the gradient class of Proposition E.4 (candidate gradients with* $\mathbf{U}^{\top}\mathbf{x}_{i,k} = 0$ *for all* k*), suppose the momentum spectrum is perfectly isotropic, i.e.* $\sigma_1 = \cdots = \sigma_r = \sigma$*. Then for every* $\mathbf{w} \geq 0$*,*

$$\mathcal{U}_t^{\langle\mu\rangle}(\mathbf{w}, \mathcal{B}_t; \mathbf{z}_{\mathsf{val}}; \eta_t) \;=\; \mathcal{U}_t^{\langle\varepsilon\rangle}(\mathbf{w}, \mathcal{B}_t; \mathbf{z}_{\mathsf{val}}; \eta_t/\sigma),$$

i.e. the μ-SMAS *objective at step size* η_t *is* pointwise identical*, as a function of* $\mathbf{w}$*, to the vanilla* GREATS *objective at rescaled step size* η_t/σ*. Consequently the optimal weights, the entire* GREEDY-OMP *selection trajectory, and the returned subset* $\widehat{\mathcal{S}}_t$ *coincide exactly between the two methods.*

Proof. Under the stated gradient class with $\sigma_k \equiv \sigma$ for all k, Proposition E.4 gives $a_i^{\mu} = a_i^{\langle\varepsilon\rangle}/\sigma$ for every candidate, and Corollary E.3 gives $K_{ij}^{\mu} = K_{ij}^{\langle\varepsilon\rangle}/\sigma^2$ for all i, j (both by specializing their general σ_k-dependence to a constant σ). Substituting into $\mu_i^{\mu} = a_i^{\mu} + \frac{\eta_t}{2}K_{ii}^{\mu} = \frac{1}{\sigma}\big(a_i^{\langle\varepsilon\rangle} + \frac{\eta_t/\sigma}{2}K_{ii}^{\langle\varepsilon\rangle}\big) = \frac{1}{\sigma}\,\mu_i^{\langle\varepsilon\rangle}(\eta_t/\sigma)$ and then into

$$\mathcal{U}_t^{\langle\mu\rangle}(\mathbf{w}; \eta_t) = \eta_t\langle\mathbf{w}, \boldsymbol{\mu}^{\mu}\rangle - \tfrac{\eta_t^2}{2}\langle\mathbf{w}, \mathbf{K}^{\mu}\mathbf{w}\rangle = \tfrac{\eta_t}{\sigma}\langle\mathbf{w}, \boldsymbol{\mu}^{\langle\varepsilon\rangle}(\eta_t/\sigma)\rangle - \tfrac{1}{2}\left(\tfrac{\eta_t}{\sigma}\right)^2\langle\mathbf{w}, \mathbf{K}^{\langle\varepsilon\rangle}\mathbf{w}\rangle = \mathcal{U}_t^{\langle\varepsilon\rangle}(\mathbf{w}; \eta_t/\sigma)$$

gives the claimed identity. Since the two objectives coincide as functions of $\mathbf{w}$ (not merely up to a monotone reparametrization of the score), every downstream quantity computed from them, the GREEDYOMP residuals of Proposition E.3, and hence the selected support at every iteration is identical. □

Corollary E.5 makes precise the sense in which μ-SMAS *strictly generalizes* GREATS: the two methods are not different heuristics that happen to agree in a limit, but literally the same optimization problem, with GREATS corresponding exactly to the isotropic ($\sigma_1 = \cdots = \sigma_r$) special case of the more general Muon-geometric objective at a correspondingly rescaled step size. Every point of divergence between them is therefore attributable, exactly and only, to the spectral anisotropy of $\boldsymbol{m}_{\boldsymbol{\theta}_t}$.

Corollary E.6 (Budget inflation under geometric mismatch)**.** *Fix a target fraction $\epsilon \in (0,1)$ of the optimal κ-sparse value $\max_{|\mathcal{S}|\le\kappa} f(\mathcal{S})$. By the standard bicriteria greedy guarantee for γ-weakly submodular functions (see e.g. Das & Kempe, 2018; Elenberg et al., 2018), a greedy budget of $\kappa' = \lceil (\kappa/\gamma)\ln(1/\epsilon)\rceil$ suffices to reach $(1-\epsilon)\max_{|\mathcal{S}|\le\kappa} f(\mathcal{S})$. Substituting the guarantees of Corollary E.4, matching μ-SMAS's $(1-\epsilon)$-value at budget κ requires vanilla* GREATS *to select a budget of order*

$$\kappa'_{\langle\varepsilon\rangle} \;=\; \Theta\big(\kappa\cdot \mathrm{cond}(\mathbf{\Sigma}^{\langle\varepsilon\rangle})\big),$$

i.e., GREATS *must select on the order of $\mathrm{cond}(\mathbf{\Sigma}^{\langle\varepsilon\rangle})$ times more examples per step than μ-*SMAS *to reach comparable guaranteed utility. Equivalently, at matched selection budget κ,* GREATS*'s guaranteed value fraction relative to μ-*SMAS*'s is $O(1/\mathrm{cond}(\mathbf{\Sigma}^{\langle\varepsilon\rangle}))$.*

Proof. Immediate from the cited bicriteria template with $\gamma = \gamma_\mu \ge 1 - O(\varepsilon)$ (so $\kappa'_\mu = \Theta(\kappa)$) versus $\gamma = \gamma_{\langle\varepsilon\rangle} \ge \mathrm{cond}(\mathbf{\Sigma}^{\langle\varepsilon\rangle})^{-1} - O(\varepsilon)$ (so $\kappa'_{\langle\varepsilon\rangle} = \Theta(\kappa\cdot\mathrm{cond}(\mathbf{\Sigma}^{\langle\varepsilon\rangle}))$ to leading order as $\mathrm{cond}(\mathbf{\Sigma}^{\langle\varepsilon\rangle})\to\infty$). □

Corollary E.6 gives the geometric-mismatch problem a directly interpretable resource-cost reading, connecting back to the memory-constrained motivation of Section 1: under anisotropic gradient spectra, using the wrong (Euclidean) selection geometry does not merely produce a lower-quality subset at fixed budget κ — it forfeits precisely the memory savings that motivated subset selection in the first place, since matching μ-SMAS's guarantee would require inflating the candidate/selection budget by the same anisotropy factor $\mathrm{cond}(\mathbf{\Sigma}^{\langle\varepsilon\rangle})$ that Muon's orthogonalization removes for free.

Corollary E.7 (Effective sample size gain)**.** *Applying the classical design-effect relation for κ equicorrelated observations with common pairwise correlation ρ, $\mathrm{ESS}(\kappa,\rho) = \kappa/\big(1+(\kappa-1)\rho\big)$, as a heuristic translation of* **??** *into an effective-sample-size statement gives*

$$\mathrm{ESS}^{\mu} \;=\; \frac{\kappa}{1+(\kappa-1)\rho^{\mu}} \;\ge\; \frac{\kappa}{1+(\kappa-1)\rho^{\langle\varepsilon\rangle}} \;=\; \mathrm{ESS}^{\langle\varepsilon\rangle},$$

*i.e., a size-κ subset selected under μ-*SMAS*'s geometry carries at least as much independent information, in the classical design-effect sense, as a size-κ subset selected under* GREATS*'s — with the gap growing as $\rho^{\langle\varepsilon\rangle}\to 1$ (extreme anisotropy). We flag this as a heuristic translation (the equicorrelated model is an approximation to the actual, adaptively-selected support) rather than a fully distribution-free guarantee*

E.4 Additional Pretraining Details

E.5 LoRA Finetuing

To show the effect of orthogonalization on minibatch coreset quality and model convergence, we evaluate **??** during the LoRA fine-tuning of Qwen2.5-0.5B on the Mol-Instructions dataset. By applying the selection function $f_{\langle\mu\rangle}(\cdot)$ (Equation (12)) strictly for subset selection while keeping the underlying optimizer unchanged, we decouple the geometric selection benefits from the optimization dynamics.

E.5.1 Setup

Model. We fine-tuned Qwen2.5-0.5B in using LoRA ($r = 16$, $\alpha = 32$, dropout 0.05). Adapters are applied to all six linear projections (`q`, `k`, `v`, `gate`, `up`, `down`) across all blocks, yielding 8.11M trainable parameters (1.62% of the total model).

Dataset We used the molecule-oriented split of MOL-INSTRUCTIONS, treating its $D = 4$ sub-tasks — description-guided molecule design, forward reaction prediction, reagent prediction, and retrosynthesis — as distinct source domains.

optimization. We train with AdamW ($\eta = 2\times10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$, weight decay 0.01, gradient clipping at 1.0), using a 3% linear warmup followed by linear decay to zero, for 1500 steps. We used Validation anchor points $\kappa = 16$ for subset selection that covers all the four tasks.

Selection. For gradient matching and scoring of datapoints for subset selection we used LoRA factors $\mathbf{A}$ and $\mathbf{B}$ of the last three blocks (21–23), which gives 36 parameter blocks. Per-example

Table 3: Final validation loss at step 1500 by selection ratio, $\kappa = 16$ throughout. Positive gap favours μ-SMAS.

	batch size n	GREATS	μ-SMAS	gap
$\rho = 0.125$	128	0.2983	0.2959	$+0.0024$
$\rho = 0.250$	64	0.2969	**0.2934**	$+0.0034$
$\rho = 0.372$	43	0.2970	0.2963	$+0.0007$
$\rho = 0.500$	32	0.2976	0.2957	$+0.0019$

gradients are recovered exactly from a single backward pass using the ghost construction of Wang et al. (2024) and normalised to unit ℓ_2 norm. At each step, a fresh validation batch of $|\mathcal{B}_t^{\text{val}}| = 8$ examples is drawn from the held-out split. The selection-only momentum buffers ($\beta = 0.95$, one per block) are defined in Equation (21).

Budget. We fix $\kappa = 16$ and vary the candidate batch size $|\mathcal{B}_t| \in \{128, 64, 43, 32\}$, giving selection ratios $\rho = \kappa/|\mathcal{B}_t| \in \{0.125, 0.250, 0.372, 0.500\}$.

E.5.2 Main result

Selection-only geometry and scoring. To isolate the effect of selection geometry from the optimizer, all models are updated using standard AdamW, confining the orthogonalized geometry entirely to a separate momentum buffer used exclusively for scoring. We score the LoRA factors across all six adapted projections in the final three blocks (21–23), yielding 36 parameter blocks. Candidate gradients are exacted via the ghost construction of Wang et al. (2024) and maintain independent momentum buffers per block. Gradients are mapped block-by-block using the Fréchet derivative of a 10-step Newton–Schulz orthogonalization surrogate, then renormalized to construct the Gram matrix and anchor-alignment vector for greedy selection. Detailed momentum recursions and AdamW update equations are deferred to Appendix **??**.

Main results. Following subset selection ($\kappa = 16$), the parameters are updated on the chosen examples. As detailed in Table 3, μ-SMAS consistently attains a lower validation loss than GREATS across all evaluated selection ratios. Specifically, μ-SMAS outperforms GREATS by a margin of $+0.0024$, $+0.0034$, $+0.0007$, and $+0.0019$ at selection ratios $\rho \in \{0.125, 0.250, 0.372, 0.500\}$. At its optimal ratio ($\rho = 0.250$), μ-SMAS reaches a validation loss of 0.2934, comprehensively beating Random (0.2990) and GradNorm (0.3186). The substantial underperformance of GradNorm demonstrates that ranking by raw gradient norm suffers from magnitude bias by selecting high-loss outliers rather than validation-relevant samples.

E.5.3 Ablation I: Geometry of the selected batch

To empirically validate that orthogonalized scoring inherently yields more diverse and less redundant coresets than standard Euclidean scoring, this ablation analyzes the internal geometric properties of the selected batches. Specifically, we evaluate the selected subsets based on two core metrics: Redundancy (the mean alignment of the selected gradients) and Effective Rank (the dimensional span of the Gram spectrum).

Comparing these set-level statistics across two separately trained runs introduces critical confounding variables, as the runs will eventually sit at different parameters and evaluate within mismatched geometries (polar for μ-SMAS, Euclidean for GREATS). We therefore employ a strictly paired design to isolate the selection mechanism. At every step of a single training run, we compute both Gram matrices from the same ghost pass, apply both selection rules to the identical candidate batch $\mathcal{B}_t$, and evaluate the two resulting sets on a shared Euclidean Gram matrix. Consequently, any differences in batch geometry are definitively attributable to the selection rule alone.

Redundancy of the selected batch. We define the redundancy of a selected batch $\mathcal{S}$ of size κ as the squared Frobenius norm of the mean normalized gradient: For $\mathbf{z}_i \in \mathcal{B}_t$, let $\hat{\mathbf{g}}_i := \boldsymbol{g}_{\boldsymbol{\theta}_t}(\mathbf{z}_i)/\|\boldsymbol{g}_{\boldsymbol{\theta}_t}(\mathbf{z}_i)\|_{\mathrm{F}}$. For a selected set $\mathcal{S} \subseteq \mathcal{B}_t$ with $|\mathcal{S}| = \kappa$, let $\mathbf{K}_{\mathcal{S}} \in \mathbb{R}^{\kappa\times\kappa}$ have entries $\langle \hat{\mathbf{g}}_i, \hat{\mathbf{g}}_j \rangle_F$ for $i, j \in \mathcal{S}$. We define

$$\mathrm{Red}(\mathcal{S}) \;:=\; \frac{1}{\kappa^2}\,\mathbf{1}^\top \mathbf{K}_{\mathcal{S}} \mathbf{1} \;=\; \Big\| \frac{1}{\kappa} \sum_{i\in\mathcal{S}} \hat{\mathbf{g}}_i \Big\|_F^2. \tag{43}$$

$\mathrm{Red}(\mathcal{S})$ is the mean of all κ^2 entries of $\mathbf{K}_{\mathcal{S}}$, and equivalently the squared norm of the mean normalised selected gradient. Because $\mathrm{diag}(\mathbf{K}_{\mathcal{S}}) = \mathbf{1}$, it separates into a constant and a data-dependent term:

$$\mathrm{Red}(\mathcal{S}) = \frac{1}{\kappa} + \frac{\kappa - 1}{\kappa}\,\bar{c}_{\mathcal{S}}, \qquad \bar{c}_{\mathcal{S}} := \frac{1}{\kappa(\kappa-1)} \sum_{i \neq j} \langle \hat{\mathbf{g}}_i,\, \hat{\mathbf{g}}_j \rangle_F, \tag{44}$$

Redundancy measures the mutual alignment of selected gradients. Because it is linear in the pairwise cosines, it distinguishes mutually orthogonal gradients from opposing ones, directly reflecting the redundancy penalty implicit in the GREATS objective.

At each step, we evaluate sets selected by both (μ-SMAS) and GREATS on a shared Euclidean Gram matrix to isolate the geometric effect of the selection rule.

As shown in Section 6.1, μ-SMAS consistently selects less mutually aligned batches than GREATS. Over steps 900–1500, the ratio of mean off-diagonal cosines ($\bar{c}_{\mathcal{S}}^{\mathrm{G}} / \bar{c}_{\mathcal{S}}^{\mu}$) ranges from 1.49 at $\rho = 0.125$ to 1.30 at $\rho = 0.500$. At the tightest budget ($\rho = 0.125$), the mean normalized selected gradient under μ-SMAS has a norm of 0.358 compared to 0.401 for GREATS—meaning the examples chosen by μ-SMAS reinforce one another 11% less.

We conclude that orthogonalized scoring intrinsically suppresses redundant sample selection. This effect is most pronounced under strict budgets and widens over training, effectively translating cross-domain coupling into more compatible training updates.

Effective rank of the selected batch. We define the effective rank of the selected batch via the participation ratio of the Gram spectrum:

$$\mathrm{ERank}(\mathcal{S}) \;:=\; \frac{\left(\sum_a \lambda_a\right)^2}{\sum_a \lambda_a^2} \;=\; \left(\sum_a p_a^2\right)^{-1}, \tag{45}$$

where $p_a := \lambda_a / \sum_b \lambda_b$ and λ_a are the eigenvalues of $\mathbf{K}_{\mathcal{S}}$.

Effective rank measures the number of distinct dimensions the selected batch effectively spans. By remaining insensitive to the sign of pairwise cosines, it complements the redundancy metric by quantifying dimensional diversity.

Like redundancy, ERank is evaluated using a strictly paired design on the shared Euclidean Gram matrix across all selection ratios.

Section 6.1 demonstrates that sets chosen by μ-SMAS maintain a higher effective rank at every evaluated ratio. Over steps 900–1500, μ-SMAS increases the effective rank over GREATS by margins of $+5.1\%$, $+2.1\%$, $+1.5\%$, and $+0.6\%$ for $\rho \in \{0.125, 0.250, 0.372, 0.500\}$, respectively.

Because both μ-SMAS and GREATS filter candidates based on equivalent relevance formulations, the superior effective rank achieved by μ-SMAS confirms that its polar geometry naturally preserves dimensional diversity without blindly sacrificing relevance (as random or maximum-diversity rules would).

E.5.4 Ablation II: Dependence on the selection ratio

To isolate how candidate pool size impacts the orthogonalized selection mechanism, this ablation fixes the selection rule to μ-SMAS and varies the selection ratio ρ over a 1500-step run. By keeping the subset budget fixed ($\kappa = 16$) and scaling the candidate batch size $|\mathcal{B}_t|$, we can directly compare set-level statistics across ratios. All reported curves utilize bias-corrected exponential moving averages, omitting the first 25 warmup steps.

Table 4: μ-SMAS diagnostics, averaged over steps 25–1500 ($\kappa = 16$).

	$\lvert\mathcal{B}_t\rvert$	ERank	Disagr	Disagr$/(1-\rho)$	DomEnt	Aniso
$\rho = 0.125$	128	13.90	0.411	0.469	0.846	8.48
$\rho = 0.250$	64	14.29	0.310	0.414	0.860	7.56
$\rho = 0.372$	43	14.36	0.237	0.378	0.871	6.92
$\rho = 0.500$	32	14.45	0.184	0.368	0.871	6.62

(a) Effective rank. We evaluate the effective rank $\mathrm{ERank}(\mathcal{S})$ via the participation ratio of the Gram spectrum equation 45, computed here exclusively on the polar Gram matrix to reflect the geometry in

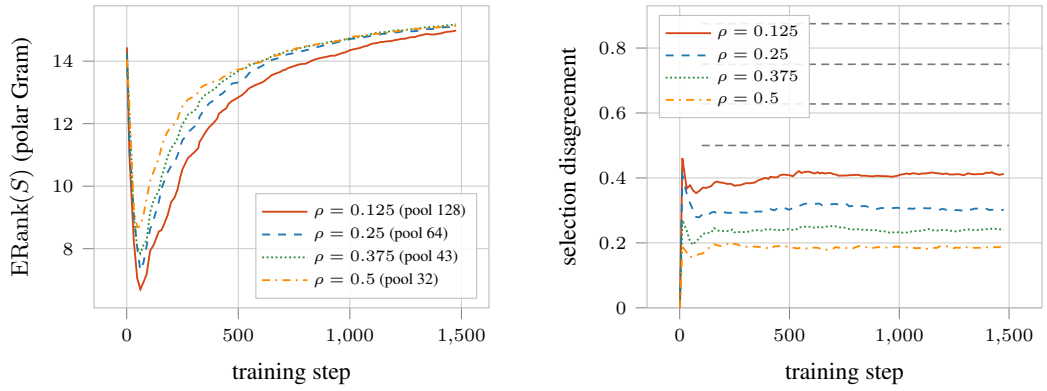


Figure 8: **Left:** selected-batch effective rank. **Right:** disagreement between the GREATS and μ-SMAS selections, with the chance level $1 - \kappa/|\mathcal{B}_t|$ dashed. Bias-corrected exponential moving averages ($\alpha = 0.99$).

which μ-SMAS selects. This metric measures the dimensional span of the selected subset. Tracking this across four ratios ($\rho \in \{0.125, 0.250, 0.372, 0.500\}$) while holding $\kappa = 16$, we observe that ERank rises monotonically with ρ, from 13.90 to 14.45. However, this increase reflects weaker selection pressure rather than improved subset quality. A wider relevance cap (higher ρ) admits more candidates, progressively approaching the uninformative rank of a random subset as $\rho \to 1$.

(b) Selection disagreement. To quantify the direct divergence between subsets chosen by standard versus orthogonalized scoring, we define disagreement as the normalized symmetric difference between sets selected by Euclidean ($\mathcal{S}_{\mathrm{G}}$) and polar ($\mathcal{S}_{\mu}$) scores:

$$\mathrm{Disagr} := \frac{|\mathcal{S}_{\mathrm{G}} \triangle \mathcal{S}_{\mu}|}{2\kappa} \in [0, 1]. \tag{46}$$

We measure this against the expected chance level $1 - \kappa/|\mathcal{B}_t|$ for two independent uniform κ-subsets. Normalized by chance, disagreement falls from 0.469 at $\rho = 0.125$ to 0.368 at $\rho = 0.500$. We conclude that the two selection rules genuinely converge by approximately 22% as the selection ratio grows, tracking closely with the drop in the selection buffer's anisotropy.

(c) Domain entropy. We measure domain entropy via the order-1 Hill number of the domain histogram over $D = 4$ source domains:

$$\mathrm{DomEnt}(\mathcal{S}) := -\frac{1}{\log D} \sum_{d=1}^{D} q_d \log q_d \in [0, 1]. \tag{47}$$

This metric serves as an empirical proxy for cross-domain coverage by measuring the effective number of domains represented. We observe that domain entropy rises from 0.846 to 0.871 as ρ increases. Because relevance naturally correlates with domain origin, a looser budget (higher ρ) admits a domain mixture closer to that of the entire candidate batch. Since estimator bias remains constant due to the fixed budget κ, it cannot explain this genuine trend.

(d) Score correlation. We define score correlation as the Pearson correlation between the Euclidean ($\mathbf{a}^{\mathrm{G}}$) and polar ($\mathbf{a}^{\mu}$) relevance scores over the candidate batch, using the centering matrix $\mathbf{C}$:

$$\mathrm{ScoreCorr} := \frac{\langle \mathbf{C}\mathbf{a}^{\mathrm{G}}, \mathbf{C}\mathbf{a}^{\mu} \rangle}{\|\mathbf{C}\mathbf{a}^{\mathrm{G}}\|_2 \, \|\mathbf{C}\mathbf{a}^{\mu}\|_2} \in [-1, 1]. \tag{48}$$

This metric measures how closely the polar ranking mimics standard Euclidean ranking, where $1 - \mathrm{ScoreCorr}^2$ isolates the unexplained variance of the polar score. We find that correlation increases with ρ, rising from 0.784 to 0.831, meaning unexplained variance drops from 39.8% to 31.3%. Smaller candidate batches severely degrade the momentum estimate's signal-to-noise ratio, which scales with $\sqrt{|\mathcal{B}_t|}$. As buffer estimation deteriorates at higher ρ, μ-SMAS loses its orthogonalizing distinctiveness and increasingly behaves like GREATS.

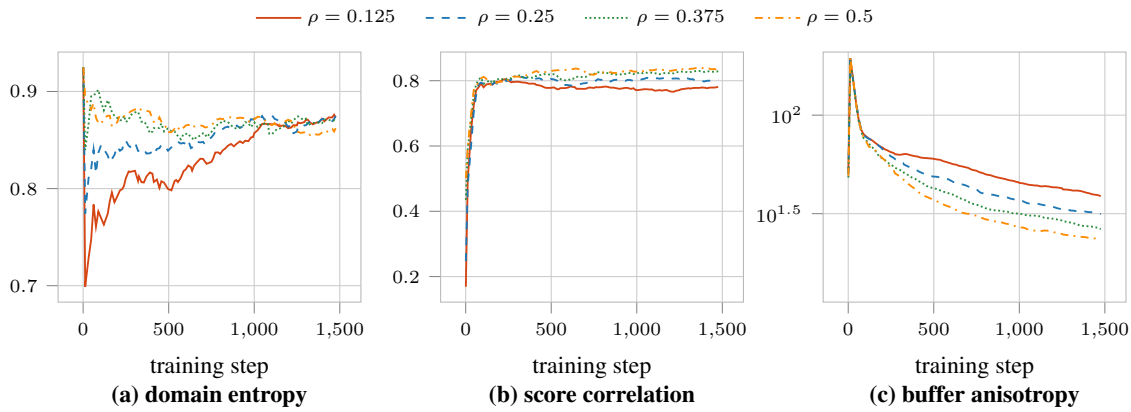


Figure 9: μ-SMAS across selection ratios. **Left:** normalised domain entropy ($\alpha = 0.997$). **Centre:** score correlation ($\alpha = 0.997$). **Right:** selection-buffer anisotropy, log scale ($\alpha = 0.995$).

(e) Selection-buffer anisotropy. We define the selection-buffer anisotropy as the robust condition number computed via the thin SVD of each block's selection buffer, averaged across 36 blocks:

$$\text{Aniso} := \frac{1}{36} \sum_{k=1}^{36} \frac{\sigma_1^{(k)}}{\sigma_6^{(k)}}. \tag{49}$$

This metric represents the ratio of weights the whitening assigns to tail versus head singular vectors, dictating the exact strength of the polar derivative's correction. Anisotropy decreases by 22% (from 8.48 to 6.62) as ρ increases. A flatter momentum spectrum forces the polar derivative toward a fixed map, driving the polar and Euclidean selection rules to converge. This convergence is confirmed by a strong negative correlation (-0.868) between anisotropy and score correlation across varying ratios.

(f) Bias toward the dominant momentum direction. Finally, we measure the correlation between a candidate's relevance score and its alignment with the buffer's leading singular direction $\mathbf{u}_1\mathbf{v}_1^\top$. This directly validates the core theoretical mechanism: Euclidean scoring favors the dominant direction while polar scoring mathematically debiases it by σ_1^{-1}. Evaluated over steps 900–1500, Table 5 shows that GREATS strongly correlates with the dominant direction (up to 0.1356), whereas μ-SMAS suppresses this bias entirely at strict budgets (0.0022 at $\rho = 0.125$). Because candidates aligned with the dominant direction are consistently ranked lower by μ-SMAS, they remain in subsequent candidate pools. The buffer then accumulates these unselected components, successfully debiasing selection and explaining the structurally higher anisotropy observed in μ-SMAS runs compared to GREATS.

Table 5: Correlation between each score and a candidate's alignment with the dominant momentum direction $\mathbf{u}_1\mathbf{v}_1^\top$, steps 900–1500.

	GREATS	μ-SMAS
$\rho = 0.125$	0.0633	**0.0022**
$\rho = 0.250$	0.1315	**0.0548**
$\rho = 0.372$	0.1330	**0.0642**
$\rho = 0.500$	0.1356	**0.0704**

E.5.5 SUMMARY

This experiment evaluates μ-SMAS end to end at LLM scale, with the model trained by AdamW and the Muon geometry used only for selection. The results support the proposed mechanism and contradict one conclusion from the controlled studies. They also show that the resulting gain in validation loss is small and may not persist with longer training.

E.6 PRETRAINING: SETUP

We pretrain three models from scratch and compare selection rules on each. The three are chosen so that they differ in ways that matter for the method: a dense decoder, a second dense decoder from a different family, and a sparse mixture-of-experts model where most of the weights are not visible to the selector.

1. **Qwen2.5-0.5B**, a dense decoder with grouped-query attention and tied embeddings.
2. **GPT-2 medium (355M)**, a second dense decoder with a different tokenizer and vocabulary, which checks that nothing depends on the Qwen family.
3. **Mixtral-style MoE (1.14B total, 0.38B active)**, where the experts hold most of the parameters and the optimizer has to handle batched three-dimensional expert tensors.

Each experiment is described in its own subsection below. The settings shared by all three are collected here.

Selection protocol. Every step loads a candidate pool of $|\mathcal{B}_t| = 64$ sequences and trains on the $\kappa = 16$ that the rule keeps, a selection ratio of 0.25. Since κ is fixed, every rule takes the same sized gradient step, including `full`, which simply loads 16 rows. Any difference between rules is therefore due to which rows were picked and not to how many. Anchors for the gradient-matching rules are redrawn at every step from the held-out split. The Taylor weight η inside the selection objective is refreshed from the live learning rate each step, so warm-up and decay reach the selector as well as the update.

optimizers. We run every arm under both AdamW (Loshchilov & Hutter) and Muon (Jordan et al., 2024). Under Muon, the embeddings and the LM head stay on the Muon branch instead of being moved to a separate AdamW optimizer, because the selection step reads their momentum buffer; AdamW then handles only the 1-D parameters at a fixed 3×10^{-4}. Learning rates are set per arm and per optimizer, as described in Section E.7.

Baselines. Each arm compares six rules under both optimizers, so twelve runs per model and thirty-six in total: RANDOM (uniform subset), GRADNORM (Katharopoulos & Fleuret, 2018) (top-κ by per-example gradient norm), FACLOC (greedy facility location on gradient similarity), TRACIN (Pruthi et al., 2020a) (influence on the anchors), GREATS (Wang et al., 2024) (validation-guided matching in the Euclidean geometry) and μ-SMAS.

Training settings. All runs use a cosine schedule with 100 warm-up steps, weight decay 0.1, gradient clipping at 1.0, bf16 for the forward pass with fp32 master weights and optimizer state, activation checkpointing, and seed 42. Within an arm every rule shares the initialisation, the data order, the anchor stream and the schedule, so the rule is the only thing that changes. Each configuration is run once, so the numbers are single-seed point estimates; we report them without error bars, and differences of a few thousandths of a nat should be read as ties. Validation loss is the token-weighted mean cross-entropy on the held-out split, measured every 100 steps.

All three arms use a sequence length of 512 tokens. This is worth stating because the GPT-2 configuration file asks for 1024: the dataset is built without a `max_length` argument and falls back to its 512-token default, so the GPT-2 runs are 512-token runs. We report the effective value.

E.7 LEARNING-RATE CALIBRATION

Comparing two optimizers is only fair at each one's own best setting, so Muon and AdamW were swept separately before any selection run. Both sweeps use the `full` rule, so the rate is tuned on the optimizer alone and no selection effect can leak into it.

For the two dense models the sweeps give $\eta_{\mathrm{Muon}} = 10^{-3}$ in both cases, with $\eta_{\mathrm{AdamW}} = 5 \times 10^{-4}$ for Qwen and 2×10^{-4} for GPT-2. The best rates differ between the two optimizers by a factor of two on Qwen and five on GPT-2, which is why a single shared rate would put one of them at a disadvantage.

The MoE rates, $\eta_{\mathrm{Muon}} = 4 \times 10^{-4}$ and $\eta_{\mathrm{AdamW}} = 2 \times 10^{-4}$, are taken from a separate sweep run on that model and are not repeated here. Every selection rule within an arm then runs at its optimizer's calibrated rate.

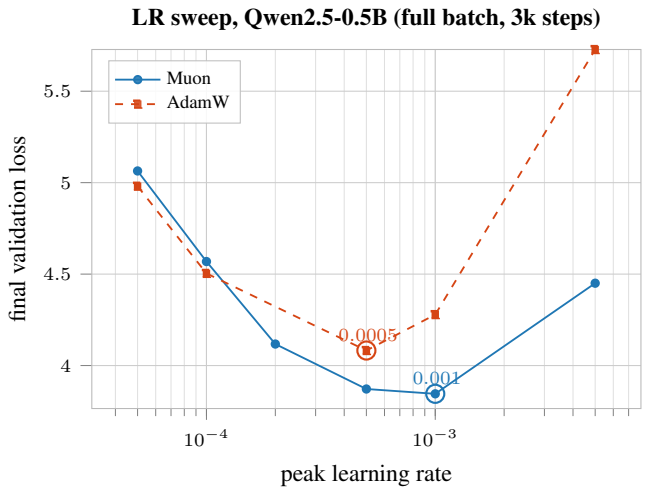


Figure 10: Learning-rate sweep for Qwen2.5-0.5B, Muon against AdamW, trained on the full batch for one epoch at batch 32×512. Circled points mark each optimizer's best setting: Muon at $\eta = 10^{-3}$ (final validation loss 3.8455), AdamW at $\eta = 5 \times 10^{-4}$ (4.0822).

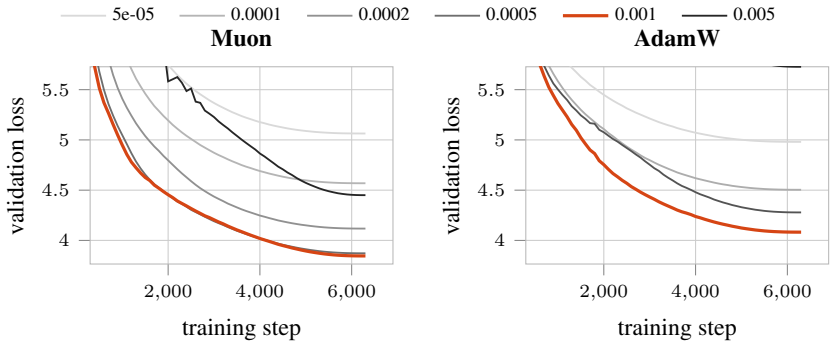


Figure 11: Qwen sweep as validation curves. The chosen rate is drawn in colour, the rest in grey. The loss has a single minimum in $\log \eta$ for both optimizers, so neither optimum sits at the edge of the grid.

E.8 QWEN2.5-0.5B

Model. The exact published Qwen2.5-0.5B configuration, with weights initialised at random.

Table 6: Qwen2.5-0.5B architecture.

blocks	24
hidden size	896
attention heads	14
key/value heads (GQA)	2
FFN width	4864
vocabulary	151,936
tied input/output embeddings	yes
parameters	0.49B

Data. OpenWebText-100K[3], a single-domain corpus of web text. It is tokenised once with the Qwen tokenizer, concatenated into one stream and cut into fixed 512-token chunks. The last 5% of chunks are held out; validation loss and the selection anchors both come from that split.

[3] https://huggingface.co/datasets/Elriggs/openwebtext-100k

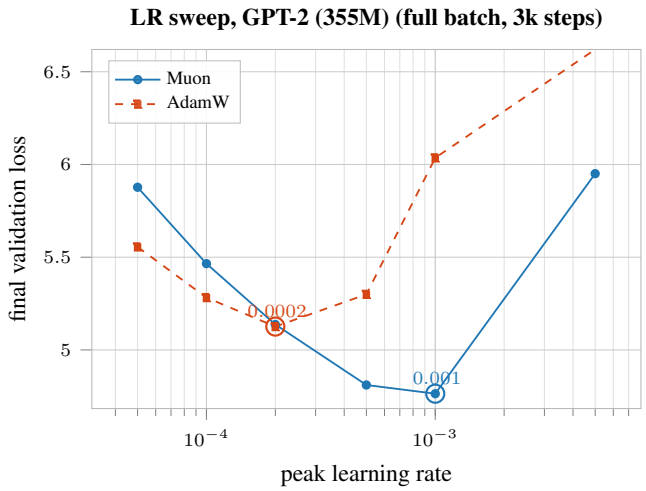


Figure 12: Learning-rate sweep for GPT-2 medium, Muon against AdamW, trained on the full batch for 3000 steps at batch 16×512. Muon peaks at $\eta = 10^{-3}$ (final validation loss 4.764), AdamW at $\eta = 2 \times 10^{-4}$ (5.1259).

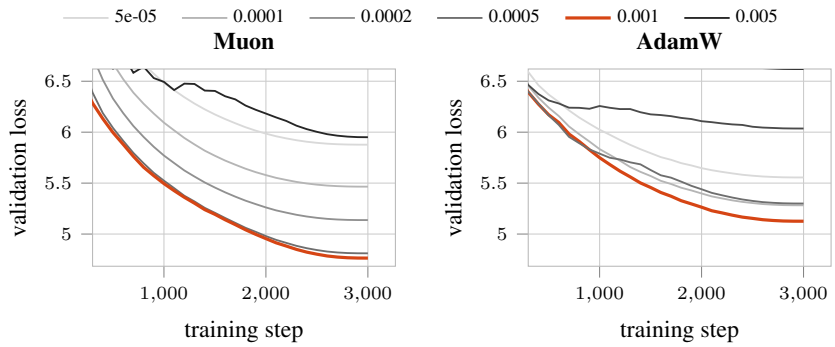


Figure 13: GPT-2 sweep as validation curves, with the chosen rate in colour.

Training. One epoch, 3159 steps, at 16 trained rows of 512 tokens per step. Muon runs at $\eta = 10^{-3}$ and AdamW at 5×10^{-4}, both with the shared schedule described above.

Selection. Pool 64, $\kappa = 16$, 16 anchors redrawn each step from the held-out split. Scoring uses the last linear layer inside the final block, which for this model is an MLP projection. Gradients are taken at that layer for every candidate and every anchor, and the selection score is built from their inner products.

Table 7: Final validation loss, Qwen2.5-0.5B. Best per column in bold.

selection rule	Muon (1×10^{-3})	AdamW (5×10^{-4})
Random	4.6757	5.1010
GREATS	4.6535	5.0540
FacLoc	4.6634	5.0613
GradNorm	4.7590	5.1019
TracIn	4.6936	**5.0456**
μ-SMAS	**4.6356**	5.0683

μ-SMAS gives the lowest loss under Muon. Under AdamW it sits in the middle of the field, behind TracIn and GREATS, which is the expected pattern: without Muon there is no mismatch between the update geometry and the Euclidean score, so the correction has nothing to fix.

E.9 GPT-2 MEDIUM (355M)

Model. GPT-2 medium built to the nanoGPT recipe, initialised at random.

Table 8: GPT-2 medium architecture.

blocks	24
hidden size	1024
attention heads	16
FFN width	4096
vocabulary	50,304
dropout	0
parameters	0.355B

Data. The same OpenWebText-100K corpus as the Qwen arm, but tokenised with the GPT-2 tokenizer and cached separately, so the two arms see the same text under different vocabularies. Chunks are 512 tokens, and the last 5% are held out for validation and anchors.

Training. 3000 steps at 16 trained rows of 512 tokens per step. Muon runs at $\eta = 10^{-3}$ and AdamW at 2×10^{-4}.

Selection. Pool 64, $\kappa = 16$, 16 anchors per step. As with Qwen, scoring uses the last linear layer of the final block.

Table 9: Final validation loss, GPT-2 medium. Best per column in bold.

selection rule	Muon (1×10^{-3})	AdamW (2×10^{-4})
Random	4.7559	5.1268
GREATS	4.7966	5.1424
FacLoc	4.8408	5.1665
GradNorm	4.9890	5.3248
TracIn	4.8088	5.1401
μ-SMAS	**4.7521**	**5.0865**

μ-SMAS is best under both optimizers here. Its margin over Random under Muon is small (0.0038) and should be treated as a tie; the clearer result is that it is the only rule to beat Random under both, while GREATS, FacLoc and GradNorm all fall behind it.

E.10 MIXTRAL MOE (1.14B TOTAL, 0.38B ACTIVE)

Model. A Mixtral-style sparse model with eight experts per block and top-2 routing, initialised at random.

Table 10: Mixtral MoE architecture.

blocks	10
hidden size	1024
attention heads	16
key/value heads (GQA)	4
FFN width	4096
experts per block	8
experts used per token	2
router load-balancing coefficient	0.01
tokenizer / vocabulary	GPT-2 / 50,257
parameters (total / active)	1.14B / 0.38B

Data. A single-domain corpus is no use here: with every token drawn from the same distribution the router has nothing to specialise on and the eight experts drift towards copies of each other. We therefore build a six-domain corpus of 110M tokens (215,092 chunks of 512 tokens), streamed and tokenised from six real sources.

Table 11: Six-domain corpus. Shares are token budgets, deliberately imbalanced.

domain	source	share
web	`HuggingFaceFW/fineweb-edu` (`sample-10BT`)(Penedo et al., 2024)	30%
code	`codeparrot/codeparrot-clean-valid` (CodeParrot)	25%
wiki	`wikimedia/wikipedia` (`20231101.en`) (Wikimedia Foundation, 2023)	15%
math	`open-web-math/open-web-math` (Paster et al., 2023)	12%
arXiv	`ccdv/arxiv-summarization` (Cohan et al., 2018)	10%
PubMed	`ccdv/pubmed-summarization` (Cohan et al., 2018)	8%

Two of the six are not prose. Web, wiki, arXiv and PubMed differ in register but share broad token statistics, whereas source code and LaTeX do not, and that contrast is what makes routing visible. Chunks never cross a domain boundary and each carries its domain label.

The imbalance would distort a held-out split taken from a shuffled pool, so we carve two balanced hold-outs instead, each with the same number of chunks from every domain: an evaluation set of $6 \times 200 = 1200$ chunks and, disjoint from it, an anchor set of $6 \times 100 = 600$ chunks. The remaining 213,292 chunks form the training split and keep the natural imbalanced mixture. The reported loss is therefore the unweighted mean over the six domains, so a model cannot hide a weak domain behind web's 30% share, and the anchor gradient means "do well on all six domains equally".

Training. 3000 steps at 16 trained rows of 512 tokens per step, Muon at $\eta = 4 \times 10^{-4}$ and AdamW at 2×10^{-4}.

Muon needs one change for this model. Mixtral stores each block's experts as two batched three-dimensional tensors rather than `Linear` modules (`gate_up_proj` of shape $(E, 2I, H)$ and `down_proj` of shape (E, H, I)), holding 1.01B of the 1.14B parameters, and the standard implementation rejects anything that is not two-dimensional. We orthogonalise each expert's own (out, in) matrix separately and scale its learning rate by that shape, so an expert gets the update a dense layer of the same size would. Flattening the slab to $(E, \text{out} \cdot \text{in})$ instead would make the E experts orthogonal to each other as vectors, which is a different object; we use the per-expert form throughout.

The reported loss is the language-modelling term only. With router logits enabled the optimised loss is $\mathcal{L}_{\text{LM}} + \lambda \mathcal{L}_{\text{aux}}$, and the auxiliary term measures how evenly the batch routed rather than anything about the data, so including it would partly compare rules on routing luck. It is logged separately.

Selection. Pool 64, $\kappa = 16$, and 18 anchors per step so that the six domains divide evenly. Anchors are drawn by a stratified sampler that puts exactly three chunks of each domain in *every* anchor batch, not just in the anchor pool.

Scoring resolves to the attention output projection of the final block. This is the right layer to land on: attention is dense, so every token passes through it and a per-example gradient there is a clean per-example quantity. Had the experts been ordinary linear layers, an expert would only see the roughly two-eighths of tokens routed to it and its gradient would track routing frequency rather than usefulness. The cost is that selection never sees the experts, which is about 90% of the model.

Table 12: Final validation loss, Mixtral MoE, averaged over the six domains. Best per column in bold.

selection rule	Muon (4×10^{-4})	AdamW (2×10^{-4})
Random	3.9798	4.2375
GREATS	3.9106	4.1937
FacLoc	3.9148	4.1800
GradNorm	4.0768	4.1966
TracIn	3.9255	4.2570
μ-SMAS	**3.9105**	**4.1367**

μ-SMAS is best under both optimizers, though its lead over GREATS under Muon (0.0001) is a tie. The per-domain table shows where the AdamW gain comes from: μ-SMAS is best on arXiv and PubMed, the two rarest domains, under both optimizers, and does so without giving up much on web.

GradNorm is the opposite case, winning on code while losing badly on web and wiki, which gives it the widest spread of any rule.

E.11 Summary

μ-SMAS gives the lowest final validation loss in five of the six model-optimizer cells. Two of those five are ties (GPT-2 under Muon, ahead of Random by 0.0038; MoE under Muon, ahead of GREATS by 0.0001). The exception is Qwen under AdamW, where it places fourth.

That exception fits the explanation rather than working against it. The correction μ-SMAS applies exists to close a gap between the Euclidean score and Muon's update geometry. Under AdamW that gap is not there, so the correction has nothing to do, and the rule should be no better than a Euclidean one. What matters is that the gain appears under Muon on all three architectures, including the MoE, where the selector can only see about a tenth of the parameters.

The gap between optimizers is larger than the gap between selection rules in every arm: switching from AdamW to Muon is worth between 0.23 and 0.43 nats, while the spread across all six selection rules within one optimizer is under 0.13. Selection is a second-order effect on top of the optimizer choice, not a substitute for it.

E.12 Ablation: Why the Selected Subsets Are Better

The three experiments above compare selection rules by their end result, the validation loss. This ablation asks a different question: what is it about the subsets that μ-SMAS picks that makes them better? We measure two properties of the selection problem itself, at fixed points along a real training run, without looking at the loss at all.

The two quantities are the *submodularity ratio* γ, which says how reliably a greedy rule can solve its own selection objective, and the *condition number* of the candidate Gram matrix, which explains why γ differs between the two rules. We also report the certified lower bound on γ, which is exact and needs no search.

E.12.1 Submodularity ratio

The two rules do not share an objective. Both rules score a candidate pool with a vector $\boldsymbol{\mu}$, whose entry μ_i measures how well candidate i aligns with the validation anchors, and a Gram matrix $\mathbf{K}$, whose entry K_{ij} measures how much candidates i and j overlap. They differ in the inner product used to build $\mathbf{K}$ — ordinary for GREATS, polar-mapped for μ-SMAS — but also, and more importantly here, in the set function they maximise.

GREATS maximises the *unweighted* form

$$f_{\mathrm{bin}}(\mathcal{S}) \;=\; \boldsymbol{\mu}^{\top}\mathbf{1}_{\mathcal{S}} \;-\; \tfrac{1}{2}\,\mathbf{1}_{\mathcal{S}}^{\top}\mathbf{K}\mathbf{1}_{\mathcal{S}}, \tag{50}$$

which its greedy rule optimises by repeatedly taking the largest marginal gain, with no weight vector anywhere. μ-SMAS maximises the value of a non-negative quadratic program restricted to $\mathcal{S}$,

$$f(\mathcal{S}) \;=\; \max_{\substack{\mathbf{w}\,\geq\,0\\ \mathrm{supp}(\mathbf{w})\subseteq\mathcal{S}}} \;\boldsymbol{\mu}^{\top}\mathbf{w} \;-\; \tfrac{1}{2}\,\mathbf{w}^{\top}\mathbf{K}\mathbf{w}, \tag{51}$$

growing the support one element at a time and re-solving $\mathbf{w}$ on it at each round.

Definition. For a chosen set $\mathcal{S}$ and a group $\mathcal{A}$ disjoint from it, compare adding the elements of $\mathcal{A}$ one at a time against adding them together. The submodularity ratio is the worst such comparison,

$$\gamma \;=\; \min_{\substack{\mathcal{S},\ \mathcal{A}:\ \mathcal{A}\cap\mathcal{S}=\emptyset\\ |\mathcal{A}|\leq\kappa}} \frac{\sum_{u\in\mathcal{A}}\big[f(\mathcal{S}\cup\{u\})-f(\mathcal{S})\big]}{f(\mathcal{S}\cup\mathcal{A})-f(\mathcal{S})}. \tag{52}$$

A value of $\gamma < 1$ means some group is worth more together than the sum of its parts. A greedy rule adds one element at a time and only ever sees the individual gains, so it cannot detect such a group. The standard guarantee is that greedy returns at least a $1-e^{-\gamma}$ fraction of the best possible subset, so larger γ means a more reliable selection step.

Which f we measure, and why the same one for both. Definition equation 52 needs a monotone f, or the denominator can vanish or change sign and the ratio is not defined. The value function equation 51 is monotone, because enlarging $\mathcal{S}$ only enlarges the feasible set for $\mathbf{w}$. The unweighted form equation 50 is not: adding a redundant candidate raises the penalty while the weight stays

pinned at one, so f_{bin} can decrease. A submodularity ratio for GREATS' own objective is therefore not a well-defined quantity.

We consequently evaluate equation 52 for *both* arms on the same function, equation 51, supplying each arm with its own $\boldsymbol{\mu}$ and $\mathbf{K}$. Only the geometry then differs between the two measurements, which is what this ablation is about, and equation 51 is the object that Theorem 5.1 is stated for. In practice we use the closed form $f(\mathcal{S}) = \frac{1}{2}\,\boldsymbol{\mu}_{\mathcal{S}}^{\top}\mathbf{K}_{\mathcal{S}\mathcal{S}}^{-1}\boldsymbol{\mu}_{\mathcal{S}}$, obtained by dropping $\mathbf{w} \geq 0$, and re-check the worst pair found with the constrained solver; the two agree throughout.

What follows should therefore be read as the submodularity ratio of the reformulated objective under two geometries, not as a property of the GREATS algorithm itself. Note also that γ is unchanged if $\boldsymbol{\mu}$ or $\mathbf{K}$ is rescaled, so it does not depend on the learning rate or on the size of the gradients, only on the shape of $\mathbf{K}$ and the direction of $\boldsymbol{\mu}$.

Setup. We train each of the three models for 500 steps with Muon, using each model's own configuration from the experiments above, and stop every 20 steps to measure γ. This gives 26 measurement points per model.

At each measurement point the candidate pool holds 32 examples and the budget is $\kappa = 8$, a selection ratio of 0.25. This is the same ratio the pretraining runs use, at half the pool size; 32 is chosen because $\binom{32}{8} = 10{,}518{,}300$, which is small enough to enumerate exhaustively.

Both rules are measured at the same point. At a given step they see the same model weights, the same 32 candidates and the same validation anchors, and each produces its own $\boldsymbol{\mu}$ and $\mathbf{K}$. The only difference between them is the inner product used to build $\mathbf{K}$: GREATS uses the ordinary one, μ-SMAS the polar-mapped one. Training is driven by μ-SMAS's selections throughout, so both rules are measured along one trajectory rather than two.

The minimisation in equation 52 runs over a domain built the same way for both rules. The base sets $\mathcal{S}$ are the first 16 prefixes of each rule's *own* greedy path: GREATS contributes the sets its marginal-gain rule on equation 50 actually visits, μ-SMAS the sets its own rule on equation 51 visits. No random sets are used, so every $\mathcal{S}$ is a state that rule genuinely passes through. For each base set, every group $\mathcal{A}$ of every size from 2 to κ is enumerated in full, with no sampling. That is 58,024,892 pairs per rule per measurement point.

Results. Figure 14 shows γ over training for both rules on all three models. Averaged over the 25 measurement points after the first, μ-SMAS reaches $\gamma = 0.75$ against 0.71 on Qwen, 0.91 against 0.71 on GPT-2, and 0.81 against 0.66 on the MoE. Counting individual measurement points, μ-SMAS is higher at 65 of 75: $16/25$ on Qwen, $24/25$ on GPT-2 and $25/25$ on the MoE.

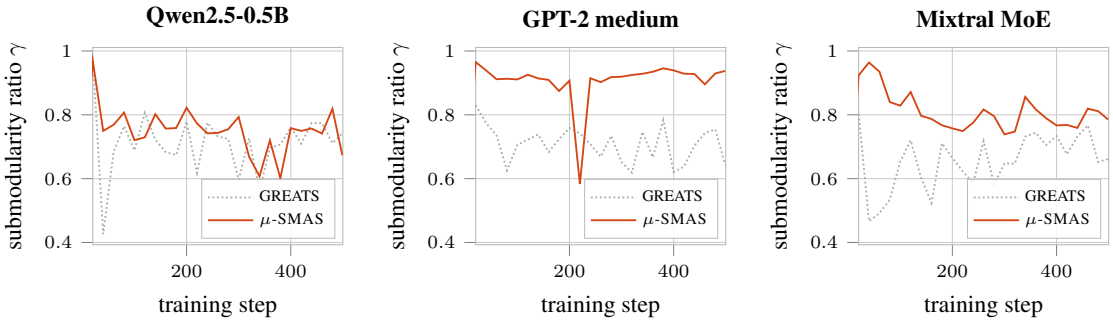


Figure 14: Submodularity ratio equation 52 over training, measured every 20 steps at pool 32 and $\kappa = 8$. Dotted grey: GREATS. Solid colour: μ-SMAS. Higher is better.

The result is clear on GPT-2 and the MoE and weak on Qwen, where the two rules are close for much of training and GREATS is ahead at nine of the 25 points. We therefore do not claim a uniform ordering on γ itself.

E.12.2 Condition number of the candidate Gram matrix

Definition. For the Gram matrix $\mathbf{K}$ of the 32 candidates, with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_{32} > 0$, we measure

$$\mathrm{cond}(\mathbf{K}) \;=\; \frac{\lambda_{\max}(\mathbf{K})}{\lambda_{\min}(\mathbf{K})}. \tag{53}$$

A large eigenvalue marks a direction that many candidate gradients share; a small one marks a direction almost none of them occupy. The condition number is therefore how lopsided the candidate gradients are as a set.

This quantity explains the previous one. The guarantee of Theorem 5.1 bounds γ from below by a ratio of restricted eigenvalues of $\mathbf{K}$, so a flatter spectrum forces γ up. The intuition is the same: a group can only be worth much more than the sum of its parts when the joint solve can reweight it along a direction that single additions cannot reach, and a flat spectrum leaves little room for that.

Setup. The same runs and the same measurement points as above. $\mathbf{K}$ is the full 32×32 Gram matrix of the candidate pool, taken before any selection, so this is a property of the scoring geometry and not of the chosen subset. It is computed from one eigendecomposition, with no search.

Results. Figure 15 shows the condition number over training. Averaged over the warm measurement points, μ-SMAS reduces it from 62.1 to 16.3 on Qwen, from 36.7 to 1.5 on GPT-2, and from 65.5 to 3.0 on the MoE. Unlike γ, the separation here holds at every measurement point on every model, and it is large: between four and twenty-four times.

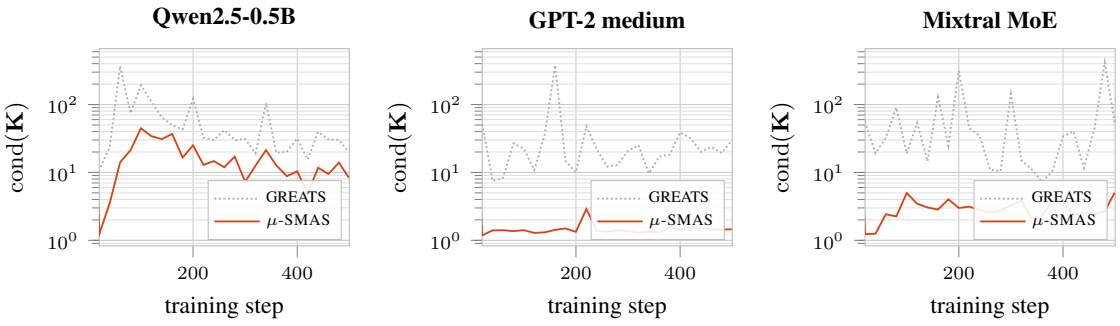


Figure 15: Condition number equation 53 of the 32×32 candidate Gram matrix, log scale, measured every 20 steps. Dotted grey: GREATS. Solid colour: μ-SMAS. Lower is better.

The polar map flattens the spectrum of the candidate gradients, and that is what raises γ.

E.12.3 Certified lower bound on γ

Definition and setup. The bound of Theorem 5.1 is $\gamma \geq c_{2\kappa}/C_1$, with $c_{2\kappa}$ and C_1 the restricted strong-concavity and smoothness constants of equation 51. For this objective those constants are restricted eigenvalues of $\mathbf{K}$, and eigenvalue interlacing gives the computable form

$$\gamma \;\geq\; \frac{\lambda_{\min}(\mathbf{K})}{\max_i K_{ii}}, \tag{54}$$

which needs one eigendecomposition and no search over subsets. It is measured on the same runs and the same points as the other two quantities.

This is worth reporting separately because the measured γ of equation 52 is a minimum over a searched domain and is therefore an upper estimate, whereas equation 54 is exact.

Results. Figure 16 shows the bound over training. Averaged over the warm points it rises from 0.071 to 0.400 on Qwen, from 0.095 to 0.837 on GPT-2, and from 0.058 to 0.629 on the MoE, a factor of between 5.7 and 10.9. As with the condition number, μ-SMAS is ahead at every measurement point on every model.

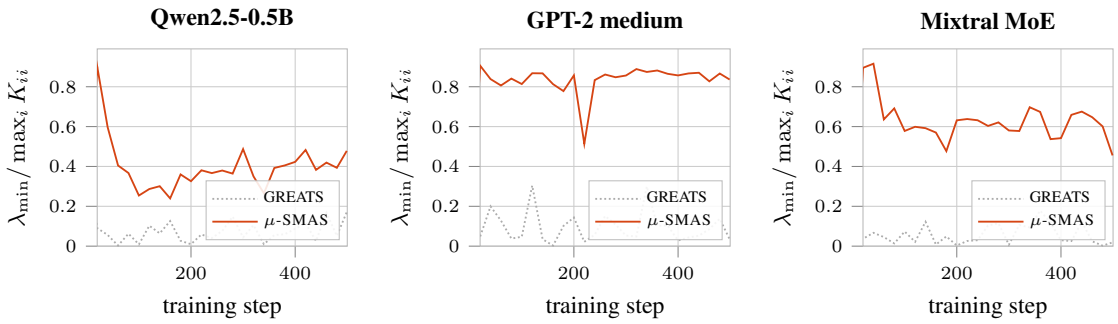


Figure 16: Certified lower bound equation 54 on the submodularity ratio, measured every 20 steps. Dotted grey: GREATS. Solid colour: μ-SMAS. Higher is better.

Table 13: Selection-problem diagnostics, averaged over the 25 measurement points after step 0. Arrows give the direction of improvement.

		Qwen2.5-0.5B	GPT-2 medium	Mixtral MoE
$\gamma \uparrow$	GREATS	0.713	0.709	0.655
	μ-SMAS	**0.752**	**0.908**	**0.811**
$\mathrm{cond}(\mathbf{K}) \downarrow$	GREATS	62.1	36.7	65.5
	μ-SMAS	**16.3**	**1.5**	**3.0**
$\lambda_{\min}/\max_i K_{ii} \uparrow$	GREATS	0.071	0.095	0.058
	μ-SMAS	**0.400**	**0.837**	**0.629**

E.12.4 What this ablation shows

Across all three models the polar geometry produces a far better conditioned selection problem: the candidate Gram matrix is between four and twenty-four times flatter, and the certified bound on γ is between six and eleven times larger. Both hold at every measurement point on every model.

The measured γ follows, but less cleanly: μ-SMAS is higher on average on all three models and at 65 of 75 individual points, with the effect strong on GPT-2 and the MoE and weak on Qwen.

One point is worth making because it rules out the obvious alternative explanation. μ-SMAS does not win by picking candidates that conflict less. The mean weight on negative off-diagonal entries of $\mathbf{K}$ is *higher* for μ-SMAS on all three models — 0.11 against 0.04 on Qwen, 0.57 against 0.14 on GPT-2, 0.39 against 0.28 on the MoE — so its candidate gradients disagree with each other more, not less. The gain comes from conditioning, not from avoiding conflict.

Two limits are worth stating. First, these numbers are properties of the selection problem, not of the trained model: a larger γ means the greedy step solves the objective it is given more reliably, and does not by itself say the resulting subset trains better. The validation losses in Sections E.8 to E.10 are the evidence for that. Second, all three quantities are computed on the reformulated objective equation 51 under two geometries. GREATS' own greedy maximises the unweighted form equation 50 and never forms equation 51, so these are not measurements of the GREATS algorithm; they measure how well behaved the selection problem becomes when the Euclidean Gram is replaced by the polar one.

F Code

We provide access to our code base at [4].

[4] https://anonymous.4open.science/r/Mu-Smas/