WSqD: A Horizon-Free Learning Rate Schedule for Large Model Training

Jianhao Ma*
University of Pennsylvania

Yuxin Chen*
University of Pennsylvania

July 14, 2026

Abstract

Standard learning rate schedules such as cosine annealing are tied to a fixed training horizon, limiting their ability to accommodate post hoc horizon extension. Warmup-stable-decay (`WSD`) partially addresses this issue by maintaining a long constant-rate phase before a short linear cooldown, allowing training to resume from a pre-decay checkpoint. However, its peak learning rate is still tuned based on the original training horizon and can become suboptimal when training is extended. Motivated by stochastic convex optimization, we propose `WSqD` (**W**armup with **Sq**uare-root base and linear **D**ecay), a learning rate schedule that replaces `WSD`'s constant stable phase with a shifted inverse-square-root base while retaining the final linear cooldown. In the stochastic convex setting, `WSqD` provably attains the minimax-optimal $O(1/\sqrt{T})$ last-iterate convergence rate. Importantly, its base learning rate schedule is horizon-independent, and the training horizon is needed only to determine when to begin the final cooldown. Empirically, on language-model pretraining using the `SlimPajama` corpus, `WSqD` matches or outperforms carefully tuned `WSD` and other baselines across multiple training horizons while reusing a single peak learning rate.

Keywords: learning rate schedule; continued training; horizon-independent training; stochastic convex optimization; large model training

1 Introduction

The learning rate schedule is a core design choice in large language model (LLM) training. A well-designed schedule can substantially improve training efficiency and stability, whereas a poorly chosen one may slow convergence or lead to degraded final performance. Modern LLM training pipelines are increasingly iterative and multi-stage. Rather than fixing the training horizon in advance, practitioners may extend training when evidence from scaling laws suggests that additional compute is likely to yield further gains (Kaplan et al., 2020; Hoffmann et al., 2022). They may also continue training on domain-specific corpora (Gururangan et al., 2020; Ke et al., 2023; Gupta et al., 2023) or organize training into multiple stages that serve different purposes (Ibrahim et al., 2024). Consequently, *mid-training* has emerged as a distinct phase of LLM development (Mo et al., 2025). Across these scenarios, it is desirable to resume training from an existing checkpoint and extend the training horizon with little or no re-tuning and without ad hoc modifications to the learning rate schedule. This raises an important question:

How should learning rate schedules be designed to remain effective under post hoc horizon extension?

1.1 Prior approaches

The aforementioned continued training requirement is not well served by cosine annealing, a standard fixed-horizon learning rate schedule used in many modern LLM training pipelines, including GPT-3 (Brown et al.,

*Department of Statistics and Data Science, University of Pennsylvania. Email: `{jianhaom,yuxinc}@wharton.upenn.edu`.

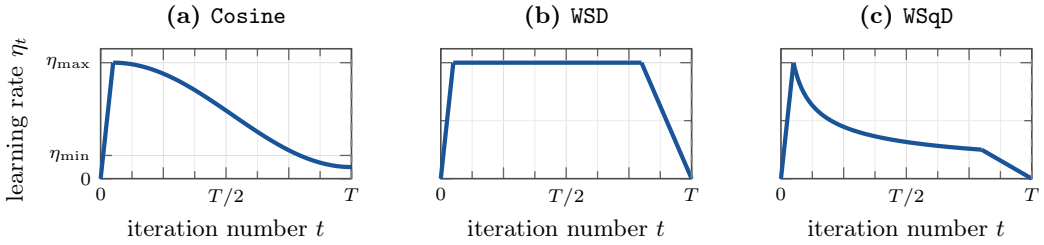


Figure 1: Illustration of cosine annealing, WSD, and the proposed WSqD learning rate schedules.

2020), LLaMA (Touvron et al., 2023), and Llama 3 (Grattafiori et al., 2024), to name a few. Cosine annealing explicitly ties the learning rate trajectory to a pre-specified endpoint by smoothly decaying the learning rate from the peak learning rate $\eta_{\max}$ to a small terminal learning rate $\eta_{\min}$ over the planned horizon:[1]

$$\eta_t^{\texttt{Cosine}} = \eta_{\min} + \frac{\eta_{\max} - \eta_{\min}}{2}\left(1 + \cos\left(\frac{\pi t}{T}\right)\right), \qquad t = 1, \ldots, T, \tag{1}$$

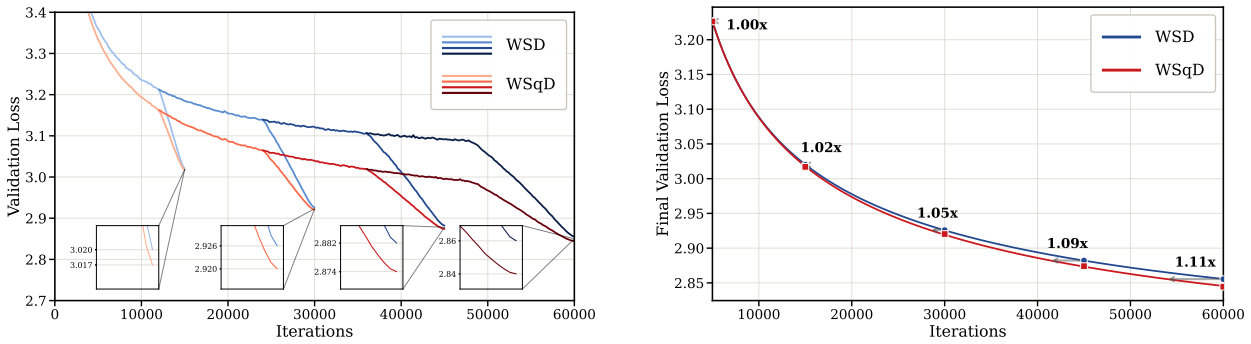
where T denotes the training horizon, i.e., the total number of training iterations. While this design works well when T is fixed and known in advance, it creates a fundamental challenge when training is extended beyond T. Once the learning rate has decayed to its terminal value, there is no natural way to keep training efficiently. One direct option is to continue with the near-zero terminal learning rate, but this results in slow subsequent progress. In practice, a more common practical choice is to restart or re-warm the schedule to a larger learning rate, but this can produce a transient loss spike known as the *stability gap* (Gupta et al., 2023; De Lange et al., 2023; Guo et al., 2024). Nor is the issue necessarily resolved by switching to a different schedule family; for example, Parmar et al. (2024) reported that, for a 15B model pretrained with cosine annealing, continuing with warmup-stable-decay (WSD) performs worse than remaining within the original cosine family. Thus, the limitation is not simply that cosine annealing requires a manual extension rule; rather, its explicit dependence on a fixed endpoint T often means that it is not naturally suited to post hoc horizon extension.

WSD learning rate schedules (Hu et al., 2024; Hägele et al., 2024) were introduced, in part, to alleviate the limitations of fully decayed fixed-horizon schedules. Instead of decaying throughout training, WSD maintains a constant learning rate during a long stable phase and applies a sharp linear cooldown only near the end of training, typically over the final 10%–20% of the training iterations. Formally, given a decay fraction $0 \le \alpha \le 1$, the WSD schedule is given by

$$\eta_t^{\texttt{WSD}} = \begin{cases} \eta_{\max}, & \text{if } t \le (1-\alpha)T, \\ \eta_{\max}\dfrac{T-t}{\alpha T}, & \text{if } (1-\alpha)T < t \le T, \end{cases} \tag{2}$$

where $\eta_{\max}$ denotes the peak learning rate. This design naturally supports training continuation: one can save a checkpoint from the stable phase as a continuation anchor, allowing training to resume and the final cooldown to be postponed until a deployable model is needed (Hägele et al., 2024; Wen et al., 2024). Consequently, extending training beyond the originally planned horizon simply amounts to resuming from this anchor, thereby avoiding learning-rate rewarming and hence addressing one of the most visible failure modes of cosine annealing. However, WSD is not truly horizon-independent, because its optimal peak learning rate still depends on the training horizon. Recent empirical evidence suggests that the optimal peak learning rate may follow a power law in the batch size and the total number of training tokens (Shen et al., 2024); equivalently, when the batch size is fixed, the optimal peak learning rate decreases as the training horizon

[1] In practice, all learning rate schedules discussed in this paper are preceded by a short linear warmup phase. We omit this warmup phase from all schedule formulas to focus on the post-warmup dynamics, and likewise exclude it from the theoretical analysis in Section 2.



(a) (b)

Figure 2: **(a)** Continued training on the `SlimPajama` dataset under `WSD` and `WSqD` across horizons of 15000, 30000, 45000, and 60000 iterations. **(b)** Fitted curves of the final validation losses after the linear cooldown phase. For both schedules, the base learning rate is selected by a 5000-iteration search; for `WSqD`, we use a shift parameter $T_0 = 5000$. The decay fraction α is set to be 0.2 for both `WSD` and `WSqD`.

increases. As a result, a peak learning rate tuned for a shorter run may become miscalibrated when training is extended, leading to suboptimal performance.

These limitations motivate the search for a more genuinely horizon-free learning rate schedule, namely one that does not rely on a pre-specified training horizon (also referred to as an *anytime* learning rate schedule in the optimization literature). Recent work has pursued this through schedule-free optimizers and related averaging or checkpoint-merging mechanisms (Defazio et al., 2024; Defazio, 2026; Apte et al., 2026; Meterez et al., 2026), or through empirically calibrated power-law schedules for transferring learning rates across token budgets and batch sizes (Shen et al., 2024). The former approach modifies the optimizer or augments the training loop with additional weight averaging or merging operations. The latter approach relies on empirically fitted scaling laws and falls short of a corresponding theoretical justification. All this motivates the development of a learning rate schedule that is horizon-free by design, requires no additional training operations, and is derived from first principles rather than empirical fitting alone. We discuss these approaches in greater detail in Section 4.

1.2 This paper

To address the above limiations, we derive a learning rate schedule from first principles in stochastic convex optimization that naturally accommodates post hoc horizon extension. A line of recent work has revealed close connections between LLM training and stochastic convex optimization (Defazio et al., 2024; Meterez et al., 2026; Pun et al., 2025). Motivated by this perspective, we revisit the canonical inverse-square-root schedule, whose learning rates are proportional to $1/\sqrt{t}$. Originating in stochastic approximation and convex optimization (Robbins and Monro, 1951; Hazan, 2016) and also adopted in the original Transformer paper (Vaswani et al., 2017), it does not explicitly depend on the training horizon, making it a natural choice for the base phase of a learning rate schedule designed for continued training.

Building on this observation, we propose the `WSqD` schedule, short for **W**armup with **Sq**uare-root base and linear **D**ecay, illustrated alongside cosine annealing and `WSD` in Figure 1. The key idea is to replace the flat stable phase of `WSD` with a shifted inverse-square-root schedule while retaining the same final linear cooldown. More concretely, given a planned horizon T and a decay fraction $\alpha \in (0, 1)$, the post-warmup schedule of `WSqD` is given by

$$\eta_t^{\texttt{WSqD}} = \begin{cases} \dfrac{c_0}{\sqrt{t + T_0}}, & \text{if } 1 \le t \le (1-\alpha)T, \\ \dfrac{c_0}{\sqrt{(1-\alpha)T + T_0}} \cdot \dfrac{T - t}{\alpha T}, & \text{if } (1-\alpha)T < t \le T, \end{cases} \tag{3}$$

where $c_0 > 0$ controls the overall learning rate scale and $T_0 \geq 0$ is a shift parameter that prevents excessively large learning rates near the beginning phase of training. Our theory shows that, unlike `WSD`, the rate-optimal scale c_0 is independent of the training horizon. Morevoer, any fixed shift T_0 is admissible once the training horizon is large enough ($T \geq 2T_0$ in our theorem). Consequently, the base phase can be reused across different training horizons without re-tuning to the eventual terminal point, making `WSqD` an "anytime" schedule in a practical sense.[2]

Although motivated by and derived from a stylized convex analysis, `WSqD` performs well empirically in continued pretraining of LLMs. Figures 2a and 2b preview our empirical results on `SlimPajama`: using a single base learning rate selected from a short 5000-iteration search and reused without further re-tuning, `WSqD` consistently matches or outperforms `WSD` after post hoc horizon extension, requiring up to 10% less compute to achieve the same validation loss.

Our main contributions are summarized as follows.

- *Theoretical guarantees.* We analyze `WSqD` under the standard convex stochastic optimization framework. We prove in Theorem 1 that `WSqD` attains an $O(1/\sqrt{T})$ last-iterate rate for stochastic mirror descent, as long as the final decay occupies a constant fraction of the training iterations. In the special case of Euclidean projected stochastic gradient descent, this rate matches the worst-case lower-bound scaling established by Agarwal et al. (2012); its constant depends only on the decay fraction and does not require knowing the final training horizon during the base phase.

- *Empirical validation.* We evaluate `WSqD` on LLM pretraining using the `SlimPajama` corpus under a continuation protocol in which the base learning rate is selected from a short pilot run and then reused across longer training horizons. `WSqD` consistently matches or improves over carefully tuned `WSD` baselines, with larger gains at longer continuation horizons. A key empirical finding is that `WSqD`'s optimal base learning rate remains remarkably stable across different training horizons, whereas the optimal peak learning rate of `WSD` varies with the total training budget. This robustness enables a base learning rate calibrated on a short run to be transferred directly to substantially longer continuations without re-tuning. We further compare `WSqD` against stronger continuation baselines, including fine-tuned two-stage `WSD` (Schaipp et al., 2025) and the power-law schedule (Shen et al., 2024); across these comparisons, `WSqD` matches or outperforms these alternatives.

Comparison with the power-law schedule. Among existing schedule-based approaches to post hoc horizon extension, the power schedule of Shen et al. (2024) is conceptually the closest to `WSqD`. We became aware of this work after the main design and theoretical development of `WSqD` had largely been completed, and we therefore view the two approaches as complementary. More specifically, both schedules adopt a warmup–base–decay structure and replace the flat stable phase of `WSD` with a decreasing base learning rate. The power schedule uses the base phase $\eta_t = \min\{\eta_{\max}, at^{-0.51}\}$, where the exponent -0.51 is fitted from large-scale experiments. In contrast, `WSqD` employs a shifted inverse-square-root base, $\eta_t = c_0/\sqrt{t+T_0}$, followed by a final linear cooldown. The two schedules also differ in how they regularize the large initial learning rates: the power schedule uses a cap $\eta_{\max}$ to truncate the initial learning rates, whereas `WSqD` uses the shift parameter T_0 to smooth the early part of the trajectory. More fundamentally, the power schedule is motivated by empirical scaling laws and is governed by a fitted power-law exponent close to $-1/2$, while `WSqD` is derived from stochastic convex optimization analysis, in which the inverse-square-root base arises as a natural horizon-independent step-size sequence and the final linear cooldown recovers the optimal last-iterate rate. In this sense, our results provide a complementary theoretical perspective on why near inverse-square-root base schedules can be effective for continued training.

[2]We use "anytime" in a practical sense: the base phase of `WSqD` is horizon-free and naturally supports resuming training from intermediate checkpoints. This differs from the stricter formal notion, which requires the entire learning rate sequence, including the final decay, to be independent of the training horizon. It turns out that this stricter requirement is incompatible with minimax-optimal last-iterate rates (Kornowski and Shamir, 2026); see Section 2 for details.

2 Theoretical analysis

In this section, we establish the convergence theory for stochastic mirror descent equipped with the `WSqD` learning rate schedule under the standard stochastic convex optimization framework.

2.1 Problem setup and assumptions

Consider the following standard convex optimization problem (Beck, 2017)

$$\underset{w\in\mathcal{W}}{\text{minimize}} \quad f(w), \tag{4}$$

where $\mathcal{W}\subseteq\mathbb{R}^d$ is a closed convex set, and $f:\mathcal{W}\to\mathbb{R}$ is a convex function. Throughout, we assume that the minimum of the problem (4) is attained at some $w^*\in\mathcal{W}$, and write $f^*=f(w^*)$.

Let ψ be a continuously differentiable, strictly convex function on a set containing $\mathcal{W}$, whose associated Bregman divergence is defined as

$$D_\psi(u,v) \coloneqq \psi(u)-\psi(v)-\langle\nabla\psi(v),u-v\rangle\,. \tag{5}$$

Starting from an initialization $w_1\in\mathcal{W}$, stochastic mirror descent follows the iterative update rule:

$$w_{t+1}=\underset{w\in\mathcal{W}}{\operatorname{argmin}}\left\{\eta_t\,\langle\widehat{g}_t,w\rangle+D_\psi(w,w_t)\right\},\qquad t\geq 1, \tag{6}$$

where $\eta_t\geq 0$ is the learning rate, and $\widehat{g}_t$ is a stochastic estimate of a subgradient of f at the point w_t. For notational convenience, we let $\mathcal{F}_1\coloneqq\sigma(w_1)$ (resp. $\mathcal{F}_t\coloneqq\sigma(w_1,\widehat{g}_1,\ldots,\widehat{g}_{t-1})$ for $t\geq 2$) denote the filtration representing the information available before sampling $\widehat{g}_1$ (resp. $\widehat{g}_t$).

Furthermore, let $\|\cdot\|$ denote a norm on $\mathbb{R}^d$, and let $\|\cdot\|_*$ denote its dual norm. Throughout the analysis, we impose the following standard assumptions.

Assumption 1 (Bounded subgradients)**.** The objective function f is convex, and G-Lipschitz on $\mathcal{W}$ with respect to $\|\cdot\|$; that is, for every $w\in\mathcal{W}$ and every subgradient $g\in\partial f(w)$, one has $\|g\|_*\leq G$. Moreover, the stochastic subgradient estimates are almost surely bounded, i.e., $\|\widehat{g}_t\|_*\leq G$ for all $t\geq 1$.

Assumption 2 (Bregman geometry and bounded diameter)**.** The function ψ is continuously differentiable and 1-strongly convex with respect to $\|\cdot\|$ on $\mathcal{W}$. The Bregman diameter of $\mathcal{W}$ is finite:

$$R_\psi^2\coloneqq\sup_{u,w\in\mathcal{W}}D_\psi(u,w)<\infty. \tag{7}$$

Assumption 3 (Unbiased stochastic oracle)**.** For each t, the stochastic subgradient estimate $\widehat{g}_t$ is conditionally unbiased: $\mathbb{E}[\widehat{g}_t\mid\mathcal{F}_t]\in\partial f(w_t)$.

Note that these assumptions generalize the classical model for nonsmooth stochastic convex optimization. In particular, we do not assume smoothness or strong convexity of f. In the Euclidean setting with $\psi(w)=\frac{1}{2}\|w\|_2^2$ and $\|\cdot\|=\|\cdot\|_2$, the stochastic mirror descent update in (6) reduces to Euclidean projected stochastic gradient descent (SGD), the Bregman divergence simplifies to $D_\psi(u,w)=\frac{1}{2}\|u-w\|_2^2$, and the Bregman diameter becomes $R_\psi^2=\frac{1}{2}\sup_{u,w\in\mathcal{W}}\|u-w\|_2^2$.

2.2 Convergence theory

With the problem setting and assumptions in place, we are positioned to present our last-iterate convergence guarantees for stochastic mirror descent under the proposed `WSqD` learning rate schedule. The proof of this theorem is postponed to Section A.

Theorem 1 (Last-iterate convergence for WSqD)**.** *Suppose Assumptions 1–3 hold. Consider stochastic mirror descent with the WSqD learning rate schedule* $\eta_t = \eta_t^{\texttt{WSqD}}$ *(cf. (3)), using any scale parameter* $c_0 > 0$*, shift paramter* $T_0 \geq 0$*, and a fixed decay fraction* $\alpha \in (0, 1/2)$*. Then we have*

$$\mathbb{E}\big[f(w_T) - f^*\big] \leq \left(\frac{R_\psi^2}{c_0} + c_0 G^2\right)\big(120 + 2\log(1/\alpha)\big)\frac{1}{\sqrt{T}}, \tag{8}$$

provided that $T \geq \max\{2T_0, 4/\alpha\}$.

Remark 1. Optimizing the upper bound in (8) w.r.t. c_0 yields the choice $c_0 = \Theta(R_\psi/G)$, under which

$$\mathbb{E}[f(w_T) - f^*] \leq O\left(\frac{R_\psi G \cdot \log(1/\alpha)}{\sqrt{T}}\right).$$

Remark 2. Note that our goal is to characterize the optimal convergence rate in T and the effect of the final linear cooldown, rather than to optimize the numerical preconstants in the convergence bound (8). Sharper preconstants may be obtainable through a more refined analysis.

We next discuss several key implications and provide further remarks concerning Theorem 1.

Rate-optimal convergence under WSqD. For any fixed decay fraction $\alpha \in (0, 1/2)$, the convergence rate in (8) exhibits the optimal $1/\sqrt{T}$ scaling on the training horizon T. In the special case of Euclidean settings, it recovers the standard minimax optimal rate for projected stochastic gradient descent, up to a constant factor depending only on α (Agarwal et al., 2012). Given that α is fixed, the logarithmic term $\log(1/\alpha)$ is absorbed into the constant and does not affect the scaling with T.

Horizon-independent learning rate parameters. Importantly, the optimal learning rate parameters are horizon-independent. Specifically, the rate-optimal scale parameter satisfies $c_0 = \Theta(R_\psi/G)$ (see Remark 1), while the shift paramter T_0 can be chosen as any fixed constant without knowing the final training horizon T. Consequently, the same parameter choice and the same convergence guarantee apply to every training horizon satisfying $T \geq \max\{2T_0, 4/\alpha\}$, eliminating the need to re-tune the base phase when training is extended.

Necessity of the final decay phase. The inverse-square-root base is horizon-free and, therefore, naturally suited for training continuation. By itself, however, it cannot achieve the desired last-iterate convergence rate. More specifically, in the Euclidean projected SGD setting with $\eta_t = c_0/\sqrt{t + T_0}$, the classical analysis of Shamir and Zhang (2013) established an $O(\log T/\sqrt{T})$ upper bound for the last iterate, while Harvey et al. (2019) showed that this logarithmic factor is unavoidable for this schedule.[3] The final linear cooldown phase is precisely what eliminates this logarithmic overhead. By committing to a final horizon only during the final decay phase, WSqD preserves the continuation flexibility of the inverse-square-root base while recovering the optimal $O(1/\sqrt{T})$ last-iterate rate (up to some horizon-independent constant). This message is consistent with the recent analysis in Schaipp et al. (2025), which showed that appending a linear cooldown to a constant learning-rate schedule (namely, WSD) similarly removes the logarithmic factor and recovers the optimal convergence rate in a standard nonsmooth stochastic optimization setting. More broadly, our theory provides further evidence for the empirical advantages of cooldown schedules in large model training.

Compatibility with anytime lower bounds. Throughout this paper, we use the term *horizon-free* (or *anytime*) in a practical sense when describing WSqD: the base phase of WSqD uses the horizon-independent schedule $c_0/\sqrt{t + T_0}$, so every pre-decay iterate can serve as a valid continuation point. In the strict formal sense, however, WSqD is not fully anytime, because the length of the final decay phase, αT, depends on the committed horizon. This distinction prevents our $O(1/\sqrt{T})$ convergence guarantee from contradicting the

[3]The results of Shamir and Zhang (2013) and Harvey et al. (2019) were stated for $\eta_t = c_0/\sqrt{t}$. The same conclusions extend directly to the shifted schedule $\eta_t = c_0/\sqrt{t + T_0}$ for any fixed T_0, since the shift only affects constant factors.

recent lower bound of Kornowski and Shamir (2026), which showed that any learning rate sequence completely independent of T must incur an additional polylogarithmic overhead in the last-iterate convergence rate, even in deterministic convex optimization. Crucially, `WSqD` introduces horizon dependence only during the final cooldown. This design aligns naturally with practical training workflows, where the additional training budget is typically known once a new continuation phase is launched, allowing the length of the next cooldown to be chosen accordingly. Moreover, extending training from T to $T' > T$ requires only discarding the original cooldown segment and resuming from the last pre-decay iterate at step $(1-\alpha)T$. Consequently, a fraction $1-\alpha$ of the original training compute is reused, while only a constant fraction is discarded.

Possible extensions of our results. Theorem 1 establishes an in-expectation last-iterate guarantee for stochastic mirror descent under general Bregman geometry. Two natural extensions are worth mentioning. First, the concentration framework of Liu and Zhou (2023) suggests that our analysis could be strengthened to yield a high-probability guarantee (so that the bound holds with probability at least $1-\delta$), at the cost of an additional logarithmic dependence on $1/\delta$. The second is to move beyond fixed mirror maps toward adaptive and momentum-based methods. Accordingly, the present theorem should be viewed as a first step toward understanding geometry-aware optimization methods, rather than as a convergence guarantee for practical optimizers such as AdamW or Muon. To better illustrate this connection without overstating the theory, Section B gives an informal mirror-descent geometric interpretation of idealized Adam- and Muon-style update directions, together with empirical diagnostics of the scale factors omitted by these idealized updates.

3 Experiments

In this section, we empirically evaluate `WSqD` on LLM pretraining. We begin by describing the experimental setup in Section 3.1. We then present continued training comparisons between `WSqD` and `WSD` in Section 3.2, analyze the dependence of the optimal learning rate on the training horizon in Section 3.3, study the sensitivity of `WSqD` to the shift parameter T_0 in Section 3.4, and finally compare `WSqD` against stronger baselines in Section 3.5. Additional experiments are provided in Section C.

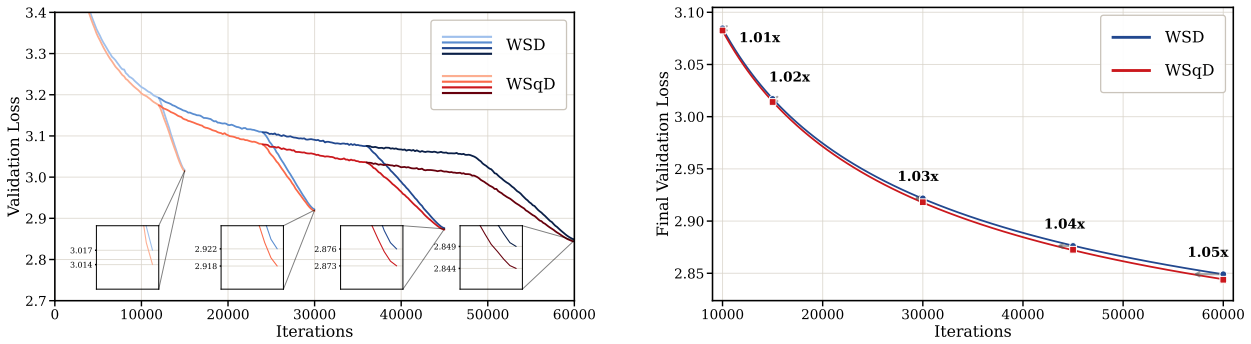
3.1 Setup

Model and data. We pretrain a 213M-parameter LLaMA-style transformer (Touvron et al., 2023) on the `SlimPajama` corpus (Soboleva et al., 2023). The model comprises 24 layers, 12 attention heads, and an embedding dimension of 768, and employs RoPE positional embeddings and RMSNorm. We tokenize the data using the GPT-2 BPE vocabulary and train with a context length of 512.

Optimizer. All main validation-loss experiments use AdamW (Loshchilov and Hutter, 2019) with weight decay 0.1, $\beta_1 = 0.9$, $\beta_2 = 0.95$, gradient clipping at 1.0, and bfloat16 mixed precision. We use an effective batch size of 200 sequences, obtained with a per-device micro-batch size of 50 and gradient accumulation over 4 steps on a single device. Each run starts with a 300-step linear warmup, after which we apply the learning rate schedule under evaluation. The Adam- and Muon-style measurements in Section B are separate scale-factor diagnostics rather than additional validation-loss baselines.

Learning rate schedules. We use the same 300-step warmup and a final linear cooldown phase for all schedules with decay fraction $\alpha = 0.2$. Our main comparison is between `WSqD` and `WSD`. We additionally consider two stronger baselines to assess whether the gains of `WSqD` persist against more carefully tuned alternatives:

- *Two-stage tuned `WSD` (Schaipp et al., 2025).* We first train `WSD` to the initial horizon using the peak learning rate that minimizes validation loss at that horizon. We then save the pre-decay checkpoint and resume training to each extended horizon, while performing a grid search over a new peak learning rate for the training continuation phase.



(a) (b)

Figure 3: **(a)** Continued training on the `SlimPajama` dataset under `WSD` and `WSqD` across horizons of 15000, 30000, 45000, and 60000 iterations. **(b)** Fitted curves of the final validation losses after the linear cooldown phase. For both schedules, the base learning rate is selected by a 10000-iteration search; for `WSqD`, we use the shift parameter $T_0 = 10000$. We set the decay fraction α as 0.2 for both `WSD` and `WSqD`.

- *Power schedule (Shen et al., 2024).* The power schedule has the form $\eta_t = \min\{\eta_{\max}, \beta a t^b\}$, where β is the batch size. Following the authors' recommendation, we fix $b = -0.51$ and tune $\eta_{\max}$ and a via a short-horizon hyperparameter search.

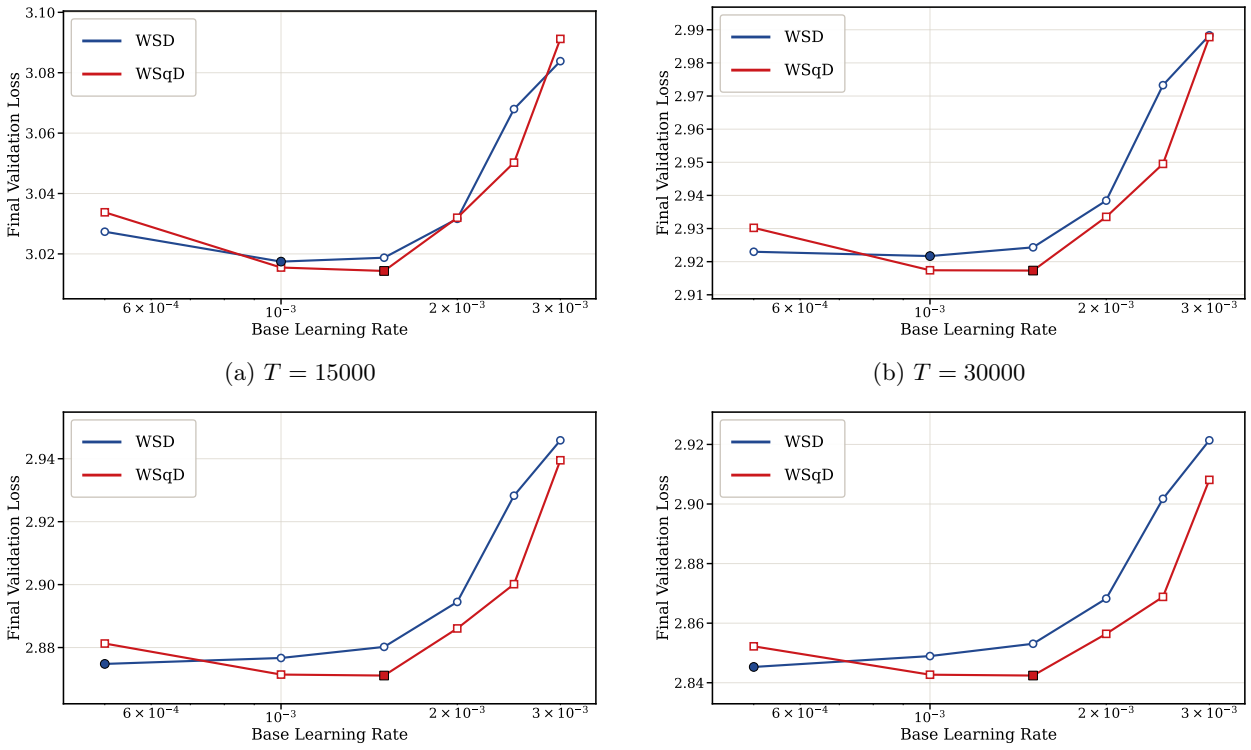
Continuation protocol. We evaluate all learning rate schedules in a setting where the final training horizon is not known in advance and training may extend beyond its originally planned budget. For `WSqD`, the power schedule, and untuned `WSD`, we use a single continuation trajectory across all reported horizons. Specifically, training resumes from the pre-decay checkpoint of the previous horizon and continues to the next horizon without re-tuning the base or peak learning rate. This protocol reflects the practical scenario in which a practitioner reaches the originally planned training budget and subsequently decides to continue training. The two-stage horizon-tuned `WSD` baseline is the only method allowed to re-tune the peak learning rate separately for each extended horizon.

3.2 Continued training across horizons

We evaluate `WSqD` in a training continuation setting and compare it with `WSD`. In contrast to the experiments previewed in Figure 2b, where the base learning rate is selected from a short 5000-iteration search, we now select the base learning rate from a longer 10000-iteration run and reuse the resulting value for all subsequent horizons. Starting from the pre-decay checkpoint, both schedules are extended to training horizons of $T \in \{15000, 30000, 45000, 60000\}$ with the same decay fraction $\alpha = 0.2$. As shown in Figure 3, `WSqD` again matches or improves upon `WSD` at every horizon. The improvement is smaller than in Figure 2b, because the learning rate is tuned at a longer initial horizon, leaving a shorter continuation range. Nevertheless, the qualitative trend is unchanged: the relative advantage of `WSqD` grows with the training horizon, increasing from approximately $1.01\times$ at the initial horizon to about $1.05\times$ at 60000 iterations. These empirical results support the central motivation for `WSqD`: its inverse-square-root base often provides a more effective trajectory for continued training than the flat stable phase of `WSD`, while the final linear cooldown still produces a strong endpoint model.

3.3 Horizon dependence of the optimal peak learning rate

A desirable anytime schedule should not require re-tuning its hyperparameters when training is extended. In this subsection, we examine how the optimal peak learning rate varies with the training horizon for `WSD` and `WSqD`.



(a) $T = 15000$ (b) $T = 30000$

(c) $T = 45000$ (d) $T = 60000$

Figure 4: Final validation loss versus base learning rate for `WSD` (blue) and `WSqD` (red, with shift $T_0 = 10000$) at horizons $T \in \{15000, 30000, 45000, 60000\}$. Filled markers indicate the optimum at each horizon. `WSqD`'s optimum remains fixed at $\eta^\star_{\max} = 0.0015$ across all four horizons, whereas `WSD`'s optimum drifts toward smaller values as T grows.

Specifically, we sweep the base learning rate $\eta_{\max} \in \{0.0005, 0.001, 0.0015, 0.002, 0.0025, 0.003\}$ for both `WSD` and `WSqD` over four training horizons $T \in \{15000, 30000, 45000, 60000\}$, with the shift parameter T_0 of `WSqD` fixed at $T_0 = 10000$ across all horizons. Figure 4 reports the final validation loss for each $(\eta_{\max}, T)$ pair, with filled markers indicating the optimal learning rate at each horizon. The two schedules exhibit qualitatively different sensitivity to the horizon. For `WSqD`, the optimum remains fixed at $\eta^\star_{\max} = 0.0015$ across all four horizons, and the loss curve as a function of $\eta_{\max}$ is nearly horizon-invariant in shape, which is a desirable property in practice, since a practitioner can tune $\eta_{\max}$ once on a short run and reuse it for substantially longer training without re-tuning. By contrast, the optimum of `WSD` drifts toward smaller values as the horizon increases. This horizon-dependent shift is consistent with prior observations (e.g., Shen et al., 2024; Schaipp et al., 2025), and confirms that `WSD` is not truly anytime: a base learning rate tuned at one horizon is generally suboptimal at another, so extending the training budget requires another round of hyperparameter tuning.

3.4 Ablation study of shift parameter T_0

The shift parameter T_0 in (3) is a hyperparameter that determines the shape of the `WSqD` schedule. We therefore examine how sensitive the final loss is to this choice and whether a simple default suffices. For each value of $T_0 \in \{500, 5000, 10000\}$, we first select the optimal peak learning rate using a $T_1 = 10000$-iteration search, and then extend training to longer horizons $T \in \{15000, 30000, 45000, 60000\}$. Figure 5 reports the resulting validation loss for each shift value across these horizons. A very small shift value like $T_0 = 500$

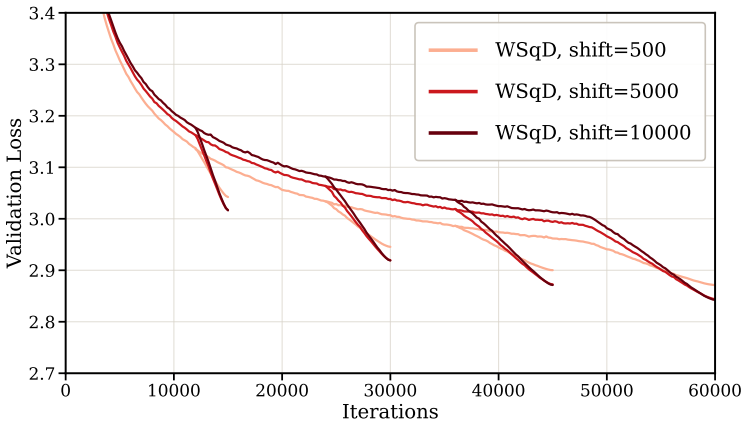


Figure 5: Ablation study over the WSqD shift parameter T_0.

causes the inverse-square-root base to decay too aggressively during the early stages of training, leading to degraded final performance. In contrast, once T_0 is moderately large, performance becomes relatively insensitive to its precise value. Based on these empirical results, we recommend the simple default $T_0 = T_1$, setting the shift equal to the length of the initial pilot horizon.

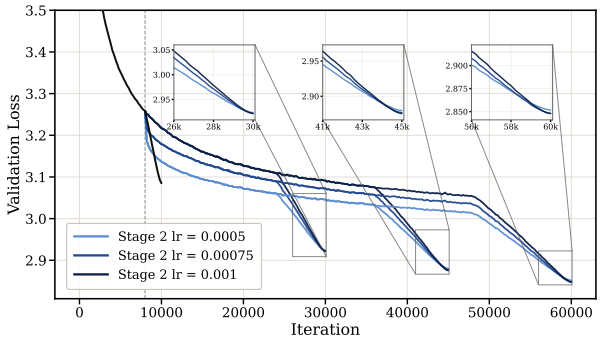
3.5 Comparison with stronger baselines

We next compare WSqD against two stronger baselines designed to mitigate the limitations of WSD. First, we study whether the horizon dependence of WSD can be addressed by re-tuning the learning rate after the horizon is extended. Motivated by Schaipp et al. (2025), who showed that re-optimizing the learning rate can improve continued WSD training, we consider a two-stage tuned WSD baseline: after resuming from the pre-decay checkpoint, we select a fresh peak learning rate for the continuation phase. As shown in Figure 6a, this additional tuning does not substantially improve performance in our setting. Across the swept second-stage peak rates, the two-stage tuned WSD does not outperform the standard single-stage WSD, suggesting that the benefit of second-stage tuning is sensitive to the continuation regime and tuning range. In our experiments, recalibrating the peak learning rate alone is not sufficient; the flat stable phase of WSD remains less effective for extended training than a gradually decaying base schedule of WSqD.

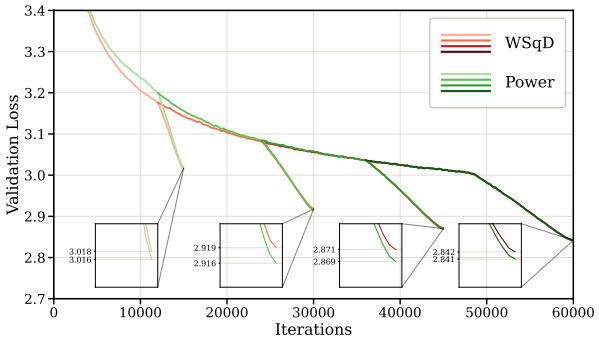
Another baseline we consider is the power schedule of Shen et al. (2024), which is conceptually closest to WSqD because it also employs a horizon-agnostic decaying base phase. We tune its hyperparameters using the same short-horizon protocol and then continue training without further re-tuning. As shown in Figure 6b, the power schedule slightly outperforms WSqD at the reported horizons. However, the gap is small and decreases as the horizon grows, with the two schedules nearly matching after 60000 iterations. This trend suggests that WSqD becomes increasingly competitive with the power schedule at longer horizons. Thus, although the power schedule is evidently a strong empirical baseline, WSqD achieves comparable performance while using a simpler inverse-square-root base, rather than a tuned power-law exponent.

4 Related work

Continued pretraining and horizon extension. We now discuss related work that addressed horizon extension at the level of *schedule shape* rather than via re-warming heuristics. A first line of work modified the WSD skeleton itself: infinite-cycle schedules stitch together repeated warmup–cooldown cycles (Singh et al., 2025), WSD-S (Wen et al., 2024) recycles previous decay phases into a single trunk, and the *microannealing* protocol used by Llama 3 and OLMo 2 (Grattafiori et al., 2024; OLMo et al., 2024) runs a short linear cooldown over the final fraction of tokens combined with high-quality data upsampling, a practice now central



(a) Two stage Tuned-WSD.



(b) WSqD vs. Power schedule.

Figure 6: **(a)** Validation-loss trajectories for two-stage tuned WSD with different second-phase peak learning rates. **(b)** Comparison of WSqD and the power schedule under continued training.

to the emerging "mid-training" phase of LLM development (Mo et al., 2025). At the cooldown extreme, Bergsma et al. (2025) argued empirically that linear decay-to-zero is optimal under a tuned peak learning rate, while Tian et al. (2025) eliminated the decay phase entirely and emulate it via post hoc checkpoint merging. A second line addressed the fact that WSD's optimal peak learning rate itself depends on the horizon: Shen et al. (2024) rescaled the peak as a power-law function of token budget and batch size, and Bjorck et al. (2025) fit an empirical scaling law for peak-rate transfer; both committed to a tuned, horizon-specific peak. A third line dispensed with the schedule altogether and relied on iterate or weight averaging: Schedule-Free optimizers (Defazio et al., 2024; Pun et al., 2025), recently extended to LLM-scale variants such as ScheduleFree+ (Defazio, 2026) and schedule-free spectral optimization with SF-NorMuon (Apte et al., 2026); horizon-free polynomial schedules with weight averaging (Meterez et al., 2026); and classical checkpoint averaging (Wortsman et al., 2022; Sanyal et al., 2024).

Last-iterate convergence of SGD for nonsmooth convex stochastic optimization. Classical SGD theory achieves the minimax-optimal $O(DG/\sqrt{T})$ rate for nonsmooth convex stochastic optimization through iterate averaging, where D denotes the Euclidean diameter. Examples include Polyak–Ruppert averaging (Polyak and Juditsky, 1992), suffix averaging (Rakhlin et al., 2011), and non-uniform averaging (Lacoste-Julien et al., 2012), all of which match the lower bound of Agarwal et al. (2012). By contrast, last-iterate convergence is more subtle. Shamir and Zhang (2013) proved an $O(\log T/\sqrt{T})$ bound for SGD with $\eta_t = c/\sqrt{t}$, and Harvey et al. (2019) showed that the logarithmic factor $\log T$ is unavoidable for inverse-square-root learning rates, even in the deterministic setting. More recently, Kornowski and Shamir (2026) showed that any anytime learning rate schedule, without prior knowledge of T, must incur an excess poly-logarithmic factor in the last-iterate suboptimality gap. When the horizon is known *a priori*, this gap can be closed; for instance, Jain et al. (2021) designed a horizon-dependent learning rate schedule whose resulting last iterate attains the optimal $O(1/\sqrt{T})$ convergence rate. Related known-horizon decay schedules, including linear decay (Defazio et al., 2023) and WSD (Schaipp et al., 2025), likewise achieve the minimax-optimal last-iterate convergence rate. Our WSqD schedule builds upon this line of work but is mainly designed to support training continuation: the inverse-square-root base enables training continuation without re-tuning, whereas the final linear cooldown recovers the optimal last-iterate rate. In the Euclidean projected-SGD special case, WSqD attains the same minimax-optimal last-iterate convergence rate while allowing the pre-decay iterates to be reused across extended horizons. Our main theorem also extends this guarantee for the Euclidean setting to accommodate general Bregman geometry.

5 Conclusion

In this work, we have introduced WSqD, a learning rate schedule motivated by stochastic convex optimization and designed to accommodate post hoc horizon extension in large model training. By combining a horizon-independent inverse-square-root base with a final linear cooldown, WSqD decouples training continuation from last-iterate optimization: the base phase provides a horizon-free trajectory suitable for training continuation, while the final cooldown recovers the optimal last-iterate convergence predicted by theory. Although our analysis has been carried out in a stylized convex setting, our experiments have demonstrated the effectiveness of WSqD for practical LLM pretraining. In particular, WSqD can reuse a base learning rate tuned on a short pilot run, continue along the same trajectory when the budget is extended, and outperform existing baselines across longer horizons.

Our empirical evaluation is necessarily limited by computational resources. Thus far, we have tested a single 213M-parameter LLaMA-style model on one pretraining corpus, SlimPajama, using AdamW as the optimizer. An important next step is to evaluate whether the same advantages persist at larger model scales, across different datasets and architectures, and under broader pretraining regimes. It would also be of interest to study how WSqD interacts with optimizers beyond AdamW, including recent alternatives such as Muon. While Section B provides a preliminary geometric diagnostic for idealized Adam- and Muon-style directions, both a rigorous theoretical treatment of their momentum and adaptive-state dynamics and large-scale empirical evaluations remain open. On the theoretical side, extending our convergence analysis beyond the classical stochastic convex setting to nonconvex objectives and adaptive or momentum-based optimization forms another important direction for future work.

Acknowledgments

Y. Chen is supported in part by the Alfred P. Sloan Research Fellowship, the ONR grant N00014-25-1-2344, the NSF grants 2221009 and 2218773, the Wharton AI & Analytics Initiative's AI Research Fund, and the Amazon Research Award.

A Convergence analysis (proof of Theorem 1)

In this section, we present the proof of Theorem 1. For ease of presentation, we define

$$T_1 \coloneqq (1-\alpha)T \qquad \text{and} \qquad T_2 \coloneqq T - T_1. \tag{9}$$

Without loss of generality, assume that $(1-\alpha)T$ is an integer. Recall that the proposed WSqD schedule

$$\eta_t^{\texttt{WSqD}} = \begin{cases} \dfrac{c_0}{\sqrt{t+T_0}}, & \text{if } 1 \le t \le T_1 \\ \dfrac{c_0}{\sqrt{T_1+T_0}} \cdot \dfrac{T-t}{T_2}, & \text{if } T_1 < t \le T \end{cases} \tag{10}$$

consists of two phases: (i) *Phase 1*, comprising the first T_1 iterations, follows a shifted inverse-square-root schedule; (ii) *Phase 2*, comprising the remaining T_2 iterations, linearly decays the learning rate to zero. In particular, Phase 2 starts from the terminal value of the inverse-square-root phase, $c_0/\sqrt{T_1+T_0}$, and reaches zero at the final iteration.

At a high level, our proof of Theorem 1 consists of the following two main parts.

- **Part I: existence of a good checkpoint in Phase 1 (Lemma 2).** We first show that the final segment (or *suffix*) of Phase 1 contains at least a good checkpoint. More precisely, among the last αT iterates of Phase 1, namely those in the time window $[(1-2\alpha)T, (1-\alpha)T]$, there exists an iterate w_{τ_0} whose expected suboptimality is at most $O\big(\log(1/\alpha)/\sqrt{T}\big)$. Thus, although the final iterate of the inverse-square-root phase may still incur the classical $O(\log T)$ overhead, the suffix of this phase contains

at least one iterate that attains the optimal convergence rate. This step follows from the suffix-averaging argument of Shamir and Zhang (2013), adapted here to identify a single good checkpoint rather than to bound the performance of an averaged iterate.

- **Part II: propagation through Phase 2 to the final iterate (Lemma 3).** We next show that the linear-decay phase propagates the performance guarantee from the checkpoint w_{τ_0} to the final iterate w_T. The proof applies a block decomposition argument inspired by Jain et al. (2021). Specifically, we partition Phase 2 into $m = \lceil \log_2 T_2 \rceil$ blocks whose sizes decrease geometrically. Given the linear cooldown in Phase 2, the effective learning rate scale decreases geometrically from one block to the next, allowing us to control the additional suboptimality accumulated across all these blocks. Since the final block consists solely of the last iterate w_T, the rate-optimal guarantee established in Phase 1 for the checkpoint can be shown to carried through to the final iterate.

Note that while multiple key components of the proof follow standard arguments, we include the complete analysis here to keep the proof self-contained.

A.1 Preliminaries

Before proceeding, we first record the following basic inequalities for stochastic mirror descent that bound the (weighted) cumulative suboptimality gap — relative to an arbitrary reference point u — under a general non-increasing learning rate schedule. Such classical inequalities play an important role in the subsequent analysis.

Lemma 1. *Let $\{w_t\}_{t\geq 1}$ be the iterates of stochastic mirror descent (6), generated using any non-increasing learning rate sequence $\{\eta_t\}_{t\geq 1}$. Consider any given $1 \leq s \leq r \leq T$, and let u be any $\mathcal{F}_s$-measurable random vector taking values in $\mathcal{W}$. Assume $\eta_t > 0$ for every $t = s, \ldots, r$. Then, we have*

$$\sum_{t=s}^{r} \eta_t \mathbb{E}\big[f(w_t) - f(u)\big] \leq \mathbb{E}\left[D_\psi(u, w_s)\right] + \frac{G^2}{2} \sum_{t=s}^{r} \eta_t^2 \tag{11}$$

and

$$\sum_{t=s}^{r} \mathbb{E}\big[f(w_t) - f(u)\big] \leq \frac{\mathbb{E}\left[D_\psi(u, w_s)\right]}{\eta_s} + R_\psi^2 \sum_{t=s+1}^{r} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + \frac{G^2}{2} \sum_{t=s}^{r} \eta_t. \tag{12}$$

A.1.1 Proof of Lemma 1

Consider any given $t \geq s$. We begin with the optimality condition (see, e.g., Beck (2017)) for the subproblem in each mirror descent update (6), which asserts that, for every $u \in \mathcal{W}$,

$$\left\langle \eta_t \widehat{g}_t + \nabla\psi(w_{t+1}) - \nabla\psi(w_t),\, u - w_{t+1} \right\rangle \geq 0,$$

where we use the basic fact that $\nabla_w D_\psi(w, w_t) = \nabla\psi(w) - \nabla\psi(w_t)$. Rearranging terms, we are left with

$$\eta_t \left\langle \widehat{g}_t, w_{t+1} - u \right\rangle \leq \left\langle \nabla\psi(w_{t+1}) - \nabla\psi(w_t), u - w_{t+1} \right\rangle. \tag{13}$$

We now recall the three-point identity for the Bregman divergence (Beck, 2017): for every $a, b, c \in \mathcal{W}$,

$$\langle \nabla\psi(b) - \nabla\psi(c), a - b \rangle = D_\psi(a, c) - D_\psi(a, b) - D_\psi(b, c).$$

Taking $(a, b, c) = (u, w_{t+1}, w_t)$ in this three-point identity leads to

$$\langle \nabla\psi(w_{t+1}) - \nabla\psi(w_t), u - w_{t+1} \rangle = D_\psi(u, w_t) - D_\psi(u, w_{t+1}) - D_\psi(w_{t+1}, w_t).$$

Substituting it into (13), we obtain

$$\eta_t \left\langle \widehat{g}_t, w_{t+1} - u \right\rangle \leq D_\psi(u, w_t) - D_\psi(u, w_{t+1}) - D_\psi(w_{t+1}, w_t).$$

This allows one to demonstrate that

$$\begin{aligned}\eta_t \langle \widehat{g}_t, w_t - u\rangle &= \eta_t \langle \widehat{g}_t, w_{t+1} - u\rangle + \eta_t \langle \widehat{g}_t, w_t - w_{t+1}\rangle \\ &\le D_\psi(u, w_t) - D_\psi(u, w_{t+1}) + \eta_t \langle \widehat{g}_t, w_t - w_{t+1}\rangle - D_\psi(w_{t+1}, w_t) \\ &\le D_\psi(u, w_t) - D_\psi(u, w_{t+1}) + \eta_t \langle \widehat{g}_t, w_t - w_{t+1}\rangle - \frac{1}{2}\|w_{t+1} - w_t\|^2 \\ &\le D_\psi(u, w_t) - D_\psi(u, w_{t+1}) + \frac{\eta_t^2}{2}\|\widehat{g}_t\|_*^2, \end{aligned} \tag{14}$$

where the penultimate step relies on 1-strong convexity of ψ w.r.t. $\|\cdot\|$ (see Assumption 2) and hence $D_\psi(w_{t+1}, w_t) \ge \frac{1}{2}\|w_{t+1} - w_t\|^2$, and the last step follows from Hölder's and Young's inequalities. As a result, setting $d_t = w_t - w_{t+1}$ in the above display (with u taken to be w_{t+1}) yields

$$\eta_t \langle \widehat{g}_t, d_t\rangle - D_\psi(w_{t+1}, w_t) \le -D_\psi(w_{t+1}, w_{t+1}) + \frac{\eta_t^2}{2}\|\widehat{g}_t\|_*^2 = \frac{\eta_t^2}{2}\|\widehat{g}_t\|_*^2. \tag{15}$$

Taking conditional expectation given $\mathcal{F}_t$ in (14), and using $\mathbb{E}[\widehat{g}_t \mid \mathcal{F}_t] = g_t \in \partial f(w_t)$, the almost-sure bound $\|\widehat{g}_t\|_* \le G$, and convexity of f, we obtain

$$\begin{aligned}\eta_t \mathbb{E}[f(w_t) - f(u)] \le \eta_t \mathbb{E}\left[\langle g_t, w_t - u\rangle\right] &= \mathbb{E}\left[\mathbb{E}\left[\eta_t \langle \widehat{g}_t, w_t - u\rangle \mid \mathcal{F}_t\right]\right] \\ &\le \mathbb{E}\left[D_\psi(u, w_t) - D_\psi(u, w_{t+1})\right] + \frac{\eta_t^2 G^2}{2}. \end{aligned} \tag{16}$$

Summing (16) from $t = s$ to r and discarding the nonnegative terminal term $\mathbb{E}[D_\psi(u, w_{r+1})]$, we establish (11).

We now turn to the proof of inequality (12). Dividing (16) by η_t, and letting $a_t = \mathbb{E}\left[D_\psi(u, w_t)\right]$ and $b_t = 1/\eta_t$, we arrive at

$$\mathbb{E}[f(w_t) - f(u)] \le b_t(a_t - a_{t+1}) + G^2\eta_t/2.$$

Summing from $t = s$ to $t = r$ gives

$$\sum_{t=s}^{r} \mathbb{E}[f(w_t) - f(u)] \le \sum_{t=s}^{r} b_t(a_t - a_{t+1}) + \frac{G^2}{2}\sum_{t=s}^{r}\eta_t. \tag{17}$$

To proceed, we rearrange the first sum on the right-hand side of (17) using Abel's summation formula, i.e.,

$$\sum_{t=s}^{r} b_t(a_t - a_{t+1}) = b_s a_s - b_r a_{r+1} + \sum_{t=s+1}^{r} a_t(b_t - b_{t-1}) \le b_s a_s + \sum_{t=s+1}^{r} a_t(b_t - b_{t-1}),$$

which results in

$$\sum_{t=s}^{r} \mathbb{E}[f(w_t) - f(u)] \le b_s a_s + \sum_{t=s+1}^{r} a_t(b_t - b_{t-1}) + \frac{G^2}{2}\sum_{t=s}^{r}\eta_t. \tag{18}$$

Since $a_s = \mathbb{E}\left[D_\psi(u, w_s)\right]$ and $b_s = 1/\eta_s$, we have

$$b_s a_s = \mathbb{E}\left[D_\psi(u, w_s)\right]/\eta_s.$$

Next, since $w_t, u \in \mathcal{W}$, the Bregman-diameter assumption gives $a_t = \mathbb{E}\left[D_\psi(u, w_t)\right] \le R_\psi^2$ for every t. Moreover, $b_t - b_{t-1} = 1/\eta_t - 1/\eta_{t-1} \ge 0$ due to the assumption that the learning rates are non-increasing. Therefore, we have

$$\sum_{t=s+1}^{r} a_t(b_t - b_{t-1}) \le R_\psi^2 \sum_{t=s+1}^{r}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right). \tag{19}$$

Putting everything together completes the proof of the advertised inequality (12).

A.2 Proof of Part I

The proof of this part is based on the suffix-averaging argument — originally developed by Shamir and Zhang (2013) for strongly convex stochastic optimization by Rakhlin et al. (2011) and subsequantly extended to nonsmooth convex optimization. Specifically, the main goal of this subsection is to establish the following lemma, which guarantees the existence of a rate-optimal iterate in the suffix of Phase 1.

Lemma 2. *Assume $T_1 \geq T_0$. Let*

$$B_0 = \{T_1 - q, T_1 - q + 1, \ldots, T_1\} \qquad \text{with } q = \lfloor T_2/2 \rfloor. \tag{20}$$

Then there exists $\tau_0 \in B_0$ such that

$$\mathbb{E}[f(w_{\tau_0}) - f^*] \leq \left(\frac{R_\psi^2}{c_0} + c_0 G^2\right) \frac{\sqrt{2} + \log(3/\alpha)}{\sqrt{T_1}}. \tag{21}$$

A.2.1 Proof of Lemma 2

The proof employs the suffix-averaging technique of Shamir and Zhang (2013). Recall the learning rate $\eta_t = c_0/\sqrt{t + T_0}$ used in Phase 1, and we also define the suffix suboptimality averages as follows:

$$S_k := \frac{1}{k+1} \sum_{t=T_1-k}^{T_1} \mathbb{E}[f(w_t) - f^*], \qquad k = 0, 1, \ldots, T_1 - 1. \tag{22}$$

We proceed in a few steps below.

Step 1: establishing a recursion between suffix suboptimality averages. Consider any given $k \in \{1, \ldots, T_1 - 1\}$. From the definition (22) of S_k, we have the identity

$$(k+1)S_k = \mathbb{E}[f(w_{T_1-k}) - f^*] + kS_{k-1}.$$

Rearranging terms results in

$$k(S_{k-1} - S_k) = S_k - \mathbb{E}[f(w_{T_1-k}) - f^*] = \frac{1}{k+1} \sum_{t=T_1-k}^{T_1} \mathbb{E}\big[f(w_t) - f(w_{T_1-k})\big]. \tag{23}$$

We would like to bound the right-hand side of (23) by applying Lemma 1 on the interval $[T_1 - k, T_1]$ with reference point $u = w_{T_1-k}$. Since the starting iterate equals the reference u, the first term in (12) of Lemma 1 vanishes (i.e., $D_\psi(u, w_{T_1-k}) = 0$). With the non-decreasing learning rates $\eta_t = c_0/\sqrt{t + T_0}$, we have

$$\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} = \frac{\sqrt{t + T_0} - \sqrt{t + T_0 - 1}}{c_0} \geq 0.$$

Using telescoping, we reach

$$R_\psi^2 \sum_{t=T_1-k+1}^{T_1} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) = \frac{R_\psi^2}{c_0}\big(\sqrt{T_1 + T_0} - \sqrt{T_1 + T_0 - k}\big).$$

For the learning rate sum in the bound (12) of Lemma 1, we have

$$\frac{G^2}{2} \sum_{t=T_1-k}^{T_1} \eta_t = \frac{c_0 G^2}{2} \sum_{t=T_1-k}^{T_1} \frac{1}{\sqrt{t + T_0}} \leq c_0 G^2\big(\sqrt{T_1 + T_0} - \sqrt{T_1 + T_0 - k - 1}\big).$$

Since $\sqrt{T_1+T_0-k} \geq \sqrt{T_1+T_0-k-1}$, substituting the above bounds into Lemma 1 gives

$$\begin{aligned}\sum_{t=T_1-k}^{T_1} \mathbb{E}\big[f(w_t) - f(w_{T_1-k})\big] &\leq \left(\frac{R_\psi^2}{c_0} + c_0 G^2\right)\left(\sqrt{T_1+T_0} - \sqrt{T_1+T_0-k-1}\right) \\ &\leq \left(\frac{R_\psi^2}{c_0} + c_0 G^2\right)\frac{k+1}{\sqrt{T_1+T_0}},\end{aligned} \tag{24}$$

where the last line follows since

$$\sqrt{T_1+T_0} - \sqrt{T_1+T_0-k-1} = \frac{k+1}{\sqrt{T_1+T_0} + \sqrt{T_1+T_0-k-1}} \leq \frac{k+1}{\sqrt{T_1+T_0}}.$$

Subsituting the bound (24) into (23), we arrive at

$$k(S_{k-1} - S_k) \leq \left(\frac{R_\psi^2}{c_0} + c_0 G^2\right)\frac{1}{\sqrt{T_1+T_0}}, \tag{25}$$

and as a result,

$$S_{k-1} \leq S_k + \left(\frac{R_\psi^2}{c_0} + c_0 G^2\right)\frac{1}{k\sqrt{T_1+T_0}}. \tag{26}$$

Step 2: telescoping the recursion. Recall that $q = \lfloor T_2/2 \rfloor = \lfloor \alpha T/2 \rfloor$ and $T_1 = (1-\alpha)T$. Summing the above recursion (26) for $k = q+1, \ldots, T_1 - 1$ leads to

$$S_q \leq S_{T_1-1} + \left(\frac{R_\psi^2}{c_0} + c_0 G^2\right)\frac{1}{\sqrt{T_1+T_0}} \sum_{k=q+1}^{T_1-1} \frac{1}{k} \leq S_{T_1-1} + \left(\frac{R_\psi^2}{c_0} + c_0 G^2\right)\frac{\log(3/\alpha)}{\sqrt{T_1+T_0}}. \tag{27}$$

To bound the full-run average S_{T_1-1}, we apply Lemma 1 with $u = w^*$, $s = 1$, and $r = T_1$ to yield

$$\begin{aligned}\sum_{t=1}^{T_1} \mathbb{E}[f(w_t) - f^*] &\leq \frac{\mathbb{E}\left[D_\psi(w^*, w_1)\right]}{\eta_1} + R_\psi^2 \sum_{t=2}^{T_1}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) + \frac{G^2}{2}\sum_{t=1}^{T_1}\eta_t \\ &\leq \frac{R_\psi^2\sqrt{T_0+1}}{c_0} + \frac{R_\psi^2}{c_0}\left(\sqrt{T_1+T_0} - \sqrt{T_0+1}\right) + c_0 G^2\left(\sqrt{T_1+T_0} - \sqrt{T_0}\right) \\ &= \frac{R_\psi^2}{c_0}\sqrt{T_1+T_0} + c_0 G^2\left(\sqrt{T_1+T_0} - \sqrt{T_0}\right) \\ &\leq \left(\frac{R_\psi^2}{c_0} + c_0 G^2\right)\sqrt{T_1+T_0}.\end{aligned}$$

Dividing by T_1 gives

$$S_{T_1-1} \leq \left(\frac{R_\psi^2}{c_0} + c_0 G^2\right)\frac{\sqrt{T_1+T_0}}{T_1}. \tag{28}$$

Step 3: putting all this together. Substituting (28) into (27) gives

$$\begin{aligned}S_q &\leq \left(\frac{R_\psi^2}{c_0} + c_0 G^2\right)\frac{\sqrt{T_1+T_0}}{T_1} + \left(\frac{R_\psi^2}{c_0} + c_0 G^2\right)\frac{\log(3/\alpha)}{\sqrt{T_1+T_0}} \\ &\leq \left(\frac{R_\psi^2}{c_0} + c_0 G^2\right)\frac{\sqrt{2} + \log(3/\alpha)}{\sqrt{T_1}},\end{aligned} \tag{29}$$

provided that $T_1 \geq T_0$. Since the minimum of a finite set is at most its average, there exists $\tau_0 \in B_0$ that satisfies $\mathbb{E}[f(w_{\tau_0}) - f^*] \leq S_q$. This completes the proof.

A.3 Proof of Part II

Recall that $T_2 = \alpha T$ (cf. (9)), $B_0 = \{T_1 - q, T_1 - q + 1, \ldots, T_1\}$ with $q = \lfloor T_2/2 \rfloor$ (cf. (20)), and there exists

$$\tau_0 \in \operatorname*{argmin}_{t \in B_0} \mathbb{E}[f(w_t)] \tag{30a}$$

that forms a good checkpoint in Phase 1. We now partition Phase 2 into $m = \lceil \log_2 T_2 \rceil$ blocks with geometrically shrinking sizes as follows.

- Set $R_j = \lceil T_2/2^j \rceil$ for $j = 0, 1, \ldots, m$, so that $R_0 = T_2$ and $R_m = 1$;
- Define $U_j = T - R_j$, which gives $U_0 = T_1$ and $U_m = T - 1$;
- Let $B_j = \{U_{j-1} + 1, \ldots, U_j\}$ for $j = 1, \ldots, m$.

In the sequel, for each such j, we let

$$\tau_j \in \operatorname*{argmin}_{t \in B_j} \mathbb{E}[f(w_t)] \tag{30b}$$

denote the best-performing iterate in block B_j, and let $\eta_j^{\max} := \eta_{U_{j-1}+1}$ be the largest learning rate in B_j.

The central technical ingredient in Part II is the following lemma, which bounds the performance gap between the best-performing iterates across different blocks.

Lemma 3. *Under the block decomposition above, for every $j = 1, \ldots, m$, we have*

$$\mathbb{E}[f(w_{\tau_j}) - f(w_{\tau_{j-1}})] \le 36 G^2 \eta_j^{\max}.$$

Moreover, it holds that

$$\mathbb{E}[f(w_T) - f(w_{\tau_0})] \le \frac{73 c_0 G^2}{\sqrt{T_1 + T_0}}.$$

A.3.1 Proof of Lemma 3

Recall that throughout Phase 2, the learning rates are given by

$$\eta_t = \frac{c_0}{\sqrt{T_1 + T_0}} \cdot \frac{T - t}{T_2}, \qquad t = T_1 + 1, \ldots, T. \tag{31}$$

For any $j \in \{1, \ldots, m\}$, applying the basic inequality (11) in Lemma 1 for stochastic mirror descent on the interval $[\tau_{j-1}, U_j]$, with the reference point chosen as $u = w_{\tau_{j-1}}$, we obtain

$$2 \sum_{t=\tau_{j-1}}^{U_j} \eta_t \mathbb{E}[f(w_t) - f(w_{\tau_{j-1}})] \le G^2 \sum_{t=\tau_{j-1}}^{U_j} \eta_t^2. \tag{32}$$

This taken together with the optimality of τ_{j-1} and τ_j within their respective blocks (cf. (30b)) gives

$$\begin{aligned} 2\Big(\sum_{t \in B_j} \eta_t \Big) \mathbb{E}\left[f(w_{\tau_j}) - f(w_{\tau_{j-1}}) \right] &= 2 \left\{ \sum_{t=\tau_{j-1}}^{U_{j-1}} \eta_t \mathbb{E}[f(w_{\tau_{j-1}}) - f(w_{\tau_{j-1}})] + \sum_{t=U_{j-1}+1}^{U_j} \eta_t \mathbb{E}[f(w_{\tau_j}) - f(w_{\tau_{j-1}})] \right\} \\ &\le 2 \sum_{t=\tau_{j-1}}^{U_j} \eta_t \mathbb{E}[f(w_t) - f(w_{\tau_{j-1}})] \le G^2 \sum_{t \in B_{j-1} \cup B_j} \eta_t^2, \end{aligned} \tag{33}$$

which in turn implies that

$$\mathbb{E}[f(w_{\tau_j}) - f(w_{\tau_{j-1}})] \le \frac{G^2 \sum_{t \in B_{j-1} \cup B_j} \eta_t^2}{2 \sum_{t \in B_j} \eta_t}. \tag{34}$$

In what follows, we bound the denominator and numerator of this upper bound (34) separately.

- First, since $R_{j-1} \le 2R_j$ by construction, every $t \in B_j$ must satisfy $\eta_t \ge \frac{1}{2}\eta_j^{\max}$. Therefore, we have

$$\sum_{t \in B_j} \eta_t \ge \frac{1}{2}|B_j|\eta_j^{\max}.$$

- Next, we develop an upper bound on the numerator in the upper bound (34). Let

$$K = \frac{c_0}{\sqrt{T_1 + T_0}}, \qquad L_j = |B_j| = R_{j-1} - R_j, \qquad \text{and} \qquad M_j = R_{j-1} - 1, \tag{35}$$

so that we can express

$$\eta_j^{\max} = KM_j/T_2. \tag{36}$$

  - First, consider $j \ge 2$. Note that the condition $t \in B_{j-1} \cup B_j$ is equivalent to

$$n \coloneqq T - t \in \{R_j, \dots, R_{j-2} - 1\}.$$

    Thus, for each such $t \in B_{j-1} \cup B_j$, it holds that

$$n \le R_{j-2} - 1 \le 3(R_{j-1} - 1) = 3M_j,$$

    which follows since $R_{j-2} \le 2R_{j-1}$ and $R_{j-1} \ge 2$. We also make the observation that

$$R_j \le 2L_j.$$

    Hence, it is readily seen that every learning rate in $B_{j-1} \cup B_j$ is at most $3\eta_j^{\max}$, and

$$|B_{j-1} \cup B_j| = R_{j-2} - R_j \le 2R_{j-1} - R_j = R_j + 2L_j \le 4L_j.$$

    As a consequence, we conclude that, for each $j \ge 2$,

$$\sum_{t \in B_{j-1} \cup B_j} \eta_t^2 \le |B_{j-1} \cup B_j| \max_{t \in B_{j-1} \cup B_j} \eta_t^2 \le 36|B_j|(\eta_j^{\max})^2.$$

  - Next, we turn to the case with $j = 1$. Recall the suffix block of Phase 1: $B_0 = \{T_1 - q, \dots, T_1\}$, where $q = \lfloor T_2/2 \rfloor$. Under the theorem assumption that $T_2 = \alpha T \ge 4$, we have

$$|B_0 \cup B_1| = (q+1) + q \le 3q = 3|B_1|.$$

    Moreover, one has $q \le T_2/2 < T_1/2$ because $\alpha < 1/2$; hence, it can be easily seen that every learning rate in B_0 is at most $\sqrt{2}K$, while

$$\eta_1^{\max} = \frac{KM_1}{T_2} = \frac{K(T_2 - 1)}{T_2} \ge \frac{3K}{4}.$$

    As a result, all learning rates in $B_0 \cup B_1$ are at most $2\eta_1^{\max}$. Therefore, the same (conservative) bound holds for $j = 1$ as well, namely,

$$\sum_{t \in B_{j-1} \cup B_j} \eta_t^2 \le 36|B_j|(\eta_j^{\max})^2 \qquad \text{when } j = 1.$$

Combining the above bounds with (34) readily yields, for all $j \ge 1$,

$$\mathbb{E}[f(w_{\tau_j}) - f(w_{\tau_{j-1}})] \le 36G^2\eta_j^{\max}.$$

Moreover, observe that

$$\eta_j^{\max} = \frac{c_0}{T_2\sqrt{T_1+T_0}}(R_{j-1}-1) \le \frac{c_0}{2^{j-1}\sqrt{T_1+T_0}},$$

where we have used

$$R_{j-1} - 1 = \left\lceil \frac{T_2}{2^{j-1}} \right\rceil - 1 \le \frac{T_2}{2^{j-1}}.$$

Since $R_m = 1$, we have $B_m = \{T-1\}$, and hence $\tau_m = T-1$. Summing over $j = 1, \ldots, m$ gives

$$\mathbb{E}[f(w_{T-1}) - f(w_{\tau_0})] \le 36G^2 \sum_{j=1}^{m} \eta_j^{\max} \le \frac{36c_0G^2}{\sqrt{T_1+T_0}} \sum_{j=1}^{\infty} \frac{1}{2^{j-1}} = \frac{72c_0G^2}{\sqrt{T_1+T_0}}. \tag{37}$$

Lastly, since f is G-Lipschitz with respect to $\|\cdot\|$, one has

$$\mathbb{E}\left[f\left(w_T\right) - f\left(w_{T-1}\right)\right] \le G\mathbb{E}\big[\,\|w_T - w_{T-1}\|\,\big].$$

It remains to control the final stochastic mirror descent iteration. Note that the optimality condition for the subproblem (6) at iteration $T-1$ gives

$$\langle \nabla\psi(w_T) - \nabla\psi(w_{T-1}), w_T - w_{T-1}\rangle \le \eta_{T-1}\left\langle \widehat{g}_{T-1}, w_{T-1} - w_T\right\rangle. \tag{38}$$

By virtue of 1-strong convexity of ψ, the left-hand side of (38) is at least $\|w_T - w_{T-1}\|^2$, while the right-hand side of (38) is at most $\eta_{T-1}\|\widehat{g}_{T-1}\|_*\|w_T - w_{T-1}\|$. Hence, it holds almost surely that

$$\|w_T - w_{T-1}\| \le \eta_{T-1}\|\widehat{g}_{T-1}\|_* \le \eta_{T-1}G.$$

This immediately reveals that

$$\mathbb{E}\left[f\left(w_T\right) - f\left(w_{T-1}\right)\right] \le G^2\eta_{T-1} = \frac{c_0G^2}{T_2\sqrt{T_1+T_0}}.$$

To finish up, adding this to the above bound (37) yields

$$\mathbb{E}[f(w_T) - f(w_{\tau_0})] \le \frac{72c_0G^2}{\sqrt{T_1+T_0}} + \frac{c_0G^2}{T_2\sqrt{T_1+T_0}} \le \frac{73c_0G^2}{\sqrt{T_1+T_0}},$$

thereby concluding the proof of Lemma 3.

A.4 Proof of Theorem 1

According to Lemma 2, there exists a checkpoint w_{τ_0} with $\tau_0 \in \operatorname{argmin}_{t\in B_0}\mathbb{E}[f(w_t)]$ such that

$$\mathbb{E}[f(w_{\tau_0}) - f^*] \le \left(\frac{R_\psi^2}{c_0} + c_0G^2\right)\frac{\sqrt{2}+\log(3/\alpha)}{\sqrt{T_1}}. \tag{39}$$

Lemma 3 also tells us that the linear cooldown phase guarantees that

$$\mathbb{E}[f(w_T) - f(w_{\tau_0})] \le \frac{73c_0G^2}{\sqrt{T_1+T_0}}. \tag{40}$$

Taking these two inequalities together gives

$$\begin{aligned}
\mathbb{E}[f(w_T) - f^*] &= \mathbb{E}[f(w_T) - f(w_{\tau_0})] + \mathbb{E}[f(w_{\tau_0}) - f^*] \\
&\le \left(\frac{R_\psi^2}{c_0} + c_0G^2\right)\frac{\sqrt{2}+\log(3/\alpha)}{\sqrt{T_1}} + \frac{73c_0G^2}{\sqrt{T_1+T_0}} \\
&\le \left(\frac{R_\psi^2}{c_0} + c_0G^2\right)\big(120 + 2\log(1/\alpha)\big)\frac{1}{\sqrt{T}},
\end{aligned}$$

where we have used the elemantary facts $T_1 = (1-\alpha)T \ge T/2$, $T_1 + T_0 \ge T/2$, $c_0G^2 \le R_\psi^2/c_0 + c_0G^2$, and $\sqrt{2}\big(\sqrt{2} + \log(3/\alpha) + 73\big) \le 120 + 2\log(1/\alpha)$ for $\alpha \in (0, 1/2)$. This establishes the claimed bound in (8).

B A mirror-descent view of Adam and Muon

This section provides an informal, geometric interpretation of the update directions of two practically important optimizers through the mirror-descent viewpoint developed in the main text. The discussion is deliberately limited in scope and should not be viewed as a convergence analysis. In particular, we ignore momentum, bias correction, weight decay, and numerical stabilizers, and focus only on the geometry of the update directions. Thus, our convergence theory in the main text should not be viewed as covering the full model of AdamW or Muon. Moreover, the mirror-descent calculations below use possibly nonsmooth norm geometries and should not be confused with the differentiable mirror-map assumptions used in our convergence theory. Extending this mirror-descent perspective to the full momentum-based algorithms, including their adaptive dynamics, is an interesting direction for future work.

Consider first the unconstrained setting, and let $\Delta = w - w_t$ denote the local displacement vector. Given a norm $\|\cdot\|$, if we take the Bregman divergence to be $D_\psi(w, w_t) = \frac{1}{2}\|w - w_t\|^2$, then the unconstrained mirror-descent step (6) takes the form

$$\Delta_t \in \operatorname*{argmin}_{\Delta} \left\{ \langle \widehat{g}_t, \Delta \rangle + \frac{1}{2\eta_t} \|\Delta\|^2 \right\}, \tag{41}$$

which can be interpreted as a steepest-descent update with respect to the norm $\|\cdot\|$. To characterize the solution, write $\Delta = -rv$, where $r = \|\Delta\|$ and $\|v\| \le 1$. Minimizing over r and v separately shows that the optimal displacement has radius $r = \eta_t \|\widehat{g}_t\|_*$ and points in a dual steepest-descent direction, namely,

$$\Delta_t = -\eta_t \|\widehat{g}_t\|_* v_t, \qquad v_t \in \operatorname*{argmax}_{\|v\| \le 1} \langle \widehat{g}_t, v \rangle . \tag{42}$$

The important point is that the mirror-descent update naturally decomposes into a normalized descent direction and a scalar factor given by the corresponding dual norm of the stochastic subgradient estimate.

Adam-style sign updates. Now consider the idealized Adam-style update, motivated by the coordinatewise normalization in Adam and AdamW (Kingma and Ba, 2015; Loshchilov and Hutter, 2019),

$$w_{t+1} = w_t - \eta_t \operatorname{sign}(\widehat{g}_t), \tag{43}$$

where $\widehat{g}_t$ denotes the stochastic gradient. Under the ℓ_∞ geometry (i.e., taking the norm $\|\cdot\|$ to be the ℓ_∞ norm), the corresponding unconstrained mirror-descent update is

$$\Delta_t \in \operatorname*{argmin}_{\Delta} \left\{ \langle \widehat{g}_t, \Delta \rangle + \frac{1}{2\eta_t} \|\Delta\|_\infty^2 \right\}. \tag{44}$$

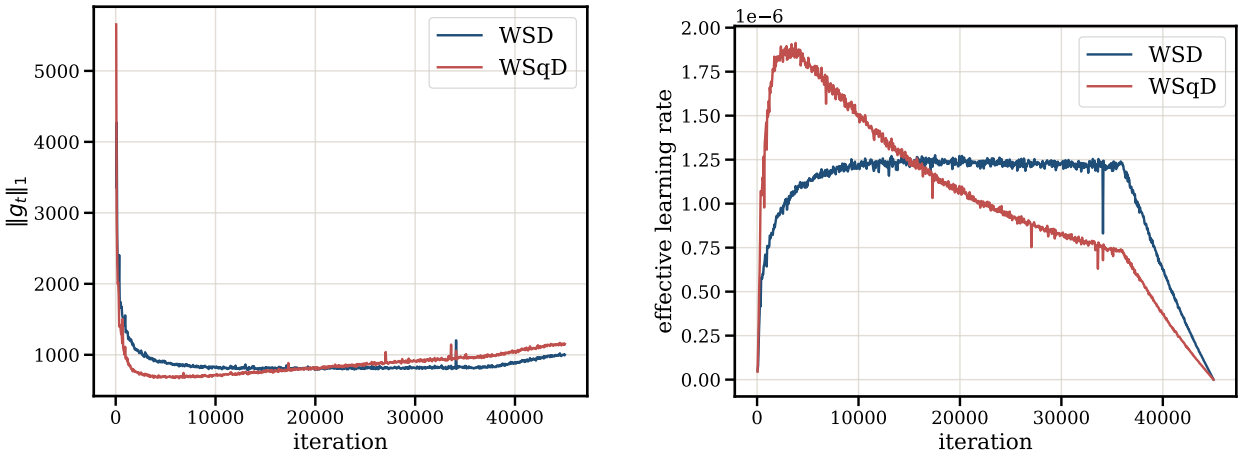
Since the dual norm of the ℓ_∞ norm is the ℓ_1 norm and $\operatorname{sign}(\widehat{g}_t)$ is a maximizer of $\langle \widehat{g}_t, v \rangle$ over $\|v\|_\infty \le 1$, a solution of (44) is

$$\Delta_t = -\eta_t \|\widehat{g}_t\|_1 \operatorname{sign}(\widehat{g}_t), \tag{45}$$

up to the non-uniqueness on zero coordinates. Therefore, the ℓ_∞-based mirror-descent update differs from the practical sign gradient update in (43) only by the scalar factor $\|\widehat{g}_t\|_1$. If this factor varies slowly during training, its effect can be approximately absorbed into the learning rate scale. This observation gives a partial explanation for why the idealized sign update can resemble a mirror-descent step, despite omitting the dual-norm scaling factor.

Muon-style block spectral updates. We next consider the Muon optimizer (Jordan, 2024). Suppose the model parameters are partitioned into matrix blocks W_i, with corresponding block stochastic gradients $\widehat{g}_{i,t}$. The idealized Muon-style update applies the polar direction blockwise:

$$W_{i,t+1} = W_{i,t} - \eta_t \operatorname{Polar}(\widehat{g}_{i,t}), \qquad \text{where } \operatorname{Polar}(G) = UV^\top \text{ for } G = U\Sigma V^\top. \tag{46}$$



(a) (b)

Figure 7: Adam-style mirror-descent scale diagnostics. **(a)** The omitted ℓ_∞ mirror-descent scale factor $\|\widehat{g}_t\|_1$ remains nearly constant during training. **(b)** Dividing the practical learning rate by this omitted scaling factor gives the corresponding mirror-descent learning-rate scale.

Equip block displacements $\Delta = (\Delta_i)_i$ with the block operator norm

$$\|\Delta\|_{\text{blk-op}} := \max_i \|\Delta_i\|_{\text{op}}.$$

Its dual norm is the sum of the block nuclear norms:

$$\|\widehat{g}\|_{\text{blk,nuc}} = \sum_i \|\widehat{g}_i\|_{\text{nuc}},$$

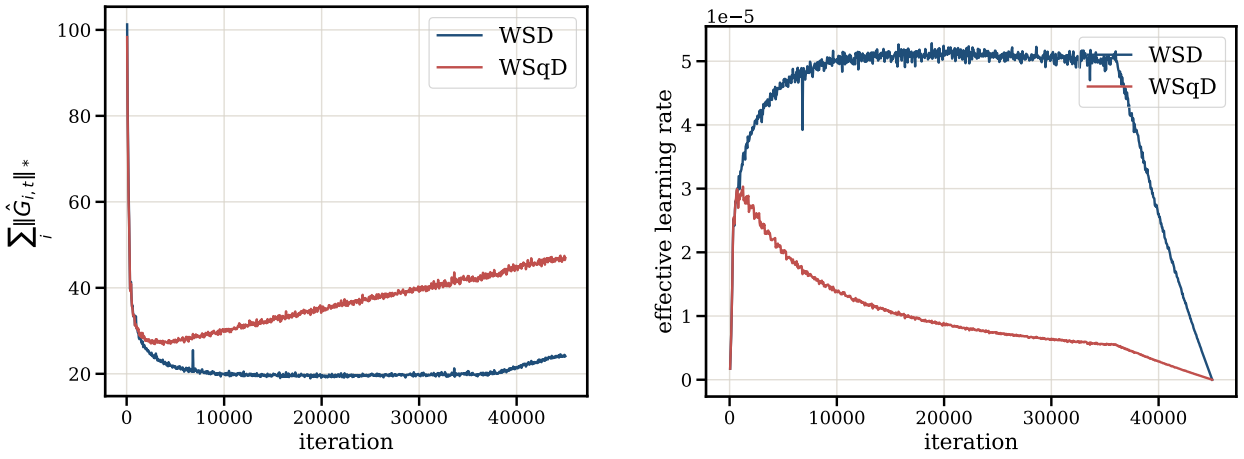
where $\|\cdot\|_{\text{nuc}}$ denotes the nuclear norm. Applying (42) with the matrix inner product gives the blockwise minimizer

$$\Delta_{i,t} = -\eta_t \left(\sum_j \|\widehat{g}_{j,t}\|_{\text{nuc}} \right) \text{Polar}(\widehat{g}_{i,t}), \tag{47}$$

up to the nonuniqueness of the polar factor for rank-deficient or zero blocks. Thus, the exact block-operator-norm mirror-descent step yields the same polar directions as the idealized Muon update, differing only by the scalar factor $\sum_i \|\widehat{g}_{i,t}\|_{\text{nuc}}$. If this factor varies slowly during training, its effect can again be approximately absorbed into the learning-rate scale.

Empirical diagnostics. The preceding calculation suggests a simple diagnostic: monitoring the omitted scaling factors during training, namely $\|\widehat{g}_t\|_1$ for Adam-style sign updates and $\sum_i \|\widehat{g}_{i,t}\|_{\text{nuc}}$ for Muon-style block spectral updates. In the AdamW experiments, $\widehat{g}_t$ denotes the diagonal-normalized update direction used in the sign-update idealization, rather than the raw stochastic gradient. In the Muon-style diagnostic experiments, the nuclear-norm sum is computed over the two-dimensional matrix blocks to which Muon-style orthogonalization is applied; scalar, vector, embedding, and output-head parameters are excluded from this block sum. Empirically, these scaling factors remain nearly constant throughout our experiments. Figures 7 and 8 report the measured factors, together with the corresponding learning-rate rescalings obtained by dividing the practical learning rate by the omitted scaling factor.

Overall, this observation helps bridge the gap a bit between practical optimizer directions and the mirror-descent geometric perspective: the main discrepancy in these idealized Adam- and Muon-style updates is an omitted scaling factor that appears fairly stable in our experiments. At the same time, our



(a) (b)

Figure 8: Muon-style mirror-descent scale diagnostics. **(a)** The omitted block-operator-norm mirror-descent scale factor $\sum_i \|\widehat{g}_{i,t}\|_{\mathrm{nuc}}$ remains nearly constant during training. **(b)** Dividing the practical learning rate by this omitted scaling factor gives the corresponding mirror-descent learning-rate scale.

empirical observations do not explain why these gradient-norm factors remain nearly constant during training. Understanding the underlying cause of this apparent stability, as well as extending the mirror-descent perspective to incorporate momentum and adaptive dynamics, is worthy of future investigation.

C Additional experiments

In addition to the numerical experiments reported in Section 3, we conduct three additional empirical studies: sensitivity of the `WSqD`–`WSD` comparison to the choice of random seeds (Section C.1), generalization of the comparison to a second pretraining corpus (Section C.2), and to a smaller LLaMA model (Section C.3). Across all three settings, the qualitative findings from the main text continue to hold: with a single base learning rate selected on a short pilot run and reused without re-tuning, `WSqD` matches or improves over `WSD` at every continuation horizon. The base learning-rate sweeps supporting the choice of $\eta_{\max} = 0.0015$ used throughout this paper are reported in Section 3.3.

C.1 Sensitivity to random seeds

To verify that the gap between `WSqD` and `WSD` in Section 3.2 is not an artifact of a particular random seed, we repeat the `SlimPajama` continuation experiment of Figure 3 using three independent seeds. Specifically, we evaluate the same four continuation horizons, $T \in 15000, 30000, 45000, 60000$, with $\eta_{\max} = 0.0015$ and shift $T_0 = 10000$, while varying both the model-initialization seed and the data-shuffling seed.

Figure 9 reports the mean validation loss over the three seeds for each (scheduler, horizon) pair, together with a band indicating the per-iteration minimum and maximum across seeds. The variability induced by random initialization at the final iterate is small for both schedules: across the three seeds, the maximum minus the minimum full-validation final loss is at most ≈ 0.01 nat at every horizon. The qualitative ordering observed in Section 3.2 is preserved at the seed-mean level: `WSqD`'s mean final loss lies below `WSD`'s at every horizon, and the gap increases from about 0.003 nat at $T = 15000$ to about 0.011 nat at $T = 60000$. The sign of the gap is consistent across seeds, indicating that the observed advantage of `WSqD` is not driven by a favorable single-seed realization.

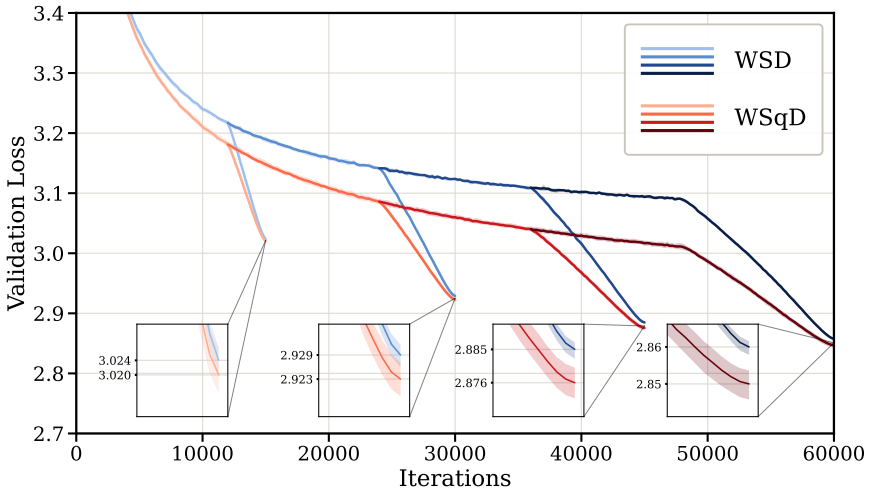


Figure 9: Seed sensitivity of the `SlimPajama` continuation experiment from Figure 3. We re-run `WSD` and `WSqD` at three independent seeds across the four horizons $T \in \{15000, 30000, 45000, 60000\}$, with $\eta_{\max} = 0.0015$ and and a shift parameter of $T_0 = 10000$ for `WSqD`. Solid lines denote the mean validation loss across seeds, while the shaded bands indicate the per-iteration minimum and maximum.

C.2 Experiments on `OpenWebText2`

To verify that the comparison is not specific to `SlimPajama`, we repeat the 4-horizon continuation experiment on `OpenWebText2` using the same 213M LLaMA model (24 layers, 12 heads, embedding dimension 768) and the same continuation protocol as in Section 3.2. The base learning rate is fixed to $\eta_{\max} = 0.0015$ across all horizons, and `WSqD` uses a shift parameter of $T_0 = 5000$. Both schedules use a decay fraction of $\alpha = 0.2$.

Figure 10 shows that the qualitative behavior on `OpenWebText2` closely matches that on `SlimPajama`: `WSqD` improves upon `WSD` at every reported horizon, and the absolute gap grows from about 0.005 nat at $T = 15000$ to 0.012 nat at $T = 60000$. These results indicate that the growing advantage of `WSqD` with increasing continuation horizon is not specific to a single pretraining corpus.

C.3 Experiments on a smaller LLaMA model

We also repeat the continuation experiment using a smaller LLaMA model with 12 layers (about 124M parameters), while keeping all other architectural choices and the optimizer configuration identical to those in Section 3.1. As before, the base learning rate is fixed to $\eta_{\max} = 0.0015$ across all horizons, `WSqD` uses a shift parameter of $T_0 = 5000$, and and both schedules employ a decay fraction of $\alpha = 0.2$.

Figure 11 reports the result on `SlimPajama`. The qualitative ordering is again preserved: `WSqD` matches `WSD` at the shortest horizon and outperforms it at every longer horizon. The absolute gap is smaller than in the 213M setting, ranging from roughly 0.003 nat at $T = 15000$ to 0.005 nat at $T = 60000$. This behavior is consistent with the smaller model being closer to its attainable validation loss at this scale, leaving less room for improvement through continued training. Overall, the advantage of `WSqD` persists across model sizes, although its magnitude depends on the headroom available at the chosen scale.

References

Agarwal, A., Bartlett, P. L., Ravikumar, P., and Wainwright, M. J. (2012). Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249.

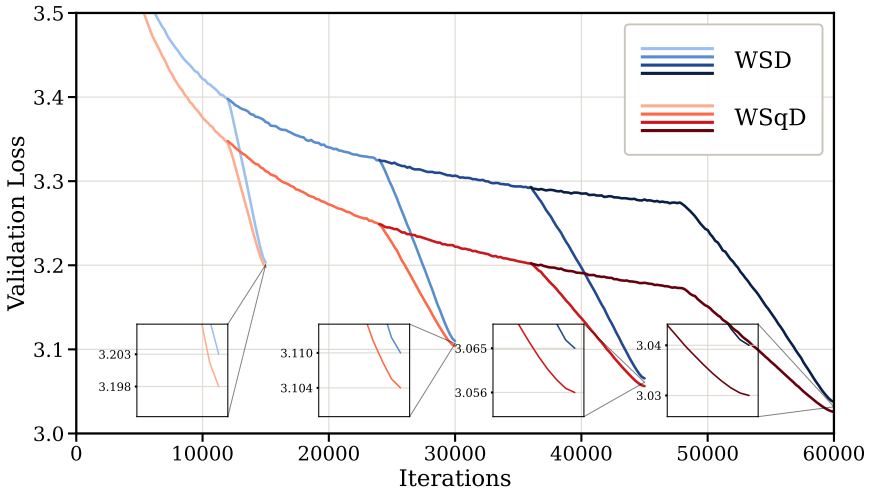


Figure 10: Continued training on `OpenWebText2` with the 213M LLaMA model across horizons $T \in \{15000, 30000, 45000, 60000\}$. The base learning rate is $\eta_{\max} = 0.0015$ for both schedules; `WSqD` uses a shift parameter of $T_0 = 5000$; and the decay fraction is $\alpha = 0.2$.

Apte, A., Deshpande, P., Kumar, N., Chakrabarti, S., and Kim, J. L. (2026). Anytime training with schedule-free spectral optimization. *arXiv preprint arXiv:2605.23061*.

Beck, A. (2017). *First-order methods in optimization*. SIAM.

Bergsma, S., Dey, N., Gosal, G., Gray, G., Soboleva, D., and Hestness, J. (2025). Straight to zero: Why linearly decaying the learning rate to zero works best for LLMs. *arXiv preprint arXiv:2502.15938*.

Bjorck, J., Benhaim, A., Chaudhary, V., Wei, F., and Song, X. (2025). Scaling optimal LR across token horizons. In *International Conference on Learning Representations (ICLR)*. arXiv:2409.19913.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

De Lange, M., van de Ven, G. M., and Tuytelaars, T. (2023). Continual evaluation for lifelong learning: Identifying the stability gap. In *International Conference on Learning Representations (ICLR)*. Spotlight; arXiv:2205.13452.

Defazio, A. (2026). ScheduleFree+: Scaling learning-rate-free & schedule-free learning to large language models. *arXiv preprint arXiv:2605.19095*.

Defazio, A., Cutkosky, A., Mehta, H., and Mishchenko, K. (2023). Optimal linear decay learning rate schedules and further refinements. *arXiv preprint arXiv:2310.07831*.

Defazio, A., Yang, X., Mehta, H., Mishchenko, K., Khaled, A., and Cutkosky, A. (2024). The road less scheduled. *Advances in Neural Information Processing Systems*, 37:9974–10007.

Grattafiori, A., Dubey, A., Jauhri, A., et al. (2024). The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Guo, Y., Fu, J., Zhang, H., Zhao, D., and Shen, Y. (2024). Efficient continual pre-training by mitigating the stability gap. *arXiv preprint arXiv:2406.14833*.

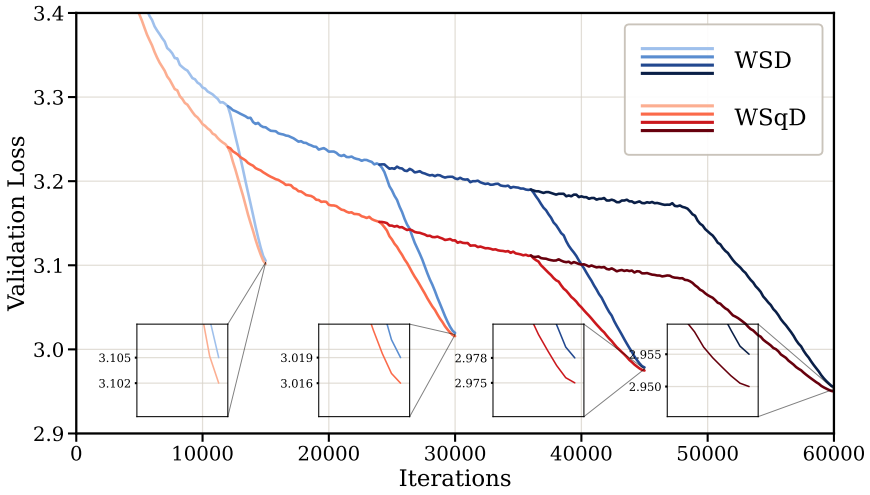


Figure 11: Continuation training on `SlimPajama` with a smaller 12-layer (124M-parameter) LLaMA model across horizons $T \in \{15000, 30000, 45000, 60000\}$. The base learning rate is $\eta_{\max} = 0.0015$ for both schedules; `WSqD` uses a shift parameter of $T_0 = 5000$; and the decay fraction is $\alpha = 0.2$.

Gupta, K., Thérien, B., Ibrahim, A., Richter, M. L., Anthony, Q., Belilovsky, E., Rish, I., and Lesort, T. (2023). Continual pre-training of large language models: How to (re)warm your model? *arXiv preprint arXiv:2308.04014*.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In *Annual Meeting of the Association for Computational Linguistics*.

Hägele, A., Bakouch, E., Kosson, A., Ben Allal, L., Von Werra, L., and Jaggi, M. (2024). Scaling laws and compute-optimal training beyond fixed training durations. *Advances in Neural Information Processing Systems*. Spotlight; arXiv:2405.18392.

Harvey, N. J. A., Liaw, C., Plan, Y., and Randhawa, S. (2019). Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory (COLT)*, pages 1579–1613.

Hazan, E. (2016). Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3–4):157–325.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., et al. (2022). Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Hu, S., Tu, Y., Han, X., He, C., Cui, G., Long, X., Zheng, Z., Fang, Y., Huang, Y., Zhao, W., Zhang, X., Thai, Z. L., Zhang, K., Wang, C., Yao, Y., Zhao, C., Zhou, J., Cai, J., Zhai, Z., Ding, N., Jia, C., Zeng, G., Li, D., Liu, Z., and Sun, M. (2024). MiniCPM: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*.

Ibrahim, A., Thérien, B., Gupta, K., Richter, M. L., Anthony, Q., Lesort, T., Belilovsky, E., and Rish, I. (2024). Simple and scalable strategies to continually pre-train large language models. *arXiv preprint arXiv:2403.08763*.

Jain, P., Nagaraj, D. M., and Netrapalli, P. (2021). Making the last iterate of SGD information theoretically optimal. *SIAM Journal on Optimization*, 31(2):1108–1130.

Jordan, K. (2024). Muon: An optimizer for hidden layers in neural networks. https://kellerjordan.github.io/posts/muon/.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Ke, Z., Shao, Y., Lin, H., Konishi, T., Kim, G., and Liu, B. (2023). Continual pre-training of language models. In *International Conference on Learning Representations*.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Kornowski, G. and Shamir, O. (2026). Gradient descent's last iterate is often (slightly) suboptimal. *arXiv preprint arXiv:2604.13870*.

Lacoste-Julien, S., Schmidt, M., and Bach, F. (2012). A simpler approach to obtaining an $o(1/t)$ convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002*.

Liu, Z. and Zhou, Z. (2023). Revisiting the last-iterate convergence of stochastic gradient methods. *arXiv preprint arXiv:2312.08531*.

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Meterez, A., Nair, P. A., Morwani, D., Pehlevan, C., and Kakade, S. (2026). Anytime pretraining: Horizon-free learning-rate schedules with weight averaging. *arXiv preprint arXiv:2602.03702*.

Mo, K., Shi, Y., Weng, W., Zhou, Z., Liu, S., Zhang, H., and Zeng, A. (2025). Mid-training of large language models: A survey. *arXiv preprint arXiv:2510.06826*.

OLMo, T., Walsh, P., Soldaini, L., Groeneveld, D., Lo, K., Arora, S., Bhagia, A., Gu, Y., Huang, S., Jordan, M., et al. (2024). 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*.

Parmar, J., Satheesh, S., Patwary, M., Shoeybi, M., and Catanzaro, B. (2024). Reuse, don't retrain: A recipe for continued pretraining of language models. *arXiv preprint arXiv:2407.07263*.

Polyak, B. T. and Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855.

Pun, Y.-M., Buchholz, M., and Gower, R. M. (2025). Schedulers for schedule-free: Theoretically inspired hyperparameters. *arXiv preprint arXiv:2511.07767*.

Rakhlin, A., Shamir, O., and Sridharan, K. (2011). Making gradient descent optimal for strongly convex stochastic optimization. *arXiv preprint arXiv:1109.5647*.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407.

Sanyal, S., Neerkaje, A. T., Kaddour, J., Kumar, A., and Sanghavi, S. (2024). Early weight averaging meets high learning rates for LLM pre-training. In *Conference on Language Modeling (COLM)*. arXiv:2306.03241, 2023.

Schaipp, F., Hägele, A., Taylor, A., Simsekli, U., and Bach, F. (2025). The surprising agreement between convex optimization theory and learning-rate scheduling for large model training. *arXiv preprint arXiv:2501.18965*.

Shamir, O. and Zhang, T. (2013). Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pages 71–79.

Shen, Y., Stallone, M., Mishra, M., Zhang, G., Tan, S., Prasad, A., Soria, A. M., Cox, D. D., and Panda, R. (2024). Power scheduler: A batch size and token number agnostic learning rate scheduler. *arXiv preprint arXiv:2408.13359*.

Singh, V., Janson, P., Mehrbod, P., Ibrahim, A., Rish, I., Belilovsky, E., and Thérien, B. (2025). Beyond cosine decay: On the effectiveness of infinite learning rate schedule for continual pre-training. *arXiv preprint arXiv:2503.02844*.

Soboleva, D., Al-Khateeb, F., Myers, R., Steeves, J. R., Hestness, J., and Dey, N. (2023). SlimPajama: A 627b token cleaned and deduplicated version of RedPajama. https://huggingface.co/datasets/cerebras/SlimPajama-627B. Cerebras Systems.

Tian, C., Wang, J., Zhao, Q., Chen, K., Liu, J., Liu, Z., Mao, J., Zhao, W. X., Zhang, Z., and Zhou, J. (2025). Wsm: decay-free learning rate schedule via checkpoint merging for llm pre-training. *arXiv preprint arXiv:2507.17634*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*.

Wen, K., Li, Z., Wang, J., Hall, D., Liang, P., and Ma, T. (2024). Understanding warmup-stable-decay learning rates: A river valley loss landscape perspective. *arXiv preprint arXiv:2410.05192*.

Wortsman, M., Ilharco, G., Gadre, S. Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A. S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., and Schmidt, L. (2022). Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International Conference on Machine Learning (ICML)*.