
A Trust-region Framework for Moment Estimation

Oluwasegun Somefun

somefuno@oregonstate.edu

Abstract

In this paper, we develop a trust-region framework for understanding the behavior of adaptive moment estimation mechanisms, such as ADAM, in stochastic gradient optimization. Specifically, the magnitude of the update step associated with each individual parameter is constrained by a finite-order p -moment trust-region, with $p \geq 1$. The resulting derivation leads to a family of learning-rate mechanisms based on second-moment estimation and normalized p -th-moment estimation. For $p = 4$, this involves kurtosis estimation. Subsequent derivations provide a unified interpretation of moment-estimation-based normalization, learning-rate scheduling, momentum as a spectral first-order lowpass regularization, and operator-level spectral-norm normalization within a common trust-region framework. Preliminary experiments on GPT2-124M trained on FineWeb-Edu and TinyStories suggest that the fourth-moment realization provides its greatest benefit when trust-region constraints are weak. As progressively stronger trust-region controls are introduced, the second-moment realization becomes increasingly competitive, often achieving slightly lower validation loss than its corresponding fourth-moment realization.

1 Introduction

A natural question arising from the popular adaptive moment estimation (ADAM, [Kingma & Ba \(2015\)](#)) mechanism in stochastic gradient optimization is whether its moment-normalized update components can be understood from a rigorous trust-region principle.

In this paper, we address this question by developing a moment-constrained trust-region control framework for stochastic gradient optimization. The central idea is to characterize update magnitudes through explicit p -moment trust-region constraints. The resulting derivation leads to a family of learning-rate mechanisms for $p \geq 1$, governed by a p -th moment trust-region constraint on the update step. In the special case $p = 4$, the mechanism involves both moment estimation, and implicit estimation of the fourth root of the kurtosis associated with the gradient process. Hence, we refer loosely to realizations of this trust-region framework as GMAKE.

Beyond the derivation of a family of learning-rate mechanisms, the framework reveals that several mechanisms traditionally studied separately can be understood within a common trust-region perspective on the update step. In particular, common learning-rate schedules arise naturally as solutions to a trust-region variational problem, while momentum as a linear iteration-invariant operator and matrix-operator spectral-norm normalization can be viewed as complementary mechanisms for progressively strengthening while preserving enforcement of an underlying trust-region control on the update step process. In this sense, the GMAKE framework unifies moment-constrained trust-region optimization, learning-rate scheduling, moment estimation, momentum, and matrix-operator spectral-norm trust-region control within a common framework.

Here, our goal is not to establish the empirical superiority of a particular optimizer, but to systematically study the role of moment-based trust-region constraints in stochastic gradient methods. The principal contributions of this paper are:

- (i.) A p -moment-based trust-region formulation of the stochastic gradient update step, for $p \geq 1$, leading to a generalized family of normalization mechanisms.
- (ii.) A unified interpretation of adaptive moment estimation, learning-rate scheduling, spectral first-order lowpass filtering as momentum, and matrix-operator spectral-norm normalization as complementary trust-region mechanisms acting on different properties of the update step, while preserving an underlying moment-based trust-region guarantees.

1.1 Related Work

The **GMAKE** mechanism draws connections between several research directions in stochastic gradient optimization, including classic trust-region optimization, adaptive moment estimation, momentum, and matrix orthogonalization.

Adaptive moment estimation. Adaptive learning-rate algorithms such as **RMSPROP** and **ADAM** normalize gradient components via second-moment estimates and has become the most popular optimizer in deep learning (Kingma & Ba, 2015; Bottou et al., 2018). In contrast, **GMAKE** is derived from explicit moment-constrained trust-region considerations. The resulting learning rule extends second-moment normalization to a family of p -moment normalization mechanisms for $p \geq 1$.

Trust-region, normalized-gradient methods, and learning-rate schedules. Trust-region methods regulate update magnitudes by constraining the optimization step to lie within a prescribed neighborhood in which a local model is considered reliable (Parikh & Boyd, 2014; Conn et al., 2000). Related ideas also appear in normalized-gradient methods, where update directions are rescaled to control step magnitudes. Learning-rate schedules (Bergsma et al., 2024; Ge et al., 2018) are also widely used to progressively reduce update magnitudes throughout training, thereby influencing the effective size of the optimization region explored by the algorithm. The formulation in this paper differs in that the trust region is characterized through moment constraints on the update itself, yielding a learning-rate mechanism directly linked to the statistical properties of the underlying gradient process. Furthermore, common learning-rate schedules emerge naturally in this framework as solutions to a trust-region variational control problem rather than as externally specified heuristics.

Momentum and filtering. Momentum methods such as Heavy-Ball and Nesterov acceleration have long been used to improve the stochastic gradient optimization process (Polyak, 2020; Sutskever et al., 2013; Polyak, 1969). More recently, momentum mechanisms have also been interpreted from a lowpass signal-processing perspective (Somefun, 2026; Somefun et al., 2024), where they introduce a smoothing effect on gradient sequences and attenuate rapidly varying gradient components (Li et al., 2025). These viewpoints provide useful intuition for understanding the noise-reduction properties of momentum-based optimization methods. In contrast, **GMAKE** incorporates a passive linear, time-invariant operator that preferentially filters out high-frequency gradient fluctuations. Distinct from previous filtering interpretations, the operator is introduced as a mechanism for strengthening satisfaction of an existing moment-based trust-region constraint.

Matrix-aware and spectral-norm-constrained updates. Recent optimizer developments have emphasized the matrix structure of neural-network parameters, leading to orthogonalized and spectral-normalized update rules (Bernstein & Newhouse, 2024). Many such methods admit interpretations in terms of alternative matrix norms or operator-norm constraints (Large et al., 2024). In contrast, spectral-norm constraints are introduced here as a secondary mechanism for progressively tightening an existing moment-based trust-region framework rather than as the primary update rule. The matrix-operator form of **GMAKE** is related to this literature in that it applies spectral-norm normalization to a matrix-valued layer or parameter group composed of p -moment-normalized gradient components.

1.2 A Trust-region problem

A iteration t , consider a stochastic gradient algorithm applied to a parameter vector $w(t) \in \mathbb{R}^{n'}$ of finite size $n' \geq 1$. For an arbitrary coordinate $i \in [1, n']$, let $w(t) \in \mathbb{R}$ denote the associated parameter $w(t, i)$, and $g(t) \in \mathbb{R}$ the associated gradient component obtained by minimizing a scalar-valued loss function $f(t)$ with respect to $w(t)$. Over a window of learning iterations $t \in [0, \tau]$, where $\tau \gg 1$, denote the update step as

$$\Delta(t+1) = w(t+1) - w(t). \quad (1)$$

Throughout, we use $\mathbb{E}\{\cdot\}$ to denote expectation with respect to the underlying stochastic process generating the gradients. No specific objective function, probabilistic structure, or distribution of the gradients is assumed. We make only the following assumptions for all t :

Assumption 1 (Lipschitz Regularity): The function $f(t)$ is at least twice continuously differentiable, and both $f(t)$ and $g(t)$ are Lipschitz continuous in $w(t)$ (Bottou et al., 2018).

Assumption 2 (Bounded p -th Moment): The stochastic signal $g(t)$ satisfies $\|g(t)\|_\infty < \infty$. Hence, for every finite $p \geq 1$, $\mathbb{E}\{|g(t)|^p\} \leq \|g(t)\|_\infty^p < \infty$. Further, we assume that each iteration of the update step (1) is constructed to at least, ensure $\mathbb{E}\{|\Delta(t+1)|\} \leq \mu$. In other words, the expected magnitude of the update step is bounded within a maximum trust region radius $\mu > 0$. The trust-region radius of the update step is characterized by its p -th moment

$$\mathbb{E}\{|\Delta(t+1)|^p\} = \delta_p^p(t) \leq \mu^p, \quad (2)$$

where μ is also the maximum allowable update step-size, while $\delta_p(t)$ is a trust-region radius shaping function governing variation of the update step, with $\delta_p(t) \leq \mu$, with $\delta_p(t) \rightarrow 0$ as $t \rightarrow \tau$.

1.3 Overview of the Trust-region Framework

The basic form of the learning algorithm, with a variational shaping function $\varsigma(t) \in [0, 1]$, and $\varsigma(t) \rightarrow 0$ as $t \rightarrow \tau$ is

$$\delta_p(t) = \mu \varsigma(t)^{\frac{1}{p}}, \quad (3)$$

$$\tilde{g}(t) = \frac{g(t)}{\mathbb{E}\{|g(t)|^2\}^{\frac{1}{2}}}, \quad (4)$$

$$\bar{g}(t) = \frac{\tilde{g}(t)}{\mathbb{E}\{|\tilde{g}(t)|^p\}^{\frac{1}{p}}}, \quad (5)$$

$$\Delta(t+1) = -\delta_p(t) \bar{g}(t). \quad (6)$$

Next, let $\mathbb{H}: \ell^2 \rightarrow \ell^2$ be a linear, iteration-invariant operator, applied on $g(t)$, with spectral-norm $\|\mathbb{H}\|_\infty \leq 1$. The spectral regularized learning algorithm can be expressed as

$$\delta_p(t) = \mu \varsigma(t)^{\frac{1}{p}}, \quad (7)$$

$$v(t) = \mathbb{H}\{g(t)\}, \quad (8)$$

$$\tilde{g}(t) = \frac{v(t)}{\mathbb{E}\{|g(t)|^2\}^{\frac{1}{2}}}, \quad (9)$$

$$\bar{g}(t) = \frac{\tilde{g}(t)}{\mathbb{E}\{|\tilde{g}(t)|^p\}^{\frac{1}{p}}}, \quad (10)$$

$$\Delta(t+1) = -\delta_p(t) \bar{g}(t). \quad (11)$$

Further, let $\tilde{\mathbf{G}}(t)$ denote a full-rank $n \times m$ matrix composed of normalized gradient components $\bar{g}(t)$ associated with a single layer or parameter group l , where $1 < n \leq m$. Denote the update step in matrix form as $\mathbf{D}(t+1)$. Now, suppose that, in order to strengthen direct enforcement of the individual trust-region constraints $\mathbb{E}\{|\Delta(t+1)|\} \leq \mu$, we additionally introduce a bound on the layer's spectral norm as $\|\mathbf{D}(t+1)\|_2 \leq \mu$. The algorithm then takes the form

$$\delta_p(t) = \mu \varsigma(t)^{\frac{1}{p}}, \quad (12)$$

$$v(t) = \mathbb{H}\{g(t)\}, \quad (13)$$

$$\tilde{g}(t) = \frac{v(t)}{\mathbb{E}\{|g(t)|^2\}^{\frac{1}{2}}}, \quad (14)$$

$$\bar{g}(t) = \frac{\tilde{g}(t)}{\mathbb{E}\{|\tilde{g}(t)|^p\}^{\frac{1}{p}}}, \quad (15)$$

$$\tilde{\mathbf{G}}(t) = (\tilde{\mathbf{G}}(t) \tilde{\mathbf{G}}(t)^\top)^{-\frac{1}{2}} \tilde{\mathbf{G}}(t), \quad (16)$$

$$\mathbf{D}(t+1) = -\delta_p(t) \tilde{\mathbf{G}}(t). \quad (17)$$

1.4 Organization of the Paper

The remainder of this paper is organized as follows. Section 2 derives the basic form of GMAKE, establishes its moment-constrained properties, and connection to learning-rate schedules. Section 3 introduces a trust-region-preserving

spectral lowpass regularization operator. Then, Section 4 develops a matrix-operator form that additionally enforces spectral-norm trust-region constraint. Section 5 notes practical implementation strategies for the required statistical estimators, and Section 6 presents numerical experiments and discussions comparing the **GMAKE** forms. Current limitations and gaps are discussed in Section 7. Finally, in Section 8, this paper is concluded.

2 **GMAKE**

In its basic form, the stochastic gradient update in (1) can be defined as

$$\Delta(t+1) = w(t+1) - w(t) = -\alpha(t) g(t), \quad (18)$$

where $\alpha(t)$ is a non-negative real-valued learning rate function of iteration t . Our goal is to design a learning-rate mechanism $\alpha(t)$ that ensures the $\Delta(t+1)$ satisfies the trust-region problem involving its p -th moment $\mathbb{E}\{|\Delta(t+1)|^p\} \leq \mu^p$, so that $\mathbb{E}\{|\Delta(t+1)|\} \leq \mu$.

2.1 Learning-rate mechanism

Recall that $\delta_p^p(t) = \mathbb{E}\{|\Delta(t+1)|^p\}$. Substituting the update step into the trust-region constraint yields

$$\mathbb{E}\{|\Delta(t+1)|^p\} = \mathbb{E}\{|-\alpha(t) g(t)|^p\} = \alpha^p(t) \mathbb{E}\{|g(t)|^p\} = \delta_p^p(t). \quad (19)$$

Solving for $\alpha(t)$ gives the learning-rate mechanism

$$\alpha(t) = \frac{\delta_p(t)}{\mathbb{E}\{|g(t)|^p\}^{\frac{1}{p}}}. \quad (20)$$

Define $\tilde{g}(t) = g(t)/\mathbb{E}\{|g(t)|^2\}^{\frac{1}{2}}$ as the root-mean-square (rms) normalized form of the input signal $g(t)$, then it follows that $\mathbb{E}\{|g(t)|^p\}^{\frac{1}{p}} = \mathbb{E}\{|\tilde{g}(t)|^p\}^{\frac{1}{p}} \mathbb{E}\{|g(t)|^2\}^{\frac{1}{2}}$. The learning rate (20) can be equivalently expressed as

$$\alpha(t) = \frac{\delta_p(t)}{\mathbb{E}\{|\tilde{g}(t)|^p\}^{\frac{1}{p}} \mathbb{E}\{|g(t)|^2\}^{\frac{1}{2}}}, \quad (21)$$

and the update step (19) becomes

$$\Delta(t+1) = -\delta_p(t) \frac{\tilde{g}(t)}{\mathbb{E}\{|\tilde{g}(t)|^p\}^{\frac{1}{p}}}. \quad (22)$$

In particular, for $p = 4$, the learning rate function has a meaningful interpretation of being composed of terms related to the second-moment and the fourth-root of the kurtosis of $g(t)$. Next, denote $\bar{g}(t) = \tilde{g}(t)/\mathbb{E}\{|\tilde{g}(t)|^p\}^{\frac{1}{p}}$, then the update step (21) can be expressed as $\Delta(t+1) = -\delta_p(t) \bar{g}(t)$, which enforces $\mathbb{E}\{|\bar{g}(t)|^p\} = 1$, and because $\delta_p^p(t) \leq \mu^p$, yields $\mathbb{E}\{|\Delta(t+1)|^p\} \leq \mu^p$. As a result, for $p \geq 1$, Lyapunov's inequality implies

$$\mathbb{E}\{|\Delta(t+1)|\} \leq \mathbb{E}\{|\Delta(t+1)|^p\}^{\frac{1}{p}} \leq \mu. \quad (23)$$

The update step $\Delta(t+1) = -\delta_p(t) \bar{g}(t)$ then directly satisfies both trust-region constraints on the p -th moment and expected maximum magnitude. The resulting update algorithm sequentially follows the outlined equations (3)–(6). Further, let $\mathbf{d}(t+1) \in \mathbb{R}^{n'}$ denote the corresponding update step vector for which $\Delta(t+1)$ is an i -th component. Since the p -moment normalization acts independently on each parameter, the trust-region constraint (23) implies that,

$$\mathbb{E}\|\mathbf{d}(t+1)\|_r^r = \sum_{i=1}^{n'} \mathbb{E}\{|\mathbf{d}(t+1, i)|^r\} \leq \sum_{i=1}^{n'} \mu^r = n' \mu^r \quad (24)$$

for any $1 \leq r \leq p$. Since $0 < \mu < 1$ and $r \geq 1$, it also follows that $\mu^r \leq \mu$, yielding a looser upper bound $\mathbb{E}\|\mathbf{d}(t+1)\|_r^r \leq n' \mu$. Therefore, the p -th moment-based trust-region constraint directly controls the expected magnitude of the vectorized update step in every ℓ_r norm, where $1 \leq r \leq p$, and in particular implies $\mathbb{E}\|\mathbf{d}(t+1)\|_p^p \leq n' \mu^p$.

For convenience, denote $\kappa_p(t) = \mathbb{E}\{|\tilde{g}(t)|^p\}^{\frac{1}{p}}$, $\lambda_2(t) = \mathbb{E}\{|g(t)|^2\}^{\frac{1}{2}}$, and note that $\kappa_p(t) \geq 1$, we can re-write the underlying learning-rate mechanism (21) as

$$\alpha(t) = \frac{\delta_p(t)}{\kappa_p(t) \lambda_2(t)} \leq \frac{\mu}{\lambda_2(t)}, \quad (25)$$

The overall algorithm in (3) – (6) can be interpreted as using $\lambda_2(t)$ to first normalize the variance scale of $g(t)$, and then applying $\kappa_p(t)$ to yield the fully normalized $\tilde{g}(t)$, whose p -th moment satisfies the trust-region constraint. Ideally, the special case, $p = 2$ yields $\kappa_2(t) = 1$, and (25) reduces to the same mechanism used in RMSProp and ADAM, where $g(t)$ is normalized solely by $\lambda_2(t)$. However, during estimation, when $\kappa_2(t) \neq 1$, the algorithm applies an additional normalization by $\kappa_2(t)$ to compensate for scaling errors that remain after the $\lambda_2(t)$ normalization. As a result, this approach can be interpreted as a principled refinement of the normalization mechanism employed by both RMSProp and ADAM.

Also, for $p \neq 2$, when the normalization factor $\kappa_p(t)$ remains close to its minimum value of one, the resulting learning-rate mechanism becomes increasingly similar to the $p = 2$ case. Conversely, when higher-order gradient fluctuations is more pronounced and $\kappa_p(t)$ departs further from one, the additional normalization acts as a stronger trust-region control mechanism.

2.2 Learning-rate Schedules as Trust-region shaping functions

The trust-region problem requires that $\delta_p^p(t) \leq \mu^p$ for all t , and that $\delta_p^p(t) \rightarrow 0$ as $t \rightarrow \tau$. This can be interpreted as progressively tightening the trust-region constraints over the learning window to enforce asymptotic stability of the update step. In order to obtain the smoothest possible variation of the trust-region radius across iterations over the learning horizon $t \in [0, \tau]$, we can define a total variational energy functional minimization problem directly on $\delta_p^p(t)$,

$$\min_{0 \leq \delta_p^p(t) \leq \mu^p} \mathcal{D}(\delta_p, \tau) = \sum_{t=1}^{\tau} [\delta_p^p(t) - \delta_p^p(t-1)]^2, \quad (26)$$

subject to fixed boundary values: the terminal boundary value $\delta_p^p(\tau) = 0$, together with at least one specified initial boundary value satisfying $\delta_p^p(0) \leq \mu^p$. The energy minimization problem (26) represents the canonical first-order smoothness functional that penalize rapid variations in $\delta_p^p(t)$, the trust-region radius between successive iterations. Equivalently, (26) encourages $\delta_p^p(t)$ to remain as close as possible to $\delta_p^p(t-1)$ throughout learning. Consequently, $\delta_p(t)$ varies no faster than necessary while remaining consistent with the asymptotic stability requirement, thereby producing the smoothest trust-region variation consistent with the boundary values.

Furthermore, (26) is strictly convex in $\delta_p^p(t)$, and therefore has a unique minimizer. Differentiating with respect to any interior index $i \in \{1, \dots, \tau-1\}$, and setting $d\mathcal{D}/d\delta_p^p(i) = 0$ results in the second-order homogeneous linear difference equation with constant coefficients

$$\delta_p^p(i+1) - 2\delta_p^p(i) + \delta_p^p(i-1) = 0. \quad (27)$$

Let $\Delta \delta_p^p(i+1) = \delta_p^p(i+1) - \delta_p^p(i)$, then (27) corresponds to the second-order finite difference operator $\Delta(\Delta \delta_p^p(i+1)) = \Delta^2 \delta_p^p(i+1) = 0$. The corresponding characteristic equation is $z^2 - 2z + 1 = (z-1)^2 = 0$ which has 1 as a root with multiplicity 2. Therefore, the general solution to (27) must be a polynomial in t , of at most degree 1 (Elaydi, 2005; Lueker, 1980), defined as

$$\delta_p^p(t) = a_0 + a_1 t. \quad (28)$$

The complete solution can then be obtained by plugging in at least $b \geq 2$ fixed boundary values: the fixed terminal value, and at least one fixed initial value, leading to $b-1$ sub-interval(s) of the learning window.

2.3 2-point boundary value

Consider the initial and terminal boundary values: $\delta_p^p(0) = \mu^p$, $\delta_p^p(\tau) = 0$ over the single sub-interval $0 \leq t \leq \tau$. Solving (28), subject to $\delta_p^p(0) = \mu^p$, $\delta_p^p(\tau) = 0$ gives $a_0 = \mu^p$, and $a_1 = -\mu^p/\tau$. Therefore, the complete solution to

(27) subject to the 2-point boundary value is

$$\delta_p^p(t) = \mu^p \left(1 - \frac{t}{\tau}\right), \quad 0 \leq t \leq \tau, \quad (29)$$

which is known as the linear decay schedule. We can simplify this to a normalized functional form. Let $u = t/\tau$, define an input function $x(u) := u \in [0, 1]$, then

$$\delta_p^p(t) = \mu^p (1 - x(u)), \quad 0 \leq u \leq 1. \quad (30)$$

2.4 3-point boundary value

Consider the three fixed boundary values: $\delta_p^p(0) = 0$, $\delta_p^p(k) = \mu^p$, $\delta_p^p(\tau) = 0$, where $0 \leq k < \tau$. This corresponds to two sub-intervals. On the first interval over $0 \leq t \leq k$, the solution is obtained by solving (28) subject to $\delta_p^p(0) = 0$, $\delta_p^p(k) = \mu^p$. On the second interval over $k \leq t \leq \tau$, the solution is obtained by solving (28) subject to $\delta_p^p(k) = \mu^p$, $\delta_p^p(\tau) = 0$. Combining the piecewise solutions, the complete solution to (27) subject to the 3-point boundary value is

$$\delta_p^p(t) = \begin{cases} \mu^p \frac{t}{k}, & 0 \leq t \leq k, \quad k > 0 \\ \mu^p \frac{\tau-t}{\tau-k}, & k \leq t \leq \tau, \quad k \geq 0. \end{cases} \quad (31)$$

Simplifying further, define $u = t/\tau \in [0, 1]$, a normalized rise-time $m = k/\tau \in [0, 1]$, and the reparameterized input function

$$x(u; m) = \begin{cases} u, & m = 0, \\ \max\left\{\frac{m-u}{m}, \frac{u-m}{1-m}\right\}, & 0 < m < 1, \end{cases} \quad (32)$$

then equivalently, (31) is

$$\delta_p^p(t) = \mu^p (1 - x(u; m)), \quad 0 \leq u \leq 1, \quad 0 \leq m < 1. \quad (33)$$

For its first sub-interval, the hat-shaped (or triangular-like) mapping (31) corresponds to a warmup schedule that reaches μ^p at time m , and then a decaying schedule to zero in the second sub-interval. Importantly, observe that both (30) and (33) share the same underlying functional form. When $m = 0$, it follows that $x(u; m) = x(u; 0) = u$, therefore (30) is a special case of (33) when $m = 0$.

2.5 4-point boundary value

Consider the four fixed boundary values: $\delta_p^p(0) = 0$, $\delta_p^p(k) = \mu^p$, $\delta_p^p(k') = \mu^p$, $\delta_p^p(\tau) = 0$, where $0 \leq k \leq k' < \tau$. This corresponds to three sub-intervals. On the first interval over $0 \leq t \leq k$, the solution is obtained by solving (28) subject to $\delta_p^p(0) = 0$, $\delta_p^p(k) = \mu^p$. On the second interval over $k \leq t \leq k'$, the solution is obtained by solving (28) subject to $\delta_p^p(k) = \mu^p$, $\delta_p^p(k') = \mu^p$. Finally, for the third interval over $k' \leq t \leq \tau$, the solution is obtained by solving (28) subject to $\delta_p^p(k') = \mu^p$, $\delta_p^p(\tau) = 0$. Combining the three piecewise solutions, the complete solution to (27) subject to the 4-point boundary value, when simplified is

$$\delta_p^p(t) = \begin{cases} \mu^p \frac{t}{k}, & 0 \leq t \leq k, \quad k > 0 \\ \mu^p, & k \leq t \leq k', \quad k, k' \geq 0, \\ \mu^p \frac{\tau-t}{\tau-k'}, & k' \leq t \leq \tau, \quad k' \geq 0. \end{cases} \quad (34)$$

Define $u = t/\tau \in [0, 1]$, and let $m = k/\tau \in [0, 1)$ indicate the rise time and subsequently, the start of a plateau of width $\varepsilon = (k' - k)/\tau \in [0, 1)$, so that $k = m\tau$, and $k' = (m + \varepsilon)\tau$. Then the reparameterized input function $x(r; m, \varepsilon) \in [0, 1]$ for (34) is

$$x(u; m) := \begin{cases} \max\left\{0, \frac{u-\varepsilon}{1-\varepsilon}\right\}, & m = 0, \\ \max\left\{\frac{m-u}{m}, 0, \frac{u-(m+\varepsilon)}{1-(m+\varepsilon)}\right\}, & 0 < m < 1, \end{cases} \quad (35)$$

and equivalently (34) can be expressed as

$$\delta_p^p(t) = \mu^p (1 - x(u; m, \varepsilon)), \quad 0 \leq u \leq 1, \quad 0 \leq m < 1. \quad (36)$$

The resulting function (34) then corresponds in the first sub-interval to a warmup schedule that rises to μ^p , and then a plateau in the second sub-interval, followed by a decaying schedule to zero in the third sub-interval. This leads a trapezoidal shape, and exactly a warmup-stable-decay schedule.

For $m = 0, \varepsilon = 0$, it follows that $x(u; m, \varepsilon) = x(u; 0, 0) = u$, therefore (30) is a special case of (36). Similarly, when $m \neq 0, \varepsilon = 0$, it follows that $x(u; m, \varepsilon) = x(u; m, 0) = x(u; m)$, therefore (33) is a special case of (36). Importantly, both (30), (33) and (36) share the same underlying functional form, with (36) having a more general input reparameterization function.

Finally, denote $\delta_p^p(t) = \mu^p \varsigma(t)$, where $\varsigma(t) = 1 - x(u; m, \varepsilon)$, then

$$\delta_p(t) = \mu \varsigma^{\frac{1}{p}}(t) \quad (37)$$

is the unique minimizer of the first-order total variational energy functional, which satisfies $\delta_p(t) \leq \mu$, $\delta_p(t) \rightarrow 0$ as $t \rightarrow \tau$ subject to $2 \leq b \leq 4$ fixed boundary points.

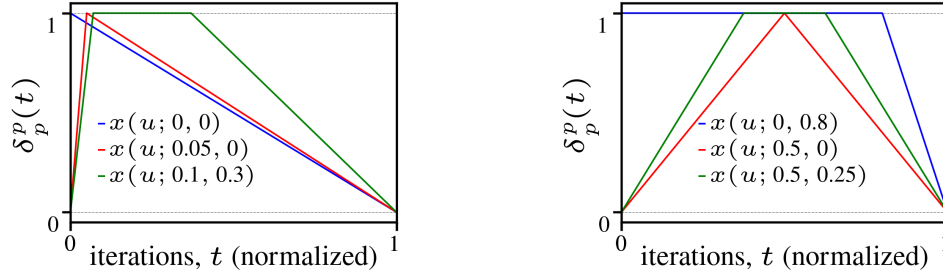


Figure 1: Different reparameterizations of the boundary-value problem on $\delta_p^p(t)$, with $\mu = 1$ lead to the shapes of common learning-rate schedules, sharing the same linear mapping $1 - x(u; m, \varepsilon)$.

Remarks: The preceding analysis shows that many commonly used learning-rate schedules arise naturally as solutions to a trust-region variational problem. By treating $\delta_p^p(t)$ as a varying trust-region radius and minimizing its total variation subject to fixed boundary values, we obtain schedules that evolve as smoothly as possible, while still satisfying the asymptotic requirement $\delta(t) \rightarrow 0$ as $t \rightarrow \tau$.

The resulting p -th root linear decay, warmup-decay, and warmup-stable-decay schedules are illustrated in Figure 1. They correspond respectively to 2-point, 3-point, and 4-point boundary-value problems, yet all share a common variational structure. Under this interpretation, schedule design is no longer an independent heuristic, but rather a principled consequence of solving a first-order variational optimization problem subject to trust-region boundary constraints. More generally, alternative variational functionals would produce different classes of schedules (Somefun, 2026), each reflecting the structure of the underlying variational principle. Since $\delta_p(t)$ directly controls the allowable update magnitude, learning-rate scheduling serves as a complementary mechanism through which the trust-region constraint of the update step is gradually tightened over the learning horizon via a variational shaping function.

2.6 Underlying Taylor-series Model

The role of the p -moment trust-region constraint can also be viewed from the perspective of a Taylor-series local approximation model. Let the update step (19) generated by (21) satisfy the p -moment trust-region constraint

$$\mathbb{E}\{|\Delta(t+1)|^p\} = \delta_p^p(t) \leq \mu^p, \quad (38)$$

where $\delta_p(t) \rightarrow 0$ as $t \rightarrow \tau$. For $p > r > 0$, Lyapunov's inequality implies

$$\mathbb{E}\{|\Delta(t+1)|^r\}^{\frac{1}{r}} \leq \mathbb{E}\{|\Delta(t+1)|^p\}^{\frac{1}{p}} = \delta_p(t). \quad (39)$$

Consequently, all lower moments up to order p are simultaneously controlled,

$$\mathbb{E}\{|\Delta(t+1)|^r\} \leq \delta_p^r(t), \quad r \leq p. \quad (40)$$

In particular, since $\Delta(t+1) = -\frac{\delta_p(t)}{\kappa_p(t)} \bar{g}(t)$, $\mathbb{E}\{|\bar{g}(t)|^2\} = 1$ and $\kappa_p(t) \geq 1$, it follows that

$$\mathbb{E}\{|\Delta(t+1)|^2\} = \frac{\delta_p^2(t)}{\kappa_p^2(t)} \leq \delta_p^2(t). \quad (41)$$

Therefore, a p -moment trust-region constraint would control every moment entering a Taylor-series expansion up to degree p . Consider the update $w(t+1) = w(t) + \Delta(t+1)$. Let $g(t)$ and $h(t)$ be the Lipschitz continuous first and second derivatives of $f(t)$ with respect to $w(t)$. By Taylor's theorem, the second-order Taylor model associated with the expected loss decrease can be expressed as

$$\mathbb{E}\{\Delta f(t+1)\} = \mathbb{E}\{g(t)\Delta(t+1)\} + \frac{1}{2} \mathbb{E}\{h(t)\Delta^2(t+1)\} + O\left(\mathbb{E}\{|\Delta(t+1)|^3\}\right), \quad (42)$$

where $\Delta f(t+1) = f(t+1) - f(t)$, and (42) follows directly from Taylor's remainder theorem. Thus, there exists a constant $c > 0$ such that the third-order remainder satisfies $|R(3)| \leq c|\Delta(t+1)|^3$. Consequently, $|E\{R(3)\}| \leq E\{|R(3)|\} \leq c E\{|\Delta(t+1)|^3\}$, and therefore $E\{R(3)\} = O(E\{|\Delta(t+1)|^3\})$.

Denote $m(t) = \mathbb{E}\{h(t)\bar{g}(t)^2\}$. Applying the moment bounds on $\Delta(t+1)$, by using (40) and (41) in (42) yields

$$\mathbb{E}\{\Delta f(t+1)\} \leq -\frac{\delta_p(t)}{\kappa_p(t)} \lambda_2(t) + \frac{1}{2} \frac{\delta_p^2(t)}{\kappa_p^2(t)} m(t) + O(\delta_p^3(t)). \quad (43)$$

Observe that the second-order Taylor model contains $\Delta(t+1)$, $\Delta^2(t+1)$, and a leading neglected remainder of order $\Delta^3(t+1)$. To simultaneously control the first-order term, second-order term, and cubic remainder, it is sufficient that $p \geq 3$. The choice $p = 4$ is also attractive because it is the smallest moment order that simultaneously controls the variance and tail heaviness behavior of the update process. Although larger values $p > 4$ provide progressively stronger higher-order moment control, they do not fundamentally improve the underlying second-order Taylor model. Moreover, the benefits of controlling moments beyond order four often diminish faster than the difficulty of reliably estimating them in practice (Ramachandran et al., 2025; Jondeau & Rockinger, 2006).

More importantly, the distinction between a fixed and a vanishing trust-region constraint can also be noted. If $\delta_p(t) = \mu$, we have $\mathbb{E}\{|\Delta(t+1)|^p\} \leq \mu^p$, then all lower moments remain bounded, and the Taylor approximation error remains controlled but generally non-vanishing. In contrast, the vanishing trust-region schedule $\delta_p(t) \rightarrow 0$ implies

$$\mathbb{E}\{|\Delta(t+1)|^r\} \rightarrow 0, \quad r \leq p, \quad (44)$$

causing every lower-order moment of the update step to vanish. As a result, $O(\delta_p^2(t))/O(\delta_p(t)) \rightarrow 0$, $O(\delta_p^3(t))/O(\delta_p(t)) \rightarrow 0$, so that both the second-order correction and higher-order Taylor remainder become asymptotically negligible relative to the leading first-order term. Consequently,

$$\mathbb{E}\{\Delta f(t+1)\} \leq -\frac{\delta_p(t)}{\kappa_p(t)} \lambda_2(t) + O(\delta_p^2(t)), \quad (45)$$

and the local Taylor model becomes progressively more accurate as learning proceeds.

Remarks: Therefore, the role of the vanishing p -moment trust-region constraint extends beyond merely controlling update magnitudes. By driving all lower moments of the update step to vanish, it simultaneously contracts the trust region and enforces asymptotic validity of the local Taylor-series model used to describe the expected decrease in the loss. As learning proceeds and $\delta_p(t) \rightarrow 0$, the expected loss decrease becomes increasingly well described by the first-order term, while higher-order terms become asymptotically negligible.

3 Trust-Region-Preserving Spectral Regularization

Denote $v(t) = \mathbb{H}\{g(t)\}$, so that the stochastic gradient update in (1) becomes

$$\Delta(t+1) = w(t+1) - w(t) = -\alpha(t) v(t), \quad (46)$$

$$\alpha(t) = \frac{\delta_p(t)}{\mathbb{E}\{|v(t)|^p\}^{\frac{1}{p}}}. \quad (47)$$

Let $\mathbb{H}: \ell^2 \rightarrow \ell^2$ be a linear iteration-invariant first-order filter characterized by a pole location $\beta \in \mathbb{R}$ and zero location $\gamma \in \mathbb{R}$, with transfer function

$$\mathbb{H}(z) = \eta \frac{1 - \gamma z^{-1}}{1 - \beta z^{-1}}, \quad \eta = \frac{1 - \beta}{1 - \gamma}, \quad (48)$$

where $z = e^{j\omega}$ over frequencies $\omega \in [-\pi, \pi]$, and $\mathbb{H}(1) = 1$. The squared magnitude, frequency response is

$$|\mathbb{H}(e^{j\omega})|^2 = \frac{(1 - \beta)^2}{(1 - \gamma)^2} \frac{1 + \gamma^2 - 2\gamma \cos \omega}{1 + \beta^2 - 2\beta \cos \omega}. \quad (49)$$

To show that the system's worst-case gain, its $\mathcal{H}_\infty$ norm (Boyd & Barratt, 1991) defined as $\|\mathbb{H}\|_\infty = \max_\omega \|\mathbb{H}(e^{j\omega})\|_2 = \max_\omega |\mathbb{H}(e^{j\omega})| \leq 1$, $\forall \omega \in [-\pi, \pi]$, it suffices to show conditions for which the denominator of (49) is always greater than or equal to its numerator. Rearranging terms, this implies

$$\begin{aligned} \Delta(\omega) &= (1 - \gamma)^2(1 + \beta^2 - 2\beta \cos \omega) - (1 - \beta)^2(1 + \gamma^2 - 2\gamma \cos \omega) \geq 0 \\ &= 2(\beta - \gamma)(1 - \beta\gamma)(1 - \cos \omega) \geq 0. \end{aligned} \quad (50)$$

Note that in (50), the following holds: $1 - \cos \omega \geq 0$ is non-negative for all ω ; factor $\beta - \gamma \geq 0$ if $|\gamma| < |\beta|$; and factor $1 - \beta\gamma > 0$ if $|\beta| < 1$ and $|\gamma| < 1$. Therefore, a sufficient condition for the system's $\|\mathbb{H}\|_\infty \leq 1$ is

$$0 \leq \beta < 1, \quad |\gamma| < |\beta|. \quad (51)$$

This design condition ensures strong attenuation of high-frequency components in $g(t)$, and ensures that the linear operator is non-expansive in its induced ℓ_2 norm. By definition of the induced norm (Boyd & Barratt, 1991), this implies $\|v(t)\|_2 \leq \|g(t)\|_2$, then, $v(t)$ denotes an output sequence, whose total finite energy can be no larger than that of its input $g(t)$.

Also, by the convolution property of the linear time-invariant operator, $v(t) = \sum_{k=0}^t \mathbf{h}(k) g(t-k)$ holds, where $\mathbf{h}(k)$ is the impulse response of $\mathbb{H}$ (Oppenheim & Schaffer, 2010) defined as $\mathbf{h}(0) = \eta$, $\mathbf{h}(k) = \eta(\beta - \gamma)\beta^{k-1}$, and due to (51) $\mathbf{h}(k) \geq 0$ for all $k \geq 0$ is satisfied. It follows that $\|\mathbf{h}\|_1 = 1$, and that $\|v(t)\|_\infty \leq \|g(t)\|_\infty$. Therefore, by definition, the linear-time invariant operator $\mathbb{H}$ is also non-expansive in its induced ℓ_∞ norm.

Using Assumption 2, if $\mathbb{E}\{|\Delta(t+1)|^p\} = \alpha^p(t) \mathbb{E}\{|g(t)|^p\} \leq \alpha^p(t) \|g(t)\|_\infty^p \leq \delta_p^p(t)$ for the unfiltered update step, then after the spectral regularization,

$$\mathbb{E}\{|\Delta(t+1)|^p\} = \alpha^p(t) \mathbb{E}\{|v(t)|^p\} \leq \alpha^p(t) \|v(t)\|_\infty^p \leq \alpha^p(t) \|g(t)\|_\infty^p \leq \delta_p^p(t). \quad (52)$$

Consequently, all trust-region guarantees regarding the allowable update magnitude, inherited from the underlying p -moment mechanism are preserved and may become strictly tighter after regularization. Therefore, $\mathbb{H}$ does not introduce an additional trust-region constraint. Rather, it improves the quality of the gradient signal used to generate the update step by smoothing high-frequency gradient fluctuations before the learning-rate mechanism is applied. We will refer to this effect providing strong attenuation of high-frequency spectral components in $g(t)$, while having a spectral norm less than one at all frequencies $\omega \in [-\pi, \pi]$, as a trust-region-preserving spectral lowpass regularization of $g(t)$.

3.1 Variance reduction factor

A useful measure of the output variance is the system $\mathcal{H}_2$ norm, its average energy over all frequencies $\omega \in [-\pi, \pi]$,

$$\|\mathbb{H}\|_2^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} \|\mathbb{H}(e^{j\omega})\|_F^2 d\omega = \frac{1}{2\pi} \int_{-\pi}^{\pi} |\mathbb{H}(e^{j\omega})|^2 d\omega = \frac{1 - \beta}{1 + \beta} \frac{1 + \gamma^2 - 2\beta\gamma}{(1 - \gamma)^2} \quad (53)$$

In particular, for a white noise input signal, this quantifies the residual noise variance after filtering (Orfanidis, 1995; Boyd & Barratt, 1991). Therefore, minimizing (53) can be interpreted as maximizing variance reduction. Given

$0 \leq \beta < 1$, (53) can be used to place γ . Differentiating with respect to γ , $\frac{d}{d\gamma} \|\mathbb{H}\|_2^2$ is strictly positive, and increasing over the admissible range $-\beta < \gamma < \beta$. Consequently, the minimum variance placement for the zero location γ is

$$\boxed{\gamma_{\text{MVR}} = -\beta}. \quad (54)$$

Substituting $\gamma = -\beta$ into $\eta = \frac{1-\beta}{1-\gamma}$, yields $\eta_{\text{MVR}} = \frac{1-\beta}{1+\beta}$. As $\beta \rightarrow 1$, $\eta_{\text{MVR}} = o(1 - \beta)$. Note that if $\eta = 1 - \beta$, then $\gamma = 0$. Either way, maximum variance reduction is obtained at the cost of an increasingly small η . The normalization gain η governs how the filter, realized in (59) responds to changes in its input signal. In this case, the input signal $g(t)$ carries information about the local optimization landscape. However as $\eta \rightarrow 0$, maximal variance reduction, smoothing effect is increased, leading to less responsiveness to meaningful changes in $g(t)$. The reverse is the case as $\eta \rightarrow 1$. Consequently, in this setting, maximizing variance reduction alone is not an efficient filter design in terms of simultaneously reducing variance while preserving ability to track changes in its input signal.

3.2 Variance reduction per unit gain

A more efficient design tradeoff is to balance the variance measure relative to its responsiveness measure, defined as

$$J(\gamma) = \frac{\|\mathbb{H}\|_2^2}{\eta} = \frac{1 + \gamma^2 - 2\beta\gamma}{(1 + \beta)(1 - \gamma)}. \quad (55)$$

Minimizing (55) is equivalent to maximizing the filter's variance reduction per unit gain. Differentiating with respect to γ yields $(1 - \gamma)^2 = 2(1 - \beta)$, whose feasible root is $\gamma = 1 - \sqrt{2(1 - \beta)}$. Subject to $-\beta \leq \gamma \leq \beta$, the optimal solution becomes

$$\boxed{\gamma_{\text{VRG}} = \max \left\{ -\beta, 1 - \sqrt{2(1 - \beta)} \right\}}. \quad (56)$$

The transition occurs when $-\beta = 1 - \sqrt{2(1 - \beta)}$, which yields a threshold $\beta = \sqrt{5} - 2 \approx 0.236$ related to the golden ratio (Pronzato et al., 2013). Therefore, for small $0 \leq \beta < \sqrt{5} - 2$, the variance-per-unit-gain optimum $\gamma_{\text{VRG}} = -\beta$ coincides with the maximum variance reduction choice. Otherwise, as $\beta \rightarrow 1$, for $\sqrt{5} - 2 \leq \beta < 1$, the variance-per-unit-gain optimum increasingly favors preservation of responsiveness, $\gamma_{\text{VRG}} = 1 - \sqrt{2(1 - \beta)}$.

3.3 Canonical Filter Realization

A direct difference equation realization of the first-order filter (48) is

$$v(t) = \beta v(t-1) + \eta (g(t) - \gamma g(t-1)). \quad (57)$$

In general, (57) admits the controllable canonical realization (Oppenheim & Schaffer, 2010; Smith, 2007),

$$\begin{aligned} q(t) &= \beta q(t-1) + g(t), \\ v(t) &= \eta (q(t) - \gamma q(t-1)), \end{aligned} \quad (58)$$

where $q(t)$ is the state generated during realization of the filter. Additional refinements, such as dividing the filter output by $1 - \beta^t$, compensate for the transient-response bias of the state during the initial iterations.

To make the lowpass regularization of $g(t)$ explicit, define $\tilde{q}(t) = (1 - \beta) q(t)$. Since $\eta = \frac{1-\beta}{1-\gamma}$ and $1 - \eta = \frac{\beta-\gamma}{1-\gamma}$, it follows that $\eta(\beta - \gamma)q(t-1) = (1 - \eta)\tilde{q}(t-1)$. Substituting into (58) yields

$$\begin{aligned} \tilde{q}(t) &= \beta \tilde{q}(t-1) + (1 - \beta) g(t), & e(t) &= \beta e(t-1) + (g(t-1) - g(t)), \\ v(t) &= \eta g(t) + (1 - \eta) \tilde{q}(t-1), & v(t) &= g(t) + (1 - \eta) e(t), \end{aligned} \quad (59)$$

where the right-hand-side representation follows from the change of variables $e(t) = \tilde{q}(t-1) - g(t)$.

3.4 Beyond Variance Reduction

From (59), recall $e(t) = \beta e(t-1) + (g(t-1) - g(t))$. Using the unit-delay operator z^{-1} , and defining the first-order finite-difference operator $\Delta = 1 - z^{-1}$, we have that $(g(t) - g(t-1)) = (1 - z^{-1})g(t) = \Delta g(t)$. The $e(t)$ state recursion with zero initialization $e(0) = 0$, and repeated substitutions can then be expressed as

$$e(t) = - \sum_{k=0}^{t-1} \beta^k z^{-k} (1 - z^{-1})g(t) = - \sum_{k=0}^{t-1} \beta^k z^{-k} \Delta g(t), \quad (60)$$

and hence $v(t) = g(t) + (1 - \eta) e(t)$ is

$$v(t) = g(t) - (1 - \eta) \sum_{k=0}^{t-1} \beta^k z^{-k} \Delta g(t). \quad (61)$$

Equation (60) is an exponentially weighted sum of the delayed first-order finite differences of $g(t)$. Each term $z^{-k}(1 - z^{-1})g(t)$, is generated by delayed applications of the first-order finite-difference operator $(1 - z^{-1})$ on $g(t)$, where $z^{-k}g(t) = g(t-k)$, and $z^{-k}(1 - z^{-1})g(t) = (1 - z^{-1})g(t-k)$. Hence, (61) can be interpreted as the gradient component $g(t)$ being regularized through a correction term formed from an exponentially weighted sum of its delayed first-order finite differences.

Consequently, recent ($0 \leq k \ll t-1$) rapidly varying terms $z^{-k} \Delta g(t)$ produce larger correction terms, whereas recent slowly varying or constant terms produce smaller corrections. Therefore, beyond typical noise variance reduction, the lowpass filter can be viewed as performing a *regularization* of the gradient sequence by selectively penalizing large iteration-to-iteration changes in each gradient component $g(t)$.

3.5 Heavy-ball and Nesterov momentum

Although, the design choice (56), makes the linear operator to efficiently balance variance reduction with responsiveness, the same filter admits other closely related configurations. More generally, given $0 \leq \beta < 1$, by appropriate choices of γ , the linear operator can be used to recover both Heavy-ball and Nesterov momentum. Algebraically, analyzing (58), Heavy-ball momentum can be recovered by selecting $\gamma_{\text{HB}} = 0$,

$$v(t) = \beta v(t-1) + (1 - \beta) g(t), \quad (62)$$

while Nesterov momentum corresponds to $\gamma_{\text{NAG}} = \frac{\beta}{1+\beta}$,

$$\begin{aligned} q(t) &= \beta q(t-1) + (1 - \beta) g(t) \\ v(t) &= \beta q(t) + (1 - \beta) g(t). \end{aligned} \quad (63)$$

Given $0 \leq \beta < 1$, four special operating points of the first-order lowpass filter are therefore

$$\boxed{\gamma_{\text{MVR}} = -\beta, \quad \gamma_{\text{HB}} = 0, \quad \gamma_{\text{NAG}} = \frac{\beta}{1+\beta}, \quad \gamma_{\text{VRG}} = \max\{-\beta, 1 - \sqrt{2(1-\beta)}\}. \quad (64)}$$

More specifically, within the same trust-region-preserving first-order filter family, both Heavy-ball and Nesterov momentum can be viewed as specific operating choices of γ . Recall that this family satisfies the spectral norm less than one design condition (51). Generally, lowpass filter design involves making tradeoffs among several competing effects (Orfanidis, 1995). Maximum variance reduction causes a simultaneous maximum reduction in the filter's normalization gain. While this improves noise suppression, it also effectively reduces the maximum step-size of the update process, which can potentially lead to slower adaptation and less responsiveness to the underlying optimization dynamics. As shown in Figure 2, Nesterov momentum achieves more variance reduction per unit gain than Heavy-ball for $\beta \gtrsim 0.71$. Overall, this unified spectral regularization interpretation of classic momentum methods through the pole-zero locations of a first-order lowpass filter, also corresponds to design tradeoffs between maximal variance reduction and responsiveness of the filtering dynamics to the local optimization landscape. Outlined in (7)–(11) is the trust-region-preserving regularized algorithm, with (8) being the only addition to the basic algorithm.

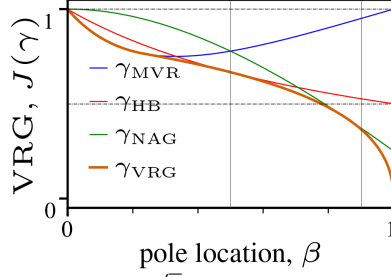


Figure 2: Variance reduction per unit gain. For $\beta > \sqrt{5} - 2$, γ_{MVR} clearly becomes inefficient. Observe that γ_{NAG} closely matches γ_{VRG} as $\beta \gg 0.5$, especially $\beta \approx 0.8 - 0.9$, whereas γ_{HB} closely matches γ_{VRG} for $\beta \approx 0.4 - 0.6$. At the cross-over point $\beta \approx 0.71$, γ_{NAG} overtakes γ_{HB} as an efficient design.

4 Matrix-operator form

In Section 2, a vectorized parameter group was considered. In contrast, now let $n' = nm$, and initially denote $\tilde{\mathbf{G}}(t) := \bar{\mathbf{G}}(t)$ as a full-rank $n \times m$ matrix composed of normalized gradient components $\bar{g}(t)$ associated with a single layer or parameter group l , where $1 < n \leq m$. Then in matrix form, the update step of (7)–(11) can be denoted as

$$\mathbf{D}(t+1) = -\delta_p(t) \tilde{\mathbf{G}}(t). \quad (65)$$

The p -moment normalization acts on each individual ij entry of the matrix, where $1 \leq i \leq n$ and $1 \leq j \leq m$, and therefore enforces $\mathbb{E}\{|\tilde{\mathbf{G}}_{ij}(t)|^r\} \leq 1$ for $1 \leq r \leq p$. It follows that $\mathbb{E}\{|\mathbf{D}_{ij}(t+1)|^r\} \leq \mu^r$ is equivalent to $\mathbb{E}\{|\Delta(t+1)|^r\} \leq \mu^r$. Viewing the update as a matrix, the corresponding operator gain satisfies,

$$\mathbb{E}\|\mathbf{D}(t+1)\|_2^2 \leq \mathbb{E}\|\mathbf{D}(t+1)\|_F^2 = \delta_p^2(t) \sum_{ij}^{mn} \mathbb{E}\{|\tilde{\mathbf{G}}_{ij}(t)|^2\} \leq \mu^2 mn. \quad (66)$$

Therefore, although the entrywise p -moment trust-region controls the magnitude of the vectorized update, it cannot directly control the operator gain of the corresponding matrix update group. In particular, the bound (66) exceeds μ^2 . An algebraic consequence is that if $\mathbb{E}\|\mathbf{D}(t+1)\|_2 \leq \mu$, then $\mathbb{E}|\mathbf{D}_{ij}(t+1)| \leq \mathbb{E}\|\mathbf{D}(t+1)\|_2 \leq \mu$. Therefore, a tighter trust-region control can be obtained by constraining the entire matrix update group at the spectral level according to

$$\|\mathbf{D}(t+1)\|_2 \leq \mu', \quad (67)$$

where $\mu \leq \mu' < \mu\sqrt{mn}$. In this range, the special case $\mu' = \mu$ corresponds to the strongest trust-region constraint relative to (66). An effective way to control the spectrum of a matrix is through an orthogonalization step (Higham, 2008), such as,

$$\tilde{\mathbf{G}}(t) = (\bar{\mathbf{G}}(t)\bar{\mathbf{G}}(t)^\top)^{-\frac{1}{2}} \bar{\mathbf{G}}(t). \quad (68)$$

By the singular-value decomposition $\bar{\mathbf{G}}(t) = U(t)\Sigma(t)V(t)^\top$, it follows that $\tilde{\mathbf{G}}(t) = U(t)V(t)^\top$, and therefore $\|\tilde{\mathbf{G}}(t)\|_2 = 1$. The layer or matrix group update step $\mathbf{D}(t+1) = -\delta_p(t) \tilde{\mathbf{G}}(t)$ then satisfies

$$\|\mathbf{D}(t+1)\|_2 = \delta_p(t) \|\tilde{\mathbf{G}}(t)\|_2 = \delta_p(t) \leq \mu. \quad (69)$$

Since $\tilde{\mathbf{G}}(t) = U(t)V(t)^\top$, each ij entry, directly satisfies $|\tilde{\mathbf{G}}_{ij}(t)| = |u_i^\top(t)v_j(t)| \leq \|u_i(t)\|_2 \|v_j(t)\|_2 \leq 1$, and therefore $|\mathbf{D}_{ij}(t+1)| \leq \mu$, equivalent to $|\Delta(t+1)| \leq \mu$ is also directly satisfied.

This matrix-operator form of the regularized algorithm is outlined in (12)–(17), with (16) introduced. Moment normalization does not directly control the operator gain (largest singular value) of the matrix group, while spectral normalization does not regulate the statistical variability of its individual entries. Therefore, the preceding p -moment normalization and the subsequent spectral normalization are complementary trust-region mechanisms that operate on different properties of the update matrix group. By directly controlling the operator gain of the entire matrix update group to be less than the maximum trust-region constant μ , spectral normalization strengthens the underlying trust-region framework.

5 Practical realizations

This section discusses practical realizations of the statistical expectations appearing in GMAKE, together with a practical realization of the principal matrix inverse square-root required by the matrix-operator form.

Matrix inverse square-root. For computing the principal matrix inverse square-root in (16), we adopt the efficient polynomial recursion of Lakić (1998).

Linear statistical estimators. GMAKE requires online estimates that ensure $\lambda_2(t) > 0$ and $\kappa_p(t) \geq 1$. Since these expectations are unavailable *a priori*, they must be replaced by recursive estimators. Here, we adopt classical estimators from the stochastic approximation literature whose statistical properties are already well established (Zoubir et al., 2018; Lehmann & Casella, 2005; James & Stein, 1961). A common practical estimator is the exponentially weighted moving average (EMA)

$$m(t) = \rho m(t-1) + (1-\rho)x(t), \quad 0 < \rho < 1, \quad (70)$$

which may be used to estimate both quantities by selecting $x(t) = |g(t)|^2$, $m(0) = 0$, or $x(t) = |\bar{g}(t)|^p$, $m(0) = 1$, respectively. For the normalized p -th moment $\kappa_p(t)$, sometimes a possible structured estimate can be unity. In such situations, a linear shrinkage estimator (LSE) such as

$$m(t) = \rho m_0 + (1-\rho)x(t), \quad 0 < \rho < 1, \quad (71)$$

with prior value $m_0 = 1$ may also provide an alternative realization (James & Stein, 1961). When the prior is accurate, the LSE can reduce estimator variance (Lédoit & Wolf, 2004) while avoiding the additional memory state required by a second EMA. Both estimators are also subject to initial transient bias from the true underlying mean value (Box et al., 2015; Goodwin & Sin, 1984). The typical bias correction may be applied by dividing the estimates by $1 - \rho^t$. For both, values of ρ very close to unity typically provide robust averaging (Haykin, 2014; Ljung & Söderström, 1983).

6 Numerical Experiments

The algorithms defined by (3)–(6), (7)–(11), and (12)–(17) will be referred to as the basic, spectrally filtered, and matrix-operator forms of GMAKE, respectively. These correspond to progressively stronger realizations of the underlying trust-region framework. As proof of concept, we compare the three forms for both $p = 2$ and $p = 4$. All experiments, were repeated three times, and the average training loss and validation loss curves are reported. The GPT2-124M model processes 8192 tokens per iteration. Several observations can be made.

6.1 FineWeb-Edu

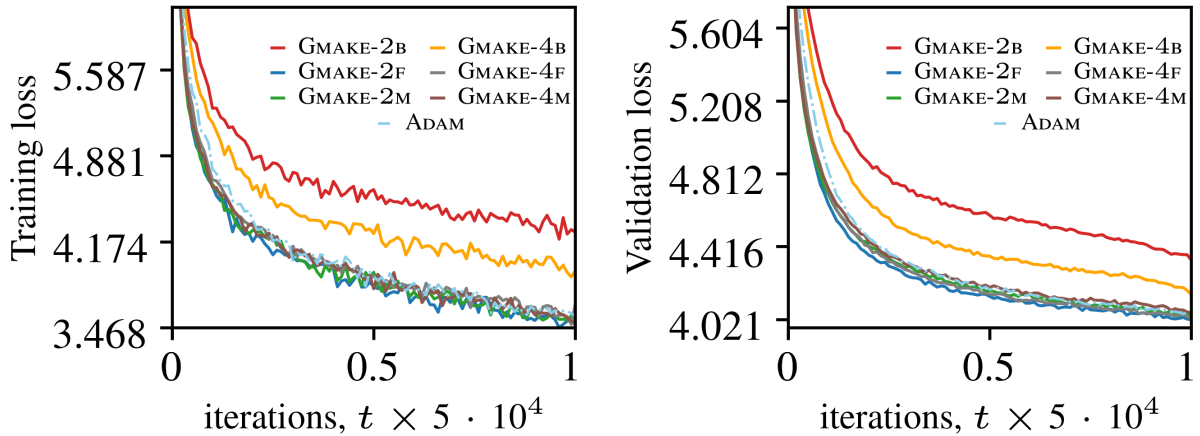


Figure 3: Training (left) and validation (right) loss for the basic (B), spectrally filtered (F), and matrix-operator (M) forms of GMAKE with $p = 2$ and $p = 4$ on GPT2-124M trained on FineWeb-Edu.

Figure 3 compares the basic (B), filtered (F), and matrix-operator (M) forms of GMAKE for both $p = 2$ and $p = 4$ on GPT2-124M trained on a 50 million token subset of FineWeb-Edu. In all experiments, the same maximum trust-region radius $\mu = 5 \times 10^{-4}$, filter parameter $\beta = 0.9$, averaging coefficient $\rho = 0.999$, and linear decay schedule were used.

First, the filtered realizations substantially improve both training and validation performance relative to their corresponding basic forms. This is consistent with the analysis of Section 3, where spectral lowpass regularization (momentum) acts as a trust-region-preserving mechanism that improves the quality of the gradient signal before update generation.

Second, within the basic realization, the fourth-moment form consistently outperforms the corresponding second-moment form. This suggests that, when a relatively large trust-region radius is permitted, the additional higher-order moment normalization provides beneficial update control beyond that obtained from second-moment normalization alone.

Third, the matrix-operator realizations remain competitive with their corresponding filtered forms despite satisfying a stronger operator-level trust-region constraint. Since all realizations use the same maximum trust-region radius μ , the matrix-operator form provides the strongest trust-region guarantees among the three realizations while maintaining comparable optimization performance.

6.2 TinyStories

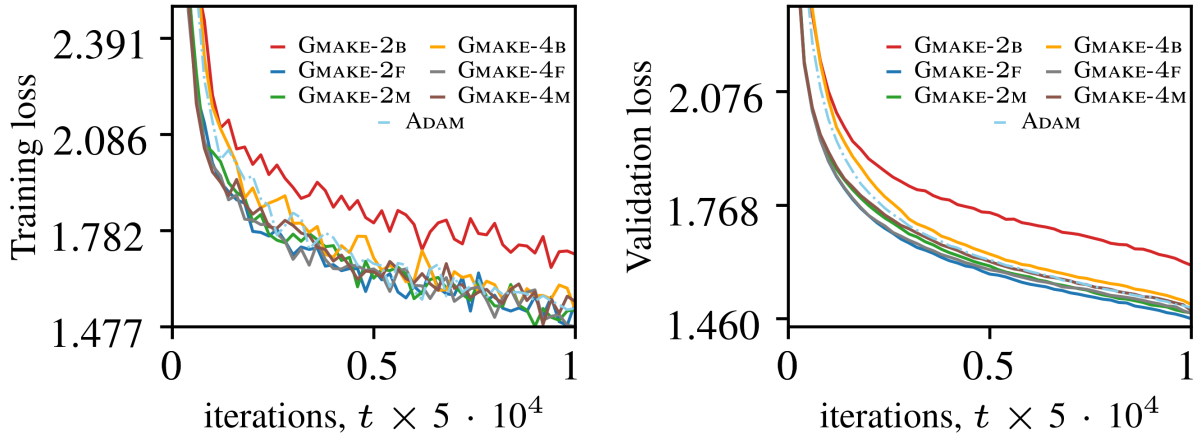


Figure 4: Training (left) and validation (right) loss for the basic (B), spectrally filtered (F), and matrix-operator (M) forms of GMAKE with $p = 2$ and $p = 4$ on GPT2-124M trained on TinyStories.

The results on TinyStories largely mirror those observed on FineWeb-Edu. The filtered realizations again outperform their corresponding basic forms, supporting the view that spectral lowpass regularization provides a useful strengthening of the underlying trust-region framework.

The fourth-moment realization again achieves lower losses than the second-moment realization in the basic form. However, after the introduction of filtering and matrix-operator normalization, the performance gap between the second- and fourth-moment realizations becomes considerably smaller.

Across both datasets, the filtered and matrix-operator training-loss trajectories are nearly indistinguishable, while their corresponding validation losses remain similarly close. These observations suggest that progressively stronger trust-region controls can be imposed without materially degrading optimization performance.

6.3 Effect of Strengthening the Trust-Region

Figure 5 investigates the effect of reducing the maximum trust-region radius from $\mu = 5 \times 10^{-4}$ to $\mu = 3 \times 10^{-4}$ in the basic realization.

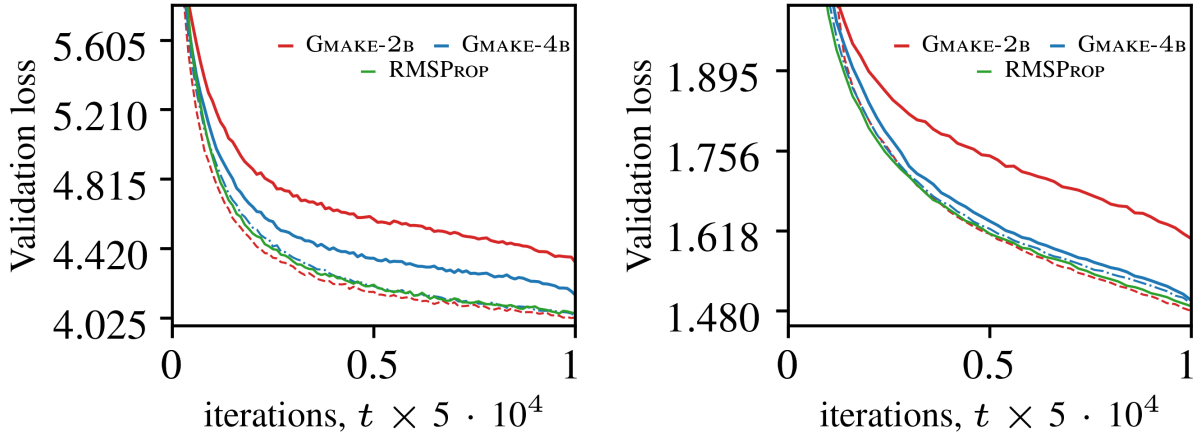


Figure 5: Validation loss on GPT2-124M trained on FineWeb-Edu (left) and TinyStories (right). Solid lines correspond to the basic GMAKE realization with trust-region radius $\mu = 5 \times 10^{-4}$, while dash-dotted lines correspond to $\mu = 3 \times 10^{-4}$. RMSProp is included only for $\mu = 3 \times 10^{-4}$ because training with $\mu = 5 \times 10^{-4}$ was unstable. The larger trust-region radius highlights the advantage of the fourth-moment realization over the second-moment realization, while reducing the trust-region radius substantially narrows the performance gap between the two forms, with the second-moment realization becoming slightly better.

A notable observation is that the performance advantage of the fourth-moment realization becomes substantially smaller as the trust-region radius is reduced. For $\mu = 5 \times 10^{-4}$, the fourth-moment realization consistently outperforms the second-moment realization on both datasets. However, at the smaller trust-region radius, the performance gap narrows considerably and the second-moment realization achieves slightly lower validation loss.

This behavior is consistent with the trust-region interpretation developed in Section 2. The higher-order normalization mechanism becomes most useful when larger update magnitudes are permitted. As the trust-region radius is reduced, update magnitudes are already more tightly constrained, leaving less opportunity for the higher-order moment normalization to provide additional benefit.

For reference, RMSProp was also evaluated at $\mu = 3 \times 10^{-4}$. Training with $\mu = 5 \times 10^{-4}$ was found to be substantially less stable and is therefore omitted from the comparison. At the smaller trust-region radius, the performance of RMSProp and the second-moment realization become comparable.

6.4 Trust-Region Insights

The preceding experiments suggest several observations that are best understood through the trust-region interpretation developed in Sections 2-4 of this paper.

First, for the larger trust-region radius $\mu = 5 \times 10^{-4}$, the fourth-moment realization consistently outperforms the corresponding second-moment realization in the basic form on both FineWeb-Edu and TinyStories. This behaviour is consistent with the interpretation that the normalized fourth-moment estimator provides additional update control beyond second-moment normalization alone when relatively large update magnitudes are permitted.

Second, the introduction of progressively stronger trust-region controls appears to reduce the performance gap between the second-moment and fourth-moment realizations. This effect is observed after the introduction of spectral lowpass regularization, matrix-operator normalization, and also when the maximum trust-region radius is reduced from $\mu = 5 \times 10^{-4}$ to $\mu = 3 \times 10^{-4}$ in the basic realization.

Figure 5 provides additional evidence for this observation. As the maximum trust-region radius is reduced, the advantage of the fourth-moment realization becomes substantially smaller, with the second-moment and fourth-moment realizations achieving comparable performance on both datasets. This suggests that the benefits of higher-order moment normalization are most pronounced when the admissible trust region is relatively large.

Taken together, these results indicate that the relative advantage of the fourth-moment realization may be inversely related to the strength of the surrounding trust-region controls. When the trust-region constraints are weak, the

additional higher-order normalization provides a more meaningful correction to the update process. As progressively stronger trust-region controls are introduced through a smaller trust-region radius, spectral filtering, or operator-level normalization, the second-moment realization appears sufficient to capture a larger fraction of the achievable performance gains.

Finally, $\text{RMSP}_{\text{PROP}}$ was observed to train less stably than its G_{MAKE} equivalent at the larger trust-region radius $\mu = 5 \times 10^{-4}$, and therefore is reported only for $\mu = 3 \times 10^{-4}$. This observation is consistent with the trust-region interpretation developed in this paper, where the additional normalization mechanisms of G_{MAKE} permit stable optimization under larger admissible update magnitudes.

7 Limitations, Gaps and Future Work

Although this trust-region framework establishes theoretical connections between moment estimation, learning-rate scheduling, spectral filtering, and matrix-operator normalization, several questions remain open.

A notable empirical observation is that the performance advantage of the fourth-moment realization is most pronounced when the trust-region constraints are relatively weak. In the basic realization, operated with a larger trust-region radius, the fourth-moment form initially outperformed its second-moment counterpart. However, this advantage diminished substantially as stronger trust-region controls were introduced, whether through spectral filtering, matrix-operator normalization, or a smaller maximum trust-region radius. This trend suggests that part of the benefit provided by fourth-moment normalization may overlap with the stabilization and regularization effects already supplied by these stronger trust-region mechanisms.

The precise cause of this behavior remains unclear. One possible explanation is that these mechanisms partially address the same update fluctuations that the higher-order moment normalization is designed to regulate, thereby reducing its incremental benefit. Alternatively, the observation may indicate that spectral lowpass regularization and spectral trust-region control naturally strengthen second-moment mechanisms more than higher-order moment mechanisms.

Nonetheless, this observation may expose a gap in the current framework. While spectral lowpass regularization appears to provide a natural strengthening of the second-moment trust-region mechanism, a corresponding regularization principle specifically aimed at higher-order moment control has not yet been identified. It therefore remains unclear whether an analogous mechanism exists and, if so, whether it could recover or further enhance the benefits of the $p = 4$ realization. More broadly, a systematic investigation of the performance and stability characteristics across the range $1 < p \leq 4$, especially the regime $p < 2$, may provide deeper insight into the role of moment order in trust-region control and accelerated learning.

Moreover, the numerical experiments were intended primarily as proof-of-concept demonstrations rather than a comprehensive empirical evaluation. For consistency across experiments, all realizations were evaluated using the same values of μ , β , and ρ , and only two representative choices of p were examined. An important direction for future work is therefore a systematic investigation of the sensitivity of the basic, filtered, and matrix-operator formulations to these hyperparameters. Such studies could also explore alternative normalization orderings, adaptive selections of p , and spectral-level trust-region constructions across a wider range of learning problems.

8 Conclusion

This paper developed a trust-region framework for studying the moment estimation mechanism in stochastic gradient optimization. The derived family of learning-rate mechanisms generalizes second-moment methods through p -th moment trust-region constraints. Within this framework, $\text{RMSP}_{\text{PROP}}$ and ADAM can be interpreted respectively, as similar to the second-moment realizations of the basic and spectrally filtered G_{MAKE} formulations.

The G_{MAKE} framework further reveals that learning-rate scheduling, momentum, and spectral-norm normalization can be interpreted as complementary trust-region mechanisms acting on different properties of the update process. By extending the basic realization through spectral lowpass filtering and matrix-operator normalization, progressively stronger trust-region controls are obtained within a common framework. The experimental results further suggest the existence of a trust-region hierarchy, in which the incremental benefit of higher-order moment normalization decreases as these progressively stronger trust-region controls are imposed on the update process.

Taken together, these developments provide a unified trust-region interpretation of several mechanisms commonly used in optimizing deep neural networks via the stochastic gradient algorithm.

References

- Shane Bergsma, Nolan Simran Dey, Gurpreet Gosal, Gavia Gray, Daria Soboleva, and Joel Hestness. Straight to zero: Why linearly decaying the learning rate to zero works best for LLMs. In *The Thirteenth International Conference on Learning Representations*, Singapore, October 2024. URL <https://openreview.net/forum?id=hr0lBgHsMI>.
- Jeremy Bernstein and Laker Newhouse. Old Optimizer, New Norm: An Anthology, September 2024. URL <https://arxiv.org/abs/2409.20325v1>.
- Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, May 2018.
- George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. *Time Series Analysis: Forecasting and Control*. Wiley, 5 edition, 2015.
- Stephen Boyd and Craig Barratt. *Linear Controller Design: Limits of Performance*. Prentice Hall, Englewood Cliffs, NJ, 1991.
- Andrew R. Conn, Nicholas I. M. Gould, and Philippe L. Toint. *Trust Region Methods*. MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics, January 2000. ISBN 978-0-89871-460-9. doi: 10.1137/1.9780898719857.
- Saber Elaydi. *An Introduction to Difference Equations*. Springer, New York, 2005. ISBN 978-0-387-27602-1.
- Rong Ge, Sham M. Kakade, Rahul Kidambi, and Praneeth Netrapalli. Rethinking learning rate schedules for stochastic optimization. In *International Conference on Learning Representations*, New Orleans, LA, September 2018. URL <https://openreview.net/forum?id=HJePy3RcF7>.
- Graham C. Goodwin and Kwai Sang Sin. *Adaptive Filtering Prediction and Control*. Dover Publications, Englewood Cliffs, N.J, 1984. ISBN 978-0-486-46932-4.
- Simon Haykin. *Adaptive Filter Theory*. Pearson, Upper Saddle River, NJ, 5th, intern. edition, 2014. ISBN 978-0-13-267145-3.
- Nicholas J. Higham. *Functions of Matrices: Theory and Computation*. Society for Industrial and Applied Mathematics, 2008. ISBN 978-0-89871-646-7. URL 10.1137/1.9780898717778.
- W. James and Charles Stein. Estimation with Quadratic Loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 4.1, pp. 361–380. University of California Press, January 1961.
- Eric Jondeau and Michael Rockinger. Optimal Portfolio Allocation under Higher Moments. *European Financial Management*, 12(1):29–55, 2006. ISSN 1468-036X. doi: 10.1111/j.1354-7798.2006.00309.x.
- Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*, San Diego, CA, USA,, May 2015.
- Slobodan Lakić. On the Computation of the Matrix k -th Root. *Journal of Applied Mathematics and Mechanics, ZAMM*, 78(3):167–172, 1998. ISSN 1521-4001.
- Tim Large, Yang Liu, Minyoung Huh, Hyojin Bahng, Phillip Isola, and Jeremy Bernstein. Scalable Optimization in the Modular Norm. *Advances in Neural Information Processing Systems*, 37:73501–73548, December 2024. doi: 10.52202/079017-2338.
- Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, February 2004. ISSN 0047-259X. doi: 10.1016/S0047-259X(03)00096-4.

-
- Erich L. Lehmann and George Casella. *Theory of Point Estimation*. Springer, New York, NY Berlin Heidelberg, 2 edition, 2005. ISBN 978-0-387-98502-2.
- Xianliang Li, Jun Luo, Zhiwei Zheng, Hanxiao Wang, Li Luo, Lingkun Wen, Linlong Wu, and Sheng Xu. On the Performance Analysis of Momentum Method: A Frequency Domain Perspective. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=tznvtmSEiN>.
- Lennart Ljung and Torsten Söderström. *Theory and Practice of Recursive Identification*. MIT Press, 1983.
- George S. Lueker. Some Techniques for Solving Recurrences. *ACM Comput. Surv.*, 12(4):419–436, December 1980. ISSN 0360-0300. doi: 10.1145/356827.356832.
- Alan V. Oppenheim and Ronald W. Schaffer. *Discrete-Time Signal Processing*. Pearson, Upper Saddle River, NJ, 3rd edition, 2010. ISBN 978-0-13-198842-2.
- Sophocles J. Orfanidis. *Introduction to Signal Processing*. Prentice-Hall, Inc., NJ, USA, 2 edition, October 1995. ISBN 978-0-13-209172-5.
- Neal Parikh and Stephen Boyd. Proximal Algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, January 2014. ISSN 2167-3888. doi: 10.1561/24000000003.
- B. T. Polyak. The conjugate gradient method in extremal problems. *USSR Computational Mathematics and Mathematical Physics*, 9(4):94–112, 1969. ISSN 0041-5553. doi: 10.1016/0041-5553(69)90035-4.
- B. T. Polyak. Accelerated gradient methods: History and properties. In *7th International Conference on Control and Optimization with Industrial Applications*, volume 1, pp. 23–25, Baku, Azerbaijan, 2020.
- L. Pronzato, A. Zhigljavsky, and E. Bukina. Estimation of Spectral Bounds in Gradient Algorithms. *Acta Applicandae Mathematicae*, 127(1):117–136, October 2013. ISSN 1572-9036. doi: 10.1007/s10440-012-9794-z.
- Sai Niranjan Ramachandran, Manish Krishan Lal, and Suvrit Sra. Cross-fluctuation phase transitions reveal sampling dynamics in diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, October 2025. URL <https://openreview.net/forum?id=b4X6cz1F9l>.
- Julius O. Smith. Pole Zero Analysis I. In *Introduction to Digital Filters with Audio Applications*. W3K Publishing, CA, USA, 2007. ISBN 978-0-9745607-1-7. URL https://ccrma.stanford.edu/~jos/fp/Pole_Zero_Analysis_I.html.
- Oluwasegun Somefun. *Fundamental signal processing elements in accelerated learning*. PhD thesis, Oregon State University, Corvallis, OR, March 2026. URL <https://ir.library.oregonstate.edu/downloads/76537923z>.
- Oluwasegun A. Somefun, Stefan Lee, and V. John Mathews. AUTOSGM: A Unified Lowpass Regularization Framework for Accelerated Learning. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, April 2024. doi: 10.1109/ICASSP48485.2024.10448203.
- Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 1139–1147, Atlanta, GA, USA, May 2013. URL <https://proceedings.mlr.press/v28/sutskever13.html>.
- Abdelhak M. Zoubir, Visa Koivunen, Esa Ollila, and Michael Muma. *Robust Statistics for Signal Processing*. Cambridge University Press, 1 edition, October 2018. doi: 10.1017/9781139084291.