A Variational Trust-Region Framework for Learning-Rate Annealing

Oluwasegun Somefun *somefuno.oregonstate.edu*
V John Mathews *mathews@oregonstate.edu*
School of Electrical Engineering and Computer Science,
Oregon State University.

Abstract

This paper develops a variational trust-region framework for deriving learning-rate schedules for stochastic gradient learning. Each iteration of the stochastic gradient algorithm is formulated as a trust-region subproblem that bounds the parameter-update step at each iteration. Under a fixed trust-region radius, this yields an iteration-dependent learning rate inversely proportional to the square-root of the second moment of the gradient, recovering the core structure of adaptive moment estimation. When the trust-region radius is allowed to vary and vanish over a finite learning window, learning-rate design becomes a variational optimization problem posed directly on the evolution of the trust-region radius. We show that linear decay and cosine annealing arise as unique minimizers of Dirichlet and regularized Dirichlet energy functionals of the trust-region radius subject to fixed boundary values. The resulting formulation generalizes naturally to multi-point boundary-value problems, yielding warmup-decay, warmup-stable-decay, and related schedules as solutions of a common first-order variational principle. We further characterize smoothness and stability properties associated with vanishing trust-region windows. Together, these results provide a unified foundation for learning-rate annealing and its coupling with moment estimation in stochastic gradient methods, showing that popular learning-rate schedules can be interpreted as boundary-constrained trust-region trajectories whose form is determined by a first-order variational energy principle.

1 Introduction

Learning-rate schedules play a central role in the practical performance of stochastic gradient methods (Umeda & Iiduka, 2024; Bergsma et al., 2024), especially in large-scale deep learning where training can exceed millions of iterations (Caron et al., 2021; Bello et al., 2021; Goujaud et al., 2022). Despite their widespread use, commonly adopted annealing rules like linear decay and cosine schedules (He et al., 2019; Hägele et al., 2024) are typically motivated heuristically and empirically (Ge et al., 2018; Agarwal et al., 2021; Bergsma et al., 2024) rather than derived from first principles. At the same time, popular stochastic gradient methods like RMSProp (Tieleman & Hinton, 2012; Graves et al., 2013) and Adam (Kingma & Ba, 2015) rely on gradient moment-based normalization and learning-rate annealing (Loshchilov & Hutter, 2017; 2019) but lack a unifying theoretical interpretation.

A substantial empirical literature already documents the influence of learning-rate annealing on large-scale training performance (Bergsma et al., 2024; Hägele et al., 2024). In contrast, this paper develops a trust-region framework that unifies second-moment normalization and annealing schedules in stochastic gradient methods. Here, we interpret each iteration of the stochastic gradient method as a trust-region minimization subproblem on the parameter update step. Under a fixed trust-region radius, this formulation yields an optimal iteration-dependent learning rate that is inversely proportional to the square-root of the second moment of the gradient, revealing the underlying structure of adaptive moment estimation in accelerated learning. We then extend this result to a more general case where the trust-region radius of the learning algorithm is varying and vanishes over a finite training window of iterations $t \in [0, \tau]$, and pose variational problems over the trust-region radius to yield its smoothest possible variation under well-defined boundary conditions. This leads to two optimal annealing schedules: linear decay, cosine annealing, together with their square-root forms, each arising as the extremal solution to a total variational problem governing the smooth contraction of the trust region.

These results provide principled derivations of some of the most widely used annealing schedules in deep learning (Huang et al., 2017; Loshchilov & Hutter, 2017; 2019; Bergsma et al., 2024), showing they are not arbitrary heuristics but arise as optimal solutions to well-posed smoothness-constrained optimization problems.

The main contributions of this paper are as follows:

(1) We develop a trust-region formulation of stochastic gradient learning in which second-moment estimation arises as a learning-rate mechanism from a fixed trust-region radius, while learning-rate annealing arises from a varying and vanishing trust-region radius.

(2) We show that popular learning-rate schedules are unique solutions of first-order variational trust-region subproblems, obtained by minimizing Dirichlet and regularized Dirichlet energies under fixed boundary values.

(3) We characterize the general boundary-value form of the resulting total variational problem, showing that warmup-decay, warmup-stable-decay, and related schedules are generated by boundary value conditions. Subsequently, we establish general smoothness and stability properties of the resulting solutions with respect to the Dirichlet energy over the trust-region window.

Beyond recovering common learning-rate schedules, this framework reveals its functional role within stochastic gradient learning. A constant trust-region radius yields bounded update magnitudes, whereas learning-rate schedules derived here based on a vanishing trust region radius induce exponential mean-square contraction of the parameter update-step vector. Therefore, learning-rate annealing arises from first principles not merely as a step-size reduction strategy, but as a variational trust-region mechanism for converting bounded mean-square update-step behavior into an exponentially vanishing update-step vector near the end of the learning window.

The remainder of the paper develops these results in detail. Section 2 formulates the vanishing trust-region problem and derives the associated optimal learning-rate function. Sections 2.1 and 2.2 derive linear decay, square-root linear decay, and cosine annealing as solutions of variational trust-region problems over a finite learning window. Section 2.4 studies structural properties of the resulting window functions, while Section 3 extends the analysis to multi-point boundary-value problems, yielding warmup-decay and warmup-stable-decay schedule families. Further, Sections 3.4 and 3.5 characterize the smoothness of window functions through a normalized Dirichlet-energy factor and establish exponential mean-square stability properties associated with the vanishing trust-region window. Finally, this paper is concluded in Section 4.

2 A Trust-region Problem

In this section, we begin by formulating a trust-region problem (Parikh & Boyd, 2014) on the parameter iterates of the learning algorithm. Let the shorthand notation (t, i) denote the i-th element of an n-dimensional vector at iteration t. Expressed coordinate-wise, the stochastic gradient update algorithm is

$$\mathrm{w}(t+1, i) = \mathrm{w}(t, i) - \alpha(t, i)\, \mathrm{g}(t, i) \tag{1}$$

where $\mathrm{w}(t, i)$ is the i-th coordinate of the parameter vector at iteration t, $\mathrm{g}(t, i)$ is the corresponding stochastic gradient, and $\alpha(t, i) > 0$ is a possibly iteration-dependent learning rate applied to that coordinate.

Throughout, we use $\mathbb{E}[\cdot]$ to denote expectation with respect to the underlying stochastic process generating the gradients, and use $\|\cdot\|_2$ to denote the Euclidean norm of a vector. No specific objective function, probabilistic structure, or distribution of the gradients is assumed. We make only the following minimal assumptions for all $t \in [0, \tau]$ within a finite learning window of total iterations $\tau > 2$, and subsequently consider the asymptotic regime $\tau \to \infty$.

Assumption 1 (Lipschitz Regularity)**.** *Both the scalar-valued objective function $f\big(\mathrm{w}(t, i)\big)$ and $\mathrm{g}(t, i)$ are Lipschitz continuous (Bottou et al., 2018).*

Assumption 2 (Bounded Second Moment)**.** *For each coordinate i, the stochastic gradient admits a bounded second moment,* i.e.,

$$\mathbb{E}[\mathrm{g}^2(t, i)] \le \bar{\xi} < \infty, \quad \textit{where} \quad \bar{\xi} := \sup_{0 \le t \le \tau} \mathbb{E}[\mathrm{g}^2(t, i)] \tag{2}$$

Under this setup for the learning algorithm, we define the parameter's step-change as $\Delta(t+1,i) = \mathrm{w}(t+1,i) - \mathrm{w}(t,i)$. From (1), for each coordinate i, the expected squared step-change (second moment),

$$\mathbb{E}[\Delta^2(t+1,i)] = \alpha^2(t,i)\,\mathbb{E}[\mathrm{g}^2(t,i)] \tag{3}$$

is well-defined for all t. This variable is used as a trust-region measure in this analysis. For notational convenience, we define $\delta^2(t) \equiv \mathbb{E}[\Delta^2(t+1,i)]$. Therefore (3) is equivalent to

$$\alpha^2(t,i)\,\mathbb{E}[\mathrm{g}^2(t,i)] = \delta^2(t). \tag{4}$$

In this trust-region formulation, it is desired that the expected step-change function $\delta(t)$ is bounded within a maximum trust-region radius $\mu > 0$, such that

$$\delta^2(t) \;\leq\; \mu^2, \tag{5}$$

and that $\delta(t) \to 0$ as $t \to \tau$. The bounded and vanishing behavior of the step changes are desirable for several reasons. For a vanishing trust-region radius, the parameter update step sequence become iteratively stabilized as $\Delta(t+1,i) \to 0$ in both the mean and mean-square sense as $t \to \tau$. Consequently, the parameter iterates do not oscillate indefinitely due to gradient noise, and the update step becomes increasingly conservative near a local optimum as the end of the learning window is approached. This is a classical stability requirement in stochastic approximation (Ge et al., 2018).

From (4), it can be observed that $\delta(t)$ directly determines the learning-rate, that is

$$\alpha(t,i) = \frac{\delta(t)}{\sqrt{\mathbb{E}[\mathrm{g}^2(t,i)]}}. \tag{6}$$

In the absence of the vanishing trust-region constraint, the function $\delta(t)$ is fixed at an initial value, specifically $\delta(t) = \mu$, and the learning rate function in (6) recovers the core structure of adaptive moment estimation. By substituting (6) in the update step, the overall trust-region constraint in the Euclidean norm,

$$\mathbb{E}\big[\|\Delta(t+1)\|_2^2\big] \leq n\,\mu^2 \tag{7}$$

is satisfied. This special-case of the learning-rate in (6) can also be derived as the solution to a first-order trust-region subproblem, analogous to the Cauchy point in classical trust-region methods (Nocedal & Wright, 2006). Under a first-order Taylor's series expansion of $f(\mathrm{w}(t))$ along the i-th coordinate, $f(\mathrm{w}(t+1,i)) - f(\mathrm{w}(t,i)) \approx \mathrm{g}(t,i)\,\Delta(t+1,i)$. Taking expectations of both sides yields $\mathbb{E}[f(\mathrm{w}(t+1,i)) - f(\mathrm{w}(t,i))] \approx \mathbb{E}[\mathrm{g}(t,i)\,\Delta(t+1,i)]$. Therefore, minimizing a surrogate of the expected local variation of the objective along the i-th coordinate under a bounded trust-region constraint leads to the first-order trust-region subproblem

$$\min_{\alpha(t,i)}\; \mathbb{E}[\Delta(t+1,i)\,\mathrm{g}(t,i)] \quad \text{s.t.} \quad \mathbb{E}[\Delta^2(t+1,i)] = \delta^2(t) \leq \mu^2. \tag{8}$$

Solving (8) leads to $\alpha(t,i) = \mu\,/\sqrt{\mathbb{E}[\mathrm{g}^2(t,i)]}$, and therefore (7) is satisfied. In (8), the direct optimization variable is $\alpha(t,i)$. Its solution recovers the special case of (6) corresponding to a constant trust-region radius $\delta(t) = \mu$. However, in (6) the direct optimization variable becomes the unknown trust-region trajectory $\delta(t)$.

Consequently, in contrast to (8), the vanishing trust-region problem $\delta(t) \to 0$ as $t \to \tau$ does not solve directly for $\alpha(t,i)$. Instead, it seeks a $\delta(t)$ trajectory satisfying at least the two fixed boundary-value conditions $\delta(0) \leq \mu$ and $\delta(\tau) = 0$. To obtain a unique closed-form solution of $\delta(t)$ that produces the smoothest possible variation of the trust-region radius, we now pose total variational minimization problems (also referred to as Dirichlet energy minimization (Iwaniec & Onninen, 2022)) directly on the vanishing trust-region radius $\delta(t)$ (or its squared radius $\delta^2(t)$). We will see that each problem enforces a different notion of smoothness that forces the variable trust-region radius not to vanish abruptly or faster than necessary during learning.

2.1 Derivation of the Linear Decay Schedule

Under fixed boundary conditions $\delta(0) = \mu$, $\delta(\tau) = 0$, we now find the sequence $\delta(t)$ that minimize an energy function on the total trust-region variation between iterations over the window of iterations $t \in [0, \tau]$. That is, for index $k \in [1,\ \tau]$, we solve for τ interior values of $\delta(t)$ via the optimization problem

$$\min_{0 \leq \delta(t) \leq \mu} \mathcal{D} \triangleq \sum_{k=1}^{\tau} \big(\delta(k) - \delta(k-1)\big)^2. \tag{9}$$

In the minimization problem (9), $\delta(t) - \delta(t-1)$ is the first-order difference of $\delta(t)$ and measures its rate of change over iterations t, therefore minimizing the total sum of squared differences in (9) forces $\delta(t)$ to be as close as possible to $\delta(t-1)$ across iterations.

The formulation in (9) is strictly convex in $\delta(k)$, and its minimizer is unique. Taking the derivative $d\,\mathcal{D}/\delta(k)$ and setting it to 0 results in $2\big(\delta(k) - \delta(k-1)\big) - 2\big(\delta(k+1) - \delta(k)\big) = 0$ which is equivalent to the second-order homogeneous linear recurrence or difference equation

$$\delta(k+1) = 2\,\delta(k) - \delta(k-1) \tag{10}$$

For the recurrence in (10), its characteristic equation is $z^2 - 2\,z + 1 = (z-1)^2 = 0$, which has 1 as a root with multiplicity 2. Therefore, $\forall\, t$, the general solution to (10) must be a polynomial in t, of at most degree 1 (Elaydi, 2005; Lueker, 1980),

$$\delta(t) = a + b\,t, \tag{11}$$

where $a, b \in \mathbb{R}$. This can be verified by plugging (11) into (10) for $t = k+1, k$ and $k-1$. Using the fixed boundary conditions $\delta(0) = \mu$, $\delta(\tau) = 0$, at $t = 0$, and $t = \tau$ in (11) yields

$$\begin{aligned} \mu &= a \\ 0 &= a + \tau\, b, \end{aligned} \tag{12}$$

which leads to both $b = -\mu/\tau$, and $a = \mu$. The complete solution to (10) is then

$$\delta(t) = \mu\left(1 - \frac{t}{\tau}\right), \tag{13}$$

and plugging (13) in (9), the minimum energy in the total variation of $\delta(t)$ is $\mu^2\,\tau^{-1}$. The unique solution to (9) in (13) is the *linear decay* function. In addition to the vanishing trust-region, the trust-region constraint $\delta^2(t) \leq \mu^2$ is always satisfied.

2.2 Derivation of the Square-root Linear Decay Schedule

In the first formulation (9), the total first-order variational problem over the window of iterations $t \in [0, \tau]$ was posed on $\delta(t)$. The same energy minimization problem may be posed directly on $\delta^2(t)$ which is equivalent to $\mathbb{E}[\Delta^2(t+1, i)]$. Specifically, given the equivalent fixed boundary conditions $\delta^2(0) = \mu^2$ and $\delta^2(\tau) = 0$, we find $\delta(t)$ by solving for τ interior values of $\delta^2(t)$ that minimize the variational problem

$$\min_{0 \leq \delta^2(t) \leq \mu^2} \sum_{k=1}^{\tau} \big(\delta^2(k) - \delta^2(k-1)\big)^2, \tag{14}$$

where $\delta^2(t) - \delta^2(t-1)$ is the first-order difference of $\delta^2(t)$. Since (14) is strictly convex in $\delta^2(k)$, so its minimizer is unique. Applying the same steps used in solving (9), we obtain the complete solution to (14) as

$$\delta^2(t) = \mu^2\left(1 - \frac{t}{\tau}\right) \tag{15}$$

to obtain

$$\delta(t) = \mu\sqrt{1 - \frac{t}{\tau}}. \tag{16}$$

The unique solution to (14) in (16) is the square-root of the *linear decay* function obtained in (13). Plugging (16) in (14), the minimum energy in the total variation of $\delta^2(t)$ is $\mu^4\,\tau^{-1}$, and similarly, the trust-region constraint $\delta^2(t) \leq \mu^2$ is always satisfied.

2.3 Derivation of Cosine Annealing via Regularized Dirichlet Energy Minimization

In the previous sections, we saw that minimizing an energy function (9) on the total trust-region variation between iterations over the window $t \in [0,\, \tau]$ under fixed boundary conditions led to the linear decay function (13) from (11). We can formulate a more general problem with solution for $\delta(t)$ in affine form

$$\delta(t) = a + b\,u(t) \tag{17}$$

where $u(t)$ is an iterative function, $a, b \in \mathbb{R}$, and the fixed boundary conditions $\delta(0) = \mu$, $\delta(\tau) = 0$ correspond to $u(0) = \frac{\mu - a}{b}$, and $u(\tau) = -\frac{a}{b}$. In this formulation, the non-trivial underlying functional form of $\delta(t)$ is captured by a shaping function $u(t)$ while a, b are trivial shift and scaling constants determined by the fixed boundary conditions. The total variational problem then becomes solving for τ interior values of $u(t)$ that minimize

$$\min_{u(t)\,\in\,[u(0),\,u(\tau)]} \quad \sum_{k=1}^{\tau} \big(u(k) - u(k-1)\big)^2. \tag{18}$$

Here, we now define a regularized minimization problem directly on $u(t)$. Due to fixed boundary conditions, the variational objective term $\big(u(t) - u(t-1)\big)^2$ is essentially defined over transition pairs $(k-1,\, k)$ within the interior of the window $k \in [1,\, \tau]$. Therefore, any added quadratic energy regularization should be expressed in a compatible pairwise form in order to maintain consistency with the transition-based structure of the objective term (Grady & Polimeni, 2010).

A natural choice is the symmetric quadratic form $\frac{1}{2}\big(u^2(t) + u^2(t-1)\big)$, which is analogous to Tikhonov regularization and introduces a shape-smoothing penalty (Yue et al., 2016) into (18) in a manner consistent with its pairwise structure. We therefore define the regularized energy minimization problem as

$$\min_{u(t)\,\in\,[u(0),\,u(\tau)]} \quad \mathcal{B} \triangleq \sum_{k=1}^{\tau} \big(u(k) - u(k-1)\big)^2 + \lambda \left[\sum_{k=1}^{\tau} \tfrac{1}{2}\big(u^2(k) + u^2(k-1)\big) - \tilde{c}\right], \tag{19}$$

where $\lambda \in \mathbb{R}$ is the penalty constant or Lagrange multiplier associated with the regularization term, and $\tilde{c} > 0$ is a fixed constant representing the allowable total quadratic energy.

Using the affine form of the general solution (17) makes (19) a regularized minimization problem of $\delta(t)$. Also, solving for $\delta(t)$ via (19) using (17) simplifies the derivation of regularized variants of (9) without altering the problem in terms of $\delta(t)$. Substituting $u(t) = \big(\delta(t) - a\big)/b$ into (19) yields

$$\min_{0 \le \delta(t) \le \mu} \quad \sum_{k=1}^{\tau} \frac{1}{b^2}\big(\delta(k) - \delta(k-1)\big)^2 + \lambda \left[\sum_{k=1}^{\tau} \frac{1}{2b^2}\big(\delta^2(k) + \delta^2(k-1)\big) + \frac{1}{b^2}\sum_{k=1}^{\tau}\big(\delta(k) + \delta(k-1)\big) + \frac{a^2}{b^2} - \tilde{c}\right]. \tag{20}$$

Because (19) is strongly convex, its minimizer is unique. For the initial boundary condition, $t = 0$, taking the derivative $d\,\mathcal{B}/u(0) = 0$ gives $-2\big(u(1) - u(0)\big) + \lambda\,u(0) = 0$, which is equivalent to

$$u(1) - u(0) = \tfrac{\lambda}{2}\,u(0). \tag{21}$$

For the final boundary condition, $t = \tau$, taking the derivative $d\,\mathcal{B}/u(\tau) = 0$ gives $2\big(u(\tau) - u(\tau-1)\big) + \lambda\,u(\tau) = 0$ which is equivalent to

$$u(\tau) - u(\tau-1) = -\tfrac{\lambda}{2}\,u(\tau). \tag{22}$$

For any interior index k, taking the derivative $d\,\mathcal{B}/u(k) = 0$ gives $2\big(u(k) - u(k-1)\big) - 2\big(u(k+1) - u(k)\big) + 2\lambda\,u(k) = 0$, which simplifies to

$$u(k+1) = (2 + \lambda)\,u(k) - u(k-1). \tag{23}$$

2.3.1 Generalized Eigenvalue Form

To proceed further, let $\mathbf{u} = [u(0), u(1), \ldots, u(\tau)] \in \mathbb{R}^{\tau+1}$ be a non-zero vector, and let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{\tau+1\times\tau+1}$. Then (19) can be written as $\mathcal{B} = \mathcal{S} + \lambda(\mathcal{Q} - \tilde{c})$, where

$$\mathcal{S} = \sum_{k=1}^{\tau}\big(u(k) - u(k-1)\big)^2 = \mathbf{u}^\mathsf{T}\mathbf{A}\mathbf{u}, \qquad \text{and} \qquad \mathcal{Q} = \sum_{k=1}^{\tau}\tfrac{1}{2}\big(u^2(k) + u^2(k-1)\big) = \mathbf{u}^\mathsf{T}\mathbf{B}\mathbf{u}, \tag{24}$$

with $\mathbf{B} = \mathrm{diag}\big([\frac{1}{2}, 1, 1, \ldots, 1, \frac{1}{2}]\big)$, and tridiagonal matrix

$$\mathbf{A} = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 \\ 0 & -1 & 2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 2 & -1 \\ 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix}. \tag{25}$$

Observe that $\mathcal{S}$ involves only the sum of squared differences. Therefore, $\mathcal{S}$ is always non-negative for any real vector $\mathbf{u}$. It is zero if and only if every difference in the sum is zero, which implies $\mathbf{u} = u(0)\,\mathbf{1}$, where $u(0)$ is a real constant, and $\mathbf{1} = [1, 1, \ldots, 1]$. However, a constant $\mathbf{u}$ is not admissible due to the endpoint constraints. Therefore, $\mathbf{A}$ is a symmetric positive semi-definite matrix with a one-dimensional null-space spanned by the all-ones vector. Similarly, $\mathcal{Q}$ is a sum of squares with strictly positive weights. $\mathcal{Q}$ is zero only if $\mathbf{u} = 0$. Converting this sum to matrix form corresponds to $\mathbf{B}$ being diagonal with strictly positive entries, making $\mathbf{B}$ a symmetric positive definite matrix.

Define $\tilde{\lambda} = -\lambda$. By taking the gradient of $\mathcal{B}$ with respect to $\mathbf{u}$ and setting the result to zero, we have $\mathbf{A}\mathbf{u} = \tilde{\lambda}\,\mathbf{B}\mathbf{u}$, which is equivalent to $\mathbf{B}^{-1}\mathbf{A}\mathbf{u} = \tilde{\lambda}\mathbf{u}$. This implies that

$$\tilde{\lambda} = \frac{\mathbf{u}^\mathsf{T}\mathbf{A}\mathbf{u}}{\mathbf{u}^\mathsf{T}\mathbf{B}\mathbf{u}} \tag{26}$$

is the generalized eigenvalue of $\mathbf{A}$ with respect to $\mathbf{B}$, and $\mathbf{u} \neq 0$ is the corresponding generalized eigenvector of the pair $(\mathbf{A}, \mathbf{B})$. Because $\mathbf{A}$ is symmetric positive semi-definite and $\mathbf{B}$ is symmetric positive definite, the generalized eigenvalues $\tilde{\lambda}$ of the system $\mathbf{A}\mathbf{u} = \tilde{\lambda}\,\mathbf{B}\mathbf{u}$ are real and non-negative. A constant vector of ones lies in the null-space of $\mathbf{A}$ so that $\tilde{\lambda} = 0$ is the eigenvalue corresponding to no variation (all entries in $\mathbf{u}$ are equal). Such an eigenvector violates the endpoint constraints $\delta(0) = \mu$, $\delta(\tau) = 0$, and consequently, it is not admissible.

Therefore in (19), to minimize $\mathcal{S}$ subject to $\mathcal{Q} = \tilde{c}$, the optimal $\mathbf{u}$ must be a generalized eigenvector of $(\mathbf{A}, \mathbf{B})$ associated with the smallest positive eigenvalue of $\mathbf{B}^{-1}\mathbf{A}$.

2.3.2 A bound for $\tilde{\lambda}$

We now show that for all $\mathbf{u} \neq 0$, $0 < \tilde{\lambda} < 4$. For this, we use the elementary inequality that $\forall\, x_1, x_2 \in \mathbb{R}$, $0 \leq (x_1 - x_2)^2 \leq 2(x_1^2 + x_2^2)$, with equality if and only if $x_1 = x_2 = 0$. Applying this for $x_1 = u(k)$ and $x_2 = u(k-1)$, and then summing both sides over the window yields

$$0 \leq \sum_{k=1}^{\tau}\big(u(k) - u(k-1)\big)^2 \leq 4\sum_{k=1}^{\tau}\tfrac{1}{2}\big(u^2(k) + u^2(k-1)\big) \tag{27}$$

Using the matrix form in (24), this is equivalent to $0 \leq \mathbf{u}^\mathsf{T}\mathbf{A}\mathbf{u} \leq 4\,\mathbf{u}^\mathsf{T}\mathbf{B}\mathbf{u}$. For $\mathbf{u} \neq 0$, strict inequality holds, so that $0 < \mathbf{u}^\mathsf{T}\mathbf{A}\mathbf{u} < 4\,\mathbf{u}^\mathsf{T}\mathbf{B}\mathbf{u}$. Combining this result with the expression for $\tilde{\lambda}$ in (26) gives the result we seek:

$$0 < \frac{\mathbf{u}^\mathsf{T}\mathbf{A}\mathbf{u}}{\mathbf{u}^\mathsf{T}\mathbf{B}\mathbf{u}} = \tilde{\lambda} < 4. \tag{28}$$

2.3.3 Characteristic equation for (23)

For the difference equation in (23) that defines the evolution of $u(t)$, its characteristic equation is $z^2 - (2+\lambda)\,z + 1 = 0$. The roots of this equation are $z_{1,2} = \frac{(2+\lambda)\pm\sqrt{(2+\lambda)^2-4}}{2}$, implying that the characteristic equation can be written as $z^2 - (z_1 + z_2)\,z + (z_1 z_2) = 0$. From the solution for z_1, z_2, we have that, $z_1 + z_2 = 2 + \lambda$, and $z_1 z_2 = 1$.

From (28), we know that the optimal value of the Lagrange multiplier $\lambda = -\tilde{\lambda}$ is in the range $-4 < \lambda < 0$. This corresponds to the case when the two roots of the characteristic equation $z^2 - (2+\lambda)\, z + 1 = 0$ are complex conjugates of each other. Further, because $z_1 z_2 = 1$, these roots lie on the unit circle. Thus, $z_{1,2} = e^{\pm j\theta}$, for some $\theta \in (0, \pi)$. From the the expression for the sum of the roots, we have $z_1 + z_2 = e^{j\theta} + e^{-j\theta} = 2\cos(\theta) = 2 + \lambda$. Solving for λ from this expression yields

$$\lambda = -2\big(1 - \cos(\theta)\big) = -4\sin^2\left(\frac{\theta}{2}\right). \tag{29}$$

2.3.4 Complete $u(t)$ solution

The general solution to (23) is of the form $u(t) = \bar{p}_1\, z_1^t + \bar{p}_2\, z_2^t$, which after substituting the roots, and absorbing $\bar{p}_1, \bar{p}_2$ into real constants p_1, p_2 leads to

$$u(t) = p_1\, \cos(t\,\theta) + p_2\, \sin(t\,\theta). \tag{30}$$

From (30), we have

$$\begin{aligned} u(0) &= p_1, \\ u(1) &= p_1\, \cos(\theta) + p_2\, \sin(\theta). \end{aligned} \tag{31}$$

Recall from (21) that $u(1) - u(0) = \frac{\lambda}{2}\, u(0)$. Plugging (31) and (29) into this result leads to $p_1\, \cos(\theta) + p_2\, \sin(\theta) - p_1 = p_1\, \cos(\theta) - p_1$. This is satisfied if and only if $p_2\, \sin(\theta) = 0$. For any non-trivial $\theta \neq 0$, and $\sin(\theta) \neq 0$, this implies $p_2 = 0$. Therefore, for all θ, we must have $p_2 = 0$, which simplifies (30) to

$$u(t) = p_1\, \cos(t\,\theta). \tag{32}$$

Next, plugging (32) into the final boundary condition $2\big(u(\tau) - u(\tau-1)\big) + \lambda\, u(\tau) = 0$ derived in (22) and simplifying, we obtain the identity $\sin(\theta)\, \sin(\tau\,\theta) = 0$. For any non-trivial $\theta \in (0, \pi)$, such that $\sin(\theta) \neq 0$, this implies that $\tau\,\theta = m\pi$ for any integer $m = 0, 1, 2, \ldots, \tau$. Each m corresponds to the $\tau+1$ ordered eigenvalues $0 \leq \tilde{\lambda}_0 \leq \tilde{\lambda}_1 \leq \ldots \leq \tilde{\lambda}_\tau < \infty$ representing the frequency modes of $\mathbf{B}^{-1}\mathbf{A}$. The integer $m = 1$ would correspond to $\tilde{\lambda}_1$, the smallest positive eigenvalue of $\mathbf{B}^{-1}\mathbf{A}$ which minimizes the regularized objective. Therefore, we have

$$\theta = \frac{\pi}{\tau}. \tag{33}$$

Plugging (33) into (32) and (29) leads to the complete solution

$$u(t) = u(0)\, \cos\big(\pi\, t/\tau\big), \tag{34}$$

and $\lambda = -4\sin^2\big(0.5\,\pi/\tau\big) \approx -\pi^2\tau^{-2}$. More broadly, for $m = 0, 1, 2 \ldots, \tau$, each eigenvalue of $\mathbf{B}^{-1}\mathbf{A}$ is of the form $\tilde{\lambda}_m = 4\sin^2\big(0.5\, m\pi/\tau\big)$.

From (24), recall that $\mathcal{B} = \mathcal{S} + \lambda(\mathcal{Q} - \tilde{c})$. The optimal eigenvector $\mathbf{u}$ is composed of $u(t)$ from (34), so $u(k) = u(0)\, \cos\big(\pi\, k/\tau\big)$, and $u(\tau) = u(0)\, \cos\big(\pi\big) = -u(0)$. Evaluating $\mathcal{Q}$ yields

$$\mathcal{Q} = \sum_{k=1}^{\tau} \tfrac{1}{2}\big(u^2(k) + u^2(k-1)\big) = u^2(0) \sum_{t=0}^{\tau-1} \cos^2\big(\pi\, t/\tau\big) = \tfrac{\tau}{2}\, u^2(0), \tag{35}$$

where we have used the identity $\sum_{t=0}^{\tau-1} \cos^2\big(\pi\, \frac{t}{\tau}\big) = \frac{\tau}{2}$. Hence, the constraint $\mathcal{Q} = \tilde{c}$ implies $\frac{\tau}{2}\, u^2(0) = \tilde{c}$, and therefore $u(0) = \sqrt{\frac{2\tilde{c}}{\tau}}$. Similarly, evaluating $\mathcal{S}$ simplifies to

$$\mathcal{S} = \sum_{k=1}^{\tau} \big(u(k) - u(k-1)\big)^2 = u^2(0) \sum_{k=1}^{\tau} \Big(\cos\big(\pi\, k/\tau\big) - \cos\big(\pi\,(k-1)/\tau\big)\Big)^2. \tag{36}$$

Applying the identity $\cos a - \cos b = -2\sin\left(\frac{a+b}{2}\right)\sin\left(\frac{a-b}{2}\right)$ simplifies (36) to

$$\mathcal{S} = 4\,u^2(0)\,\sin^2\left(0.5\,\tfrac{\pi}{\tau}\right)\sum_{k=1}^{\tau}\sin^2\!\left(\pi\,\tfrac{2k-1}{2\,\tau}\right). \tag{37}$$

Using the identity $\sum_{k=1}^{\tau}\sin^2\!\left(\pi\,\frac{2k-1}{2\,\tau}\right) = \frac{\tau}{2}$ with $\frac{\tau}{2}u^2(0) = \tilde{c}$, and $\lambda = -4\sin^2\left(0.5\,\frac{\pi}{\tau}\right)$, it follows that

$$\mathcal{S} = 4\,u^2(0)\tfrac{\tau}{2}\sin^2\left(0.5\,\tfrac{\pi}{\tau}\right) = -\tilde{c}\,\lambda, \tag{38}$$

and therefore, the minimum value of (19) is $\mathcal{B} = \mathcal{S} + \lambda(\tilde{c} - \tilde{c}) = -\tilde{c}\,\lambda$.

2.3.5 Complete $\delta(t)$ solution

The final step is to obtain the complete solution for $\delta(t) = a + b\,u(t)$ in (17). The corresponding boundary conditions given by $\delta(0) = \mu$, $\delta(\tau) = 0$ leads to

$$\begin{aligned}\delta(0) &= a + b\,u(0) = \mu\\ \delta(\tau) &= a + b\,u(\tau) = a - b\,u(0) = 0,\end{aligned} \tag{39}$$

where from $u(t)$ in (34), we have $u(\tau) = -u(0)$. Solving the simultaneous equation in (39) for a, b leads to the constants $a = \frac{\mu}{2}$, and $b = \frac{a}{u(0)}$. Substituting $a, b, u(t)$ in $\delta(t) = a + b\,u(t)$ leads to the final form

$$\delta(t) = \mu\,\frac{1}{2}\left[1 + \cos\left(\tfrac{\pi\,t}{\tau}\right)\right] = \mu\,\cos^2\left(\tfrac{\pi\,t}{2\tau}\right). \tag{40}$$

Therefore, the regularized minimization problem (20) leads to (40) a half-period second-order raised cosine function commonly referred to as the *cosine annealing* schedule. By inspection, the trust-region constraint $\delta^2(t) \leq \mu^2$ is always satisfied.

Similarly, applying the same regularized Dirichlet energy minimization to $\delta^2(t)$ leads to $\delta^2(t) = \mu^2\cos^2\left(\frac{\pi\,t}{2\tau}\right)$, and therefore yields the half-period first-order raised cosine function

$$\delta(t) = \mu\,\cos\left(\tfrac{\pi\,t}{2\tau}\right). \tag{41}$$

2.4 Structural Properties of the Dirichlet Energy Minimizing Window Functions

The unique solutions that arise from the three variational problems posed in (9), (14), and (19) share several properties. Their boundary conditions of the trust-region are well-defined, and the trust-region vanishes toward the end of the learning window. Each solution can also be written in a common structural form

$$\delta(t) = \mu\,\psi(t), \tag{42}$$

where $0 \leq \psi(t) \leq 1$, $\psi(0) = 1$. The function $\psi(t)$ decays monotonically over the interior of the finite window $t \in \{0, \ldots, \tau\}$ that governs the variation of $\delta^2(t) = \mathbb{E}[\Delta^2(t+1, i)]$. Intuitively, forcing $\delta(t) \to 0$ at the end of the window mirrors classical trust-region methods, where the trust-region radius is tightened as the parameter iterates approach a candidate solution. In the stochastic setting, this encourages smaller parameter updates near the end of the window, so that if $\mathrm{w}(t+1, i)$ is already in the neighborhood of a local optimum, the iterative trust-region is tightened around that neighborhood.

Finally, the average of the variable trust-region radius and its squared radius over the learning window are

$$\frac{1}{\tau}\sum_{t=0}^{\tau-1}\delta(t) = \frac{\mu}{\tau}\sum_{t=0}^{\tau-1}\psi(t), \qquad \text{and} \qquad \frac{1}{\tau}\sum_{t=0}^{\tau-1}\delta^2(t) = \frac{\mu^2}{\tau}\sum_{t=0}^{\tau-1}\psi^2(t). \tag{43}$$

For the window functions (13), (16), (40) and (41), the average $\frac{\mu}{\tau}\sum_{t=0}^{\tau-1}\psi(t)$ converges to strictly positive constant values, since $0 \leq \psi(t) \leq 1$. Specifically, as $\tau \to \infty$, these values are: $\frac{\mu}{2}$ for the linear window (13) and second-order raised cosine window (40), $\frac{2}{3}\mu$ for the square-root linear window (16), and $\frac{2}{\pi}\mu$ for the first-order raised cosine window (41). These asymptotic values follow from the Riemann-sum approximation

$$\lim_{\tau\to\infty} \frac{1}{\tau}\sum_{t=0}^{\tau-1} \psi(t) = \lim_{\tau\to\infty} \frac{1}{\tau}\sum_{t=0}^{\tau-1} \varphi\big(\tfrac{t}{\tau}\big) = \int_0^1 \varphi(x)\, dx, \tag{44}$$

where $\varphi : [0,1] \to \mathbb{R}$ denotes the continuous version of $\psi(t)$, obtained by replacing the sampled normalized index t/τ with the continuous variable $x \in [0,1]$. For example, $\varphi(x) = 1 - x$ for the linear decay window and $\varphi(x) = \cos^2(\frac{\pi}{2}\,x)$ for the cosine-annealing window.

Also, since $0 \le \psi^2(t) \le \psi(t) \le 1$, we have $0 < \frac{1}{\tau}\sum_{t=0}^{\tau-1} \psi^2(t) < \frac{1}{\tau}\sum_{t=0}^{\tau-1} \psi(t)$. Consequently, despite $\delta^2(\tau) = 0$, the average of the squared trust-region radius $\frac{1}{\tau}\sum_{t=0}^{\tau-1} \delta^2(t)$ remains bounded away from zero. In this sense, the window functions avoid an overly aggressive decay of the learning-rate while still enforcing a vanishing trust-region at the end of the window.

In summary, the variational problems and their solutions correspond to three distinct notions of smoothness for the vanishing trust-region radius. Linear decay (13) arises as the unique solution to the Dirichlet energy minimization problem defined on the trust-region radius $\delta(t)$ itself (9), and therefore corresponds to the function $\delta(t)$ with minimal total variation over the learning window. In contrast, square-root linear decay (16) is the solution to the Dirichlet energy minimization problem defined on the squared radius $\delta^2(t)$ (14), and corresponds to the function $\delta(t)$ whose squared radius exhibits minimal total variation over the learning window. As a consequence, this solution maintains larger values than the linear decay function over the whole learning window, but still vanishes to zero the end of the window. Conversely, the raised-cosine (cosine annealing) functions (40) and (41) arise as solutions to regularized Dirichlet energy minimization problem (19) posed respectively on $\delta(t)$ and $\delta^2(t)$. These window functions, therefore correspond to extremal solutions of distinct first-order variational problems, each encoding a different notion of smoothness in the evolution of the variable trust-region radius.

2.5 On the role of Boundary values

The appearance of boundary values in the present framework follows as a direct consequence of the vanishing trust-region formulation. The trust-region radius is designed to begin at some initial value and evolve to a terminal value over a finite learning window. Consequently, the unknown function is not free to vary arbitrarily, but must evolve between fixed endpoints. The variational energy functional then selects, among all trust-region trajectories satisfying these endpoint conditions, the one with minimum energy. This is analogous to the structure encountered in classical variational theory, where the solution is determined jointly by an energy functional and its associated boundary conditions (Iwaniec & Onninen, 2022; Grady & Polimeni, 2010).

In this sense, learning-rate schedules arise as solutions to trust-region variational boundary-value problems. The energy functional determines the smoothness properties of the trajectory, while the boundary values determine the class of admissible trust-region evolutions. Different schedules therefore correspond to different solutions of the same underlying variational formulation and boundary conditions.

3 Generalized forms

Let $1 \le p < \infty$, and define the first-order variational energy minimization functional

$$\min_{0 \le \delta^p(t) \le \mu^p} \sum_{k=1}^{\tau} \big(\delta^p(k) - \delta^p(k-1)\big)^2, \tag{45}$$

Differentiating with respect to $\delta^p(k)$ for any interior index $k \in [1, \tau]$, and setting to zero results in the second-order homogeneous linear difference equation with constant coefficients

$$\delta^p(k+1) - 2\,\delta^p(k) + \delta^p(k-1) = 0. \tag{46}$$

The corresponding characteristic equation is $z^2 - 2\,z + 1 = (z-1)^2 = 0$ which has 1 as a root with multiplicity 2. Therefore, the general solution to (46) must be a polynomial in t, of at most degree 1

$$\delta^p(t) = a_0 + a_1\, t. \tag{47}$$

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

3.1 2-point boundary value problem

Consider the initial and terminal boundary values: $\delta^p(0) = \mu^p$, $\delta^p(\tau) = 0$ over the interval $0 \leq t \leq \tau$, $\tau > 2$. Solving (47), subject to $\delta^p(0) = \mu^p$, $\delta^p(\tau) = 0$ gives $a_0 = \mu^p$, and $a_1 = -\mu^p/\tau$. Therefore, the complete solution to (46) subject to the 2-point boundary value is

$$\delta^p(t) = \mu^p \left(1 - \tfrac{t}{\tau}\right); \quad 0 \leq t \leq \tau, \tag{48}$$

or equivalently,

$$\delta(t) = \mu \sqrt[p]{1 - \frac{t}{\tau}}; \quad 0 \leq t \leq \tau, \tag{49}$$

which is the pth root decay schedule. We can simplify this to a normalized functional form. Let $r = t/\tau$, define an input function $x_2(r) := r \in [0, 1]$, then

$$\delta(r) = \mu \sqrt[p]{1 - x_2(r)}; \quad 0 \leq r \leq 1. \tag{50}$$

3.2 3-point boundary value problem

Consider three fixed boundary values: $\delta^p(0) = 0$, $\delta^p(j) = \mu^p$, $\delta^p(\tau) = 0$, where $0 \leq j < \tau$, $\tau > 3$. This corresponds to two sub-intervals. On the first interval over $0 \leq t \leq j$, the solution is obtained by solving (47) subject to $\delta^p(0) = 0$, $\delta^p(j) = \mu^p$. On the second interval over $j \leq t \leq \tau$, the solution is obtained by solving (47) subject to $\delta^p(j) = \mu^p$, $\delta^p(\tau) = 0$. Combining the piecewise solutions, the complete solution to (46) subject to the 3-point boundary value is

$$\delta^p(t) = \begin{cases} \mu^p \, \frac{t}{j}; & 0 \leq t \leq j, \quad j > 0 \\ \mu^p \, \frac{\tau - t}{\tau - j}; & j \leq t \leq \tau, \quad j \geq 0. \end{cases} \tag{51}$$

This can also be reformulated in the normalized functional form. Let $r = t/\tau \in [0, 1]$, a normalized rise-time $m = j/\tau \in [0, 1)$, then the reparameterized input function $x_3(r; m) \in [0, 1]$ for (51) is

$$x_3(r; m) := \begin{cases} r; & m = 0, \\ \max\{\frac{m-r}{m}, \frac{r-m}{1-m}\}; & 0 < m < 1, \end{cases} \tag{52}$$

and equivalently, (51) becomes

$$\delta(r) = \mu \sqrt[p]{1 - x_3(r; m)}; \quad 0 \leq r \leq 1, \quad 0 \leq m < 1. \tag{53}$$

For its first sub-interval, the hat-shaped (or triangular-like) mapping (51) corresponds to a warmup schedule that reaches μ at time m, and then a decaying schedule to zero in the second sub-interval. Importantly, both (50) and (53) share the same underlying functional form. When $m = 0$, it follows that $x_3(r; m) = x_3(r; 0) = r$. Hence (50) is a special case of (53) when $m = 0$.

3.3 4-point boundary value problem

Consider the four fixed boundary values: $\delta^p(0) = 0$, $\delta^p(j) = \mu^p$, $\delta^p(l) = \mu^p$, $\delta^p(\tau) = 0$, where $0 \leq j \leq l < \tau$, $\tau > 4$. This corresponds to three sub-intervals. On the first interval over $0 \leq t \leq j$, the solution is obtained by solving (47) subject to $\delta^p(0) = 0$, $\delta^p(j) = \mu^p$. On the second interval over $j \leq t \leq l$, the solution is obtained by solving (47) subject to $\delta^p(k) = \mu^p$, $\delta^p(k') = \mu^p$. Finally, for the third interval over $l \leq t \leq \tau$, the solution is obtained by solving (47) subject to $\delta^p(l) = \mu^p$, $\delta^p(\tau) = 0$. Combining the three piecewise solutions, the complete solution to (46) subject to the 4-point boundary value, results in

$$\delta^p(t) = \begin{cases} \mu^p \, \frac{t}{j}; & 0 \leq t \leq j, \quad j > 0 \\ \mu^p; & j \leq t \leq l, \quad j \geq 0 \\ \mu^p \, \frac{\tau - t}{\tau - l}; & l \leq t \leq \tau, \quad l \geq j \geq 0. \end{cases} \tag{54}$$

To derive the normalized functional form, let $r = t/\tau \in [0,1]$, $m = j/\tau \in [0,1)$ indicate the rise time, and $\varepsilon = (l-j)/\tau \in [0,1)$ represent the plateau width, so that $j = m\,\tau$, and $l = (m+\varepsilon)\,\tau$. Then the reparameterized input function $x_4(r; m, \varepsilon) \in [0,1]$ for (54) is

$$x_4(r;m,\varepsilon) := \begin{cases} \max\{0, \frac{r-\varepsilon}{1-\varepsilon}\}; & m = 0, \\ \max\{\frac{m-r}{m},\, 0,\, \frac{r-(m+\varepsilon)}{1-(m+\varepsilon)}\}; & 0 < m < 1, \end{cases} \tag{55}$$

and equivalently (54) can be expressed as

$$\delta(r) = \mu\, \sqrt[p]{1 - x_4(r;m,\varepsilon)}; \quad 0 \le r \le 1, \quad 0 \le m < 1. \tag{56}$$

The resulting function (54) 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 results in a trapezoidal shape for $\delta^p(r)$, and exactly a warmup-stable-decay schedule (Hägele et al., 2024).

Again, observe that for the case $m = 0, \varepsilon = 0$, it follows that $x_4(r;m,\varepsilon) = x_4(r;0,0) = r$, and therefore (50) is a special case of (56). Similarly, when $m \neq 0, \varepsilon = 0$, it follows that $x_4(r;m,\varepsilon) = x_4(r;m,0) = x_3(r;m)$, therefore (53) is a special case of (56). Hence, both (50), (53) and (56) share the same underlying functional form, with (56) having a more general input reparameterization function.

Finally, denote $\delta^p(t) = \mu^p\, \phi(t)$, where $\phi(t) = 1 - x_4(r(t);m,\varepsilon)$, $r(t) = t/\tau$, and $m, \varepsilon \in [0,1)$ then

$$\delta(t) = \mu\, \psi(t) = \mu\, \phi^{\frac{1}{p}}(t) \tag{57}$$

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

Likewise, instead of (45), applying the regularized first-order total variational energy functional to each subinterval of the multi-point boundary-value problem yields the raised-cosine family

$$\phi(t) = \cos^2\big(0.5\,\pi\, x_4(r(t);m,\varepsilon)\big). \tag{58}$$

Remarks: The preceding analysis shows that many commonly used learning-rate schedules arise naturally as solutions to a trust-region variational problem. By treating $\delta^p(t)$ 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 end point requirement $\delta(t) \to 0$ as $t \to \tau$. Consequently, learning-rate schedules may be interpreted as vanishing trust-region functions that progressively tighten the admissible trust-region of the update step over the learning horizon.

The resulting p-th root decay, warmup-decay, and warmup-stable-decay schedules are illustrated in Figure 1 of Appendix A. They correspond respectively to 2-point, 3-point, and 4-point boundary-value problems, yet all share a common variational structure. In particular, the 4-point formulation may be viewed as two adjacent 2-point BVPs connected by a constant plateau. Algorithmic implementations of the normalized BVP profile with the associated Dirichlet-energy and Tikhonov-regularized window functions are given in Appendix B. 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 boundary-value problem formulations or variational functionals would produce different classes of schedules, each reflecting the structure of the underlying variational principle.

Notably, the warmup-stable-decay schedules are also closely related to classical tapering functions used in spectral analysis (Prabhu, 2018; Nuttall, 1981; Blackman & Tukey, 1958). In that setting, the shape of a window determines its spectral smoothing properties. This follows from the fact that multiplying an iteration-dependent quantity by $\delta(t)$ in the time domain corresponds, in the Fourier domain, to convolving its spectrum with the Fourier transform of $\delta(t)$ (Harris, 1978).

3.4 Smoothness Factor

Recall that $\delta^p(t) = \mu^p\, \phi(t)$, for a finite $p \ge 1$. Consider a continuously differentiable (smooth) prototype function $\varphi : [0,1] \to [0,1]$ defined over a fixed window of iterations $t \in [0,\, \tau]$ by

$$\phi(t) = \varphi\big(\tfrac{t}{\tau}\big)\,. \tag{59}$$

Let $x(t) = \frac{t}{\tau}$ denote an ordered uniform discretization of $(0, 1)$ over which $\varphi(x)$ decays monotonically. Assume further that $\varphi'(x)$, the first-derivative of φ with respect to x exists for all $x \in (0, 1)$ and is bounded. The first-order total variational energy functional (45) over the interior $k \in [1, \tau]$ yields

$$\mathcal{D}_{\tau,\phi(t)} = \mu^{2p} \sum_{k=1}^{\tau} \big(\phi(k) - \phi(k-1)\big)^2. \tag{60}$$

By the mean-value theorem, there exists a $\bar{x}_k \in \left[\frac{k-1}{\tau}, \frac{k}{\tau}\right]$, such that $\phi(k) - \phi(k-1) = \varphi\left(\frac{k}{\tau}\right) - \varphi\left(\frac{t-1}{\tau}\right) = \varphi'(\bar{x}_k)\left(\frac{t}{\tau} - \frac{t-1}{\tau}\right) = \frac{1}{\tau}\,\varphi'(\bar{x}_k)$. Plugging this into (60), we obtain

$$\mathcal{D}_{\tau,\phi(t)} = \left(\frac{\mu^p}{\tau}\right)^2 \sum_{k=1}^{\tau} \big(\varphi'(\bar{x}_k)\big)^2, \tag{61}$$

Let

$$L = \frac{1}{\tau} \sum_{t=1}^{\tau} \big(\varphi'(\bar{x}_k)\big)^2. \tag{62}$$

Then (61) can be re-expressed in terms of (62) as

$$\mathcal{D}_{\tau,\phi(t)} = \mu^{2p}\,\tau^{-1}\,L, \tag{63}$$

where the factor L separates the contribution of $\phi(t)$ to the Dirichlet energy from the multiplicative scaling term $\mu^{2p}\,\tau^{-1}$. Smaller values of L correspond to smoother variations, whereas larger values indicate more rapid variations over the window. Furthermore,

$$\lim_{\tau\to\infty} L = \int_0^1 |\varphi'(x)|^2\,dx, \tag{64}$$

showing that L converges to the continuous Dirichlet energy of φ. In the limit $\tau \to \infty$, L becomes independent of the particular choice of scaling parameters μ, p, and τ. Consequently, the factor L may be viewed as the normalized Dirichlet-energy measure of the smoothness of $\phi(t)$ over the interval $[0, \tau]$.

Because $\mathcal{D}_{\tau,\phi(t)} = \mu^{2p}\tau^{-1}L$, the minimization of the Dirichlet energy is equivalent to the minimization of the smoothness factor L. Consequently, among all admissible window functions satisfying the same boundary-values, the variational solutions derived earlier attain the smallest possible value of L and therefore represent the smoothest trust-region radius trajectories under their corresponding variational formulations.

3.5 Vanishing Window Ratios

All trust-region schedules derived so far, including the p-th root decay, warmup-decay, and warmup-stable-decay schedules, satisfy $\delta(t) \to 0$ as $t \to \tau$. In other words, after some $t > t'$ within the learning window, they yield a strictly monotonic decaying sequence converging to the terminal value $\delta(\tau) = 0$. Here, we now study their asymptotic behavior obtained in the limit $\tau \to \infty$.

First, for $t > t'$, we will show that if the ratio $\delta(t)/\delta(t-1) \to 0$ then we must have $\delta(t) \to 0$. Next, we show that $\delta(t) \to 0$ faster than $\delta(t)/\delta(t-1) \to 0$. Finally, we show that a vanishing trust-region induces a stronger exponential contraction of the update step than the fixed trust-region bound in (7).

For this proof, assume that $0 \le \delta(t) \le 1$ and that $\delta(t)/\delta(t-1) \to 0$ as $t \to \infty$. Therefore, by definition of a limit, there exists a fixed $\epsilon \in (0, 1)$ and an index t' such that $\delta(t)/\delta(t-1) \le \epsilon, \forall\, t > t'$. This inequality implies $\delta(t'+1) \le \epsilon\,\delta(t')$. Iterating forward, for any integer $k \ge 1$ we obtain

$$\delta(t'+k) \le \epsilon^k\,\delta(t'). \tag{65}$$

Because $\epsilon \in (0, 1)$, we must have $\epsilon^k \to 0$ as $k \to \infty$, and so in (65) $\delta(t'+k) \to 0$ as $t'+k \to \infty$. Therefore, for $t > t'$, it follows that $\delta(t)/\delta(t-1) \to 0$ implies $\delta(t) \to 0$.

In addition, because $0 \leq \delta(t) \leq 1$ is a monotonically decreasing sequence for $t > t'$, it follows that $0 \leq \delta(t) \leq \delta(t-1) \leq 1$. This directly implies that $0 \leq \delta(t)/\delta(t-1) \leq 1$, and that $\delta(t)\,\delta(t-1) < \delta(t)$ for $t > t'$, so that

$$\delta(t) \leq \frac{\delta(t)}{\delta(t-1)} \leq \epsilon; \quad t > t'. \tag{66}$$

The inequality in (66) shows that the ratio $\delta(t)/\delta(t-1)$ cannot be smaller than $\delta(t)\ \forall t > t'$. From (65), for any $t = t' + k$, we have that

$$\delta(t) \leq \epsilon^{t-t'}\delta(t'); \quad t > t', \tag{67}$$

which implies that $\delta(t)$ decays at an exponential rate proportional to $\epsilon^{t-t'}$.

Also, for $t > t'$, repeated forward iteration of the identity $\delta(t'+1) = \delta(t')\,\frac{\delta(t'+1)}{\delta(t')}$ yields

$$\delta(t) = \delta(t')\prod_{j=t'+1}^{t}\frac{\delta(j)}{\delta(j-1)}; \quad t > t'. \tag{68}$$

From (68), $\delta(t)$ is the product of the past $k = t - t'$ ratios scaled by $0 \leq \delta(t') \leq 1$, where each ratio $\delta(j)/\delta(j-1) < \epsilon$. Since the product decays faster than its individual factors, it follows that $\delta(t) \to 0$ faster than $\delta(t)/\delta(t-1) \to 0, \forall t > t'$.

Furthermore, squaring both sides of (67), results in

$$\delta^2(t) \leq \epsilon^{2\,(t-t')}\delta^2(t'); \quad t > t'. \tag{69}$$

Recall that $\mathbb{E}[\Delta^2(t+1,i)] = \delta^2(t)$, therefore (69) is equivalent to

$$\mathbb{E}[\Delta^2(t+1,i)] \leq \epsilon^{2\,(t-t')}\,\mathbb{E}[\Delta^2(t'+1,i)]; \quad t > t'. \tag{70}$$

Since $0 < \epsilon < 1$, this implies the update step contracts exponentially towards zero in the mean-square sense.

Finally, summing both sides of (70) over the i coordinates results in

$$\mathbb{E}\big[\|\Delta(t+1)\|_2^2\big] \leq \epsilon^{2\,(t-t')}\,\mathbb{E}\big[\|\Delta(t'+1)\|_2^2\big]; \quad t > t'. \tag{71}$$

Therefore, the vanishing trust-region window induces exponential mean-square stability of the update step dynamics.

4 Conclusion

This paper developed a variational trust-region framework for stochastic gradient learning. By formulating the update step as a trust-region problem, we obtained an iteration-dependent learning rate that recovers the core structure of adaptive moment estimation. Extending the framework to a varying trust-region radius led naturally to variational problems posed on the evolution of the trust-region itself.

Within this setting, linear decay, cosine annealing, and their corresponding p-th root variants arise as unique solutions of Dirichlet and regularized Dirichlet energy minimization problems. More generally, the derivations reveal that learning-rate schedules can be interpreted as solutions of trust-region variational boundary-value problems. The resulting framework extends naturally to multi-point boundary-value formulations, yielding warmup-decay, warmup-stable-decay, and related hybrid schedules within the same mathematical structure.

More broadly, learning-rate annealing can be interpreted within the class of total variational boundary-value problems. This places learning-rate scheduling within a well-established mathematical framework in which solutions are determined jointly by an energy functional and fixed boundary values. Hence, this work provides a unified formulation connecting adaptive moment estimation, learning-rate annealing, and trust-region radius variation in stochastic gradient methods.

Overall, the framework is model-agnostic and recovers both existing and previously unexplored learning-rate schedules through design parameters with well-defined ranges on the unit interval. Consequently, learning-rate schedule design becomes a trust-region variational optimization problem rather than a collection of heuristic choices. This provides a principled foundation for the automated construction, analysis, and tuning of learning-rate schedules.

References

Naman Agarwal, Surbhi Goel, and Cyril Zhang. Acceleration via Fractal Learning Rate Schedules. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 87–99, Virtual, USA, July 2021. PMLR. URL `https://proceedings.mlr.press/v139/agarwal21a.html`.

Irwan Bello, William Fedus, Xianzhi Du, Ekin Dogus Cubuk, Aravind Srinivas, Tsung-Yi Lin, Jonathon Shlens, and Barret Zoph. Revisiting ResNets: Improved training and scaling strategies. In *35th Annual Conference on Neural Information Processing Systems*, Virtual, USA, December 2021. Curran Associates, Inc. URL `https://openreview.net/forum?id=dsmxf7FKiaY`.

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=hrOlBgHsMI`.

R. B. Blackman and J. W. Tukey. The measurement of power spectra from the point of view of communications engineering — Part I. *The Bell System Technical Journal*, 37(1):185–282, January 1958. ISSN 0005-8580. doi: 10.1002/j.1538-7305.1958.tb03874.x.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, May 2018.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, NashviIle, TN, 2021. URL `https://openaccess.thecvf.com/content/ICCV2021/html/Caron_Emerging_Properties_in_Self-Supervised_Vision_Transformers_ICCV_2021_paper.html`.

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`.

Baptiste Goujaud, Damien Scieur, Aymeric Dieuleveut, Adrien B. Taylor, and Fabian Pedregosa. Super-acceleration with cyclical step-sizes. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pp. 3028–3065, Virtual, 2022. PMLR. URL `https://proceedings.mlr.press/v151/goujaud22a.html`.

Leo J. Grady and Jonathan R. Polimeni. *Discrete Calculus*. Springer, London, 2010. ISBN 978-1-84996-289-6 978-1-84996-290-2. doi: 10.1007/978-1-84996-290-2.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6645–6649, Vancouver, BC, Canada, 2013. IEEE. doi: 10.1109/ICASSP.2013.6638947.

Alexander Hägele, Elie Bakouch, Atli Kosson, Loubna Ben Allal, Leandro Von Werra, and Martin Jaggi. Scaling Laws and Compute-Optimal Training Beyond Fixed Training Durations. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, Vancouver, Canada, November 2024. URL `https://openreview.net/forum?id=Y13gSfTjGr&noteId=jbPUoMWzI4`.

F.J. Harris. On the use of windows for harmonic analysis with the discrete Fourier transform. *Proceedings of the IEEE*, 66(1):51–83, January 1978. ISSN 1558-2256. doi: 10.1109/PROC.1978.10837.

Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 558–567, Long Beach, CA, 2019. URL `https://openaccess.thecvf.com/content_CVPR_2019/html/He_Bag_of_Tricks_for_Image_Classification_with_Convolutional_Neural_Networks_CVPR_2019_paper.html`.

Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E. Hopcroft, and Kilian Q. Weinberger. Snapshot ensembles: Train 1, get M for free. In *5th International Conference on Learning Representations*, Toulon, France, April 2017. URL `https://openreview.net/forum?id=BJYwwY9ll`.

Tadeusz Iwaniec and Jani Onninen. The Dirichlet principle for inner variations. *Mathematische Annalen*, 383(1-2): 315–351, June 2022. ISSN 0025-5831, 1432-1807. doi: 10.1007/s00208-020-02133-y.

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.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations*, Toulon, France, April 2017. URL `https://openreview.net/forum?id=Skq89Scxx`.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations*, New Orleans, LA, USA, May 2019. URL `https://openreview.net/forum?id=Bkg6RiCqY7`.

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.

Jorge Nocedal and Stephen J. Wright. Trust-Region Methods. In *Numerical Optimization*, pp. 66–100. Springer, New York, NY, 2006. ISBN 978-0-387-40065-5. doi: 10.1007/978-0-387-40065-5_4.

A. Nuttall. Some windows with very good sidelobe behavior. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(1):84–91, February 1981. ISSN 0096-3518. doi: 10.1109/TASSP.1981.1163506.

Neal Parikh and Stephen Boyd. Proximal Algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, January 2014. ISSN 2167-3888. doi: 10.1561/2400000003.

K. M. M. Prabhu. *Window Functions and Their Applications in Signal Processing*. CRC Press, Boca Raton, September 2018. ISBN 978-1-315-21638-6. doi: 10.1201/9781315216386.

Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. Technical report, COURSERA: Neural Networks for Machine Learning, 2012.

Hikaru Umeda and Hideaki Iiduka. Increasing Both Batch Size and Learning Rate Accelerates Stochastic Gradient Descent. *Transactions on Machine Learning Research*, November 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=sbmp55k6iE`.

Linwei Yue, Huanfeng Shen, Jie Li, Qiangqiang Yuan, Hongyan Zhang, and Liangpei Zhang. Image super-resolution: The techniques, applications, and future. *Signal Processing*, 128:389–408, November 2016. ISSN 0165-1684. doi: 10.1016/j.sigpro.2016.05.002.

A Appendix: Window Shapes from a single function

Figure 1 demonstrates how the normalized input function $x_4(r; m, \varepsilon)$ generates a wide range of schedules including: decay, warmup-decay, constant-cooldown, and warmup-stable-decay schedules. By varying m and ε and applying either linear or raised-cosine mappings (or their square-root variants), the resulting windows transition smoothly between standard decay, triangular, half-trapezoidal, and full-trapezoidal profiles.

B Appendix: 2-point to 4-point Boundary-Value Problem

The 2-point, 3-point, and 4-point boundary-value solutions derived in Section 3 are only the simplest members of a broader class of variational trust-region windows that can be formulated.

Note that the 2-point boundary-value formulation has no design parameter to tune. The 3-point boundary-value formulation introduces a single tunable parameter, $0 \leq m \leq 1$. In contrast, the 4-point boundary-value formulation can generate a richer variety of shapes through two tunable design parameters $0 \leq m,\, \varepsilon \leq 1$.

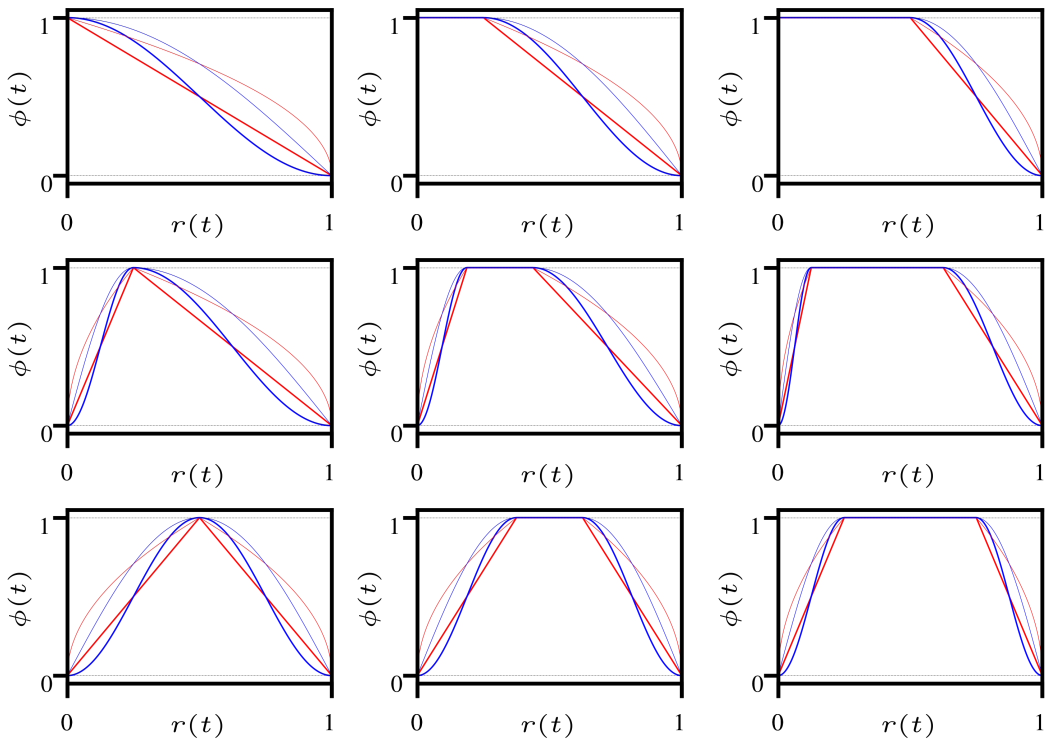

Figure 1: Window shapes obtained by passing $x_4(r(t); m, \varepsilon)$ into the linear function $1 - x_4$ (thick, red), raised-cosine function $\cos^2(0.5\,\pi\, x_4)$ (thick, blue) and their square-root forms (thin lines). **First row**: $m = 0$, $\varepsilon = \{0, 0.25, 0.5\}$ (Left to Right). **Second row**: $m = 0.25$, $\varepsilon = \{0, 0.25, 0.5\}$ (Left to Right). **Third row**: $m = 0.5$, $\varepsilon = \{0, 0.25, 0.5\}$ (Left to Right). Observe the progression from the standard linear decay shape (first row, first column) to the half-trapezoidal shapes, triangular shapes, and full-trapezoidal shapes as m and ε increase.

Specifically, the 4-point design was

$$\delta^p(r) = \begin{cases} 0; & r(t) = 0 \\ \mu^p; & m \le r(t) \le m + \varepsilon \\ 0; & r(t) = 1. \end{cases} \tag{72}$$

where $r \equiv r(t) = t/\tau$, $\tau > 4$ and yielded

$$x_4(r; m, \varepsilon) := \begin{cases} \max\{0, \frac{r-\varepsilon}{1-\varepsilon}\}; & m = 0, \\ \max\{\frac{m-r}{m},\ 0,\ \frac{r-(m+\varepsilon)}{1-(m+\varepsilon)}\}; & 0 < m < 1, \end{cases} \tag{73}$$

Since increasing the number of boundary values introduces additional sub-intervals, more complex trust-region window shapes can be generated without modifying the underlying Dirichlet-energy formulation. Although more variety of shapes can be generated, the trade-off here is that the higher the number of boundary values, the higher the number of tunable design parameters that need to be managed for the resulting schedule function.

B.1 Algorithms

Using the identity $1 - \min\{a, b, c\} = \max\{1 - a,\ 1 - b,\ 1 - c\}$, the BVP profile in (73) can be rewritten as the equivalent min-based implementation given in Algorithm 1.

The Dirichlet-energy solution and its Tikhonov-regularized solution can be implemented as Algorithms 2 and 3, respectively.

Algorithm 1 NORMALIZED BVP PROFILE, $x(r(t); m, \varepsilon)$

Require: $0 \le r(t) \le 1$, $0 \le m, \varepsilon \le 1$

1: $$\ell \leftarrow \begin{cases} \min\left(\dfrac{r(t)}{m},\, 1\right); & m > 0, \\ 1; & m = 0. \end{cases}$$
$$x \leftarrow 1 - \begin{cases} \min\left(\dfrac{1 - r(t)}{1 - (m + \varepsilon)},\, \ell\right); & m + \varepsilon < 1, \\ \ell; & m + \varepsilon \ge 1. \end{cases}$$

2: **return** x

Algorithm 2 P-TH ROOT LINEAR WINDOW, $\phi_{\text{lin}}(x; p)$

Require: $0 \le x \le 1$, $p \ge 1$

1: $y \leftarrow (1 - x)^{1/p}$

2: **return** y

Algorithm 3 P-TH ROOT RAISED-COSINE WINDOW, $\phi_{\text{rcos}}(x; p)$

Require: $0 \le x \le 1$, $p \ge 1$

1: $y \leftarrow \cos^{2/p}(0.5\,\pi\, x)$

2: **return** y

Consequently, several widely used schedule families in the deep learning literature can be viewed as compositions of the 2-point boundary-value-constrained Dirichlet-energy solution. The 3-point boundary-value solution corresponds to two adjacent 2-point solutions, thereby producing a warmup-decay schedule family. Similarly, the 4-point boundary-value solution corresponds to two adjacent 2-point solutions separated by a constant plateau, thereby producing the warmup-stable-decay schedule family. Triangular, half-trapezoidal, and full-trapezoidal shapes therefore arise from the placement of these additional boundary values rather than from a change in the underlying variational principle.