OPT2025: 17th Annual Workshop on Optimization for Machine Learning

Analysis of Schedule-Free Non-Convex Optimization

Connor Brown CONNORBROWN@PRINCETON.EDU
Chi Jin CHIJ@PRINCETON.EDU
Ahmed Khaled AHMED.KHALED@PRINCETON.EDU
Department of Electrical and Computer Engineering, Princeton University

Abstract

The Schedule-Free (SF) method [4] promises optimal performance with hyperparameters that are independent of a known training time horizon T. Nevertheless, non-convex analysis of SF has been limited or reliant on strong global assumptions. Under minimal assumptions of smoothness lower-boundedness, and bounded variance assumptions, we introduce a robust Lyapunov framework for analyzing SF in the non-convex setting, yielding a family of horizon-agnostic $\mathcal{O}(1/\log T)$, $\mathcal{O}(1/T)$, and $O\big(T^{-(1-\alpha)}\big)$ rates under a variety of conditions. Our results — complemented by Performance Estimation Problem (PEP) experiments [5] — extend SF's horizon-free guarantees to smooth non-convex optimization and charts future directions for optimal non-convex rates.

1. Introduction

First-order methods remain the workhorses of modern machine-learning pipelines because each step costs just one gradient while delivering competitive wall-clock performance. Classical theory, however, ties their convergence rates to carefully scheduled step sizes that depend on the total training horizon T. Gradient descent with stepsize $\eta_t \propto 1/\sqrt{T}$ attains the optimal $f(x_T) - f^* = \mathcal{O}(1/\sqrt{T})$ rate on smooth convex objectives [12] and $\min_{0\leq t<T} ||\nabla f(x_t)||^2 = \mathcal{O}(1/T)$ rate for non-convex objectives [8, 10]. Yet in practice T is rarely known in advance — training may stop early for validation, resume for fine-tuning, or continue on a new dataset. To remedy this, Defazio et al. [4] introduced a three-sequence method, Schedule-Free, that proposes to maintain optimality while also being horizon-agnostic.

Our contribution is a Lyapunov-style framework which reduces convergence analysis of the three-sequence Schedule-Free algorithm to a single-step descent inequality with minimal deterministic assumptions, i.e., smoothness and function lower-boundedness. In doing so, we upper-bound the algorithm's performance in non-convex smooth settings under a broad class of horizon-agnostic hyperparameter values. We also utilize the Performance Estimation Problem (PEP) framework for empirically validating our mathematical results on such problem classes, justifying additional boundedness assumptions between Schedule-Free's iterate sequences, and presenting potential future directions for related work. We note that, because most immediately related results are typically presented in the deterministic regime, we state our headline theorems under zero-noise dynamics. Readers interested in the more general stochastic case can find extended proofs in the Appendices.

2. Related Work

2.1. Overview of Schedule-Free

The general Schedule-Free (SF) method maintains three sequences with the following updates:

$$\begin{aligned} y_t &= (1-\beta)z_t + \beta x_t, \\ z_{t+1} &= z_t - \eta_t \nabla f(y_t, \zeta_t), \\ x_{t+1} &= (1-c_{t+1})x_t + c_{t+1}z_{t+1}, \end{aligned} \tag{1}$$

where $x_0 = z_0$ and $\beta \in [0,1]$ controls the interpolation between Stochastic Gradient Descent with Momentum (SGD+M) and Polyak-Ruppert averaging. Specifically, the intuition behind SF can be expressed through the following special cases: $\beta = 1$, $c_{t+1} = 1t+1) \rightarrow$ Primal Averaging (equivalent to SGD+M (Theorem 2) and Stochastic Heavy-Ball Momentum (Theorem 3)) and for $\beta = 0$, $c_{t+1} = 1t+1) \rightarrow$ Polyak-Ruppert (arithmetic iterate averaging). By interpolating between momentum and arithmetic averaging, SF seeks to combine the acceleration effects of momentum on bias decay with the variance reduction properties of averaging.

2.2. Polyak-Ruppert averaging

Polyak–Ruppert (PR) averaging employs a simple arithmetic average of iterates produced by SGD. Given iterates $z_{t+1} = z_t - \eta \nabla f(z_t, \zeta_t)$, the averaged solution is defined as $\bar{z}_T = \frac{1}{T}\sum_{t=1}^{T} z_t$. This can be represented through recursive updates

$$\begin{aligned} z_{t+1} &= z_t - \eta_t \nabla f(z_t, \zeta_t) \\ x_{t+1} &= (1-c_{t+1})x_t + c_{t+1}z_{t+1}, \end{aligned}$$

where choosing $c_{t+1} = 1t+1)$, as in the original SF method, yields the classical arithmetic average. Non-asymptotic analyses show that PR averaging smooths out gradient noise — achieving $f(\bar{x}_T) - f^* = \mathcal{O}(1/\sqrt{T})$ in convex settings and $\mathcal{O}(1/T)$ under strong convexity — without requiring a vanishing stepsize [2, 14]. In practice, these effects translate to reduced sensitivity to stepsize mis-specification, improved training stability, and enhanced generalization in large-scale machine learning. In the case of non-convex problems, PR averaging has a $\mathcal{O}(1/\sqrt{T})$ convergence rate to a stationary point; but for both convex and non-convex problems satisfying additional assumptions (e.g., the Kurdyka–Łojasiewicz (KL) Condition), a $\min_{0\le t<T} ||\nabla f(x_t)||^2 = \mathcal{O}(1/T)$ rate can be achieved [6].

2.3. SGD with Momentum

Stochastic Gradient Descent with Momentum (SGD+M) maintains an auxiliary velocity buffer $m_{t+1} = \lambda_t m_t + \nabla f(x_t, \zeta_t)$ and updates the iterate by $x_{t+1} = x_t - \alpha_t m_{t+1}$. Defazio et al. and Sebbouh et al. independently observed that exactly the same $\{x_t\}$ sequence can be generated by a Stochastic Primal Averaging (SPA) scheme that is algebraically more convenient for analysis:

$$\begin{aligned} z_{t+1} &= z_t - \eta_t \nabla f(x_t, \zeta_t), \\ x_{t+1} &= (1-c_{t+1})x_t + c_{t+1}z_{t+1}, \end{aligned}$$

which is equivalent to SF when $\beta = 1$. The exact correspondence between SPA and SGD+M is denoted by Theorem 2 in Appendix A.

If we consider the parameter identities from Theorem 2, then the SPA iterates coincide with those of SGD+M for all $t \geq 0$. Because the z_t sequence appears naturally in Lyapunov arguments (and is expressed directly in SF) we conduct our non-convex analysis using the SPA form, noting that every result transfers verbatim to momentum via this mapping. The existing theory establishes three benchmark rates for SPA/SGD+M. When f is strongly convex, $f(x_T) - f^\star = O(1/T)$. For purely convex L-smooth objectives, the minimax optimal $f(x_T) - f^\star = O(1/\sqrt{T})$ is achieved. In the non-convex smooth setting, Defazio et al. [3] showed that constant hyperparameters are already sufficient for $\min_{0 \leq t < T} ||\nabla f(x_t)||^2 = O\big(1/\sqrt{T}\big)$, matching the best known stochastic rate without requiring a vanishing schedule. For gradually geometrically-decaying c_{t+1} and η_t, the same $\mathcal{O}(1/T)$ non-convex smooth rate is also achieved. Similarly, Liu et al. also provide optimal convergence rates for SGD+M, conditioned on gradually changing hyperparameters [11]. In fact, for the case of $c_{t+1} = 1t + 1)$ described in Theorem 1 of this paper, the averaging weight c_{t+1} is considered to decay "too fast" under both Defazio and Liu's analyses.

2.4. Stochastic Heavy-Ball Momentum

The stochastic Heavy-Ball method updates the current point by adding a scaled gradient step and a momentum term that re-uses the last displacement: $x_{t+1} = x_t - \lambda_t \nabla f(x_t, \zeta_t) + \theta_t\big(x_t - x_{t-1}\big),$, where $\lambda_t > 0$ is the stepsize and $\theta_t \in [0, 1)$ is the momentum weight; both are usually kept constant in practice. Heavy-Ball can be written in the SPA/averaging form used by SF, so every guarantee we derive for SPA carries over to Heavy-Ball through the algebraic mapping outlined in Theorem 3 in Appendix B. Because SF specializes SPA to $\beta = 1$, this mapping places our non-convex analysis in direct correspondence with the Heavy-Ball literature. For convex Lipschitz objectives, Heavy-Ball attains the optimal $\mathcal{O}(1/T)$ rate for the running average of iterates, and — with a carefully tapered stepsize — the last iterate enjoys the same $\mathcal{O}(1/T)$ guarantee [7].

2.5. Back to Schedule-Free

Schedule–Free (SF) unifies PA (momentum) and PR averaging, as defined above. Moreover, for strongly convex problems, it recovers the standard $\mathcal{O}(1/T)$ sub-optimality rate; and general convex problems, the standard $\mathcal{O}\big(1/\sqrt{T}\big)$ rate [4]. In both cases, SF does this without ever tuning its stepsize to the horizon T.

The performance of SF in the non-convex regime is less widely studied, with developments only recently initiated by Ahn et al. [1]. Assuming f has G-Lipschitz gradients ($\|\nabla f(x)\| \leq G$) and is "well-behaved"[1], Ahn et al. show that SF achieves the optimal $\mathcal{O}(\lambda^{1/2}\epsilon^{-7/2})$ rate for finding (λ, ϵ)-Goldstein stationary points. Moreover, their analysis explains why setting β close to one and using large base optimizer stepsizes — choices that empirically worked well in [4] — are theoretically justified for non-convex optimization. Nevertheless, while this provides the first non-convex guarantees for SF methods, the analysis requires strong global assumptions. Moreover, the parameter choices achieving optimal rates depend on the target accuracy ϵ [1].

Our analysis tackles the non-convex landscape of SF while dispensing with the additional assumptions required in [1] by assuming only that f is L smooth and lower-bounded. Furthermore, our analysis provides a simple and flexible groundwork for analyzing SF in non-convex regimes, in the sense that with a single Lyapunov potential we prove a family of rates for an expanded variety of SF settings.

2.6. The Performance Estimation Problem (PEP) framework

The PEP framework, introduced by Drori and Teboulle [5], is a rigorous framework for analyzing the worst-case convergence rates of gradient-descent style algorithms. To do so, PEP transforms the convergence analysis into a structured semi-definite programming (SDP) formulation which maximizes a worst-case performance metric (e.g., $f(x_T) - f^*$, $||x_T - x_0||^2$, etc.) over a feasible set of functions constrained by their function class (convexity, smoothness properties, etc.). Solving this SDP yields a provably tight worst-case performance bound for the given optimization algorithm over the specified function class [9, 16]. A significant advantage of the PEP framework is its ability to validate theoretical convergence proofs. Indeed, many well-known convergence inequalities, such as those found in Nesterov's classical analyses [13], naturally arise as feasibility conditions within the PEP formulation. Since its introduction, the PEP framework has been successfully applied to a variety of optimization scenarios, including strongly convex, convex, and more recently, certain non-convex optimization settings [17, 18]. These extensions to non-convex problems typically involve adapting the objective function to measure squared gradient norms [15]. For example, Appendix E outlines the full application of PEP to our own analysis in more detail.

3. Main Result

Theorem 1 *Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a non-convex L-smooth function lower-bounded by a finite constant, and $\zeta_1, \ldots, \zeta_T$ is an i.i.d. sequence of random variables such that $\mathbb{E}[\nabla f(y_t, \zeta_t)] = \nabla f(y_t)$ and $\mathbb{E}||\nabla f(y_t) - \nabla f(y_t, \zeta_t)||^2 \leq \sigma^2$. Consider the updates defined in Equation 1. Let $w_{t+1} = \frac{c_t}{2\eta(1-c_t)^2}$. Let t^* be the smallest integer such that for all $t \geq t^*$, $Lc_t(1+\eta) \leq 1/4$ and:*

$$w_t \geq \frac{c_t}{2\eta}(1 + 2\eta L(1-\beta)) + \frac{\eta c_t L^2(1-\beta)^2}{2} + \frac{Lc_t^2}{2}(1+\eta)(1 + 2\eta L^2(1-\beta)^2).$$

Then, for $\{c_{t+1},\ \eta_t\} = \{(t+1)^{-\alpha},\ \eta\}$:

$$\min_{t^* \leq t \leq T-1} \mathbb{E}\|\nabla f(x_t)\|^2 \leq \begin{cases} \frac{4V_{t^*}(1-\alpha)}{\eta(T^{1-\alpha}-(t^*)^{1-\alpha})} + 2\sigma^2\left(1 + \frac{L\eta(1-\alpha)}{1-2\alpha}T^{-\alpha}\right), & \alpha \in [0, 0.5) \\ \frac{4V_{t^*}(1-\alpha)}{\eta(T^{1-\alpha}-(t^*)^{1-\alpha})} + 2\sigma^2\left(1 + \frac{L\eta C_\alpha(1-\alpha)}{T^{1-\alpha}-(t^*)^{1-\alpha}}\right), & \alpha \in [0.5, 1) \\ \frac{4V_{t^*}}{\eta \log(T/t^*)} + 2\sigma^2\left(1 + \frac{L\eta\pi^2/6}{\log(T/t^*)}\right), & \alpha = 1. \end{cases}$$

where $V_{t^*} = f(x_{t^*}) - f^* + w_{t^*}\mathbb{E}\|z_{t^*} - x_{t^*}\|^2$. Under the same conditions, but for $\{c_{t+1},\ \eta_t\} = \{t^\alpha(t+1)^{-\alpha},\ \eta\}$:

$$\min_{t^* \leq t \leq T-1} \mathbb{E}\|\nabla f(x_t)\|^2 \leq \frac{4V_{t^*}}{\eta(T-t^*)-\alpha\eta\log(T/t^*)} + 2\sigma^2(1 + L\eta).$$

Theorem 1 notably spans the specific case of SF, as originally proposed [4], when $\alpha = 1$ and c_{t+1} equals $1t + 1)$. Our analyses suggest that "classical" SF converges to a stationary point at a non-optimal rate of $\mathcal{O}(1/\log T)$ in a non-convex, smooth setting under bounded variance. Our proposed modifications to SF dramatically enhance this rate. Specifically, for $c_{t+1} = (t+1)^{-\alpha}$ and $0 \leq \alpha < 1$, we improve the rate to $\mathcal{O}(1/T^{1-\alpha})$; and for $c_{t+1} = t^\alpha(t+1)^{-\alpha}$ the rate is improved to $\mathcal{O}(1/T)$.

We also note that Theorem 1 shows minimal performance dependence on the interpolation parameter β. In fact, β only affects the time threshold t^* for which convergence is guaranteed to be upper-bounded by the rates we've shown. In this manner, an "optimal" choice of β is one which — conditioned on L and η, achieves a t^* closest to 0. In the PEP experiments that follow, we default to letting $L\eta \leq 1$ and $\beta = 1$ such that $t^* = 2$.

4. PEP Experiments

We validated these results via numerical performance estimation (PEP) studies with results shown in Figures 1 and 2 below. In Figure 2, we note that when c_{t+1} is set to $(t+1)^{-\alpha}$, for $\alpha \in \{0.01, 0.1, 0.5\}$, the maximum value of $||\nabla f(x_t)||^2$ multiplied by $t^{1-\alpha}$ is bounded above by a constant for all t up to $T = 100$ steps. Likewise, Figure 1 demonstrates that for c_{t+1} equaling $t^\alpha(t+1)^{-\alpha}$, for the same values of α, the maximum value of $||\nabla f(x_t)||^2$ multiplied by t is also bounded above by a constant for all $T = 100$ steps. In this way, our numerical PEP experiments validate our analytical results from Theorem 1 up to $T = 100$ steps for a discrete range of α hyperparameter values.

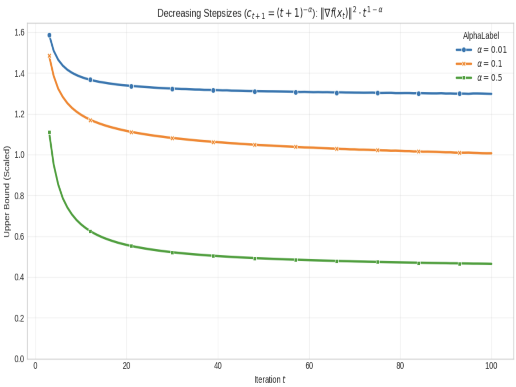

Figure 1: Worst-case bias curves for Theorem 1, $c_{t+1} = \left(\frac{1}{t+1}\right)^\alpha$ for several α values, validating that $||\nabla f(x_t)||^2 \leq \mathcal{O}(1/T^{1-\alpha})$ up to $T = 100$.

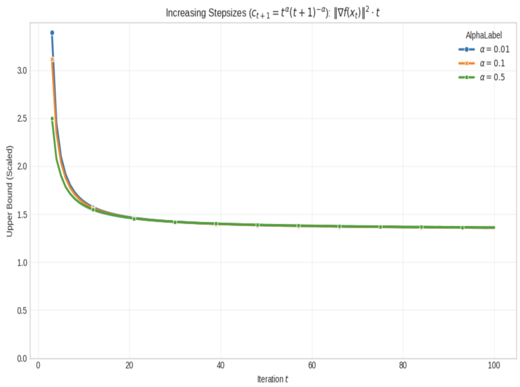

Figure 2: Worst-case bias curves for Theorem 1, $c_{t+1} = \left(\frac{t}{t+1}\right)^\alpha$ for several α values, validating that $||\nabla f(x_t)||^2 \leq \mathcal{O}(1/T)$ up to $T = 100$.

Interestingly, a numerical performance estimate (PEP) study of the "classic" case for SF — i.e., when c_{t+1} equals $(t+1)^{-1}$ — suggests that the true rate may not be improved to the optimal $\mathcal{O}(1/T)$ rate, since Figure 4 insofar shows an unbounded-ness of $||\nabla f(x_t)||^2$ multiplied by t. Nevertheless, the rate provided for this classic SF case is validated up to $T = 100$ steps in Figure 3, where multiplying $||\nabla f(x_T)||^2$ by $\log t$ results in a curve seemingly bounded above by a constant.

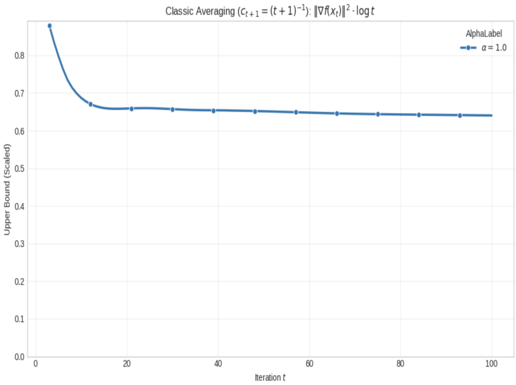

Figure 3: Worst-case bias curves for Theorem 1, $c_{t+1} = (t+1)^{-1}$, validating that $||\nabla f(x_t)||^2 \leq \mathcal{O}(1/\log T)$ up to $T = 100$.

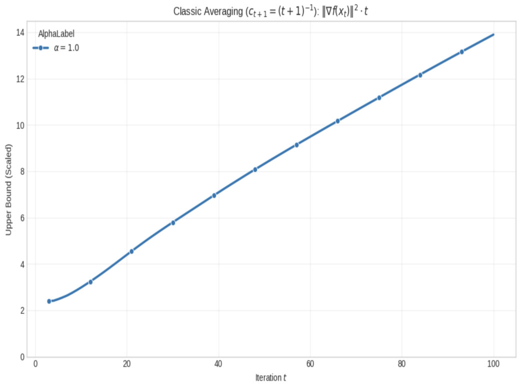

Figure 4: Worst-case bias curves for Theorem 1, $c_{t+1} = (t+1)^{-1}$, suggesting that $||\nabla f(x_t)||^2 \not\leq \mathcal{O}(1/T)$ up to $T = 100$.

5. Discussion and future work

We have developed a unified Lyapunov framework for the Schedule-Free algorithm in the non-convex smooth setting under minimal L-smoothness, lower-boundedness, and bounded variance assumptions. Our main theoretical results from Theorem 1 show that with the classic SF choice $(c_{t+1} = 1t + 1)$, $\eta_t = \eta, \beta_t = 1)$ one attains $\min_{2 \leq t < T} \|\nabla f(x_t)\|^2 = O\big(1/\log T\big)$. More generally, with polynomial-averaging weights $c_{t+1} = (t+1)^{-\alpha}$ and $t^{\alpha}(t+1)^{-\alpha}$, we derive respective $O\big(T^{\alpha-1}\big)$ and $O\big(T^{-1}\big)$ rates for $\alpha \in [0, 1)$.

To support our theoretical bounds, we used the Performance Estimation Problem (PEP) framework [5, 16] to compute worst-case values of $||\nabla f(x_t)||^2$ under the various SF modifications presented in Theorem 1. Across all regimes up to $T = 100$, these PEP SDPs support the rates we provided. In fact, these experiments suggest possible lower-bounds of $\Omega(1)$ on classic SF when $c_{t+1} = (t+1)^{-1}$ since for this specific case, Figure 4 suggests an unbounded curve for $||\nabla f(x_t)||^2$ multiplied by t. Optimistically, while PEP itself faces these finite-horizon drawbacks, its trends support the conjecture that optimal $\mathcal{O}(1/T)$ upper bounds on the convergence of $||\nabla f(x_t)||^2$ can only be approximately achieved through the proposed modifications (via α) to SF's averaging scheme.

Moreover, we have also shown that the interpolation parameter β only modifies the convergence rates of the modified SF algorithm up to a constant factor reflecting a critical point in time t^*, from which convergence is upper-bounded for all $t \geq t^*$. Likewise, for $c_{t+1} = (t+1)^{-\alpha}$, the proposed hyperparameter α embodies a tradeoff between bias convergence of $|||\nabla f(x_t)||^2$ and the persistent noise term σ^2, wherein decreasing α improves the $\mathcal{O}(1/T^{1-\alpha})$ rate, but worsens the multiplicative factor on σ^2. Nevertheless, for this case of $c_{t+1} = (t+1)^{-\alpha}$, the asymptotic noise contribution is lower-bounded by a factor of $2\sigma^2$, and thus no noise "reduction" is readily apparent. Likewise, for $c_{t+1} = t^{\alpha}(t+1)^{-\alpha}$, the asymptotic noise contribution is also lower-bounded by $2\sigma^2$, despite achieving an $\mathcal{O}(1/T)$ rate of bias convergence.

References

[1] Kwangjun Ahn, Gagik Magakyan, and Ashok Cutkosky. General framework for online-to-nonconvex conversion: Schedule-free SGD is also effective for nonconvex optimization, 2024. URL https://arxiv.org/abs/2411.07061.

[2] Francis Bach and Eric Moulines. Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 24:451–459, 2011.

[3] Aaron Defazio. Momentum via Primal Averaging: Theoretical Insights and Learning Rate Schedules for Non-Convex Optimization, 2021. URL https://arxiv.org/abs/2010.00406.

[4] Aaron Defazio, Xingyu Yang, Harsh Mehta, Konstantin Mishchenko, Ahmed Khaled, and Ashok Cutkosky. The Road Less Scheduled, 2024. URL https://arxiv.org/abs/2405.15682v4.

[5] Yoel Drori and Marc Teboulle. Performance of First-Order Methods for Convex Optimization. *Mathematical Programming*, 145(1-2):451–482, 2014.

[6] Sébastien Gadat and Fabien Panloup. Optimal Non-Asymptotic Bound of the Ruppert–Polyak Averaging without Strong Convexity. *Bernoulli*, 23(3):1991–2021, 2017.

[7] Euhanna Ghadimi, Hamid Reza Feyzmahdavian, and Mikael Johansson. Global convergence of the Heavy-ball method for convex optimization, 2014. URL https://arxiv.org/abs/1412.7457.

[8] Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[9] Donghwan Kim and Jeffrey A. Fessler. Exact Worst-case Performance of First-Order Methods for Smooth Convex Optimization. *SIAM Journal on Optimization*, 26(2):1142–1172, 2016.

[10] Yingbin Lei, Michael I. Jordan, and Noureddine El Karoui. Stochastic gradient descent for nonconvex learning without variance reduction. *Machine Learning*, 108(12):2313–2338, 2019.

[11] Yanli Liu, Yuan Gao, and Wotao Yin. An Improved Analysis of Stochastic Gradient Descent with Momentum, 2020. URL https://arxiv.org/abs/2007.07989.

[12] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4): 1574–1609, 2009.

[13] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87 of *Applied Optimization*. Springer, 2013. ISBN 978-1-4020-7553-7.

[14] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning*, pages 249–256, 2012.

[15] Adrien Taylor and Baptiste Goujaud. On worst-case analyses for first-order optimization methods, 2024. URL https://trade-opt-itn.eu/workshop.html. Lecture notes from TraDE-OPT workshop.

[16] Adrien B. Taylor, Julien M. Hendrickx, and Francis Glineur. Exact Worst-Case Performance of First-Order Methods for Composite Convex Optimization. *Mathematical Programming*, 161(1-2):1–33, 2017.

[17] Adrien B. Taylor, Francis Bach, Julien M. Hendrickx, and Francis Glineur. Smooth Strongly Convex Interpolation and Exact Worst-case Performance of First-order Methods. *Mathematical Programming*, 178(1-2):393–418, 2019.

[18] Adrien B. Taylor, Julien M. Hendrickx, and Francis Glineur. Stochastic Performance Estimation Problem: Tight Convergence Guarantees for Stochastic Gradient Methods. *SIAM Journal on Optimization*, 31(3):2323–2353, 2021.

Appendix A. SGD+M and SPA Equivalence [3]

Theorem 2 *Define the SGD+M method by the two sequences:*

$$\begin{aligned} m_{t+1} &= \theta_t m_t + \nabla f(x_t, \zeta_t), \\ x_{t+1} &= x_t - \lambda_t m_{t+1}, \end{aligned}$$

and the SPA sequences as:

$$\begin{aligned} z_{t+1} &= z_t - \eta_t \nabla f(x_t, \zeta_t), \\ x_{t+1} &= (1 - c_{t+1}) x_t + c_{t+1} z_{t+1}. \end{aligned}$$

Consider the case where $m_0 = 0$ for SGD+M and $z_0 = 0$ for SPA. Then if $c_1 = \lambda_0/\eta_0$ and for $t \geq 0$

$$\eta_{t+1} = \frac{\eta_t - \lambda_t}{\theta_{t+1}}, \quad c_{t+1} = \frac{\lambda_t}{\eta_t},$$

the x sequence produced by the SPA method is identical to the x sequence produced by the SGD+M method.

Proof Consider the base case where $x_0 = z_0$. Then for SGD+M:

$$\begin{aligned} m_1 =& \nabla f(x_0, \zeta_t) \\ \therefore x_1 = x_0 &- \lambda_0 \nabla f(x_0, \zeta_t) \end{aligned} \tag{2}$$

and for the SPA form:

$$\begin{aligned} z_1 &= x_0 - \eta_0 \nabla f(x_0, \zeta_0) \\ x_1 &= (1 - c_0) x_0 + c_0 (x_0 - \eta_0 \nabla f(x_0, \zeta_0)) \\ &= x_0 - c_0 \eta_0 \nabla f(x_0, \zeta_0) \end{aligned} \tag{3}$$

Clearly, Equation 2 is equivalent to Equation 7 when $\lambda_0 = c_0 \eta_0$.

Now consider $t > 0$. We will define z_t in terms of quantities in the SGD+M method, then show that with this definition the step-to-step changes in z correspond exactly to the SPA method. In particular, let:

$$z_t = x_t - \left(\frac{1}{c_t} - 1\right) \lambda_{t-1} m_t. \tag{4}$$

Then

$$\begin{aligned} z_{t+1} &= x_{t+1} - \left(\frac{1}{c_{t+1}} - 1\right) \lambda_t m_{t+1} \\ &= x_t - \lambda_t m_{t+1} - \left(\frac{1}{c_{t+1}} - 1\right) \lambda_t m_{t+1} \\ &= z_t + \left(\frac{1}{c_t} - 1\right) \lambda_{t-1} m_t - \frac{\lambda_t}{c_{t+1}} (\theta_t m_t + \nabla f(x_t, \zeta_t)) \\ &= z_t + \left[\left(\frac{1}{c_t} - 1\right) \lambda_{t-1} - \frac{\lambda_t}{c_{t+1}} \theta_t\right] m_t - \frac{\lambda_t}{c_{t+1}} \nabla f(x_t, \zeta_t). \end{aligned} \tag{5}$$

This is equivalent to the SPA step

$$z_{t+1} = z_t - \eta_t \nabla f(x_t, \zeta_t),$$

as long as $\frac{\lambda_t}{c_{t+1}} = \eta_t$ and

$$\begin{aligned} 0 &= \left(\frac{1}{c_t} - 1\right)\lambda_{t-1} - \frac{\lambda_t}{c_{t+1}}\theta_t \\ &= (\eta_{t-1} - \lambda_{t-1}) - \eta_t\theta_t, \\ \text{i.e., } \eta_t &= \frac{\eta_{t-1} - \lambda_{t-1}}{\theta_t}. \end{aligned}$$

Using this definition of the z sequence, we can rewrite the SGD+M x sequence using a rearrangement of Equation 4:

$$\begin{aligned} m_{t+1} &= \left(\frac{1}{c_{t+1}} - 1\right)^{-1}\lambda_t^{-1}(x_{t+1} - z_{t+1}) \\ &= \frac{c_{t+1}}{1 - c_{t+1}}\lambda_t^{-1}(x_{t+1} - z_{t+1}), \end{aligned}$$

as

$$\begin{aligned} x_{t+1} &= x_t - \lambda_t m_{t+1} \\ &= x_t - \frac{c_{t+1}}{1 - c_{t+1}}(x_{t+1} - z_{t+1}) \\ &= x_t - \frac{c_{t+1}}{1 - c_{t+1}}x_{t+1} + \frac{c_{t+1}}{1 - c_{t+1}}z_{t+1} \\ &= (1 - c_{t+1})x_{t+1} + c_{t+1}z_{t+1}, \end{aligned}$$

matching the SPA update. ■

Appendix B. SHBM and SPA Equivalence

Theorem 3 *Define the SHBM method by the sequence:*

$$x_{t+1} = x_t - \lambda_t \nabla f(x_t, \zeta_t) + \theta_t (x_t - x_{t-1}),$$

and the SPA sequences as:

$$\begin{aligned} z_{t+1} &= z_t - \eta_t \nabla f(x_t, \zeta_t), \\ x_{t+1} &= (1 - c_{t+1}) x_t + c_{t+1} z_{t+1}. \end{aligned}$$

Consider the case where $x_0 = 0$ for SHBM and $z_0 = 0$ for SPA. Then if for all $t \geq 0$

$$\lambda_t = c_{t+1}\eta_t, \quad \theta_t = \frac{c_{t+1}(1 - c_t)}{c_t},$$

the x sequence produced by the SPA method is identical to the x sequence produced by the SHBM method.

Proof Consider the base case where $t = 0$. For SHBM with $x_0 = 0$:

$$\begin{aligned} x_1 &= x_0 - \lambda_0 \nabla f(x_0, \zeta_0) + \theta_0 (x_0 - x_{-1}) \\ &= x_0 - \lambda_0 \nabla f(x_0, \zeta_0) \end{aligned} \tag{6}$$

For the SPA form with $z_0 = x_0$:

$$\begin{aligned} z_1 &= z_0 - \eta_0 \nabla f(x_0, \zeta_0) \\ &= x_0 - \eta_0 \nabla f(x_0, \zeta_0) \\ x_1 &= (1 - c_1) x_0 + c_1 z_1 \\ &= x_0 - c_1 \eta_0 \nabla f(x_0, \zeta_0) \end{aligned} \tag{7}$$

Equations 6 and Equation 7 are identical when $\lambda_0 = c_1 \eta_0$.

Now consider $t > 0$. We will define z_t in terms of quantities in the SHBM method, then show that with this definition the step-to-step changes in z correspond exactly to the SPA method. Let:

$$z_t = x_t + \frac{1 - c_t}{c_t}(x_t - x_{t-1}). \tag{8}$$

Then:

$$\begin{aligned} z_{t+1} &= x_{t+1} + \frac{1 - c_{t+1}}{c_{t+1}}(x_{t+1} - x_t) \\ &= \left(1 + \frac{1 - c_{t+1}}{c_{t+1}}\right) x_{t+1} - \frac{1 - c_{t+1}}{c_{t+1}} x_t \\ &= \frac{1}{c_{t+1}} x_{t+1} - \frac{1 - c_{t+1}}{c_{t+1}} x_t \end{aligned} \tag{9}$$

Substituting the SHBM update for x_{t+1}:

$$\begin{aligned} z_{t+1} &= \frac{1}{c_{t+1}} \left[x_t - \lambda_t \nabla f(x_t, \zeta_t) + \theta_t (x_t - x_{t-1}) \right] - \frac{1 - c_{t+1}}{c_{t+1}} x_t \\ &= z_t + \frac{1 - c_t}{c_t}(x_t - x_{t-1}) - \frac{\lambda_t}{c_{t+1}} \nabla f(x_t, \zeta_t) \\ &\quad + \frac{\theta_t}{c_{t+1}}(x_t - x_{t-1}) - \frac{1 - c_{t+1}}{c_{t+1}}(x_t - z_t) \end{aligned} \tag{10}$$

This simplifies to the SPA update $z_{t+1} = z_t - \eta_t \nabla f(x_t, \zeta_t)$ when:

$$\frac{\lambda_t}{c_{t+1}} = \eta_t \quad \text{and} \quad \theta_t = \frac{c_{t+1}(1-c_t)}{c_t}.$$

Finally, the SHBM x update can be rewritten using Equation 8:

$$\begin{aligned} x_{t+1} &= (1-c_{t+1})x_t + c_{t+1}z_{t+1} \\ &= (1-c_{t+1})x_t + c_{t+1}\left(z_t - \eta_t \nabla f(x_t, \zeta_t)\right), \end{aligned}$$

which matches the SPA update by construction. ■

Appendix C. Main Theorem Proof

Theorem 1 *Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a non-convex L-smooth function lower-bounded by a finite constant, and $\zeta_1, \dots, \zeta_T$ is an i.i.d. sequence of random variables such that $\mathbb{E}[\nabla f(y_t, \zeta_t)] = \nabla f(y_t)$ and $\mathbb{E}||\nabla f(y_t) - \nabla f(y_t, \zeta_t)||^2 \le \sigma^2$. Consider the updates defined in Equation 1. Let $w_{t+1} = \frac{c_t}{2\eta(1-c_t)^2}$. Let t^* be the smallest integer such that for all $t \ge t^*$, $Lc_t(1+\eta) \le 1/4$ and:*

$$w_t \ge \frac{c_t}{2\eta}(1 + 2\eta L(1-\beta)) + \frac{\eta c_t L^2(1-\beta)^2}{2} + \frac{Lc_t^2}{2}(1+\eta)(1 + 2\eta L^2(1-\beta)^2).$$

Then, for $\{c_{t+1},\ \eta_t\} = \{(t+1)^{-\alpha},\ \eta\}$:

$$\min_{t^* \le t \le T-1} \mathbb{E}\|\nabla f(x_t)\|^2 \ \le \ \begin{cases} \frac{4V_{t^*}(1-\alpha)}{\eta(T^{1-\alpha}-(t^*)^{1-\alpha})} + 2\sigma^2\left(1 + \frac{L\eta(1-\alpha)}{1-2\alpha}T^{-\alpha}\right), & \alpha \in [0, 0.5) \\ \frac{4V_{t^*}(1-\alpha)}{\eta(T^{1-\alpha}-(t^*)^{1-\alpha})} + 2\sigma^2\left(1 + \frac{L\eta C_\alpha(1-\alpha)}{T^{1-\alpha}-(t^*)^{1-\alpha}}\right), & \alpha \in [0.5, 1) \\ \frac{4V_{t^*}}{\eta\log(T/t^*)} + 2\sigma^2\left(1 + \frac{L\eta\pi^2/6}{\log(T/t^*)}\right), & \alpha = 1. \end{cases}$$

where $V_{t^*} = f(x_{t^*}) - f^* + w_{t^*}\mathbb{E}\|z_{t^*} - x_{t^*}\|^2$. Under the same conditions, but for $\{c_{t+1},\ \eta_t\} = \{t^\alpha(t+1)^{-\alpha},\ \eta\}$:

$$\min_{t^* \le t \le T-1} \mathbb{E}\|\nabla f(x_t)\|^2 \le \frac{4V_{t^*}}{\eta(T-t^*)-\alpha\eta\log(T/t^*)} + 2\sigma^2(1 + L\eta).$$

Proof We analyze the algorithm defined by $y_t = (1-\beta)z_t + \beta x_t$, $z_{t+1} = z_t - \eta\nabla f(y_t, \zeta_t)$, and $x_{t+1} = (1-c_t)x_t + c_t z_{t+1}$. We employ the Lyapunov function $V_t = f(x_t) - f^* + w_t\|e_t\|^2$, where $e_t = z_t - x_t$. We define the weight sequence recursively as $w_{t+1} = \frac{c_t}{2\eta(1-c_t)^2}$.

Step 1: Objective Descent. By the update rule, $x_{t+1} - x_t = c_t(z_{t+1} - x_t) = c_t(e_t - \eta\nabla f(y_t, \zeta_t))$. Applying L-smoothness of f:

$$f(x_{t+1}) \le f(x_t) + \langle\nabla f(x_t), x_{t+1} - x_t\rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2.$$

Taking expectations conditioned on time t ($\mathbb{E}_t$) and letting $g_t = \nabla f(y_t, \zeta_t)$:

$$\begin{aligned}\mathbb{E}_t[f(x_{t+1})] \le f(x_t) &+ c_t\langle\nabla f(x_t), e_t\rangle - \eta c_t\langle\nabla f(x_t), \nabla f(y_t)\rangle \\ &+ \frac{Lc_t^2}{2}\left(\|e_t\|^2 - 2\eta\langle e_t, \nabla f(y_t)\rangle + \eta^2\|\nabla f(y_t)\|^2 + \eta^2\sigma^2\right).\end{aligned}$$

We expand the inner product $-\eta c_t\langle\nabla f(x_t), \nabla f(y_t)\rangle = -\frac{\eta c_t}{2}(\|\nabla f(x_t)\|^2 + \|\nabla f(y_t)\|^2 - \|\nabla f(x_t) - \nabla f(y_t)\|^2)$. Using $y_t - x_t = (1-\beta)e_t$ and smoothness, $\|\nabla f(x_t) - \nabla f(y_t)\|^2 \le L^2(1-\beta)^2\|e_t\|^2$. Using Young's inequality, $-2\eta\langle e_t, \nabla f(y_t)\rangle \le \eta\|e_t\|^2 + \eta\|\nabla f(y_t)\|^2$. Substituting these into the objective bound:

$$\begin{aligned}\mathbb{E}_t[f(x_{t+1})] \le f(x_t) + c_t\langle\nabla f(x_t), e_t\rangle - \frac{\eta c_t}{2}\|\nabla f(x_t)\|^2 - \frac{\eta c_t}{2}\|\nabla f(y_t)\|^2 + \frac{\eta c_t L^2(1-\beta)^2}{2}\|e_t\|^2 \\ + \frac{Lc_t^2}{2}\left((1+\eta)\|e_t\|^2 + \eta(1+\eta)\|\nabla f(y_t)\|^2 + \eta^2\sigma^2\right). \qquad (11)\end{aligned}$$

Step 2: Error Contraction. We observe $e_{t+1} = z_{t+1} - x_{t+1} = (1-c_t)(z_{t+1} - x_t) = (1-c_t)(e_t - \eta g_t)$. Squaring and taking expectations:

$$\mathbb{E}_t\|e_{t+1}\|^2 = (1-c_t)^2\left(\|e_t\|^2 - 2\eta\langle e_t, \nabla f(y_t)\rangle + \eta^2\|\nabla f(y_t)\|^2 + \eta^2\sigma^2\right).$$

Decomposing the inner product via $\nabla f(y_t) = \nabla f(x_t) + (\nabla f(y_t) - \nabla f(x_t))$ yields $-2\eta\langle e_t, \nabla f(y_t)\rangle \leq -2\eta\langle e_t, \nabla f(x_t)\rangle + 2\eta L(1-\beta)\|e_t\|^2$. Multiplying by w_{t+1}:

$$w_{t+1}\mathbb{E}_t\|e_{t+1}\|^2 \leq w_{t+1}(1-c_t)^2\left((1+2\eta L(1-\beta))\|e_t\|^2 - 2\eta\langle e_t, \nabla f(x_t)\rangle + \eta^2\|\nabla f(y_t)\|^2 + \eta^2\sigma^2\right). \tag{12}$$

Step 3: Lyapunov Combination. The choice $w_{t+1} = \frac{c_t}{2\eta(1-c_t)^2}$ ensures the term $-w_{t+1}(1-c_t)^2 2\eta\langle e_t, \nabla f(x_t)\rangle$ cancels $c_t\langle \nabla f(x_t), e_t\rangle$ from (11). We aggregate the coefficients of $\|\nabla f(y_t)\|^2$:

$$C_{\nabla y} = -\frac{\eta c_t}{2} + \frac{Lc_t^2\eta}{2}(1+\eta) + w_{t+1}(1-c_t)^2\eta^2 = -\frac{\eta c_t}{2} + \frac{Lc_t^2\eta}{2}(1+\eta) + \frac{\eta c_t}{2} = \frac{Lc_t^2\eta}{2}(1+\eta).$$

Since $C_{\nabla y} > 0$, we use $\|\nabla f(y_t)\|^2 \leq 2\|\nabla f(x_t)\|^2 + 2L^2(1-\beta)^2\|e_t\|^2$. The total coefficient for $\|\nabla f(x_t)\|^2$ is $-\frac{\eta c_t}{2} + 2C_{\nabla y} = -\frac{\eta c_t}{2} + L\eta c_t^2(1+\eta) = -\frac{\eta c_t}{2}[1 - 2Lc_t(1+\eta)]$. For $t \geq t^*$, we have $Lc_t(1+\eta) \leq \frac{1}{4}$, ensuring the coefficient is $\leq -\frac{\eta c_t}{4}$. The coefficient for $\|e_t\|^2$, denoted C_e, collects terms from f-descent, e-contraction, and the substitution of $\|\nabla f(y_t)\|^2$. For $\mathbb{E}[V_{t+1}] \leq V_t$, we require $C_e \leq 0$, which yields the condition defining t^*:

$$w_t \geq \frac{c_t}{2\eta}(1+2\eta L(1-\beta)) + \frac{\eta c_t L^2(1-\beta)^2}{2} + \frac{Lc_t^2}{2}(1+\eta)(1+2\eta L^2(1-\beta)^2).$$

Substituting $w_t = \frac{c_{t-1}}{2\eta(1-c_{t-1})^2}$, this defines t^* explicitly. Since w_t grows while the RHS is bounded for decaying c_t, such a t^* exists. For $t \geq t^*$, the descent is:

$$\frac{\eta c_t}{4}\mathbb{E}\|\nabla f(x_t)\|^2 \leq V_t - \mathbb{E}_t[V_{t+1}] + \frac{\sigma^2}{2}(Lc_t^2\eta^2 + \eta c_t).$$

Summing from $t = t^*$ to $T-1$:

$$\min_{t^* \leq t < T}\mathbb{E}\|\nabla f(x_t)\|^2 \leq \frac{4V_{t^*}}{\eta\sum_{t=t^*}^{T-1} c_t} + \frac{2\sigma^2\sum_{t=t^*}^{T-1}(L\eta^2 c_t^2 + \eta c_t)}{\eta\sum_{t=t^*}^{T-1} c_t}.$$

Case 1: $c_t = t^{-\alpha}$**.** The denominator is bounded by $\sum_{t=t^*}^{T-1} c_t \geq \int_{t^*}^{T} x^{-\alpha}dx$. We analyze the overall convergence rate by explicitly combining the bias and variance terms:

- For $\alpha \in [0, 0.5)$, using $\sum_{t=t^*}^{T-1} c_t^2 \leq \frac{T^{1-2\alpha}}{1-2\alpha}$:

$$\min\mathbb{E}\|\nabla f(x_t)\|^2 \leq \frac{4V_{t^*}(1-\alpha)}{\eta(T^{1-\alpha} - (t^*)^{1-\alpha})} + 2\sigma^2\left(1 + \frac{L\eta(1-\alpha)}{1-2\alpha}T^{-\alpha}\right).$$

- For $\alpha \in (0.5, 1)$, noting $\sum c_t^2 \leq C_\alpha$:

$$\min\mathbb{E}\|\nabla f(x_t)\|^2 \leq \frac{4V_{t^*}(1-\alpha)}{\eta(T^{1-\alpha} - (t^*)^{1-\alpha})} + 2\sigma^2\left(1 + \frac{L\eta C_\alpha(1-\alpha)}{T^{1-\alpha} - (t^*)^{1-\alpha}}\right).$$

- For $\alpha = 1$, where $\sum c_t \geq \log(T/t^*)$ and $\sum c_t^2 \leq \pi^2/6$:

$$\min\mathbb{E}\|\nabla f(x_t)\|^2 \leq \frac{4V_{t^*}}{\eta\log(T/t^*)} + 2\sigma^2\left(1 + \frac{L\eta\pi^2/6}{\log(T/t^*)}\right).$$

Case 2: $c_t = t^\alpha(t+1)^{-\alpha}$**.** Here $c_t = (1+\frac{1}{t})^{-\alpha} \geq 1 - \frac{\alpha}{t}$. Summing this explicitly: $\sum_{t=t^*}^{T-1} c_t \geq (T-t^*) - \alpha \sum_{t=t^*}^{T-1} \frac{1}{t} \geq T - t^* - \alpha \log(T/t^*)$. For the variance term, we simply bound $c_t \leq 1$, yielding $\sum(L\eta^2 c_t^2 + \eta c_t) \leq (L\eta^2 + \eta)\sum c_t$. Substituting this into the second term of the bound:

$$\frac{2\sigma^2(L\eta^2+\eta)\sum c_t}{\eta \sum c_t} = 2\sigma^2(L\eta+1).$$

This variance bound is constant and tight for large T since $c_t \to 1$. Combining with the bias denominator:

$$\min \mathbb{E}\|\nabla f(x_t)\|^2 \leq \frac{4V_{t^*}}{\eta(T - t^* - \alpha\log(T/t^*))} + 2\sigma^2(1+L\eta).$$

■

Appendix D. Experimental Setup and Code

Using the SDP formulations presented in Appendix E, we applied the Python 3.8 PEPit library to validate our analysis of Schedule-Free under the various choices of hyperparameters we considered.

For each choice of hyperparameter, we ran PEP for $n = 100$ iterations, with $\beta = 1$ and $L\eta \leq 1$ to yield $t^* = 2$ (see proofs in earlier Appendices), sweeping values of $\alpha = \{0.01,\ 0.1\, 0.5\}$ for both $c_{t+1} = (t+1)^{-\alpha}$ and $t^{\alpha}(t+1)^{-\alpha}$. We also ran a PEP analysis for classic SF, when $c_{t+1} = (t+1)^{-1}$. The full code can be found at https://github.com/cbrownaz24/SF-non-convex.