Safely Exploring Large Momentum Steps with Stochastic Curve Searches

Anonymous authors
Paper under double-blind review

Abstract

The use of stochastic line searches has emerged as an effective safeguard strategy for employing large learning rates in the training of deep models via stochastic gradient descent. However, exploiting this approach with different search directions is not straightforward; momentum type directions, in particular, pose several challenges in this regard both from the theoretical and the computational sides. In this work, we present stochastic curve search (SCS) as a generalization of the stochastic line search. SCS allows to evaluate updates along directions that may not be of descent, while still ensuring the sufficient decrease of the mini-batch objective at each iteration. We show that the proposed framework is well-defined and that, under standard assumptions, the method converges in expectation. We also empirically establish that using SCS alongside several momentum based algorithms allows the employment of aggressive hyperparameters, improving either the stability or the speed of the training process. The resulting algorithmic framework is demonstrated to perform competitively against state-of-the-art methods, achieving interesting results in terms of both efficiency and effectiveness across a diverse set of learning benchmarks.

1 Introduction

Training of deep learning models amounts to the solution of a nonlinear, nonconvex, unconstrained finite-sum optimization problem of the form

$$\min_{x\in\mathbb{R}^n} \; f(x) = \frac{1}{N}\sum_{i=1}^{N} f_i(x), \tag{1}$$

where each term $f_i : \mathbb{R}^n \to \mathbb{R}$ in the sum represents the contribution to the overall loss function of each training data point and the number N, denoting the number of data points, is thus extremely large in modern applications (Bottou et al., 2018; Goodfellow et al., 2016).

In this scenario, Stochastic Gradient Descent (SGD) method and its variants have established themselves as preferred choices for carrying out optimization (Bottou et al., 2018), especially because of their very limited per-iteration cost compared to the so called full-batch algorithms; moreover, convergence rates comparable to full-batch methods have been shown to hold under regularity assumptions linking the full f and the individual samples f_i (Schmidt & Roux, 2013; Vaswani et al., 2019a; Khaled & Richtárik, 2023). In fact, fast (linear) convergence rates have even been shown to hold for the over-parametrized setting in the interpolation regime (Schmidt & Roux, 2013; Ma et al., 2018; Vaswani et al., 2019a). In this latter scenario, which is common when dealing with the very large learning architectures employed in recent years, models are powerful enough to perfectly fit all training data. In mathematical terms, this is known as the *interpolation condition* (Mishkin, 2020), which states that if $x^\star \in \arg\min_x f(x)$, then $x^\star \in \arg\min_x f_i(x)$ for all $i = 1, \dots, N$. In other terms, at the optimal solution every single component of the finite-sum is simultaneously minimized.

The choice of the learning rate schedule is well-known to be crucial for obtaining properly trained networks (Schmidt et al., 2021): although constant or descending sequences could be theoretically sound, in practice such approaches require multiple runs of trial and error to tune hyperparameters, making the entire process

not only slow but also extremely expensive. Algorithms based on adaptive diagonal pre-conditioners (Duchi et al., 2011; Kingma & Ba, 2014; Loshchilov & Hutter, 2017) have thus emerged as reference optimizers for deep learning, as they exhibit fast convergence and training stability without the need of a particularly careful tuning. More recent research dealt with the specificity of the over-parametrized case to devise tailored strategies for the adaptive selection of the stepsize in SGD according to model-based (Loizou et al., 2021; Orvieto et al., 2022) or line search strategies (Vaswani et al., 2019b; Galli et al., 2023).

In the perspective of accelerating the performance of SGD, the addition of a momentum term (Polyak, 1964) has long been known to be extremely beneficial (Sutskever et al., 2013; Sebbouh et al., 2021; Tseng, 1998; Jelassi & Li, 2022; Gitman et al., 2019). Most often, momentum appears in popular frameworks - such as in Adam (Kingma & Ba, 2014) - as a moving average of past gradients, rather than as a plain heavy-ball term; the search direction at iteration k is therefore commonly given by one of the following rules

$$d_k = -\nabla f_k(x^k) + \beta_k(x^k - x^{k-1}), \tag{2}$$

$$d_k = -(1-\beta_k)\nabla f_k(x^k) + \beta_k d_{k-1}, \tag{3}$$

where f_k denotes the stochastic unbiased approximation of f sampled at iteration k.

Adaptive stepsize selection methods however have mainly been devised for the vanilla SGD direction. Only recently some studies have appeared pointing at suitable ways to integrate momentum terms with adaptive learning rates (Sebbouh et al., 2021). In particular, generalized Polyak's stepsizes for momentum-type directions were proposed by Topollai & Choromanska (2025); Wang et al. (2023); Oikonomou & Loizou (2024); Schaipp et al. (2023), based on different points of view and under different sets of assumptions. For these approaches, however, we have convergence results under convexity assumptions (Schaipp et al., 2023; Oikonomou & Loizou, 2024; Wang et al., 2023) or for bounded stochastic gradients (Topollai & Choromanska, 2025). Moreover, for most of these methods, clipping techniques or conservative choices for the hyperparameter need to be used in practice, so that aggressive setups, which could result in particularly efficient training, are avoided for the sake of a more likely stable behavior.

Meanwhile, the employment of momentum terms within stochastic line search based frameworks working in the over-parametrized regime has been studied by Fan et al. (2023); Lapucci & Pucci (2025; 2026). This integration is computationally intriguing as, similarly to the vanilla SGD case, (non-monotone) line searches can be employed to devise fast and provably convergent methods able to simultaneously exploit aggressive adaptive stepsizes (e.g., Polyak's stepsize as done by Galli et al., 2023) and the momentum term (which was empirically shown to provide further speedup by Berrada et al., 2021).

Unfortunately, as noted by Lapucci & Pucci (2026), stochastic line searches and momentum terms are somewhat in contrast with each other: in order to ensure that line searches are well-defined, at each iteration the search direction should be of descent with respect to the sampled approximation f_k at x^k; however, since momentum carries information about previous iterations, which are in practice based on previously sampled approximations of f (i.e., different minibatches of data points), it rarely is a significantly descent direction for f_k; therefore, a correct use of line searches requires to often heavily modify the direction of the form in equation 3 (or to select very small values for β_k), up to the point that the benefits of the momentum term vanish altogether. While Lapucci & Pucci (2026) proposed a strategy based on minibatch persistency to alleviate this issue, this solution is arguably impractical with very large models.

In this paper we therefore propose an alternative approach, leveraging stochastic curve searches, to define adaptive learning rate momentum methods that do not require the first tentative update to define a descent direction. The proposed stochastic curve search generalizes the technique proposed by Donnini et al. (2025) for full-batch algorithms: arbitrary updates can be tried, and only in case of failure of a sufficient decrease condition, backtrack is performed. Further evaluations follow a curve that is tangent to the negative gradient at the current iterate, so that for small steps gradient descent updates are roughly recovered.

The introduction of the stochastic curve search, which is provably sound and does not spoil the convergence guarantees of stochastic line search approaches, allows us to devise momentum-based optimizers coupling theoretical guarantees under standard assumptions and good computational performance compared to other related strategies and in general to state-of-the-art optimizers. In particular, we are presenting numerical evidence that non-monotone curve searches make other momentum-based approaches from the literature

more robust, allowing to leverage more aggressive choices of hyperparameters, which in turn lead to faster convergence. The resulting algorithmic framework is then shown to be competitive w.r.t. state-of-the-art methods both in terms of efficiency and effectiveness on a diverse benchmark of learning tasks.

The rest of the paper is organized as follows: in Section 2 we present the main concepts and the conditions that are commonly used to ensure convergence of stochastic gradient based methods for learning tasks. In Section 3 we introduce the stochastic curve search (SCS) framework and theoretically analyze its properties, showing that a) the curve search procedure is well-defined and b) under reasonable assumptions the overall framework is provably convergent in expectation with classes of nonconvex functions. In Section 4 we empirically validate the SCS framework in large scale supervised learning problems, showing its practical effectiveness. Finally, in Section 5 we give some concluding remarks.

2 Preliminaries

Finite-sum problems of the form (1) are commonly solved by iterative methods that define the update vector to be applied at solution x^k based on the information provided by stochastic estimators of f and ∇f, which we denote respectively by f_k and g_k. The estimators are assumed to be conditionally unbiased, i.e., if we denote the conditional expectation by $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot|x^k]$, we have $\mathbb{E}_k[f_k] = f$ and $\mathbb{E}_k[g_k] = \nabla f$. Typically, the estimators are obtained by uniform sampling of the terms f_i in the sum, i.e., drawing samples from the training set, so that $f_k(x) = \frac{1}{|B_k|}\sum_{i\in B_k} f_i(x)$, with $B_k \subset \{1,\dots,N\}$, and $g_k(x) = \nabla f_k(x)$. Throughout this work, we assume that $f_k(x^k)$ and $g_k(x^k)$ are obtained accordingly; moreover we assume that f is L-smooth, i.e., $\|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\|$ for any $x, y \in \mathbb{R}^n$, and also that each function f_i is L_i-smooth. We denote by $L_{\max} = \max_{i\in\{1,\dots,N\}} L_i$, and it can be easily verified that $L \leq L_{\max}$. Moreover, we shall note that f_k as defined earlier is L_k-smooth, with $L_k \leq \frac{1}{|B_k|}\sum_{i\in B_k} L_i$.

We now introduce recurrent concepts that are often assumed to hold in related literature works.

Definition 1. *A function $f : \mathbb{R}^n \to \mathbb{R}$ of the form (1) satisfies the strong growth condition (SGC) if there exists $\rho > 0$ such that $\mathbb{E}_k[\|g_k(x)\|^2] \leq \rho\|\nabla f(x)\|^2$ for any $x \in \mathbb{R}^n$ and for any estimator g_k.*

Definition 2. *A function $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the Polyak-Lojasiewicz (PL) condition if there exists $\mu > 0$ such that $2\mu(f(x) - f(x^\star)) \leq \|\nabla f(x)\|^2$ for all $x \in \mathbb{R}^n$, with $x^\star \in \arg\min_x f(x)$.*

Definition 3. *A function $f : \mathbb{R}^n \to \mathbb{R}$ of the form (1) satisfies the (minimizer) interpolation condition if*

$$x^\star \in \arg\min_x f(x) \implies x^\star \in \arg\min_x f_i(x), \qquad \text{for all } i = 1,\dots,N.$$

In this kind of scenarios, stochastic gradient descent (SGD) algorithms have been designed that make use of stochastic line searches to determine the stepsize in the update rule $x^{k+1} = x^k - \alpha_k g_k(x^k)$. In particular, stochastic versions of the monotone (Vaswani et al., 2019b) and non-monotone (Galli et al., 2023) Armijo rules have been considered, imposing a sufficient decrease condition w.r.t. current f_k; the stepsize α_k is selected via a backtracking procedure that stops as soon as the following condition is met:

$$f_k(x^k - \alpha_k g_k(x^k)) \leq C_k - \sigma\alpha_k\|g_k(x^k)\|^2,$$

where $\sigma \in (0,1)$, $C_k = f_k(x^k)$ for the monotone line search and $C_k \geq f_k(x^k)$ is a suitable safeguard (see, e.g., the rule of Zhang & Hager 2004 used by Galli et al. 2023) in non-monotone approaches. For SGD with monotone stochastic line searches, Vaswani et al. (2019b) proved convergence of the gradient norm in expectation at a sublinear rate assuming SGC, whereas Galli et al. (2023) proved that both monotone and non-monotone rules can lead to a linear convergence rate for some classes of functions satisfying the PL and interpolation conditions.

However, the use of stochastic line searches with arbitrary directions d_k needs to be carefully pondered. Lapucci & Pucci (2025) indeed showed how the stochastic-gradient-related conditions

$$\|d_k\| \leq c_1\|g_k(x^k)\|, \quad d_k^\top g_k(x^k) \leq -c_2\|g_k(x^k)\|^2, \qquad (c_1, c_2 > 0) \tag{4}$$

should hold in order for the line search to result well-defined; moreover, the stronger conditions of the form

$$\|\mathbb{E}_k[d_k]\| \leq \Gamma_1 \|\nabla f(x^k)\|, \quad \mathbb{E}_k[d_k]^\top \nabla f(x^k) \leq -\Gamma_2 \|\nabla f(x^k)\|^2, \qquad (\Gamma_1, \Gamma_2 > 0) \tag{5}$$

are needed for the linear convergence of the whole framework under PL and interpolation; in fact, the same conditions also allow to prove the (sublinear) convergence result in the nonconvex case under the SGC, as proved in Appendix C of this manuscript. Conditions from equation 5 are satisfied, for example, if the direction is obtained by an uncorrelated positive definite preconditioning of the stochastic gradient direction, or if the directions are stochastic-gradient-related and also have conditional covariance with $-g_k(x^k)$ bounded by the variance of $g_k(x^k)$ itself. When d_k is a momentum-type direction like in equation 2 or equation 3, the conditions in equation 4 can only be enforced carrying out safeguard checks and suitable adjustments, often dampening the value of β_k and thus canceling out a large part of contribution of momentum altogether. Moreover, the condition concerning covariance, although reasonable, is numerically uncheckable at runtime. These considerations motivate the search for an alternative route.

3 The stochastic curve search framework

In this work, we propose a new strategy, based on curve searches, to safely carry out possibly aggressive steps along momentum-type directions. Without the need of assuming or checking any special properties for the main search direction, which might even be not of descent w.r.t. the current f_k, the proposed approach allows to test vanilla updates with aggressive stepsizes and revert to more cautious updates only if a stochastic sufficient decrease condition is not met.

The (stochastic) curve search method iteratively defines the new iterate according to

$$x^{k+1} = \gamma_k(t_k), \tag{6}$$

where $\gamma_k : [0, 1] \to \mathbb{R}^n$ is a differentiable curve and $t_k > 0$ is a parameter that determines how far to move along the curve. In the deterministic scenario, curve search methods have been extensively studied, for example, by Goldfarb (1980); Ben-Tal et al. (1990); Donnini et al. (2025). In this paper, we particularly focus on the methodology proposed by Donnini et al. (2025) for unconstrained problems, and we consider quadratic search curves parametrized as:

$$\gamma_k(t) = \gamma\left(t; x^k, \tau_k g_k(x^k), s_k\right) = x^k - t\tau_k g_k(x^k) + t^2(s_k + \tau_k g_k(x^k)), \tag{7}$$

where $x^k \in \mathbb{R}^n$ is the current solution, the stochastic gradient direction serves as an anchor direction, τ_k is a scaling factor and $s_k \in \mathbb{R}^n$ represents the update that we ideally would like to apply. In fact, while in this work we are particularly motivated by momentum-type updates, the same technique could be employed to deal with any update rule (for instance, the ones devised by Kingma & Ba 2014; Foret et al. 2021; Jordan et al. 2024). Such a curve is differentiable and, at $t = 0$, we have $\gamma_k'(0) = -\tau_k g_k(x^k)$.

For a stepsize $t = 1$ along the curve defined in equation 7, the evaluated point is $\gamma_k(1) = x^k + s_k$, i.e., the outcome of the "ideal" update suggested by any iterative rule. On the other hand, for small values of t the second order term in the curve equation almost vanishes, so that in the limit we revert to steps along the negative stochastic gradient direction $-g_k(x^k)$. We can thus think of using 1 as first tentative stepsize, and then backtracking along the curve as long as some sufficient decrease condition is not met. Since we revert to the baseline direction in bad cases, the mechanism is guaranteed to produce a meaningful update at each iteration.

Concerning the (stochastic) sufficient decrease condition, we can generalize the rule defined by Donnini et al. (2025) in a similar way as Vaswani et al. (2019b) and Galli et al. (2023) generalize the Armijo-type conditions for deterministic line searches to the stochastic setting; hence, we can require that the stepsize t_k used at iteration k for the update described by equation 6 satisfies

$$f_k(\gamma_k(t_k)) \leq C_k - \sigma\tau_k t_k \|g_k(x^k)\|^2, \tag{8}$$

where $\sigma \in (0, 1)$, $C_k = f_k(x^k)$ in the monotone setup and $C_k \geq f_k(x^k)$ in the non-monotone case; for instance, a Grippo-type non-monotone reference value (Grippo et al., 1986) can be defined for the stochastic

setting as

$$C_k = \max_{0 \leq j \leq m(k)} f_{k-j}(x^{k-j}),$$

where $m(0) = 0$ and $0 \leq m(k) \leq \min\{m(k-1)+1, M\}$, with $M > 0$, for $k \geq 1$.

Operationally, given an initial stepsize $t_{\max} \in (0, 1]$ and $\delta \in (0, 1)$, we can use a backtracking method synthesized by the following expression:

$$t_k = \max_{j \in \mathbb{N}} \{\delta^j t_{\max} \mid f_k(\gamma_k(\delta^j t_{\max})) \leq C_k - \sigma \tau_k \delta^j t_{\max} \|g_k(x^k)\|^2\}. \tag{9}$$

As we formalize in the following section, the stepsize t_k obtained accordingly is always well-defined.

3.1 Theoretical analysis

In this section, we present some theoretical results concerning the correctness of the stochastic curve search framework and the convergence properties of the overall resulting algorithm.

We begin by stating that the stochastic curve search procedure is well-defined at each iteration, producing a stepsize t_k satisfying equation 8 within a finite number of backtracks.

Lemma 1. *Let $f_k : \mathbb{R}^n \to \mathbb{R}$ and $g_k : \mathbb{R}^n \to \mathbb{R}^n$ be the random estimators of f and ∇f used at iteration k. Let $x^k \in \mathbb{R}^n$ such that $\|g_k(x^k)\| > 0$, $s_k \in \mathbb{R}^n$, $\tau_k > 0$, $C_k \geq f_k(x^k)$, $t_{max} \in (0, 1]$, $\sigma \in (0, 1)$, $\delta \in (0, 1)$ and γ_k defined according to equation 7. Then, there exists $\ell(k) \in \mathbb{N}$ such that $t_k = \delta^{\ell(k)} t_{max}$ satisfies equation 8.*

Proof. The proof is postponed to Appendix B. □

Under the L-smoothness assumptions made for f and all f_is, and adding a further hypothesis concerning the directions used to construct the curves at each iteration, we can also prove a uniform lower bound for the stepsizes employed throughout the entire optimization process.

Assumption 1. *There exists $c_1 > 0$ and $\tau_{min} > 0$ such that conditions $\|s_k\| \leq c_1 \|g_k(x^k)\|$ and $\tau_{min} \leq \tau_k \leq c_1$ hold for all $k \geq 0$.*

Remark 1. *Note that the assumption above can be enforced at each iteration, if necessary, by suitable clipping operations on the target vector s_k or the desired value of τ_k.*

Before stating the result, we need to state a property of functions parametrized by curves of the form (7), similar to Proposition 6 by Donnini et al. (2025).

Lemma 2. *Let Assumption 1 hold, f_k be L_k-smooth and γ_k be defined as in equation 7. Then, the composite function $\bar{\varphi}_k = f_k \circ \gamma_k : [0, 1] \to \mathbb{R}$ has Lipschitz continuous gradient in the interval $[0, 1]$ with Lipschitz constant $\bar{L}_k \leq (2(c_1 + \tau_k) + L_k(6c_1^2 + 13\tau_k^2 + 18c_1\tau_k))\|g_k(x^k)\|^2$.*

Proof. The proof is postponed to Appendix B. □

Lemma 3. *Let Assumption 1 hold, all estimators f_k be L_k-smooth with $L_k \leq L_{max}$ for all k, $t_{max} \in (0, 1]$, $\delta \in (0, 1)$ and $\sigma \in (0, 1)$. Then, at each iteration k, the Armijo-type curve search condition (8) is satisfied for all $t \in [0, t_{low}^k]$, with $t_{low}^k = \frac{2\tau_k(1-\sigma)}{2(c_1+\tau_k)+L_k(6c_1^2+13\tau_k^2+18c_1\tau_k)}$, and a uniform lower bound exists for the sequence of stepsizes $\{t_k\}$ so that, for all k,*

$$t_k \geq \min\{t_{max}, \delta t_{low}\},$$

with $t_{low} = \frac{2\tau_{min}(1-\sigma)}{4c_1+37c_1^2 L_{max}}$.

Proof. The proof is postponed to Appendix B. □

In order to state a convergence result for the overall iterative scheme, we need to state a tighter bond linking the directions used to construct the search curves and the true gradients $\nabla f(x^k)$. The following assumption is in fact related to a condition considered in the past by Tseng (1998) and then by Schmidt & Roux (2013), which can be seen as a stronger version of the SGC and is indeed also referred to as *maximal strong growth condition* (Khaled & Richtárik, 2023).

Assumption 2. *There exists $c > 0$ such that conditions $\tau_k \|g_k(x^k)\| \leq c\|\nabla f(x^k)\|$ and $\|s_k\| \leq c\|\nabla f(x^k)\|$ hold for all $k \geq 0$.*

Remark 2. *Assumption 2 is not particularly restrictive in practice. For example, suppose Assumption 1 holds, and f is a finite-sum objective that satisfies both the PL condition and the minimizer interpolation property. Let f_k be L_k-smooth and defined as $f_k(x^k) = \frac{1}{|B_k|}\sum_{i \in B_k} f_i(x^k)$, where the batch size $B = |B_k|$ is constant across all iterations k, as is commonly used in practice. Under these conditions, Assumption 2 is readily satisfied, as established in Lemma 5 in Appendix B.*

Before turning to the main convergence theorem, we need to state one last preliminary result.

Lemma 4. *Let Assumption 2 hold, f be L-smooth and γ_k be defined as in equation 7. Then, the composite function $\hat{\varphi}_k = f \circ \gamma_k : [0,1] \to \mathbb{R}$ has Lipschitz continuous gradient in the interval $[0,1]$ with Lipschitz constant $\hat{L}_k \leq (4c + 37c^2 L)\|\nabla f(x^k)\|^2$.*

Proof. The thesis follows by similar reasoning as in the proof of Lemma 2, with Assumption 2 used in place of Assumption 1. □

We are finally able to state the convergence result for the stochastic curve search algorithmic framework.

Proposition 1. *Let Assumptions 1-2 hold, $f : \mathbb{R}^n \to \mathbb{R}$ be an L-smooth function and the randomly drawn functions f_k considered at any iteration k be L_k-smooth functions, with $L_k \leq L_{max}$. Let the sequence $\{(x^k, \gamma_k)\}$ be produced by iterating updates of the form $x^{k+1} = \gamma_k(t_k)$, where, for all k, γ_k is defined as in equation 7 and t_k is chosen by the Armijo-type curve search (9) where $\delta \in (0,1)$, $\sigma \in \left(1 - \frac{\bar{L}_{max}\left(1+\tilde{c}^2+\frac{c_1}{\tau_{min}}(1-\tilde{c}^2)\right)}{2\delta\hat{L}}, 1\right)$ and $t_{max} \in \left(0, \frac{c_1+\tau_{min}-\tilde{c}^2(c_1-\tau_{min})}{\hat{L}}\right)$, with $\bar{L}_{max} = 4c_1 + 37c_1^2 L_{max}$, $\hat{L} = 4c + 37c^2 L$ and $\tilde{c} = \frac{c}{\tau_{min}}$. Then,*

$$\min_{k \in \{0,\dots,K-1\}} \mathbb{E}[\|\nabla f(x^k)\|^2] \leq \frac{2}{\nu K}(f(x^0) - f(x^\star)),$$

with $\nu = \theta_{max} + \theta_{min} - t_{max}^2 \hat{L} - \tilde{c}^2(\theta_{max} - \theta_{min}) > 0$, $\theta_{min} = \min\left\{\tau_{min} t_{max}, \frac{2\tau_{min}^2 \delta(1-\sigma)}{\bar{L}_{max}}\right\}$ and $\theta_{max} = c_1 t_{max}$.

Proof. The proof is postponed to Appendix A. □

Remark 3. *The convergence result holds assuming a quite strict upper bound on the stepsize. Clearly, this requirement limits the validity of theoretical guarantees in most concrete settings, as with the stochastic curve search method we would like to always start with the unit stepsize to test the "ideal" update. Still, Proposition 1 at least ensures that for sufficiently small stepsizes the framework is sound. In fact, the same limitation in the analysis is highlighted by Vaswani et al. (2019b) for the case of stochastic line searches, where convergence follows only for possibly small initial stepsizes, which is not what works well in practice. The issue of proving convergence for backtracking frameworks with arbitrary upper bounds on the initial stepsize is thus an open problem.*

Remark 4. *With the proof techniques currently available from the literature (in particular, the one used by Galli et al. 2023), and with the (possibly weak) bounds on the L-smoothness constants for $f_k \circ \gamma_k$ and $f \circ \gamma_k$ used in this work, it is unfortunately impossible to prove linear convergence rates under PL and interpolation for the stochastic curve search framework. In fact, by a careful check on the original proof by Galli et al. (2023) and making some fixes to small algebraic errors, we can note that even for SGD with stochastic line searches consistent choices of constants leading to linear rate only exist if $L < 4\mu$. The corrected result is reported and proved in Appendix D, where we also underline the adjusted intervals for constants and the need for a stronger interpolation assumption in the case of the non-monotone rule. If we carried out the analogous reasoning for SCS, we would end up with the requirement $\mu > 37/4L$, which is inconsistent as we know $L \geq \mu$.*

4 Numerical results

In this section, we report in detail experimental results demonstrating the effectiveness of the stochastic curve search method. The code implementing the experiments is available in the Supplementary Material. All experiments were conducted on a computer equipped with the following specifications: 13th Gen Intel(R) Core(TM) i5-13400F CPU, NVIDIA GeForce RTX 4060 GPU, 32 GB RAM and a 512 GB SSD.

4.1 Experimental settings

The stochastic curve search (SCS) approach was evaluated using the following parameters configuration, chosen based on preliminary experiments (omitted here for brevity): τ_k (equation 7) was set to the initial stepsize proposed by Loizou et al. (2021); $m(k) = \min\{\max\{1, \lfloor k/10 \rfloor\}, M\}$ with $M = 400$; $t_{\max} = 1$; $\delta = 0.5$; and $\sigma = 0.5$.

Regarding the trial update s_k used to define the quadratic curve γ_k in SCS (equation 7), we considered multiple momentum-type update rules, each one corresponding to a recent contribution from the literature. As noted in Section 1, each of these methods involves hyper-parameters whose aggressive tuning may yield faster convergence but can also lead to instability. This issue is typically mitigated in the original works through clipping or conservative strategies to ensure robustness. While all other hyper-parameters are set to their default values, as specified in the corresponding references, we study the influence of these critical hyper-parameters both within the original methodologies and when the update is encapsulated in the SCS framework. The update rules considered in the experiments are the following.

- **Stochastic Polyak's Heavy Ball** (SGDM) (Polyak, 1964) with constant β, so that the direction in (2) can be equivalently expressed as $-\sum_{j=0}^{k} \beta^j \nabla f_{k-j}(x^{k-j})$. The (constant) learning rate lr is treated as the primary hyperparameter, whereas we set $\beta = 0.9$.
- **ALR-MAG** (Wang et al., 2023), where we studied the hyperparameters $\eta_{\max}$ and c_p, which determine the Polyak's stepsize along the momentum-type search direction.
- **MOM-SPS-MAX** (Oikonomou & Loizou, 2024), where the parameter c_p influences Polyak's stepsize.
- **MOMO** (Schaipp et al., 2023), where the learning rate lr is the main hyperparameter under investigation.
- **AM-MSGD** (Topollai & Choromanska, 2025), where, similarly, the learning rate lr is considered as the key parameter.

Experiments were conducted on the following network architectures trained on well-known datasets using standard training/validation splits commonly adopted in the literature.

- A multilayer perceptron (Goodfellow et al., 2016) with 1000 hidden units and ReLU activations, trained on the **MNIST** dataset (Lecun et al., 1998).
- A convolutional neural network consisting of three convolutional layers with 32 channels each, followed by a fully connected layer with 512 units, trained on the **Fashion-MNIST** dataset (Xiao et al., 2017).
- A ResNet-34 architecture (He et al., 2015), trained on **CIFAR-10** (Krizhevsky, 2009).
- A ResNet-18 architecture without batch normalization (making the underlying optimization problem harder), as proposed by Civitelli et al. (2025), hereafter referred to as ResNet-18*, trained on **CIFAR-100** (Krizhevsky, 2009).
- The Vision Transformer ViT-B/32 (Dosovitskiy et al., 2021), trained on the **SVHN** dataset (Netzer et al., 2011).

- The Vision Transformer ViT-B/16 (Dosovitskiy et al., 2021), trained on the **EuroSAT** dataset (Helber et al., 2019).

All architectures were initialized according to standard guidelines and as implemented in PyTorch (Paszke et al., 2017), with the exception of ResNet-18*, which follows the initialization scheme of Civitelli et al. (2025). Since all tasks are classification problems, training was performed using the cross-entropy loss with a batch size $|B| = 128$. Standard techniques for data augmentation are used in all six problems.

4.2 Effect of curve searches in momentum methods

We begin the analysis by showing - in Figure 1 - the effect of incorporating the curve search methodology within the simple SGDM algorithm, throughout all six test problems.

We can observe that an aggressive choice of the learning rate heavily degrades the performance of vanilla SGDM and even causes training to fail altogether in most cases. On the other hand, once curve searches are employed, larger learning rates perform similarly to more cautious choices, and the onset of training divergence is at least delayed in the few remaining bad cases.

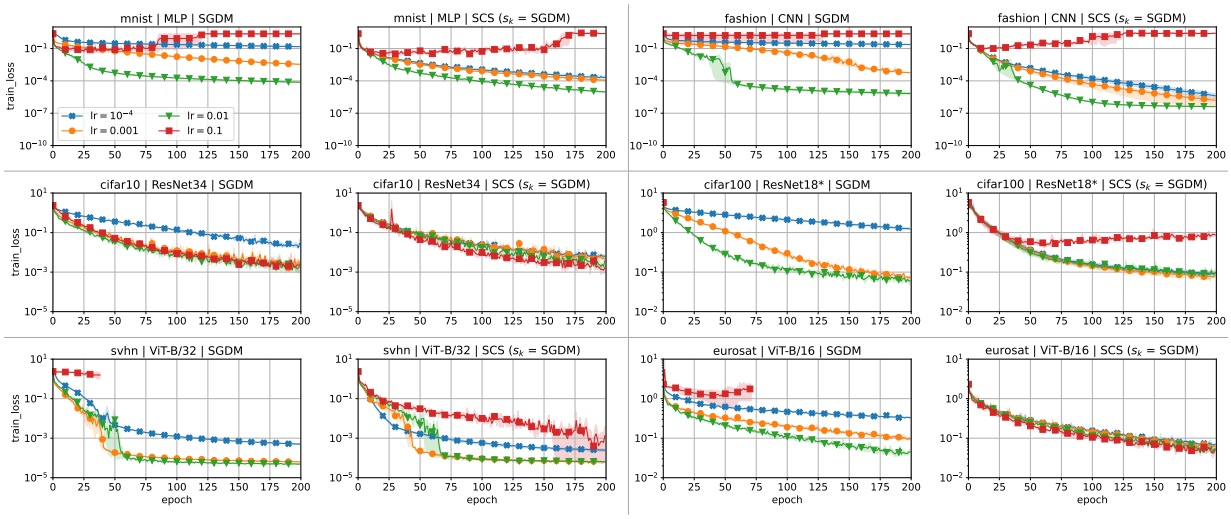

Figure 1: Training loss trend for Polyak's Heavy-Ball method with and without stochastic curve searches in selected training tasks for different learning rate (lr) values. Each pair of plots corresponds to a training problem. For each algorithm, the mean over three runs is reported, with a shaded region indicating the minimum and maximum values.

We now analyze the behavior of SCS coupled with more sophisticated rules for momentum-type updates that employ adaptive step sizes. In the easier MNIST (Figure 2) and Fashion-MNIST (Figure 3) problems, we can observe that SCSs allow not only to intercept the majority of divergent behaviors, but also to quite consistently drive the loss to lower values w.r.t. vanilla methods. Notably, MOMO appears to be extremely effective and robust with extreme choices of lr in these simpler tasks. Interestingly, the introduction of SCSs does not spoil such a good behavior.

In more challenging scenarios, such as those presented by the CIFAR10 (Figure 4) and EuroSAT (Figure 5) problems, MOMO begins to show a less stable behavior. In these tasks, SCSs allow to prevent all degenerate trainings, leading to good training performance almost independently of the parameter configuration. In Appendix E, we report the same plots for CIFAR100 and SVHN tasks, from which we can draw essentially similar conclusions.

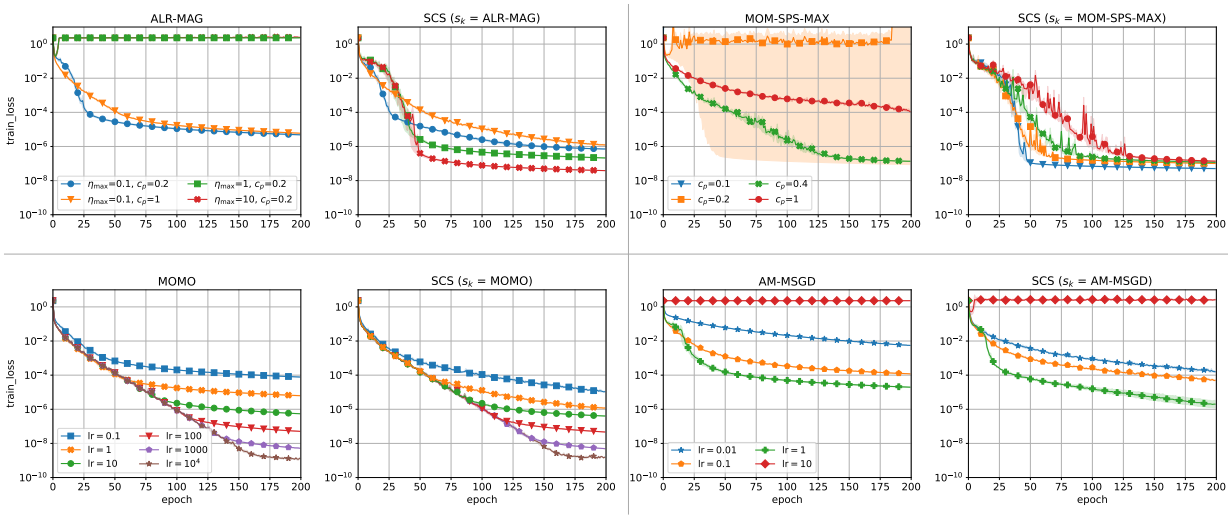

Figure 2: Training loss trends for a MLP on the MNIST dataset. Each pair of plots corresponds to a momentum method with and without stochastic curve searches, for different values of the critical hyperparameters. For each algorithm, the mean over three runs is reported, with a shaded region indicating the minimum and maximum values.

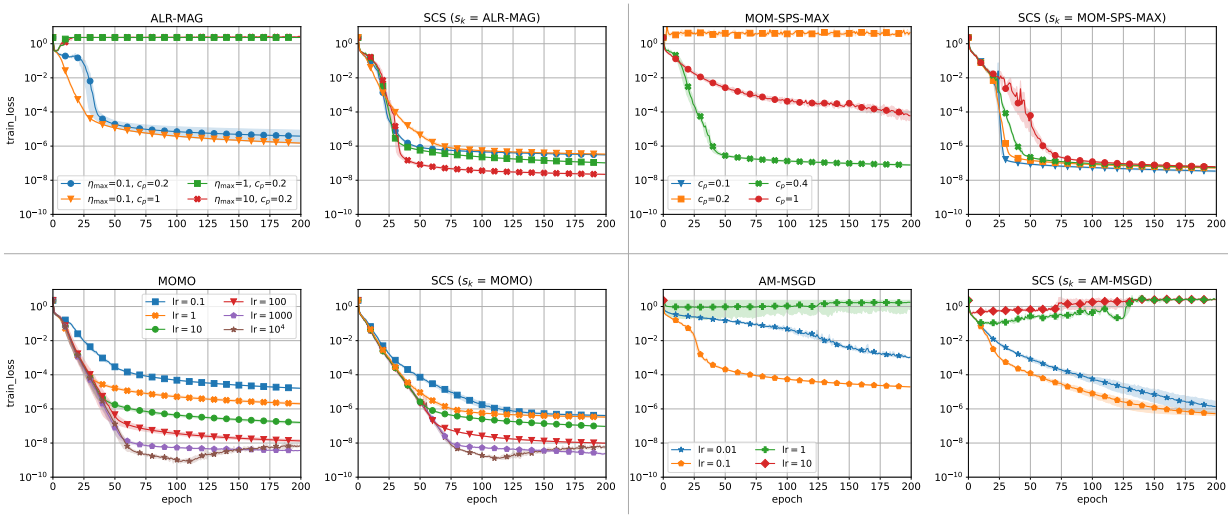

Figure 3: Training loss trends for a CNN on the Fashion-MNIST dataset. Each pair of plots corresponds to a momentum method with and without stochastic curve searches, for different values of the critical hyperparameters. For each algorithm, the mean over three runs is reported, with a shaded region indicating the minimum and maximum values.

4.3 Comparison with the state-of-the-art

We finally identify two precise instances of the *curve search* framework, based on the previously observed performance, to understand the overall viability of the approach w.r.t. state-of-the-art approaches. Namely, we consider SCS with s_k computed as in ALR-MAG ($\eta_{\max} = 10$ and $c_p = 0.2$) and SCS with s_k computed as in MOMO ($lr = 100$).

The proposed methods are in particular compared to **Adam** optimizer (Kingma & Ba, 2014) and the **PoNoS** method (Galli et al., 2023), the latter arguably representing the most closely related approach - it consists of a non-monotone Armijo-type line search along the stochastic gradient direction, coupled with Polyak's

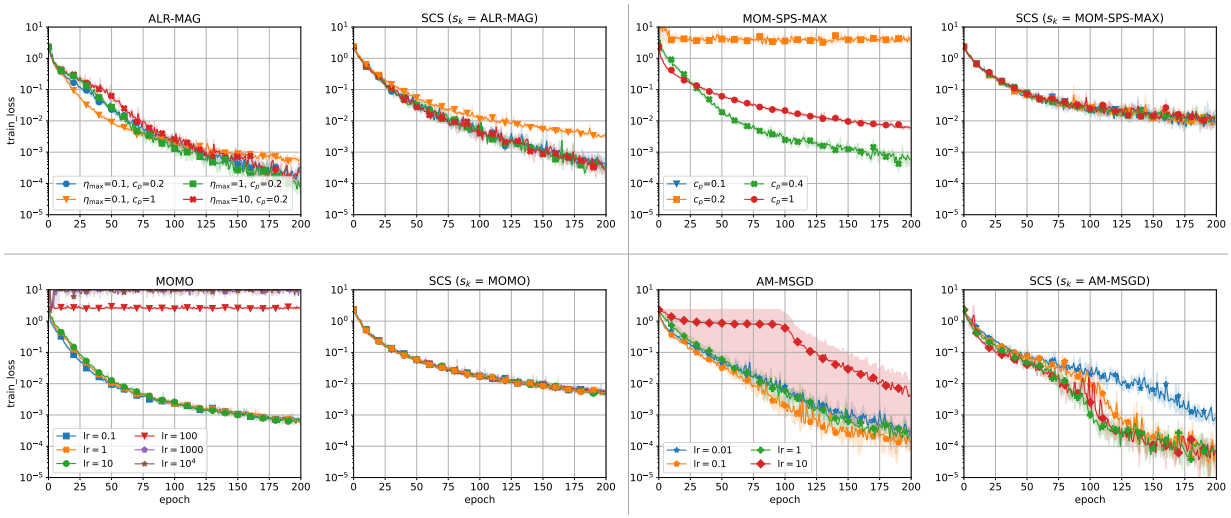

Figure 4: Training loss trends for ResNet-34 on the CIFAR10 dataset. Each pair of plots corresponds to a momentum method with and without stochastic curve searches, for different values of the critical hyperparameters. For each algorithm, the mean over three runs is reported, with a shaded region indicating the minimum and maximum values.

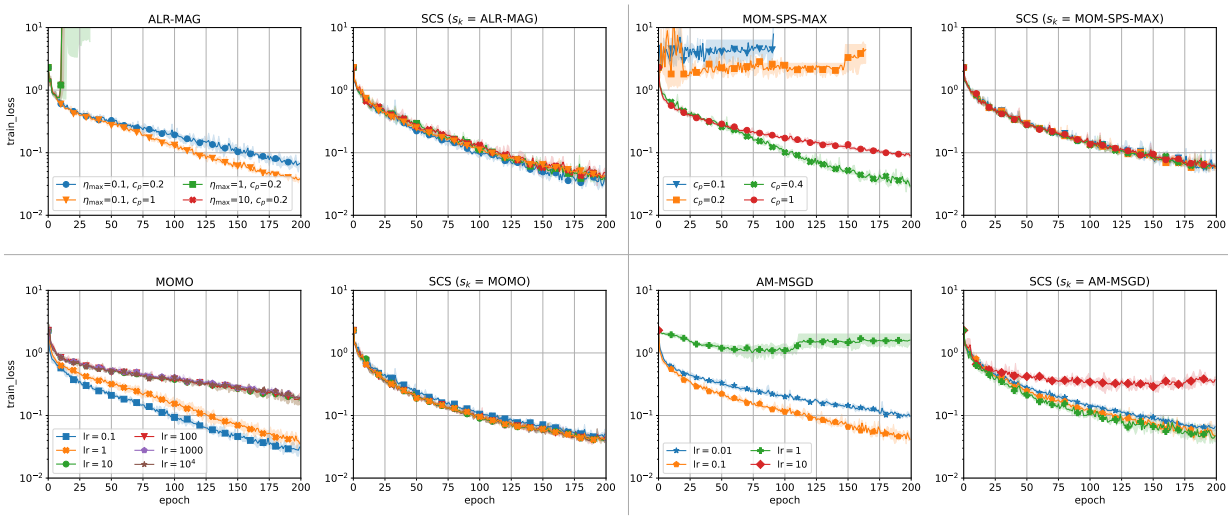

Figure 5: Training loss trends for ViT-B/16 on the EuroSAT dataset. Each pair of plots corresponds to a momentum method with and without stochastic curve searches, for different values of the critical hyperparameters. For each algorithm, the mean over three runs is reported, with a shaded region indicating the minimum and maximum values.

stepsize estimation. For Adam we employ $lr = 10^{-4}$, as it proved to be the most stable configuration in the six problems considered in preliminary experiments. For PoNoS the default settings are used. The analysis is conducted under equal time budgets; in Figures 6–7, we report the evolution of the training loss both as a function of epochs and elapsed time.

In terms of epochs, the two stochastic curve search methods proved to be generally competitive, both outperforming Adam and PoNoS in 4 out of 6 cases and matching the best in another one. Similar conclusions can be drawn when considering elapsed time, although both SCSs and PoNoS exhibited some loss of efficiency due to the overhead coming from backtracking steps. All in all, curve search methods still outperformed

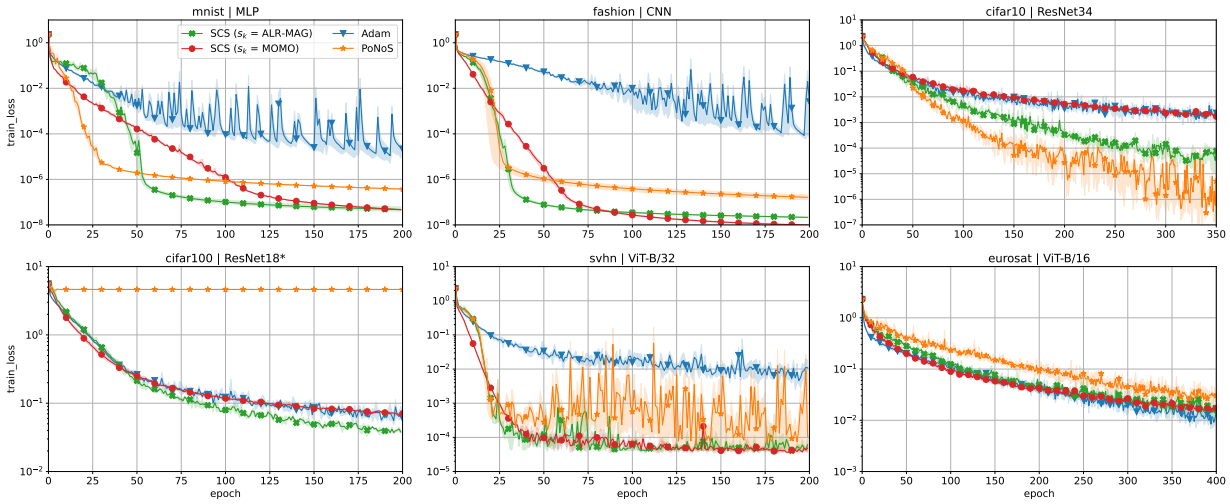

Figure 6: Comparison of the training loss as a function of the number of epochs between SCS with s_k computed using ALR-MAG or MOMO, Adam and PoNoS. Each subfigure corresponds to a specific training task. For each algorithm, the mean over three runs is reported, with a shaded region indicating the minimum and maximum values.

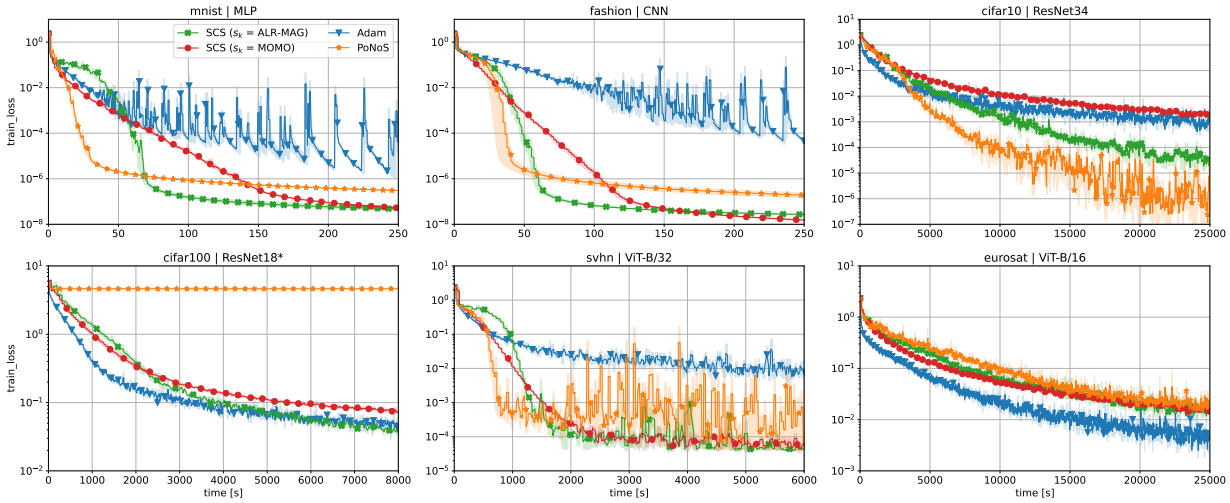

Figure 7: Comparison of the training loss as a function of the elapsed time between SCS with s_k computed using ALR-MAG or MOMO, Adam and PoNoS. Each subfigure corresponds to a specific training task. For each algorithm, the mean over three runs is reported, with a shaded region indicating the minimum and maximum values.

Adam and PoNoS in 3 out of 6 scenarios, while remaining competitive in the others, and showing a generally more consistent behavior.

These findings are also reflected in the validation accuracies achieved at the end of training (Table 1). Overall, the proposed curve search approach achieves strong performance, with the ALR-MAG-based variant standing out as the most robust in terms of out-of-sample evaluation.

Problem	*Algorithm*			
	SCS (s_k = ALR-MAG)	SCS (s_k = MOMO)	Adam	PoNoS
mnist | MLP	98.35 (±0.07)	98.29 (±0.03)	98.27 (±0.04)	**98.54 (±0.02)**
fashion | CNN	90.42 (±0.08)	90.77 (±0.26)	**91.19 (±0.12)**	89.45 (±0.59)
cifar10 | ResNet-34	94.26 (±0.28)	90.64 (±0.29)	93.51 (±0.31)	**94.35 (±0.24)**
cifar100 | ResNet-18*	**59.84 (±0.23)**	57.48 (±0.17)	56.92 (±0.44)	1.00 (±0.00)
svhn | ViT-B/32	**86.91 (±0.11)**	85.15 (±0.07)	85.23 (±0.70)	86.57 (±0.09)
eurosat | ViT-B/16	94.41 (±0.13)	94.16 (±0.12)	**94.87 (±0.14)**	94.41 (±0.21)

Table 1: Final validation accuracies (mean ± standard deviation) for SCS with s_k computed using ALR-MAG or MOMO, Adam and PoNoS. For each problem, the best value is highlighted in bold, while the second-best is underlined.

5 Conclusions

In this work we have proposed stochastic curve searches as an alternative approach to control, ensuring a sufficient decrement on the mini-batch objective, the optimization process even when possibly ascent directions are considered. We proved that the curve search approach is well-defined in the mini-batch setting and that, in the case of non-convex smooth objectives, (sublinear) convergence is guaranteed under classical assumptions - even in the case of non-monotone acceptance conditions. We also showed that momentum-based algorithms from the literature become more robust if equipped with stochastic curve searches, as the safe employment of aggressive hyperparameters is enabled. In addition, we collected numerical evidence that stochastic curve searches can be competitive w.r.t. state-of-the-art methods both in terms of efficiency and effectiveness on a diverse benchmark of learning tasks.

In future research, an investigation about the possibility of proving a linear rate of convergence for the proposed stochastic curve search framework under the PL condition would certainly be of interest, as we have remarked that current proof techniques are substantially not helpful in this regards. Moreover, the effectiveness of stochastic curve searches could be explored also for other popular algorithms, such as Adam (Kingma & Ba, 2014), SAM (Foret et al., 2021) or Muon (Jordan et al., 2024): making it possible to employ aggressive hyperparameters for these methods could be of significant practical interest, especially when considering LLM pretraining tasks.

References

Aharon Ben-Tal, Aharon Melman, and Jochem Zowe. Curved search methods for unconstrained optimization. *Optimization*, 21(5):669–695, 1990.

Leonard Berrada, Andrew Zisserman, and M. Pawan Kumar. Comment on stochastic polyak step-size: Performance of ALI-G, 2021.

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

Enrico Civitelli, Alessio Sortino, Matteo Lapucci, Francesco Bagattini, and Giulio Galvan. A robust initialization of residual blocks for effective resnet training without batch normalization. *IEEE Transactions on Neural Networks and Learning Systems*, 36(1):1947–1952, 2025.

Federica Donnini, Matteo Lapucci, and Pierluigi Mansueto. Efficient globalization of heavy-ball type methods for unconstrained optimization based on curve searches, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

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

Chen Fan, Sharan Vaswani, Christos Thrampoulidis, and Mark Schmidt. Msl: An adaptive momentem-based stochastic line-search framework. In *OPT 2023: Optimization for Machine Learning*, 2023.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=6Tm1mposlrM`.

Leonardo Galli, Holger Rauhut, and Mark Schmidt. Don't be so monotone: Relaxing stochastic line search in over-parameterized models. *Advances in Neural Information Processing Systems*, 36:34752–34764, 2023.

Igor Gitman, Hunter Lang, Pengchuan Zhang, and Lin Xiao. Understanding the role of momentum in stochastic gradient methods. *Advances in neural information processing systems*, 32, 2019.

Donald Goldfarb. Curvilinear path steplength algorithms for minimization which use directions of negative curvature. *Mathematical programming*, 18(1):31–40, 1980.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

L. Grippo, F. Lampariello, and S. Lucidi. A nonmonotone line search technique for newton's method. *SIAM Journal on Numerical Analysis*, 23(4):707–716, 1986.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Samy Jelassi and Yuanzhi Li. Towards understanding how momentum improves generalization in deep learning. In *International Conference on Machine Learning*, pp. 9965–10040. PMLR, 2022.

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

Ahmed Khaled and Peter Richtárik. Better theory for SGD in the nonconvex world. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL `https://openreview.net/forum?id=AU4qHN2VkS`. Survey Certification.

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

Alex Krizhevsky. Learning multiple layers of features from tiny images, 2009.

Matteo Lapucci and Davide Pucci. Convergence conditions for stochastic line search based optimization of over-parametrized models. *Optimization*, pp. 1–20, 2025.

Matteo Lapucci and Davide Pucci. Effectively leveraging momentum terms in stochastic line search frameworks for fast optimization of finite-sum problems. *Computational Optimization and Applications*, Mar 2026. ISSN 1573-2894.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Nicolas Loizou, Sharan Vaswani, Issam Hadj Laradji, and Simon Lacoste-Julien. Stochastic polyak step-size for sgd: An adaptive learning rate for fast convergence. In *International Conference on Artificial Intelligence and Statistics*, pp. 1306–1314. PMLR, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2017.

Siyuan Ma, Raef Bassily, and Mikhail Belkin. The power of interpolation: Understanding the effectiveness of sgd in modern over-parametrized learning. In *International Conference on Machine Learning*, pp. 3325–3334. PMLR, 2018.

Aaron Mishkin. *Interpolation, growth conditions, and stochastic gradient descent*. PhD thesis, University of British Columbia, 2020.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, pp. 4. Granada, 2011.

Dimitris Oikonomou and Nicolas Loizou. Stochastic polyak step-sizes and momentum: Convergence guarantees and practical performance, 2024.

Antonio Orvieto, Simon Lacoste-Julien, and Nicolas Loizou. Dynamics of sgd with stochastic polyak stepsizes: Truly adaptive variants and convergence to exact solution. *Advances in Neural Information Processing Systems*, 35:26943–26954, 2022.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *USSR computational mathematics and mathematical physics*, 4(5):1–17, 1964.

Fabian Schaipp, Ruben Ohana, Michael Eickenberg, Aaron Defazio, and Robert M. Gower. Momo: Momentum models for adaptive learning rates, 2023.

Mark Schmidt and Nicolas Le Roux. Fast convergence of stochastic gradient descent under a strong growth condition, 2013.

Robin M Schmidt, Frank Schneider, and Philipp Hennig. Descending through a crowded valley-benchmarking deep learning optimizers. In *International Conference on Machine Learning*, pp. 9367–9376. PMLR, 2021.

Othmane Sebbouh, Robert M Gower, and Aaron Defazio. Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball. In *Conference on Learning Theory*, pp. 3935–3971. PMLR, 2021.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pp. 1139–1147. pmlr, 2013.

Kristi Topollai and Anna Choromanska. Adaptive memory momentum via a model-based framework for deep learning optimization, 2025.

Paul Tseng. An incremental gradient (-projection) method with momentum term and adaptive stepsize rule. *SIAM Journal on Optimization*, 8(2):506–531, 1998.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd international conference on artificial intelligence and statistics*, pp. 1195–1204. PMLR, 2019a.

Sharan Vaswani, Aaron Mishkin, Issam Laradji, Mark Schmidt, Gauthier Gidel, and Simon Lacoste-Julien. Painless stochastic gradient: Interpolation, line-search, and convergence rates. *Advances in neural information processing systems*, 32, 2019b.

Xiaoyu Wang, Mikael Johansson, and Tong Zhang. Generalized polyak step size for first order optimization with momentum. In *International Conference on Machine Learning*, pp. 35836–35863. PMLR, 2023.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms, 2017.

Hongchao Zhang and William W. Hager. A nonmonotone line search technique and its application to unconstrained optimization. *SIAM Journal on Optimization*, 14(4):1043–1056, 2004.

A Proof of proposition 1

Proposition 1. *Let Assumptions 1-2 hold, $f : \mathbb{R}^n \to \mathbb{R}$ be an L-smooth function and the randomly drawn functions f_k considered at any iteration k be L_k-smooth functions, with $L_k \leq L_{max}$. Let the sequence $\{(x^k, \gamma_k)\}$ be produced by iterating updates of the form $x^{k+1} = \gamma_k(t_k)$, where, for all k, γ_k is defined as in equation 7 and t_k is chosen by the Armijo-type curve search (9) where $\delta \in (0,1)$, $\sigma \in \left(1 - \frac{\bar{L}_{max}\left(1+\tilde{c}^2+\frac{c_1}{\tau_{min}}(1-\tilde{c}^2)\right)}{2\delta\hat{L}}, 1\right)$ and $t_{max} \in \left(0, \frac{c_1+\tau_{min}-\tilde{c}^2(c_1-\tau_{min})}{\hat{L}}\right)$, with $\bar{L}_{max} = 4c_1 + 37c_1^2 L_{max}$, $\hat{L} = 4c + 37c^2 L$ and $\tilde{c} = \frac{c}{\tau_{min}}$. Then,*

$$\min_{k\in\{0,\dots,K-1\}} \mathbb{E}[\|\nabla f(x^k)\|^2] \leq \frac{2}{\nu K}(f(x^0) - f(x^\star)),$$

with $\nu = \theta_{max} + \theta_{min} - t_{max}^2\hat{L} - \tilde{c}^2(\theta_{max} - \theta_{min}) > 0$, $\theta_{min} = \min\left\{\tau_{min}t_{max}, \frac{2\tau_{min}^2\delta(1-\sigma)}{\bar{L}_{max}}\right\}$ and $\theta_{max} = c_1 t_{max}$.

Proof. By the $\hat{L}_k$-smoothness of the composite function $\hat{\varphi}_k := f \circ \gamma_k : [0,1] \to \mathbb{R}$ (Lemma 4), with $\hat{L}_k \leq (4c + 37c^2L)\|\nabla f(x^k)\|^2 = \hat{L}\|\nabla f(x^k)\|^2$, and the descent lemma on $\hat{\varphi}_k$, we get that

$$f(x^{k+1}) - f(x^k) = \hat{\varphi}_k(t_k) - \hat{\varphi}_k(0) \leq t_k\hat{\varphi}_k'(0) + t_k^2\frac{\hat{L}_k}{2} \leq -t_k\tau_k\nabla f(x^k)^\top g_k(x^k) + \frac{t_k^2\hat{L}}{2}\|\nabla f(x^k)\|^2.$$

Recalling that $-\nabla f(x^k)^\top g_k(x^k) = \frac{1}{2}(\|\nabla f(x^k) - g_k(x^k)\|^2 - \|\nabla f(x^k)\|^2 - \|g_k(x^k)\|^2)$, we obtain

$$f(x^{k+1}) - f(x^k) \leq \frac{t_k\tau_k}{2}\left(\|\nabla f(x^k) - g_k(x^k)\|^2 - \|\nabla f(x^k)\|^2 - \|g_k(x^k)\|^2\right) + \frac{t_k^2\hat{L}}{2}\|\nabla f(x^k)\|^2.$$

Rearranging the terms we get

$$2(f(x^{k+1}) - f(x^k)) \leq t_k\tau_k\|\nabla f(x^k) - g_k(x^k)\|^2 - t_k\tau_k\|\nabla f(x^k)\|^2 - t_k\tau_k\|g_k(x^k)\|^2 + t_k^2\hat{L}\|\nabla f(x^k)\|^2.$$

By Lemma 3 we define $t_{\min} = \min\left\{t_{\max}, \frac{2\tau_{\min}\delta(1-\sigma)}{\bar{L}_{\max}}\right\} \leq t_k \leq t_{\max}$, with $\bar{L}_{\max} = 4c_1 + 37c_1^2L_{\max}$. Moreover, recalling that from Assumption 1 it holds $\tau_{\min} \leq \tau_k \leq c_1$, we define $\theta_{\min} = t_{\min}\tau_{\min} = \min\left\{\tau_{\min}t_{\max}, \frac{2\tau_{\min}^2\delta(1-\sigma)}{\bar{L}_{\max}}\right\} \leq t_k\tau_k \leq t_{\max}c_1 = \theta_{\max}$. We thus obtain

$$\begin{aligned}2(f(x^{k+1}) - f(x^k)) &\leq \theta_{\max}\|\nabla f(x^k) - g_k(x^k)\|^2 - \theta_{\min}\|\nabla f(x^k)\|^2 - \theta_{\min}\|g_k(x^k)\|^2 + t_{\max}^2\hat{L}\|\nabla f(x^k)\|^2 \\ &= (\theta_{\max} - \theta_{\min} + t_{\max}^2\hat{L})\|\nabla f(x^k)\|^2 + (\theta_{\max} - \theta_{\min})\|g_k(x^k)\|^2 - 2\theta_{\max}\nabla f(x^k)^\top g_k(x^k).\end{aligned}$$

By Assumption 2, we know that $\|g_k(x^k)\| \leq \tilde{c}\|\nabla f(x^k)\|$ with $\tilde{c} = \frac{c}{\tau_{\min}}$; therefore, it holds

$$2(f(x^{k+1}) - f(x^k)) \leq (\theta_{\max} - \theta_{\min} + t_{\max}^2\hat{L} + \tilde{c}^2(\theta_{\max} - \theta_{\min}))\|\nabla f(x^k)\|^2 - 2\theta_{\max}\nabla f(x^k)^\top g_k(x^k).$$

Taking the expectation w.r.t. x^k and recalling that $\mathbb{E}_k[\nabla f(x^k)^\top g_k(x^k)] = \|\nabla f(x^k)\|^2$, from the unbiasedness of g_k, we get

$$\begin{aligned}2\mathbb{E}_k[f(x^{k+1}) - f(x^k)] &\leq (\theta_{\max} - \theta_{\min} + t_{\max}^2\hat{L} + \tilde{c}^2(\theta_{\max} - \theta_{\min}))\|\nabla f(x^k)\|^2 - 2\theta_{\max}\mathbb{E}_k[\nabla f(x^k)^\top g_k(x^k)] \\ &= -\left(\theta_{\max} + \theta_{\min} - t_{\max}^2\hat{L} - \tilde{c}^2(\theta_{\max} - \theta_{\min})\right)\|\nabla f(x^k)\|^2 = -\nu\|\nabla f(x^k)\|^2,\end{aligned}$$

where we set $\nu = (\theta_{\max} + \theta_{\min} - t_{\max}^2 \hat{L} - \tilde{c}^2(\theta_{\max} - \theta_{\min}))$.
Assuming that $\nu > 0$, we have

$$\|\nabla f(x^k)\|^2 \leq \frac{2}{\nu}\mathbb{E}_k[f(x^k) - f(x^{k+1})].$$

By taking the total expectation, summing the terms corresponding to $k = 0, \ldots, K-1$ and dividing both sides by K, we obtain

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(x^k)\|^2] \leq \frac{2}{\nu K}\sum_{k=0}^{K-1}\mathbb{E}[f(x^k) - f(x^{k+1})] = \frac{2}{\nu K}\mathbb{E}[f(x^0) - f(x^K)] \leq \frac{2}{\nu K}\left(f(x^0) - f(x^\star)\right).$$

Therefore, we conclude that

$$\min_{k\in\{0,\ldots,K-1\}}\mathbb{E}[\|\nabla f(x^k)\|^2] \leq \frac{2}{\nu K}(f(x^0) - f(x^\star)).$$

Now, in order to show that $\nu > 0$ holds, we consider two cases.

1. If $t_{\max} \leq \frac{2\tau_{\min}\delta(1-\sigma)}{\bar{L}_{\max}}$, then $t_{\min} = t_{\max}$. In this case, we have

   $$\nu = \theta_{\max} + \theta_{\min} - t_{\max}^2\hat{L} - \tilde{c}^2(\theta_{\max} - \theta_{\min}) = t_{\max}(c_1 + \tau_{\min} - t_{\max}\hat{L} - \tilde{c}^2(c_1 - \tau_{\min})),$$

   therefore, $\nu > 0$, since $0 < t_{\max} < \frac{c_1 + \tau_{\min} - \tilde{c}^2(c_1 - \tau_{\min})}{\hat{L}}$ by assumption. Here, we implicitly require $c_1 + \tau_{\min} - \tilde{c}^2(c_1 - \tau_{\min}) > 0$. In the worst case, the condition can always be satisfied fixing $\tau_{\min}$ close enough to c_1, i.e., having less freedom in the choice of τ_k.

2. If $t_{\max} > \frac{2\tau_{\min}\delta(1-\sigma)}{\bar{L}_{\max}}$, then $t_{\min} = \frac{2\tau_{\min}\delta(1-\sigma)}{\bar{L}_{\max}}$. Hence,

   $$\begin{aligned}\nu &= \theta_{\max} + \frac{2\tau_{\min}^2\delta(1-\sigma)}{\bar{L}_{\max}} - t_{\max}^2\hat{L} - \tilde{c}^2\theta_{\max} + \tilde{c}^2\frac{2\tau_{\min}^2\delta(1-\sigma)}{\bar{L}_{\max}} \\ &= -\hat{L}t_{\max}^2 + (1-\tilde{c}^2)c_1 t_{\max} + (1+\tilde{c}^2)\frac{2\tau_{\min}^2\delta(1-\sigma)}{\bar{L}_{\max}}.\end{aligned}$$

   We have that ν is concave quadratic in $t_{\max}$. Therefore, given

   $$\Delta = (1-\tilde{c}^2)^2 c_1^2 + 4\hat{L}(1+\tilde{c}^2)\frac{2\tau_{\min}^2\delta(1-\sigma)}{\bar{L}_{\max}} > 0,$$

   we have that $\nu > 0$ if $t_{\max} \in \left(\frac{(1-\tilde{c}^2)c_1 - \sqrt{\Delta}}{2\hat{L}}, \frac{(1-\tilde{c}^2)c_1 + \sqrt{\Delta}}{2\hat{L}}\right)$. We note that $\sqrt{\Delta} \geq |(1-\tilde{c}^2)c_1|$ and, thus, $t_{\max} > t_{\min} > \frac{(1-\tilde{c}^2)c_1 - \sqrt{\Delta}}{2\hat{L}}$, since $\frac{(1-\tilde{c}^2)c_1 - \sqrt{\Delta}}{2\hat{L}} < 0$ and $t_{\max} > t_{\min} > 0$ by assumption. Hence, for any choice of $t_{\max} \in \left(t_{\min}, \frac{(1-\tilde{c}^2)c_1 + \sqrt{\Delta}}{2\hat{L}}\right)$ it holds $\nu > 0$. Let us verify that this interval is nonempty, that is, $t_{\min} = \frac{2\tau_{\min}\delta(1-\sigma)}{\bar{L}_{\max}} < \frac{(1-\tilde{c}^2)c_1 + \sqrt{\Delta}}{2\hat{L}}$, which is equivalent to

   $$\sqrt{\Delta} > \frac{4\tau_{\min}\delta\hat{L}(1-\sigma)}{\bar{L}_{\max}} - (1-\tilde{c}^2)c_1.$$

   Squaring both terms, we get that

   $$8\hat{L}(1+\tilde{c}^2)\frac{\tau_{\min}^2\delta(1-\sigma)}{\bar{L}_{\max}} > \frac{16\tau_{\min}^2\delta^2\hat{L}^2(1-\sigma)^2}{\bar{L}_{\max}^2} - \frac{8\tau_{\min}\delta\hat{L}(1-\sigma)}{\bar{L}_{\max}}(1-\tilde{c}^2)c_1,$$

   which turns into

   $$\tau_{\min}(1+\tilde{c}^2) > \frac{2\tau_{\min}\delta\hat{L}(1-\sigma)}{\bar{L}_{\max}} - (1-\tilde{c}^2)c_1.$$

By rearranging, we can write

$$\delta\hat{L}(1-\sigma) < \frac{\bar{L}_{\max}}{2}\left(1+\tilde{c}^2+\frac{c_1}{\tau_{\min}}(1-\tilde{c}^2)\right),$$

which is satisfied since $\sigma > 1 - \frac{\bar{L}_{\max}}{2\delta\hat{L}}\left(1+\tilde{c}^2+\frac{c_1}{\tau_{\min}}(1-\tilde{c}^2)\right)$ by assumption. Similarly to case 1, we implicitly require $1+\tilde{c}^2+\frac{c_1}{\tau_{\min}}(1-\tilde{c}^2) > 0$, which, again, in the worst case it can always be satisfied fixing $\tau_{\min}$ close enough to c_1, i.e., having less freedom in the choice of τ_k.

Now, substituting the maximum value for σ into the upper-bound of $t_{\max}$, we get that

$$\frac{(1-\tilde{c}^2)c_1 + \sqrt{(1-\tilde{c}^2)^2c_1^2 + 4(\tau_{\min}^2(1+\tilde{c}^2)^2 + \tau_{\min}c_1(1-\tilde{c}^4))}}{2\hat{L}} = \frac{c_1+\tau_{\min}-\tilde{c}^2(c_1-\tau_{\min})}{\hat{L}},$$

i.e.,

$$t_{\max} \in \left(t_{\min}, \frac{c_1+\tau_{\min}-\tilde{c}^2(c_1-\tau_{\min})}{\hat{L}}\right).$$

Since $t_{\min} < t_{\max} < \frac{c_1+\tau_{\min}-\tilde{c}^2(c_1-\tau_{\min})}{\hat{L}}$ by assumption, it finally follows that $\nu > 0$.

In both cases, we have that $\nu > 0$, completing the proof. □

B Supplementary mathematical proofs

In this appendix, we provide mathematical proofs of lemmas that did not find space in the main body of the manuscript.

Lemma 1. *Let $f_k : \mathbb{R}^n \to \mathbb{R}$ and $g_k : \mathbb{R}^n \to \mathbb{R}^n$ be the random estimators of f and ∇f used at iteration k. Let $x^k \in \mathbb{R}^n$ such that $\|g_k(x^k)\| > 0$, $s_k \in \mathbb{R}^n$, $\tau_k > 0$, $C_k \geq f_k(x^k)$, $t_{max} \in (0,1]$, $\sigma \in (0,1)$, $\delta \in (0,1)$ and γ_k defined according to equation 7. Then, there exists $\ell(k) \in \mathbb{N}$ such that $t_k = \delta^{\ell(k)}t_{max}$ satisfies equation 8.*

Proof. Assume by contradiction that, for all $j \in \mathbb{N}$, equation 8 is not satisfied by the stepsize $\delta^j t_{\max}$, i.e.,

$$f_k(\gamma_k(\delta^j t_{\max})) > C_k - \sigma\tau_k\delta^j t_{\max}\|g_k(x^k)\|^2 \geq f_k(x^k) - \sigma\tau_k\delta^j t_{\max}\|g_k(x^k)\|^2.$$

Recalling that $\gamma_k(0) = x^k$ and rearranging, we can further write

$$\frac{f_k(\gamma_k(\delta^j t_{\max})) - f_k(\gamma_k(0))}{\delta^j t_{\max}} > -\sigma\tau_k\|g_k(x^k)\|^2.$$

Taking the limits for $j \to \infty$, recalling that $\delta \in (0,1)$ and that γ_k is differentiable with $\gamma_k'(0) = -\tau_k g_k(x^k)$, we get

$$\nabla f_k(\gamma_k(0))^\top \gamma_k'(0) = -\tau_k g_k(x^k)^\top g_k(x^k) = -\tau_k\|g_k(x^k)\|^2 \geq -\sigma\tau_k\|g_k(x^k)\|^2,$$

which is a contradiction as $\tau_k > 0$, $\sigma \in (0,1)$ and $\|g_k(x^k)\| > 0$ by hypothesis. □

Lemma 2. *Let Assumption 1 hold, f_k be L_k-smooth and γ_k be defined as in equation 7. Then, the composite function $\bar{\varphi}_k = f_k \circ \gamma_k : [0,1] \to \mathbb{R}$ has Lipschitz continuous gradients in the interval $[0,1]$ with Lipschitz constant $\bar{L}_k \leq (2(c_1+\tau_k) + L_k(6c_1^2 + 13\tau_k^2 + 18c_1\tau_k))\|g_k(x^k)\|^2$.*

Proof. We note that all the assumptions of Proposition 6 in Donnini et al. (2025) are verified: indeed, f_k is an L_k-smooth function, γ_k is a quadratic curve and Assumption 4 from Donnini et al. (2025) is satisfied since, recalling that $g_k(x^k) = \nabla f_k(x^k)$ and Assumption 1, it holds that $\|d_k\| = \tau_k\|g_k(x^k)\| \leq c_1\|\nabla f_k(x^k)\|$ and $\|s_k\| \leq c_1\|\nabla f_k(x^k)\|$. Following then a similar reasoning of the proof of Proposition 6 in Donnini et al. (2025), we get that

$$|\bar{\varphi}_k'(t_1) - \bar{\varphi}_k'(t_2)| \leq (2(c_1+\tau_k) + L_k(6c_1^2 + 13\tau_k^2 + 18c_1\tau_k))\|g_k(x^k)\|^2|t_1 - t_2|.$$

Hence, $\bar{\varphi}_k$ is $\bar{L}_k$-smooth with $\bar{L}_k \leq (2(c_1+\tau_k) + L_k(6c_1^2 + 13\tau_k^2 + 18c_1\tau_k))\|g_k(x^k)\|^2$. □

Lemma 3. *Let Assumption 1 hold, all estimators f_k be L_k-smooth with $L_k \leq L_{max}$ for all k, $t_{max} \in (0,1]$, $\delta \in (0,1)$ and $\sigma \in (0,1)$. Then, at each iteration k, the Armijo-type curve search condition (8) is satisfied for all $t \in [0, t^k_{low}]$, with $t^k_{low} = \frac{2\tau_k(1-\sigma)}{2(c_1+\tau_k)+L_k(6c_1^2+13\tau_k^2+18c_1\tau_k)}$, and a uniform lower bound exists for the sequence of stepsizes $\{t_k\}$ so that, for all k,*

$$t_k \geq \min\{t_{max}, \delta t_{low}\},$$

with $t_{low} = \frac{2\tau_{min}(1-\sigma)}{4c_1+37c_1^2L_{max}}$.

Proof. Let us consider iteration k and assume that the stepsize t does not satisfy equation 8. We thus have

$$f_k(\gamma_k(t)) - f_k(x^k) > -\sigma t \tau_k \|g_k(x^k)\|^2.$$

On the other hand, recalling the L_k-smoothness of f_k and Lemma 2, we get that

$$f_k(\gamma_k(t)) - f_k(x^k) \leq t g_k(x^k)^\top \gamma_k'(0) + t^2 \frac{\bar{L}_k}{2} \leq -t\tau_k \|g_k(x^k)\|^2 + t^2 \frac{(2(c_1+\tau_k) + L_k(6c_1^2+13\tau_k^2+18c_1\tau_k))\|g_k(x^k)\|^2}{2}.$$

Combining the two inequalities we get

$$-\sigma t\tau_k \|g_k(x^k)\|^2 < -t\tau_k\|g_k(x^k)\|^2 + t^2 \frac{(2(c_1+\tau_k) + L_k(6c_1^2+13\tau_k^2+18c_1\tau_k))\|g_k(x^k)\|^2}{2}.$$

Dividing both sides by $t\|g_k(x^k)\|^2$, we have that

$$-\sigma\tau_k < -\tau_k + t\frac{2(c_1+\tau_k) + L_k(6c_1^2+13\tau_k^2+18c_1\tau_k)}{2},$$

which, rearranging the terms, turns into

$$t > \frac{2\tau_k(1-\sigma)}{2(c_1+\tau_k)+L_k(6c_1^2+13\tau_k^2+18c_1\tau_k)}.$$

We thus get the first part of the proof.

Next, we can notice by Assumption 1 and $L_k \leq L_{\max}$ that

$$t^k_{\text{low}} = \frac{2\tau_k(1-\sigma)}{2(c_1+\tau_k)+L_k(6c_1^2+13\tau_k^2+18c_1\tau_k)} \geq \frac{2\tau_{\min}(1-\sigma)}{4c_1+37c_1^2L_{\max}}.$$

Therefore, $t_{\text{low}} = \frac{2\tau_{\min}(1-\sigma)}{4c_1+37c_1^2L_{\max}}$ surely satisfies the Armijo-type curve search condition (equation 8) for all k. Hence, either the initial step is accepted ($t_k = t_{\max}$), or the last rejected step t_k/δ exceeded the guaranteed acceptance threshold t_{low}, i.e., $t_k > \delta t_{\text{low}}$. We can thus conclude that, for all k,

$$t_k \geq \min\{t_{\max}, \delta t_{\text{low}}\}.$$

□

Lemma 5. *Let Assumption 1 hold, f be a finite-sum function of the form (1) satisfying the PL condition and the minimizer interpolation property, and the randomly drawn functions f_k considered at any iteration k be L_k-smooth, with $L_k \leq L_{max}$, and defined as $f_k(x^k) = \frac{1}{|B_k|}\sum_{i\in B_k} f_i(x^k)$, with $|B_k| = B$ for all k. Thus, Assumption 2 is satisfied with $c = c_1\sqrt{\frac{NL_{max}}{B\mu}}$.*

Proof. By the descent lemma, we know that

$$f_k(x^\star) \leq f_k(x^k) - \frac{1}{2L_k}\|g_k(x^k)\|^2,$$

while, by the PL condition, it holds

$$|B_k|(f_k(x^k)-f_k(x^\star)) = \sum_{i\in B_k}\big(f_i(x^k)-f_i(x^\star)\big) \leq \sum_{i=1}^{N}\big(f_i(x^k)-f_i(x^\star)\big) = N(f(x^k)-f(x^\star)) \leq \frac{N}{2\mu}\|\nabla f(x^k)\|^2,$$

where the first inequality comes by the minimizer interpolation property. Combining then the two results and using that $L_k \leq L_{\max}$, we get that

$$\|g_k(x^k)\| \leq \sqrt{\frac{NL_{\max}}{|B_k|\mu}}\|\nabla f(x^k)\| = \sqrt{\frac{NL_{\max}}{B\mu}}\|\nabla f(x^k)\|. \tag{10}$$

Recalling that $\tau_k \leq c_1$ by Assumption 1, we obtain that

$$\tau_k\|g_k(x^k)\| \leq c_1\sqrt{\frac{NL_{\max}}{B\mu}}\|\nabla f(x^k)\|.$$

Finally, by equation 10 and Assumption 1, we get that

$$\|s_k\| \leq c_1\|g_k(x^k)\| \leq c_1\sqrt{\frac{NL_{\max}}{B\mu}}\|\nabla f(x^k)\|,$$

which concludes the proof. □

C Sublinear convergence rate under SGC for general stochastic line search based methods

In this appendix, we state and prove a sublinear convergence rate under the SGC condition (Definition 1) in the nonconvex setting for stochastic algorithms of the form $x^{k+1} = x^k + \alpha_k d_k$, where d_k is an arbitrary direction and α_k is determined by a stochastic line search.

The following result complements the analysis by Lapucci & Pucci (2025) considering the aforementioned different setup, which leverages a milder set of assumptions. From the referenced work we still need to borrow the upcoming definition, assumption and lemma.

Definition 4. *We define the variance of a random vector $v \in \mathbb{R}^n$ and the covariance between two random vectors $u, v \in \mathbb{R}^n$ as*

$$\mathrm{Var}(v) = \mathbb{E}[\|v\|^2] - \|\mathbb{E}[v]\|^2,$$
$$\mathrm{Cov}(v) = \mathbb{E}[u^\top v] - \mathbb{E}[u]^\top\mathbb{E}[v].$$

Assumption 3. *There exists $c_3 > 0$ such that the sequence of search directions $\{d_k\}$ satisfies the following property:*

$$\mathrm{Cov}_k(d_k, g_k(x^k)) \geq -c_3\,\mathrm{Var}_k(g_k(x^k)).$$

Lemma 6. *Let the SGC condition and Assumption 3 hold, and the sequence of search directions $\{d_k\}$ be stochastic-gradient-related, with $c_2 > c_3(1-\frac{1}{\rho})$. Then, we have:*

$$\|E_k[d_k]\| \leq c_1\sqrt{\rho}\|\nabla f(x^k)\|,$$
$$E_k[d_k]^\top\nabla f(x^k) \leq -\left(c_2 - c_3\left(1-\frac{1}{\rho}\right)\right)\|\nabla f(x^k)\|^2.$$

As already mentioned in Section 2, the stochastic-gradient-related conditions and Assumption 3 (which seems reasonable in most settings, but still appears to be uncheckable at runtime) implies that the conditions on the expected direction of equation 5 are satisfied.

Proposition 2. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be an L-smooth function satisfying the SGC condition and the randomly drawn functions f_k considered at any iteration k be L_k-smooth functions, with $L_k \leq L_{max}$. Let $x^{k+1} = x^k + \alpha_k d_k$, where, for all k, $d_k \in \mathbb{R}^n$ is a stochastic-gradient-related direction such that Assumption 3 is satisfied with $c_2 > c_3(1 - \frac{1}{\rho})$, and the stepsize α_k is chosen by Armijo-type line search, i.e.,*

$$\alpha_k = \max_{j \in \mathbb{N}} \left\{ \delta^j \alpha_0^k \mid f_k(x^k + \delta^j \alpha_0^k d_k) \leq f_k(x^k) + \sigma \delta^j \alpha_0^k d_k^\top g_k(x^k) \right\},$$

where $\delta \in (0,1)$, $\sigma \in \left(1 - \frac{(\rho c_2 - \rho c_3 + c_3) L_{max}}{\delta \rho^2 c_2 L}, 1\right)$, and $\frac{2\delta c_2(1-\sigma)}{c_1^2 L_{max}} < \alpha_0^k \leq \alpha_{max} \in \left(\frac{2\delta c_2(1-\sigma)}{c_1^2 L_{max}}, \frac{2(\rho c_2 - \rho c_3 + c_3)}{\rho^2 c_1^2 L}\right)$ for all k. Then,

$$\min_{k \in \{0,\dots,K-1\}} \mathbb{E}[\|\nabla f(x^k)\|^2] \leq \frac{2}{\nu K} \left(f(x^0) - f(x^\star)\right),$$

with $\nu = \left(\alpha_{max}(2(c_2 - c_3(1 - \frac{1}{\rho})) - 1 - \rho c_1^2 (L\alpha_{max} + 1)) + \frac{2\delta c_2(1-\sigma)}{c_1^2 L_{max}}(1 + \rho c_1^2)\right) > 0$.

Proof. By the L-smoothness of f and $x^{k+1} = x^k + \alpha_k d_k$, it follows

$$f(x^{k+1}) - f(x^k) \leq \nabla f(x^k)^\top (x^{k+1} - x^k) + \frac{L}{2}\|x^{k+1} - x^k\|^2 = \alpha_k \nabla f(x^k)^\top d_k + \frac{L}{2}\alpha_k^2 \|d_k\|^2.$$

Now, recalling that $\nabla f(x^k)^\top d_k = \frac{1}{2}(\|\nabla f(x^k) + d_k\|^2 - \|\nabla f(x^k)\|^2 - \|d_k\|^2)$, we obtain

$$f(x^{k+1}) - f(x^k) \leq \frac{\alpha_k}{2}\|\nabla f(x^k) + d_k\|^2 - \frac{\alpha_k}{2}\|\nabla f(x^k)\|^2 + \left(\frac{L\alpha_k^2}{2} - \frac{\alpha_k}{2}\right)\|d_k\|^2. \tag{11}$$

By hypothesis and Proposition 4.3 in Lapucci & Pucci (2025) we can define $\alpha_{\min} = \frac{2\delta c_2(1-\sigma)}{c_1^2 L_{\max}} = \min\left\{\alpha_0^k, \frac{2\delta c_2(1-\sigma)}{c_1^2 L_{\max}}\right\} \leq \alpha_k \leq \alpha_{\max}$. Then, by rearranging the terms in equation 11, we get

$$\begin{aligned}
2(f(x^{k+1}) - f(x^k)) &\leq \alpha_{\max}\|\nabla f(x^k) + d_k\|^2 - \alpha_{\min}\|\nabla f(x^k)\|^2 + (L\alpha_{\max}^2 - \alpha_{\min})\|d_k\|^2 \\
&= (\alpha_{\max} - \alpha_{\min})\|\nabla f(x^k)\|^2 + (L\alpha_{\max}^2 + \alpha_{\max} - \alpha_{\min})\|d_k\|^2 + 2\alpha_{\max}\nabla f(x^k)^\top d_k \\
&\leq (\alpha_{\max} - \alpha_{\min})\|\nabla f(x^k)\|^2 + (L\alpha_{\max}^2 + \alpha_{\max} - \alpha_{\min})c_1^2\|g_k(x^k)\|^2 + 2\alpha_{\max}\nabla f(x^k)^\top d_k,
\end{aligned}$$

where the last inequality follows from d_k being stochastic-gradient-related. By hypotheses, Lemma 6, and by taking the conditional expected value on both sides, we can then write

$$\begin{aligned}
2\mathbb{E}_k[f(x^{k+1}) - f(x^k)] &\leq (\alpha_{\max} - \alpha_{\min})\|\nabla f(x^k)\|^2 + (L\alpha_{\max}^2 + \alpha_{\max} - \alpha_{\min})c_1^2\mathbb{E}_k[\|g_k(x^k)\|^2] + 2\alpha_{\max}\nabla f(x^k)^\top \mathbb{E}_k[d_k] \\
&\leq (\alpha_{\max} - \alpha_{\min})\|\nabla f(x^k)\|^2 + (L\alpha_{\max}^2 + \alpha_{\max} - \alpha_{\min})c_1^2\mathbb{E}_k[\|g_k(x^k)\|^2] - 2\alpha_{\max}\left(c_2 - c_3\left(1 - \frac{1}{\rho}\right)\right)\|\nabla f(x^k)\|^2.
\end{aligned}$$

By recalling that the SGC condition is also satisfied and by suitably rearranging the terms, we get

$$\begin{aligned}
2\mathbb{E}_k[f(x^{k+1}) - f(x^k)] &\leq \left(\alpha_{\max} - \alpha_{\min} - 2\alpha_{\max}\left(c_2 - c_3\left(1 - \frac{1}{\rho}\right)\right)\right)\|\nabla f(x^k)\|^2 + (L\alpha_{\max}^2 + \alpha_{\max} - \alpha_{\min})c_1^2\rho\|\nabla f(x^k)\|^2 \\
&= -\left(2\alpha_{\max}\left(c_2 - c_3\left(1 - \frac{1}{\rho}\right)\right) + \alpha_{\min} - \alpha_{\max} - \rho c_1^2(L\alpha_{\max}^2 + \alpha_{\max} - \alpha_{\min})\right)\|\nabla f(x^k)\|^2.
\end{aligned}$$

Hence, if we set $\nu = \left(2\alpha_{\max}(c_2 - c_3(1 - \frac{1}{\rho})) + \alpha_{\min} - \alpha_{\max} - \rho c_1^2(L\alpha_{\max}^2 + \alpha_{\max} - \alpha_{\min})\right) = \left(\alpha_{\max}(2(c_2 - c_3(1 - \frac{1}{\rho})) - 1 - \rho c_1^2(L\alpha_{\max} + 1)) + \frac{2\delta c_2(1-\sigma)}{c_1^2 L_{\max}}(1 + \rho c_1^2)\right)$, we obtain

$$\nu\|\nabla f(x^k)\|^2 \leq 2\mathbb{E}_k\left[f(x^k) - f(x^{k+1})\right].$$

If $\nu > 0$, then the above inequality is equivalent to

$$\|\nabla f(x^k)\|^2 \leq \frac{2}{\nu}\mathbb{E}_k\left[f(x^k) - f(x^{k+1})\right].$$

By taking the total expectation, summing the terms corresponding to $k = 0, \ldots, K-1$ and dividing both sides by K, we have

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla f(x^k)\|^2] \leq \frac{2}{\nu K}\sum_{k=0}^{K-1}\mathbb{E}[f(x^k) - f(x^{k+1})] = \frac{2}{\nu K}\mathbb{E}[f(x^0) - f(x^K)] \leq \frac{2}{\nu K}(f(x^0) - f(x^\star)).$$

Therefore, we conclude that

$$\min_{k\in\{0,\ldots,K-1\}}\mathbb{E}[\|\nabla f(x^k)\|^2] \leq \frac{2}{\nu K}\left(f(x^0) - f(x^\star)\right),$$

which proves the result.

We now show that it indeed holds that $\nu > 0$.

Substituting $\alpha_{\min}$ in the definition of ν, we get

$$\begin{aligned}\nu &= 2\alpha_{\max}\left(c_2 - c_3\left(1 - \frac{1}{\rho}\right)\right) + \frac{2\delta c_2(1-\sigma)}{c_1^2 L_{\max}} - \alpha_{\max} - \rho c_1^2\left(L\alpha_{\max}^2 + \alpha_{\max} - \frac{2\delta c_2(1-\sigma)}{c_1^2 L_{\max}}\right)\\ &= -\rho c_1^2 L\alpha_{\max}^2 + \left(2\left(c_2 - c_3\left(1-\frac{1}{\rho}\right)\right) - 1 - \rho c_1^2\right)\alpha_{\max} + (1+\rho c_1^2)\frac{2\delta c_2(1-\sigma)}{c_1^2 L_{\max}}.\end{aligned}$$

Therefore, ν is concave quadratic in $\alpha_{\max}$. By computing

$$\Delta = \left(2\left(c_2 - c_3\left(1-\frac{1}{\rho}\right)\right) - 1 - \rho c_1^2\right)^2 + 4\rho c_1^2 L(1+\rho c_1^2)\frac{2\delta c_2(1-\sigma)}{c_1^2 L_{\max}} > 0,$$

we have that $\nu > 0$ if $\alpha_{\max} \in \left(\frac{(2(c_2 - c_3(1-\frac{1}{\rho})) - 1 - \rho c_1^2) - \sqrt{\Delta}}{2\rho c_1^2 L}, \frac{(2(c_2 - c_3(1-\frac{1}{\rho})) - 1 - \rho c_1^2) + \sqrt{\Delta}}{2\rho c_1^2 L}\right)$. Note that $\sqrt{\Delta} \geq |2(c_2 - c_3(1-\frac{1}{\rho})) - 1 - \rho c_1^2|$; hence, $\alpha_{\max} > \alpha_{\min} > \frac{(2(c_2 - c_3(1-\frac{1}{\rho})) - 1 - \rho c_1^2) - \sqrt{\Delta}}{2\rho c_1^2 L}$, since $\frac{(2(c_2 - c_3(1-\frac{1}{\rho})) - 1 - \rho c_1^2) - \sqrt{\Delta}}{2\rho c_1^2 L} < 0$ and $\alpha_{\max} > \alpha_{\min} > 0$ by assumption.

Hence, for any choice of $\alpha_{\max} \in \left(\alpha_{\min}, \frac{(2(c_2 - c_3(1-\frac{1}{\rho})) - 1 - \rho c_1^2) + \sqrt{\Delta}}{2\rho c_1^2 L}\right)$ it holds $\nu > 0$. Let us verify that this interval is nonempty, i.e., $\alpha_{\min} = \frac{2\delta c_2(1-\sigma)}{c_1^2 L_{\max}} < \frac{(2(c_2 - c_3(1-\frac{1}{\rho})) - 1 - \rho c_1^2) + \sqrt{\Delta}}{2\rho c_1^2 L}$. Equivalently, we want to show that

$$\sqrt{\Delta} > \frac{4\delta\rho c_2(1-\sigma)L}{L_{\max}} - \left(2\left(c_2 - c_3\left(1-\frac{1}{\rho}\right)\right) - 1 - \rho c_1^2\right).$$

Squaring both terms, we get

$$8\rho L(1+\rho c_1^2)\frac{\delta c_2(1-\sigma)}{L_{\max}} > \frac{16\delta^2\rho^2 c_2^2(1-\sigma)^2 L^2}{L_{\max}^2} - \frac{8\delta\rho c_2(1-\sigma)L}{L_{\max}}\left(2\left(c_2 - c_3\left(1-\frac{1}{\rho}\right)\right) - 1 - \rho c_1^2\right),$$

which is equivalent to

$$1 + \rho c_1^2 > \frac{2\delta\rho c_2(1-\sigma)L}{L_{\max}} - 2\left(c_2 - c_3\left(1-\frac{1}{\rho}\right)\right) + 1 + \rho c_1^2.$$

By rearranging, we can write

$$\delta\rho c_2(1-\sigma)L < \left(c_2 - c_3\left(1-\frac{1}{\rho}\right)\right)L_{\max},$$

which holds since $\sigma > 1 - \frac{(\rho c_2 - \rho c_3 + c_3)L_{\max}}{\delta\rho^2 c_2 L}$ by assumption.

Now, substituting the maximum value for σ into the upper-bound on $\alpha_{\max}$ yields

$$\frac{(2(c_2 - c_3(1-\frac{1}{\rho})) - 1 - \rho c_1^2) + \sqrt{(2(c_2 - c_3(1-\frac{1}{\rho})) - 1 - \rho c_1^2)^2 + 8(1+\rho c_1^2)(c_2 - c_3(1-\frac{1}{\rho}))}}{2\rho c_1^2 L} = \frac{2(\rho c_2 - \rho c_3 + c_3)}{\rho^2 c_1^2 L},$$

i.e.,

$$\alpha_{\max} \in \left(\alpha_{\min}, \frac{2(\rho c_2 - \rho c_3 + c_3)}{\rho^2 c_1^2 L}\right).$$

Since $\alpha_{\min} < \alpha_{\max} < \frac{2(\rho c_2 - \rho c_3 + c_3)}{\rho^2 c_1^2 L}$ by assumption, it thus follows that $\nu > 0$. □

Note that the result reported here also covers the case of non-monotone line search, since the same bound on $\alpha_{\min}$ holds; consequently, the same (sublinear) rate of convergence is obtained in that case as well.

D Revisiting the proof of Galli et al. (2023) under PL and interpolation

The non-monotone line search used in Galli et al. (2023) is, in essence, a direct extension to the stochastic setting of the procedure proposed by Zhang & Hager (2004), that is,

$$\begin{aligned} f_k(x^k - \alpha_k g_k(x^k)) &\leq C_k - \sigma\alpha_k \|g_k(x^k)\|^2, \\ C_k = \max\left\{\tilde{C}_k, f_k(x^k)\right\}, \quad \tilde{C}_k &= \frac{\xi Q_k C_{k-1} + f_k(x^k)}{Q_{k+1}}, \quad Q_{k+1} = \xi Q_k + 1, \end{aligned} \tag{12}$$

where $\xi \in [0, 1]$, $Q_0 = 0$ and $C_{-1} = f_0(x^0)$.
Galli et al. (2023) prove a linear convergence rate for SGD using this line search in the case of f satisfying the PL condition in the interpolation regime. We remark that the convergence result is valid only for some classes of L-smooth functions that satisfy the PL condition, where the smoothness constant L is not too large compared to the μ of the PL condition. In addition, we note that minimizer interpolation itself is not enough to prove the convergence result with the non-monotone rule, as the family of sampled functions $\{f_k\}$ needs to have a common optimal value, which is not directly implied by the former condition.

In the following, we revisit the proof of the results by Galli et al. (2023), starting from their Lemma 3, with the common optimal value assumption being now explicitly included. Compared to the proof by Galli et al. (2023) we do not make use of induction.

Lemma 7. *Let C_k be defined in equation 12. Assuming that the minimizer interpolation condition holds and that $f^\star = f_k(x^\star)$ for all k, the following bounds hold for all $k \in \mathbb{N}$,*

$$\begin{aligned} C_k - f_k(x^\star) &\geq 0, \quad \forall k, \\ C_{k-1} - f_k(x^\star) &\geq 0, \quad \forall k. \end{aligned}$$

Proof. We begin by noting that $C_k = \max\left\{\tilde{C}_k, f_k(x^k)\right\} \geq f_k(x^k)$ for all k; therefore, we have that

$$C_k - f_k(x^\star) \geq f_k(x^k) - f_k(x^\star) \geq 0,$$

since, from the minimizer interpolation condition, the solution $x^\star$ is a minimizer of f_k.

Similarly, we know that $C_{k-1} = \max\left\{\tilde{C}_{k-1}, f_{k-1}(x^{k-1})\right\} \geq f_{k-1}(x^{k-1})$ for all k and that $x^\star$ is a minimizer of f_{k-1}. From the assumption on the common optimal value we have that $f^\star = f_k(x^\star) = f_{k-1}(x^\star)$ and, thus,

$$C_{k-1} - f_k(x^\star) \geq f_{k-1}(x^{k-1}) - f_k(x^\star) \geq f_{k-1}(x^\star) - f_k(x^\star) = 0,$$

which concludes the proof. □

At this point, we can formally state the convergence result for SGD with the non-monotone line search of equation 12 under interpolation for f satisfying the PL condition. We remark that the rate of convergence obtained in this case is not general, as suitable algorithmic constants making it hold can only be set when $L < 4\mu$.

Proposition 3. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be an L-smooth function satisfying the PL condition, the minimizer interpolation property and that $f^\star = f_k(x^\star)$ for all k. Let the randomly drawn functions f_k considered at any iteration k be L_k-smooth, with $L_k \leq L_{max}$. Let the sequence $\{x^k\}$ be produced by iterating updates of the form $x^{k+1} = x^k - \alpha_k g_k(x^k)$, where, for all k, the stepsize α_k is chosen by non-monotone Armijo-type line search, i.e.,*

$$\alpha_k = \max_{j \in \mathbb{N}} \left\{ \delta^j \alpha_0^k \mid f_k(x^k - \delta^j \alpha_0^k g_k(x^k)) \leq C_k - \sigma \delta^j \alpha_0^k \|g_k(x^k)\|^2 \right\},$$

where C_k is defined according to equation 12, $\delta \in (0,1)$, $\sigma \in \left(\frac{L}{4\mu}, 1\right)$, and $\bar{\alpha}_{min} \leq \alpha_0^k \leq \alpha_{max} \in \left(\bar{\alpha}_{min}, \frac{2\delta\sigma(1-\sigma)}{\sigma L_{max} + \delta(1-\sigma)L - 4\mu\delta\sigma(1-\sigma)}\right)$ for all k, with $\bar{\alpha}_{min} > \frac{2\delta(1-\sigma)}{L_{max}}$. Then, defining $\nu = \alpha_{max}\left(\max\left\{\frac{1}{\bar{\alpha}_{min}}, \frac{L_{max}}{2\delta(1-\sigma)}\right\} + \frac{L}{2\sigma} + a_1 - 2\mu\right) \in (0,1)$ with $a_1 = \frac{4\mu\delta\sigma(1-\sigma) - \sigma L_{max} - \delta(1-\sigma)L}{4\delta\sigma(1-\sigma)} + \frac{1}{2\alpha_{max}} > 0$, and $a_2 = \xi\left(1 + \frac{L}{2\sigma a_1}\right) \in (0,1)$ with $\xi < \frac{2\sigma a_1}{2\sigma a_1 + L}$, we have that

$$\mathbb{E}[f(x^{k+1}) - f(x^\star) + a_1\alpha_{max}(C_k - f(x^\star))] \leq (a_3)^{k+1}(1 + a_1\alpha_{max})(f(x^0) - f(x^\star)),$$

where $a_3 = \max\{\nu, a_2\} \in (0,1)$.

Proof. By the L-smoothness of f, we have

$$f(x^{k+1}) \leq f(x^k) + \nabla f(x^k)^\top (x^{k+1} - x^k) + \frac{L}{2}\|x^{k+1} - x^k\|^2 = f(x^k) - \alpha_k \nabla f(x^k)^\top g_k(x^k) + \frac{L\alpha_k^2}{2}\|g_k(x^k)\|^2.$$

By rearranging the terms, summing the quantity $a_1(C_k - f_k(x^\star))$ on both sides, and using the stochastic Armijo-type condition and interpolation, we obtain

$$\begin{aligned}
\frac{f(x^{k+1}) - f(x^k)}{\alpha_k} + a_1(C_k - f_k(x^\star)) &\leq -\nabla f(x^k)^\top g_k(x^k) + \frac{L\alpha_k}{2}\|g_k(x^k)\|^2 + a_1(C_k - f_k(x^\star)) \\
&\leq -\nabla f(x^k)^\top g_k(x^k) + \frac{L}{2\sigma}(C_k - f_k(x^\star)) + a_1(C_k - f_k(x^\star)) \qquad (13) \\
&= -\nabla f(x^k)^\top g_k(x^k) + \left(\frac{L}{2\sigma} + a_1\right)(C_k - f_k(x^\star)).
\end{aligned}$$

Let us now distinguish two cases.

1. If $C_k = f_k(x^k)$, assuming $a_1, a_2 > 0$, we can obtain

$$\begin{aligned}
\frac{f(x^{k+1}) - f(x^k)}{\alpha_k} + a_1(C_k - f_k(x^\star)) &\leq -\nabla f(x^k)^\top g_k(x^k) + \left(\frac{L}{2\sigma} + a_1\right)(f_k(x^k) - f_k(x^\star)) \\
&\leq -\nabla f(x^k)^\top g_k(x^k) + \left(\frac{L}{2\sigma} + a_1\right)(f_k(x^k) - f_k(x^\star)) + a_1 a_2(C_{k-1} - f_k(x^\star)),
\end{aligned}$$

   where the last inequality follows from Lemma 7.

2. If $C_k = \tilde{C}_k$, by equation 13 we have

$$\begin{aligned}
\frac{f(x^{k+1}) - f(x^k)}{\alpha_k} + a_1(C_k - f_k(x^\star)) &\leq -\nabla f(x^k)^\top g_k(x^k) + \left(\frac{L}{2\sigma} + a_1\right)(\tilde{C}_k - f_k(x^\star)) \\
&= -\nabla f(x^k)^\top g_k(x^k) + \left(\frac{L}{2\sigma} + a_1\right)\frac{1}{\xi Q_k + 1}(f_k(x^k) - f_k(x^\star)) + \left(\frac{L}{2\sigma} + a_1\right)\frac{\xi Q_k}{\xi Q_k + 1}(C_{k-1} - f_k(x^\star)) \\
&\leq -\nabla f(x^k)^\top g_k(x^k) + \left(\frac{L}{2\sigma} + a_1\right)(f_k(x^k) - f_k(x^\star)) + \left(\frac{L}{2\sigma} + a_1\right)\xi(C_{k-1} - f_k(x^\star)),
\end{aligned}$$

   where the equality follows by the definition of $\tilde{C}_k$, whereas the second inequality follows from the minimizer interpolation property, Lemma 7 and $\frac{\xi Q_k}{\xi Q_k + 1} \leq \xi$, as shown in Lemma 2 of Galli et al. (2023).

By defining $a_2 = \xi\left(1 + \frac{L}{2\sigma a_1}\right)$, we get the same bound in both cases. Now, by taking the expectation w.r.t. x^k and by applying the PL condition, we obtain

$$\begin{aligned}\mathbb{E}_k\left[\frac{f(x^{k+1}) - f(x^k)}{\alpha_k} + a_1(C_k - f(x^\star))\right] &\leq -\|\nabla f(x^k)\|^2 + \left(\frac{L}{2\sigma} + a_1\right)(f(x^k) - f(x^\star)) + a_1 a_2(C_{k-1} - f(x^\star)) \\ &\leq \left(\frac{L}{2\sigma} + a_1 - 2\mu\right)(f(x^k) - f(x^\star)) + a_1 a_2(C_{k-1} - f(x^\star)),\end{aligned}$$

where the first inequality derives from the fact that C_{k-1} does not depend on k and from the unbiasedness of f_k and g_k. By the linearity of $\mathbb{E}_k[\cdot]$, we can now subtract from both sides $\mathbb{E}_k\left[\frac{f(x^\star)}{\alpha_k}\right]$ and rearrange the terms to obtain

$$\begin{aligned}\mathbb{E}_k\left[\frac{f(x^{k+1}) - f(x^\star)}{\alpha_k} + a_1(C_k - f(x^\star))\right] &\leq \mathbb{E}_k\left[\frac{f(x^k) - f(x^\star)}{\alpha_k}\right] + \left(\frac{L}{2\sigma} + a_1 - 2\mu\right)(f(x^k) - f(x^\star)) + a_1 a_2(C_{k-1} - f(x^\star)) \\ &\leq \left(\frac{1}{\alpha_{\min}} + \frac{L}{2\sigma} + a_1 - 2\mu\right)(f(x^k) - f(x^\star)) + a_1 a_2(C_{k-1} - f(x^\star)),\end{aligned}$$

where we have used the fact that $\alpha_k \geq \min\left\{\bar{\alpha}_{\min}, \frac{2\delta(1-\sigma)}{L_{\max}}\right\} = \alpha_{\min}$, as shown in Lemma 1 of Galli et al. (2023).

Recalling that $\alpha_k \leq \alpha_{\max}$ and taking the total expectation, we have that

$$\begin{aligned}\mathbb{E}\left[f(x^{k+1}) - f(x^\star) + a_1\alpha_{\max}(C_k - f(x^\star))\right] &\leq \nu\mathbb{E}[f(x^k) - f(x^\star)] + a_1 a_2\alpha_{\max}\mathbb{E}[C_{k-1} - f(x^\star)] \\ &\leq \max\{\nu, a_2\}(\mathbb{E}[f(x^k) - f(x^\star)] + a_1\alpha_{\max}\mathbb{E}[C_{k-1} - f(x^\star)]),\end{aligned}$$

where we define $\nu = \alpha_{\max}\left(\frac{1}{\alpha_{\min}} + \frac{L}{2\sigma} + a_1 - 2\mu\right) = \alpha_{\max}\left(\max\left\{\frac{1}{\bar{\alpha}_{\min}}, \frac{L_{\max}}{2\delta(1-\sigma)}\right\} + \frac{L}{2\sigma} + a_1 - 2\mu\right)$. Recursively applying the last result from k to 0, we obtain

$$\begin{aligned}\mathbb{E}[f(x^{k+1}) - f(x^\star) + a_1\alpha_{\max}(C_k - f(x^\star))] &\leq (a_3)^{k+1}(\mathbb{E}[f(x^0) - f(x^\star)] + a_1\alpha_{\max}\mathbb{E}[C_{-1} - f(x^\star)]) \\ &= (a_3)^{k+1}(1 + a_1\alpha_{\max})(f(x^0) - f(x^\star)),\end{aligned}$$

where $a_3 = \max\{\nu, a_2\}$, $C_{-1} = f_0(x^0)$ and $\mathbb{E}[f_0(x^0)] = \mathbb{E}_0[f_0(x^0)] = f(x^0)$ as the total expectation at this point only concerns the choice of the first mini-batch.

We will now show that $a_2 \in (0, 1)$, $\nu \in (0, 1)$ and, as a consequence, $a_3 \in (0, 1)$.

In order to prove that $a_2 \in (0, 1)$, we first show that the quantity

$$a_1 = \frac{4\mu\delta\sigma(1-\sigma) - \sigma L_{\max} - \delta(1-\sigma)L}{4\delta\sigma(1-\sigma)} + \frac{1}{2\alpha_{\max}} \tag{14}$$

is positive if the hypothesis $\alpha_{\max} < \frac{2\delta\sigma(1-\sigma)}{\sigma L_{\max} + \delta(1-\sigma)L - 4\mu\delta\sigma(1-\sigma)}$ holds:

$$\begin{aligned}a_1 &= \frac{4\mu\delta\sigma(1-\sigma) - \sigma L_{\max} - \delta(1-\sigma)L}{4\delta\sigma(1-\sigma)} + \frac{1}{2\alpha_{\max}} \\ &> \frac{4\mu\delta\sigma(1-\sigma) - \sigma L_{\max} - \delta(1-\sigma)L}{4\delta\sigma(1-\sigma)} + \frac{\sigma L_{\max} + \delta(1-\sigma)L - 4\mu\delta\sigma(1-\sigma)}{4\delta\sigma(1-\sigma)} = 0.\end{aligned}$$

Thus, $a_1 > 0$ and, by definition of a_2, the condition $a_2 > 0$ holds. Moreover, since by assumption $\xi < \frac{2\sigma a_1}{2\sigma a_1 + L}$, it follows that $a_2 = \xi\left(1 + \frac{L}{2\sigma a_1}\right) < 1$.

Let us prove now that $\nu \in (0, 1)$. By definition of $\alpha_{\min}$ and the assumption $\bar{\alpha}_{\min} > \frac{2\delta(1-\sigma)}{L_{\max}}$, we have that $\alpha_{\min} = \frac{2\delta(1-\sigma)}{L_{\max}}$. Thus, it follows that

$$\nu = \alpha_{\max}\left(\frac{\sigma L_{\max} + \delta(1-\sigma)L - 4\mu\delta\sigma(1-\sigma)}{2\delta\sigma(1-\sigma)} + a_1\right) = \alpha_{\max}\left(\frac{\sigma L_{\max} + \delta(1-\sigma)L - 4\mu\delta\sigma(1-\sigma)}{4\delta\sigma(1-\sigma)}\right) + \frac{1}{2},$$

where the second equality comes from the definition of a_1 (equation 14). By hypothesis $\alpha_{\max} < \frac{2\delta\sigma(1-\sigma)}{\sigma L_{\max}+\delta(1-\sigma)L-4\mu\delta\sigma(1-\sigma)}$, we get that $\nu < \frac{1}{2} + \frac{1}{2} = 1$. Furthermore, since $L_{\max} \geq L$ and it is known that $L \geq \mu$, we have

$$\begin{aligned}\sigma L_{\max} + \delta(1-\sigma)L - 4\mu\delta\sigma(1-\sigma) &\geq (\sigma + \delta - \delta\sigma)L - 4\mu\delta\sigma(1-\sigma) \\ &\geq \mu(\sigma + \delta(1 - 5\sigma + 4\sigma^2)) = \mu(\sigma + \delta(4\sigma-1)(\sigma-1)).\end{aligned} \tag{15}$$

We thus note that if $\sigma \in \left(0, \frac{1}{4}\right]$ then $\mu(\sigma + \delta(4\sigma-1)(\sigma-1)) > 0$. On the other hand, if $\sigma \in \left(\frac{1}{4}, 1\right)$ the quantity $(4\sigma-1)(\sigma-1)$ is negative; therefore, recalling that $\delta < 1$, we have

$$\mu(\sigma + \delta(4\sigma-1)(\sigma-1)) > \mu(\sigma + (4\sigma-1)(\sigma-1)) = \mu(2\sigma-1)^2 \geq 0. \tag{16}$$

This result also confirms that the upper bound on $\alpha_{\max}$ is positive. Recalling that $\alpha_{\max} > 0$, we have

$$\nu = \alpha_{\max}\left(\frac{\sigma L_{\max} + \delta(1-\sigma)L - 4\mu\delta\sigma(1-\sigma)}{4\delta\sigma(1-\sigma)}\right) + \frac{1}{2} > 0,$$

for all $\sigma \in (0,1)$, then concluding that $\nu \in (0,1)$.

At this point, we only need to ensure that $\frac{2\delta(1-\sigma)}{L_{\max}} < \bar{\alpha}_{\min} < \alpha_{\max} < \frac{2\delta\sigma(1-\sigma)}{\sigma L_{\max}+\delta(1-\sigma)L-4\mu\delta\sigma(1-\sigma)}$, which is equivalent to

$$0 < \frac{1}{L_{\max}}\left(-1 + \frac{\sigma L_{\max}}{\sigma L_{\max} + \delta(1-\sigma)L - 4\mu\delta\sigma(1-\sigma)}\right) = \frac{\delta}{L_{\max}}\left(\frac{-4\mu\sigma^2 + (4\mu + L)\sigma - L}{\sigma L_{\max} + \delta(1-\sigma)L - 4\mu\delta\sigma(1-\sigma)}\right).$$

Since the denominator is positive as shown in equations 15-16, the inequality holds for $\sigma \in \left(\frac{L}{4\mu}, 1\right)$. Consequently, the analysis conducted here applies only to functions that satisfy $L < 4\mu$. □

E Supplementary experimental results

In this appendix, we report results of experiments on CIFAR100 (Figure 8) and SVHN (Figure 9) problems which did not find space in Section 4.2 of the paper.

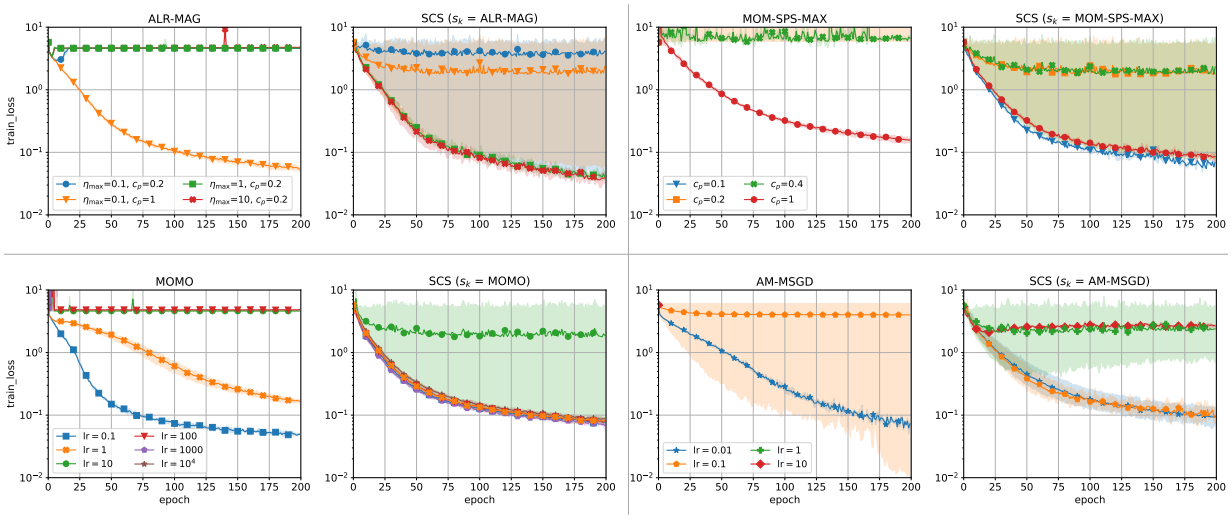

Figure 8: Training loss trends for ResNet-18* on CIFAR100 dataset. Each pair of plots corresponds to a momentum method with and without stochastic curve searches, for different values of the critical hyperparameters. For each algorithm, the mean over three runs is reported, with a shaded region indicating the minimum and maximum values.

Similarly to the results in the section, we observe that, overall, SCS is significantly more robust to aggressive choices of the hyper-parameters considered across the different methods.

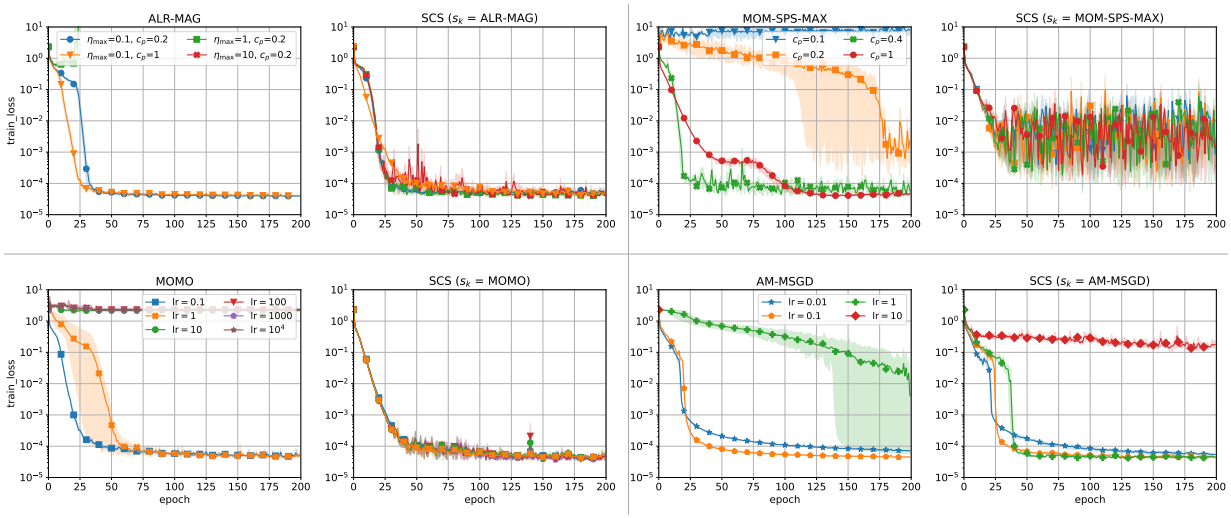

Figure 9: Training loss trends for ViT-B/32 on the SVHN dataset. Each pair of plots corresponds to a momentum method with and without stochastic curve searches, for different values of the critical hyperparameters. For each algorithm, the mean over three runs is reported, with a shaded region indicating the minimum and maximum values.