NON-CONVEX ONLINE SGD WITH DEPENDENT DATA: LAST-ITERATE CONVERGENCE AND TEMPORAL PRECONDITIONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Stochastic gradient descent (SGD) is usually analyzed under the assumption that data samples are independent and identically distributed. In practice, however, samples collected over time are often correlated, which can bias stochastic gradients and slow down the training process. We study online SGD for non-convex optimization problems when the data follow a temporally dependent ϕ-mixing process. Our analysis explains how the sampling strategy, the rate at which data correlations decay, and the optimization dynamics together determine last-iterate convergence. In particular, we investigate how subsampling and mini-batch sampling methods mitigate the effects of data dependence and how their effectiveness depends on the mixing rate. We also propose a preconditioner that uses online estimates of temporal dependence to adjust SGD updates. We study this approach in controlled settings with known additive noise. The proposed preconditioner complements our convergence analysis of sampling strategies by addressing temporal dependence through the SGD updates.

1 INTRODUCTION

We consider the unconstrained optimization problem $\min_{\theta\in\mathbb{R}^d} F(\theta) := \mathbb{E}_{\xi\sim\pi}[f(\theta;\xi)]$, where θ is the model parameter, ξ is a data sample, and π is the stationary marginal distribution of a data-generating process $\mathbf{P}$. Computing the full gradient can be expensive for a large dataset or impossible for a data stream whose future observations have not yet arrived. Stochastic gradient descent (SGD) replaces this full calculation with a gradient computed from a small amount of data (Robbins & Monro, 1951; Bottou et al., 2018). In its simplest online form, $\theta_{t+1} = \theta_t - \eta_t \nabla f(\theta_t;\xi_t)$, where $\eta_t > 0$ is the learning rate, and at each iteration, online SGD queries a sample ξ_t from the data-generating process.

SGD has well-established convergence guarantees for convex and strongly convex objectives (Polyak & Juditsky, 1992; Nemirovski et al., 2009; Moulines & Bach, 2011). For general non-convex problems, many convergence results for SGD show that the gradient becomes small, either at a randomly chosen iteration or on average over training (Ghadimi & Lan, 2013; Bottou et al., 2018). However, a small gradient does not by itself mean that the parameters are close to a minimum. We ask whether the current loss $F(\theta_t)$ approaches its minimum value, how quickly it does so, and whether the parameters θ_t converge to a single solution θ_*. Our *last-iterate* bounds track the loss of the current model directly, while our parameter convergence results describe whether the iterates approach one limiting point.

To obtain such results without convexity, it is useful to relate the size of the gradient to the loss. The Polyak–Łojasiewicz (PL) condition is a classical example (Polyak, 1963; Karimi et al., 2016). Related gradient lower bounds have been proved for sufficiently wide neural networks near suitable initializations, leading to convergence guarantees without convexity (Liu et al., 2022; Allen-Zhu et al., 2019; Du et al., 2019). More generally, in the zero-loss setting considered here, we study a *local gradient domination condition*, $|\nabla F(\theta)| \geq \alpha F(\theta)^\beta$, in a region around initialization, where $\alpha > 0$ and $\beta \in [1/2, 1)$; see also Weissmann et al. (2025). This condition ensures that the gradient cannot be too small relative to the remaining loss. The case $\beta = 1/2$ gives the local PL condition with target loss zero. To show that this condition holds in concrete models, Chatterjee (2022) ex-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

plicitly constructs feedforward neural networks with finite width and depth, together with suitable initializations, and verifies the local PL condition for their training loss. Extending this approach to SGD, An & Lu (2025) constructs network examples that also satisfy the additional assumptions needed for its convergence results. A key step in these local analyses is to show that the iterates stay in the region where the assumptions hold.

Data dependence creates another challenge. Observations collected over time are often related: consecutive driving frames show similar scenes, and observations along a reinforcement-learning trajectory depend on earlier states and actions (Geiger et al., 2012; Sutton & Barto, 2018). Such dependence can slow convergence and increase the number of observations needed to reach a given accuracy (Duchi et al., 2012; Ma et al., 2022). We study this issue using a single stationary data stream, with a ϕ-mixing coefficient $\phi_\xi(k)$ that measures how much information about the past can change the distribution of an observation k steps later. The condition $\phi_\xi(k) \to 0$ means that this influence vanishes as the time gap grows; see Doukhan (1994); Bradley (2005) for background on mixing. The main difficulty is that the current parameter θ_t depends on earlier samples and is therefore linked to the next observation. As a result, the expected stochastic gradient given the algorithm's history may differ from $\nabla F(\theta_t)$, even when averaging over the stationary distribution gives the correct gradient at every fixed parameter. This *conditional bias* is a central obstacle in the analysis of SGD with dependent data (Duchi et al., 2012; Doan et al., 2020).

We study two sampling rules that can reduce conditional bias, which were considered for convex objectives by Ma et al. (2022). *Subsampling* keeps one observation every r time steps and skips those in between. *Mini-batch sampling* which averages gradients from B consecutive observations at a fixed parameter. We analyze how these rules control conditional bias and variance under local gradient domination and loss-dependent noise. Squared bias enters our descent bounds at order η_t, while variance enters at order η_t^2. We distinguish update, observation, and gradient-evaluation costs when comparing the methods.

As a complementary approach, one can apply *diagonal preconditioning* to online SGD, using temporal correlations to choose a different step size for each coordinate. This changes how SGD responds to dependent noise. We study this approach for stochastic gradients of the form $g_t = \nabla F(\theta_t) + z_t$, where z_t is centered stationary additive noise. Errors at different times may reinforce or cancel one another. The coordinate-wise *long-run variance*, $\omega_i^{\mathrm{LRV}} = \gamma_i(0) + 2\sum_{k=1}^{\infty} \gamma_i(k)$, where $\gamma_i(k) = \mathrm{Cov}(z_{t,i}, z_{t+k,i})$, accounts for these effects when the covariance series is absolutely summable. Using the online estimator of Wu (2009), we propose an algorithm called LRV-SGD which forms normalized coordinate scales proportional to the inverse square roots of positive long-run variance estimates.

1.1 Our contributions

1. We establish quantitative last-iterate loss bounds and high-probability confinement for online SGD with ϕ-mixing data and loss-dependent noise. Our results cover subsampling and consecutive mini-batching under the local PL condition ($\beta = 1/2$) and the more general local gradient domination condition ($\beta \in (1/2, 1)$).
2. We propose LRV-SGD, which uses online long-run variance estimates to choose coordinate step sizes that account for temporal noise correlations. In controlled experiments on nonconvex problems with observable additive noise, LRV-SGD achieves lower loss than SGD, RMSProp, and Adam, showing the benefit of using temporal information in the updates.

1.2 Related work

Local convergence and loss-dependent noise. Over-parametrized neural network analyses prove convergence by controlling training near initialization (Allen-Zhu et al., 2019; Du et al., 2019). Chatterjee (2022) instead studies fixed finite width and depth, using smooth activations with positive derivatives, linearly independent inputs, and suitable initial weights. Thus, the input dimension must be at least the number of training examples. An & Lu (2025) extend this approach to SGD under an additional local assumption, proving convergence with positive probability. Dereich & Kassing (2024) study parameter convergence on the event that the iterates remain bounded, while Weissmann et al. (2025) give almost sure last-iterate loss rates under gradient domination. Fast convergence with noise that vanishes at a shared minimizer is also studied by Gower et al. (2019) and Vaswani et al.

(2019b). Our dependent data analysis requires conditional noise bounds at the current iterate, which unconditional bounds at a fixed parameter do not directly provide. In particular, fitting all training samples does not by itself show that a dependent stream satisfies these conditional bounds.

Dependent observations and sampling. Classical stochastic approximation covers Markovian noise (Kushner & Yin, 2003; Borkar, 2008); Agarwal & Duchi (2013) study generalization with mixing data. Existing work covers gradient methods driven by Markov chains (Sun et al., 2018; Even, 2023), methods that do not require knowing the mixing time (Dorfman & Levy, 2022), and constrained non-convex problems under more general dependent sampling (Alacaoglu & Lyu, 2023). Tadić (2015) proves convergence to one point under noise assumptions covering Markovian settings. Kar et al. (2026) give high probability PL-SGD loss bounds valid for all iterations; Sarkar et al. (2026) improve the dependence on mixing time and include a local-PL extension. The closest sampling comparison is Ma et al. (2022), who study subsampling and mini-batch sampling for convex objectives. Their analysis measures sample complexity, so progress per update should be distinguished from progress at a fixed observation budget. Our focus is the joint role of local geometry, loss-dependent conditional noise, and sampling in controlling last-iterate loss and the chance of leaving the local region, including exponents beyond the classical PL condition.

Adaptive steps and temporal covariance. AdaGrad, RMSProp, and Adam use squared gradients to set coordinate scales (Duchi et al., 2011; Shi et al., 2021; Kingma & Ba, 2015). Long-run variance instead includes covariances across time, a classical target in time-series estimation (Newey & West, 1987; Andrews, 1991). This quantity can be smaller than the one-time noise variance when errors are negatively correlated. We use Wu's existing constant memory estimator (Wu, 2009) and study how its output affects the loss under a chosen preconditioner. AdaShift addresses dependence between gradients and their adaptive scales (Zhou et al., 2019), while Lang et al. (2019) use statistical tests to change the learning rate. Motivated by finite step analyses (Liu et al., 2021), our quadratic formulas account for how the optimizer retains past errors; ordinary long-run variance gives a small step approximation. Estimation uses observed noise or centered residuals computed with fixed parameters during calibration.

2 PRELIMINARIES

Notations: Let $\{\mathcal{F}_t\}_{t\geq 0}$ be a filtration that contains only the observations actually used by the algorithm. We write $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_t]$. We also often use $a_t := \frac{\eta_t \sigma \kappa_t^2}{2} + \frac{C_L \eta_t^2 \sigma v_t}{2} + C_L \eta_t^2 \sigma \kappa_t^2$.

Let $\{\xi_t\}_{t\geq 0}$ be a stationary process taking values in a standard Borel space $\mathcal{X}$, with marginal law π. We focus on the standard optimization problem

$$\min_{\theta \in \mathbb{R}^d} F(\theta) := \mathbb{E}_{\xi \sim \pi}[f(\theta; \xi)]. \tag{1}$$

Throughout this work, we need $F \geq 0$ and state several assumptions, including geometry assumptions for the objective F (Assumption 2.1) and gradient noise assumptions (Assumptions 2.4-2.5).

Assumption 2.1. *For non-empty compact sets $\mathcal{U}_0 \subset \mathcal{U} \subset \mathcal{U}_1 \subset \mathbb{R}^d$, with $d_0 := \operatorname{dist}(\mathcal{U}_0, \mathcal{U}^c) > 0$, $d_1 := \operatorname{dist}(\mathcal{U}, \mathcal{U}_1^c) > 0$. Assume the following.*

- *(i) C_L-**smoothness:** There exists a constant $C_L > 0$ such that for all $\theta_1, \theta_2 \in \mathcal{U}_1$, $\|\nabla F(\theta_1) - \nabla F(\theta_2)\| \leq C_L \|\theta_1 - \theta_2\|$.*
- *(ii) **Confinement near boundary:** There exists $l_0 > 0$ such that $\inf_{\theta \in \mathcal{U} \setminus \mathcal{U}_0} F(\theta) \geq l_0$.*
- *(iii) **Local gradient domination on $\mathcal{U}$:** There exist constants $\alpha > 0$ and $\beta \in [1/2, 1)$ such that $\|\nabla F(\theta)\|^2 \geq \alpha F(\theta)^{2\beta}$ holds for every $\theta \in \mathcal{U}$.*

By compactness and continuity, there exists $\overline{C} > 0$ such that $\max_{\theta \in \mathcal{U}_1} \|\nabla F(\theta)\| = \overline{C}$.

We briefly discuss the assumptions here. (i) is standard. (iii) can be verified in several machine learning problems as discussed in Weissmann et al. (2025). When $\beta = 1/2$, it is the local Polyak–Łojasiewicz condition with target loss zero, and larger values of β describe flatter local geometry. (ii) was used in An & Lu (2025) in order to separate the region of small loss from the outer part of

$\mathcal{U}_0$. Together with control of the gradient growth, it helps bound the probability of leaving the local region. Building on the construction of Chatterjee (2022), they also give finite-width, finite-depth neural network examples satisfying both a local PL condition and this additional condition (ii), with suitable restrictions on the trainable parameters.

We also need an upper bound on the gradient in terms of the loss. The next lemma gives such a bound locally for a non-negative smooth function. (The proof is given in Appendix C.1.)

Lemma 2.2. *Suppose that $F : \mathbb{R}^d \to \mathbb{R}$ is a non-negative function, and ∇F is Lipschitz continuous in $\mathbb{R}^d$ with the constant C_L. Given a large compact set $\mathcal{U}_1 \supset \mathcal{U}$, there exists $\overline{C} > 0$ such that $\max_{\theta\in\mathcal{U}_1} \|\nabla F(\theta)\| = \overline{C}$. Then as long as $d_1 := \operatorname{dist}(\mathcal{U}, \mathcal{U}_1^c) \geq \overline{C}/C_L$, we have that, for any $\theta \in \mathcal{U}$, $\|\nabla F(\theta)\|^2 \leq 2C_L F(\theta)$.*

2.1 Dependence of data generation process

We adopt the definition from Agarwal & Duchi (2013) to quantify the dependence of the streaming data. This ϕ-mixing coefficient measures how much knowledge of the past can change the distribution of a later observation.

Definition 2.3 (ϕ-mixing process)**.** *Consider a stochastic process $\{\xi_t\}_{t\geq 0}$ be a stationary process on a standard Borel state space $\mathcal{X}$ with stationary marginal law π. Let $\mathbb{P}(\xi_{t+k} \in \cdot \mid \mathcal{F}_t)$ be the distribution of the $(t+k)$-th sample conditioned on $\mathcal{F}_t$, and let d_{TV} denote the total variation distance. Then the process $\{\xi_t\}_{t\in\mathbb{N}}$ is called ϕ-mixing if the following mixing coefficient*

$$\phi_\xi(k) := \sup_{t\in\mathbb{N},\, A\in\mathcal{F}_t} 2\, d_{\mathrm{TV}}\left(\mathbb{P}(\xi_{t+k} \in \cdot \mid A), \pi\right)$$

satisfies the property that $\phi_\xi(k) \to 0$ as $k \to \infty$.

2.2 Stochastic gradient decomposition

We write the general SGD update as $\theta_{t+1} = \theta_t - \eta_t G_t$, where G_t is a sampled stochastic gradient, which is $\mathcal{F}_t$-measurable. We define the dependence bias and centered fluctuation as

$$b_t := \mathbb{E}_t[G_t] - \nabla F(\theta_t), \qquad \varepsilon_t := G_t - \mathbb{E}_t[G_t], \qquad \mathbb{E}_t[\varepsilon_t] = 0, \tag{2}$$

so that the stochastic gradient G_t has the decomposition $G_t = \nabla F(\theta_t) + b_t + \varepsilon_t$. Throughout this work, we will also heavily utilize the local events $\Omega_t := \{\theta_k \in \mathcal{U}_0,\ \forall\ k \leq t\},\ \Omega_\infty := \bigcap_{t=0}^\infty \Omega_t$. Here, $\{\Omega_t\}_{t\geq 0}$ is a decreasing sequence of events such that $\Omega_{t+1} \subseteq \Omega_t,\ \Omega_t \in \mathcal{F}_t$.

Assumption 2.4. *There are positive deterministic sequences $\{\kappa_t\}_{t\geq 0}$ and $\{v_t\}_{t\geq 0}$ such that, almost surely, for some $\sigma > 0$, $\|b_t\|^2 \mathbb{1}_{\Omega_t} \leq \sigma\kappa_t^2 F(\theta_t)\mathbb{1}_{\Omega_t},\ \mathbb{E}_t[\|\varepsilon_t\|^2]\mathbb{1}_{\Omega_t} \leq \sigma v_t F(\theta_t)\mathbb{1}_{\Omega_t}$.*

Assumption 2.4 requires both the squared bias and the conditional variance to decrease with the loss. It is connected to the weak-growth condition studied by Vaswani et al. (2019a) and the loss-dependent noise model of Wojtowytsch (2023). Our formulation is local and conditional, and it allows the next stochastic gradient to be biased. The coefficient κ_t controls the remaining bias under the chosen sampling rule, while v_t controls the centered fluctuation.

To obtain concrete bounds on κ_t and v_t for different sampling rules, we impose a moment condition on the individual sample gradients.

Assumption 2.5. *Assume the stationary consistency identity $\int \nabla f(\theta; z)\, \pi(dz) = \nabla F(\theta)$ for all $\theta \in \mathcal{U}$. For some fixed $p \geq 2$ and $K_p < \infty$, as well as the same $\sigma > 0$ as in Assumption 2.4, we assume that, for every observation index i used in the update at time t, almost surely on Ω_t,*

$$\left(\mathbb{E}_t[\|\nabla f(\theta_t; \xi_i) - \nabla F(\theta_t)\|^p] + \int \|\nabla f(\theta_t; z) - \nabla F(\theta_t)\|^p\, \pi(dz)\right)^{2/p} \leq \sigma K_p F(\theta_t). \tag{3}$$

In (3), the first term controls the sample-gradient error under the conditional law of the next observations. The second controls the same error under the stationary law. Both are evaluated at the current parameter θ_t. The stationary consistency identity does not imply $\mathbb{E}_t[G_t] = \nabla F(\theta_t)$: the parameter θ_t depends on earlier observations, which can also affect the distribution of the samples used in G_t.

The next lemma separates the optimization argument from the sampling-specific estimates. Once κ_t and v_t have been bounded, it gives a common descent inequality for the different sampling rules.

Lemma 2.6. *Suppose Assumptions 2.1 and 2.4 hold and* $0 < \eta_t \le \frac{1}{4C_L}$. *With* $a_t = \frac{\eta_t \sigma \kappa_t^2}{2} + \frac{C_L \eta_t^2 \sigma v_t}{2} + C_L \eta_t^2 \sigma \kappa_t^2$, $d_1 = \operatorname{dist}(\mathcal{U}, \mathcal{U}_1^c) \ge \overline{C}/C_L$, *we then have the following localized one-step descent:*

$$\mathbb{E}\left[F(\theta_{t+1}) \mathbb{1}_{\Omega_{t+1}}\right] \le (1 + a_t) \mathbb{E}\left[F(\theta_t) \mathbb{1}_{\Omega_t}\right] - \frac{\alpha \eta_t}{4} \mathbb{E}\left[F(\theta_t)^{2\beta} \mathbb{1}_{\Omega_t}\right]. \tag{4}$$

3 Localized last-iterate convergence

We now apply the one-step descent bound equation 4 to online SGD with dependent data. We consider subsampling and consecutive mini-batch sampling, and explain how each scheme controls the conditional bias and variance. For each scheme, we treat two cases: the local PL case $\beta = 1/2$ and the more general case $\beta \in (1/2, 1)$. This distinction matters because a fixed sampling rule and a constant step size can suffice in the PL case, whereas our results for $\beta > 1/2$ use decreasing step sizes and growing sampling gaps or batch sizes.

3.1 Subsampling

Subsampling updates the parameter using selected observations and skips those in between. Let $r_t \ge 1$ be an integer gap, we define the observation times by $s_0 = 0$, and $s_{t+1} = s_t + r_t,\ t \ge 0$. At iteration t, the algorithm uses the next retained observation $\xi_{s_{t+1}}$:

$$G_t = \nabla f(\theta_t; \xi_{s_{t+1}}), \qquad \theta_{t+1} = \theta_t - \eta_t G_t. \tag{5}$$

The filtration is $\mathcal{F}_t = \sigma(\theta_0, \xi_{s_1}, \ldots, \xi_{s_t})$. For $t \ge 0$, the next retained observation is separated from the most recent retained observation by r_t original time steps.

The next lemma gives the resulting coefficients in the conditional bias and fluctuation bounds.

Lemma 3.1. *Suppose the data process is stationary and satisfies the ϕ-mixing condition, and Assumptions 2.4 and 2.5 hold. For the exponent $p \ge 2$ and constant $K_p < \infty$ in the moment bounds, the coefficients may be chosen as $\kappa_t^2 = K_p \phi_\xi(r_t)^{2-2/p},\ v_t = K_p$.*

Theorem 3.2. *Consider online SGD with subsampling (5)with a fixed integer gap $r_t \equiv r \ge 1$ and a constant step size $\eta_t \equiv \eta > 0$. Suppose Assumption 2.1 holds with $\beta = 1/2$, and Assumptions 2.4 and 2.5 hold. Let $0 < \eta \le \min\left\{\frac{1}{4C_L}, \frac{6}{\alpha}\right\}$, and PL condition coefficient $\alpha \ge \sigma\left(9K_p \phi_\xi(r)^{2-2/p} + \frac{3K_p}{2}\right)$, then for every $\delta \in (0,1)$, there exists $\varepsilon > 0$ such that, if $\theta_0 \in \mathcal{U}_0$ and $F(\theta_0) \le \varepsilon$, the sequence $\{\theta_t\}_{t \ge 0}$ converges to a point $\theta_* \in \mathcal{U}_0$ with probability at least $1 - \delta$. Moreover, set $\rho_r := 1 - \alpha\eta/6$. For every $\gamma \in [1, \rho_r^{-1})$,*

$$\lim_{t \to \infty} \gamma^t F(\theta_t) = 0, \qquad \text{almost surely on } \Omega_\infty. \tag{6}$$

For $\beta > 1/2$, we then keep the Robbins-Monro type learning rate η_t, and furthermore require gaps $\{r_t\}_{t \ge 0}$ to be an increasing sequence.

Theorem 3.3. *Suppose Assumption 2.1 holds with $\beta \in (1/2, 1)$, and Assumptions 2.4 and 2.5 hold. Let $\theta_0 \in \mathcal{U}_0$, and choose step sizes satisfying $0 < \eta_t \le \frac{1}{4C_L}$, $\sum_{t=0}^\infty \eta_t = \infty$, $\sum_{t=0}^\infty \eta_t^2 < \infty$. Furthermore, we choose an increasing sequence $\{r_t\}_{t \ge 0}$ such that $\sum_{t=0}^\infty \eta_t \phi_\xi(r_t)^{2-2/p} < \infty$. Then $\{a_t\}_{t \ge 0}$ is uniformly bounded and $A := \sum_{t=0}^\infty a_t < \infty$. We set $\bar{a} := \sup_{t \ge 0} a_t$, $P_t := \prod_{s=0}^{t-1}(1 + a_s)$, $e^A := \prod_{s=0}^\infty (1 + a_s) \le e^A$, $T_t := \sum_{s=0}^{t-1} \eta_s$. Then the localized expected loss satisfies*

$$\mathbb{E}[F(\theta_t) \mathbb{1}_{\Omega_t}] \le e^A \left[F(\theta_0)^{-(2\beta-1)} + \frac{\alpha(2\beta - 1)}{2(1 + \bar{a})} T_t\right]^{-1/(2\beta-1)}, \tag{7}$$

Furthermore, for every $\delta \in (0,1)$, there exists $\varepsilon > 0$ such that, if $F(\theta_0) \le \varepsilon$, the sequence $\{\theta_t\}_{t \ge 0}$ converges to a point $\theta_* \in \mathcal{U}_0$ with probability at least $1 - \delta$. For every $\gamma \in (0, 1/(2\beta - 1))$,

$$\lim_{t \to \infty} T_t^\gamma F(\theta_t) = 0, \qquad \text{almost surely on } \Omega_\infty. \tag{8}$$

These results separate the role of initialization from that of the sampling schedule. For $\theta_0 \in \mathcal{U}_0$, a sufficiently small initial loss gives the stated high-probability parameter convergence, while the loss

rates apply on the confinement event Ω_∞. On this event, the loss tends to zero; any parameter limit there is therefore a zero-loss minimum by continuity of F.

The conditions also explain why bias and variance need different controls. The one-step error coefficient a_t contains a term proportional to $\eta_t\kappa_t^2$ and a term proportional to $\eta_t^2 v_t$. For subsampling, $v_t = K_p$ is bounded, so variance is controlled by $\sum_t \eta_t^2 < \infty$. The separate condition $\sum_t \eta_t\kappa_t^2 < \infty$ controls the accumulated bias and determines how quickly the gaps must grow in the case $\beta > 1/2$.

3.2 MINI-BATCH SAMPLING

Mini-batch sampling consecutively uses every observation. At each iteration, it computes gradients from the next block at the same parameter and averages them to make one update. Let $\{B_t\}_{t\geq 0}$ be integer batch sizes and set $S_t := \sum_{s=0}^{t-1} B_s,\ S_0 = 0$. At iteration t, we use the next consecutive block to update, so that the stochastic gradient and the corresponding filtration are

$$G_t = \frac{1}{B_t}\sum_{j=1}^{B_t} \nabla f(\theta_t; \xi_{S_t+j}), \qquad \mathcal{F}_t = \sigma(\theta_0, \xi_1, \ldots, \xi_{S_t}). \tag{9}$$

The first observation in the new block is only one time step beyond this history. Later observations are farther away, so their conditional bias can be bounded using the mixing coefficient at larger gaps. Averaging these bounds reduces the bias contribution, while averaging the gradients can also reduce random fluctuations. The next lemma accounts for both effects and retains the dependence within each batch.

Lemma 3.4. *Suppose the data process is stationary and satisfies the $\phi-$mixing condition, and Assumptions 2.4 and 2.5 hold, then for some $p \geq 2$ and $K_p < \infty$, for each $B_t \geq 1$,*

$$\kappa_t^2 = K_p\left[\frac{1}{B_t}\sum_{k=1}^{B_t}\phi_\xi(k)^{1-1/p}\right]^2, \quad v_t = \frac{K_p}{B_t} + \frac{2K_p}{B_t^2}\sum_{k=1}^{B_t-1}(B_t-k)\phi_\xi(k)^{1-1/p}. \tag{10}$$

The second summation is zero when $B_t = 1$.

These formulas also apply to a fixed batch size B. The first term in v_t comes from the individual gradient fluctuations, while the second bounds the covariances between different observations in the batch. The bias coefficient averages the mixing bounds over the whole block, rather than using only the bound for its first observation.

Lemma 3.4 also gives asymptotic bounds on κ_t^2 and v_t as $B_t \to \infty$, depending on how quickly $\phi_\xi(k)^{1-1/p}$ decays. In particular, if $\sum_{k\geq 1}\phi_\xi(k)^{1-1/p} < \infty$, then $\kappa_t^2 = O(B_t^{-2})$ and $v_t = O(B_t^{-1})$. See Remark E.1 in the appendix for further details.

We now apply these mixing coefficient bounds to the convergence analysis. As with subsampling, we first consider the local PL case. A fixed batch size suffices when the resulting bias and fluctuation bounds are small enough relative to the gradient lower bound.

Theorem 3.5. *Consider online SGD with mini-batch sampling (9). Suppose Assumption 2.1 holds with $\beta = 1/2$, and Assumptions 2.4 and 2.5 hold. Use a fixed batch size $B_t \equiv B$, and let κ_B and v_B be the coefficients in Lemma 3.4. Assume. Let the fixed step size $0 < \eta \leq \min\left\{\frac{1}{4C_L}, \frac{6}{\alpha}\right\}$, and the PL-condition coefficient $\alpha \geq \sigma\left(9\kappa_B^2 + \frac{3v_B}{2}\right)$, then by choosing small $F(\theta_0)$, with high probability, Ω_∞ occurs, and for every $\lambda \in [1, \rho_B^{-1})$, $\rho_B := 1 - \frac{\alpha\eta}{6}$,*

$$\lim_{t\to\infty} \lambda^t F(\theta_t) = 0 \quad \textit{almost surely on } \Omega_\infty. \tag{11}$$

For $\beta > 1/2$, we let the batch size grow so that the accumulated bias is finite. Remark E.1 also demonstrates quantitatively how the required growth depends on the decay of the mixing coefficient.

Theorem 3.6. *Suppose Assumption 2.1 holds with $\beta \in (1/2, 1)$, and Assumptions 2.4-2.5 are satisfied. Consider growing batch sizes $\{B_t\}_{t\geq 0}$ and Robbins-Monro type learning rates η_t, such that $0 < \eta_t \leq \frac{1}{4C_L}$, $\sum_{t=0}^\infty \eta_t = \infty$, $\sum_{t=0}^\infty \eta_t^2 < \infty$, and $\sum_{t=0}^\infty \eta_t\kappa_{B_t}^2 < \infty$. Let $T_t = \sum_{s=0}^{t-1}\eta_s$. Then by choosing small $F(\theta_0)$, with high probability, Ω_∞ occurs, and for every $\lambda \in (0, 1/(2\beta-1))$,*

$$\lim_{t\to\infty} T_t^\lambda F(\theta_t) = 0 \quad \textit{almost surely on } \Omega_\infty. \tag{12}$$

We mention that Appendix A.2 presents numerical experiments on a non-convex phase retrieval problem with temporally dependent data. We compare subsampling with different gaps and mini-batch sampling with different batch sizes over the same number of parameter updates. Both sampling scheme reduce fluctuations in the test loss compared to vanilla SGD, and larger gaps and batch sizes can result in lower losses, which are consistent with our theoretical findings in this section.

4 A LONG-RUN VARIANCE PRECONDITIONER

Aside from sampling methods such as subsampling and mini-batch sampling to reduce the data dependence for better convergence, one may wonder whether there are other algorithmic approaches to reduce the dependence. In fact, for online SGD, suppose the stochastic gradient is of the structure $G_t = \nabla F(\theta_t) + z_t$, where the additive noise z_t is observable at each update. We consider the preconditioned iteration $\theta_{t+1} = \theta_t - \eta P_t G_t$, where P_t is a diagonal matrix with diagonal entries $\{p_{t,i}\}_{i=1}^d$. Although z_t is observable, its temporal dependence structure is generally unknown. We therefore estimate the long-run variance (LRV) of each coordinate from the observed noise sequence and use these estimates to choose the scales.

Assume that $\{z_t\}$ is second-order stationary. For each coordinate $i \in \{1, \dots, d\}$, define the lag-k autocovariance, marginal variance, and long-run variance by

$$\gamma_i(k) = \mathrm{Cov}(z_{t,i}, z_{t+k,i}), \qquad \sigma_i^2 = \gamma_i(0), \qquad \omega_i = \gamma_i(0) + 2\sum_{k\geq 1} \gamma_i(k). \tag{13}$$

We assume that the autocovariances are absolutely summable. The marginal variance σ_i^2 measures fluctuations at a single update, whereas the LRV ω_i describes how noise accumulates over many updates. Indeed, $\lim_{n\to\infty} \frac{1}{n} \mathrm{Var}(\sum_{t=1}^n z_{t,i}) = \omega_i$. Thus, coordinates with the same marginal variance can have different levels of accumulated noise because of their temporal correlations.

Let $\omega = (\omega_1, \dots, \omega_d)$, and assume that the marginal variances $\sigma_i^2 > 0$ are known. Given positive LRV estimates $s = (s_1, \dots, s_d)$, we set $\mathcal{P}_i(s) := \frac{\chi(s)}{\sqrt{s_i}}$, $\quad \chi(s) = \left(\frac{d}{\sum_{j=1}^d \sigma_j^2 / s_j} \right)^{1/2}$.

The factor $s_i^{-1/2}$ assigns smaller relative scales to coordinates with larger estimated LRV. The common factor $\chi(s)$ fixes the overall scale through the normalization $\sum_{i=1}^d \sigma_i^2 \mathcal{P}_i(s)^2 = d$. If every $\omega_i > 0$ and the true LRV is used, then $p_i = \mathcal{P}_i(\omega)$ satisfies $p_i^2 \omega_i = \chi(\omega)^2$ for every i. Hence, these scales equalize the long-run noise variances across coordinates. This provides a rule for balancing accumulated noise, without implying that the resulting preconditioner is optimal for minimizing F.

Estimating the LRV online. We call the resulting method LRV-SGD. It combines the scaling rule with the centered recursive LRV estimator proposed by Wu (2009). For a block-scale parameter $c > 0$, blocks start at time 1 and at the distinct values of $\lfloor ck^{3/2} \rfloor$ exceeding 1, for $k = 1, 2, \dots$, taken in increasing order. The estimator uses partial sums within these blocks and incorporates each new observation immediately, without waiting for the current block to end. Algorithm 1 requires $O(d)$ work and memory per step. Here, W and ℓ denote the sum and length of the current block, respectively, and $\odot$ denotes coordinatewise multiplication.

During warm-up, $p = \mathbf{1}$, so the method uses ordinary SGD. Thereafter, the scales are refreshed every I updates and remain fixed between refreshes. The floor τ keeps the LRV estimates positive, smoothing controls changes in the scales, and coordinatewise clipping keeps each scale in $[1/R, R]$. At a refresh, the estimator incorporates z_t before the SGD step, so the current noise observation can influence the scales used in the current update. The normalization above holds for the map $\mathcal{P}(s)$; smoothing and clipping may change it.

5 NUMERICAL EXPERIMENTS

We use numerical experiments to examine how temporal dependence affects SGD and how sampling and long-run variance preconditioning change its performance.

Algorithm 1 LRV-SGD with observable additive noise

Input: Initial iterate θ_1, number of updates T, step size η, marginal variances $\{\sigma_i^2\}_{i=1}^d$, and block scale c; warm-up length T_w, refresh interval I, smoothing parameter ρ, clipping bound R, and floor $\tau > 0$.

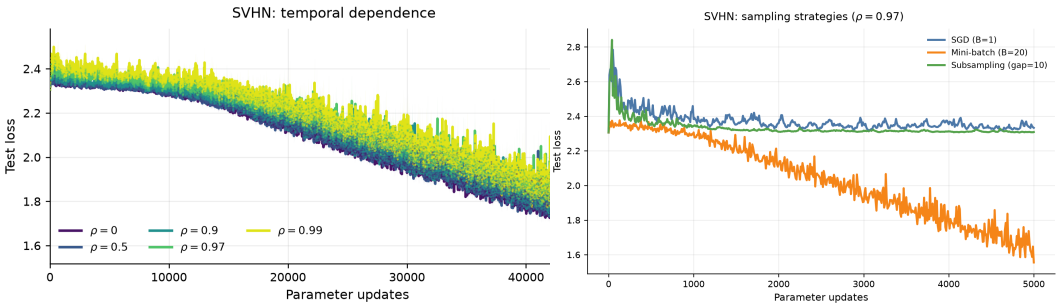
1. Initialize $p = \mathbf{1}$; $S, W, V, U = \mathbf{0}$; $\ell, Q, D = 0$.
2. **for** $t = 1, \dots, T$ **do**
3. Receive $G_t = \nabla F(\theta_t) + z_t$ and z_t; if t starts a new block, reset $W = \mathbf{0}$ and $\ell = 0$.
4. $W \leftarrow W + z_t$; $\ell \leftarrow \ell + 1$; $S \leftarrow S + z_t$.
5. $V \leftarrow V + W^{\odot 2}$; $U \leftarrow U + \ell W$; $Q \leftarrow Q + \ell^2$; $D \leftarrow D + \ell$.
6. $\bar{z} \leftarrow S/t$; $\widehat{\omega} \leftarrow (V - 2U \odot \bar{z} + Q\bar{z}^{\odot 2})/D$.
7. **if** $t > T_w$ and $(t - T_w) \bmod I = 0$ **then**
8. For each i, set $s_i \leftarrow \tau$ if $\widehat{\omega}_i$ is NaN or $\pm\infty$; otherwise, set $s_i \leftarrow \max\{\tau, \widehat{\omega}_i\}$.
9. $p \leftarrow \mathrm{clip}_{[1/R,R]}\big((1-\rho)p + \rho\mathcal{P}(s)\big)$.
10. **end if**
11. $\theta_{t+1} \leftarrow \theta_t - \eta p \odot G_t$.
12. **end for**

5.1 Real-image learning with refresh-Markov sampling

We first study image classification on SVHN using a data stream with controlled temporal dependence. The latent class follows a stationary refresh-Markov chain with transition matrix $P_\rho = \rho I + (1-\rho)\mathbb{1}\pi^\top$, $\pi = (1/10, \dots, 1/10)$. Thus, the chain keeps its current class with probability ρ and otherwise draws a class from π. At each step, we set the observed label to the latent class with probability 0.2; otherwise, we draw it uniformly from the ten classes. We then sample an image with replacement from the corresponding SVHN training class. This construction keeps the class-balanced image distribution the same for every ρ, while changing the dependence across observations. The observation process inherits a geometric mixing bound from the latent chain.

We train all parameters of a three-block CNN with channels $(16, 32, 64)$, GroupNorm, ReLU, and a classification head. We use cross-entropy loss and process observations in time order with SGD, without momentum, weight decay, data augmentation, shuffling, or gradient clipping. Test loss is evaluated on a fixed, class-balanced subset of the official test split containing $15{,}950$ images. To study the effect of dependence, we use single sample SGD with learning rate $\eta = 0.003$ for $50{,}000$ updates and vary $\rho \in \{0, 0.5, 0.9, 0.97, 0.99\}$. The case $\rho = 0$ gives i.i.d. sampling. Figure 1(a) shows that the strongest dependence leads to a higher final test loss than i.i.d. sampling. The intermediate dependence levels, however, do not follow a monotonic ordering.

We next compare sampling strategies at $\rho = 0.97$. We use the same learning rate $\eta = 0.03$ for single sample SGD, consecutive mini-batches of size $B = 20$, and subsampling with a gap of 10. Figure 1(b) compares these methods over the same number of parameter updates. Mini-batch SGD reaches a lower test loss in this comparison. The comparison measures progress per update: each mini-batch update uses 20 gradients, while subsampling uses one gradient every 10 stream observations, so the methods have different data and computation costs.



(a) Different dependence levels.

(b) Different sampling strategies at $\rho = 0.97$.

Figure 1: Experiments on refresh-Markov SVHN. (a) Performance under different dependence levels. (b) Comparison of sampling strategies at $\rho = 0.97$.

5.2 LONG-RUN VARIANCE PRECONDITIONING

We next study whether temporal covariance information can guide coordinate scaling. To compare the methods under controlled conditions, we use a shared noise stream and the 10-dimensional coupled tilted double-well objective $F(\theta) = \sum_{i=1}^{10} \left[\frac{1}{4}(\theta_i^2 - 1)^2 + 0.2\theta_i\right] + \frac{0.1}{2}\sum_{i=1}^{9}(\theta_{i+1} - \theta_i)^2$. For the scalar function $\phi(u) = \frac{1}{4}(u^2 - 1)^2 + 0.2u$, the stationary points solve $u^3 - u + 0.2 = 0$. Let u_- and u_+ be the smallest and largest roots, corresponding to the left and right local minima. The left minimum is the global minimum. Since the coupling term in double-well $F(\theta)$ is non-negative and vanishes when all coordinates are equal, the global minimum value is $F^\star = F(u_-\mathbf{e})$, where $\mathbf{e} = (1, \ldots, 1)^\top \in \mathbb{R}^{10}$.

We consider two initializations: the right local minimum $\theta_0 = u_+\mathbf{e}$ and the origin $\theta_0 = \mathbf{0}$. The origin lies between the wells but is not stationary, since $\nabla F(\mathbf{0}) = 0.2\mathbf{e}$. We report the excess loss per coordinate, $(F(\theta_t) - F^\star)/10$. The stochastic gradients have the form $g_t = \nabla F(\theta_t) + z_t$, where the coordinates of z_t follow independent stationary AR(1) processes. All coordinates have unit marginal variance, but their different temporal correlations lead to different long-run variances. LRV-SGD observes the exact noise z_t for variance estimation. This controlled setup allows us to assess the value of temporal information beyond the variance at a single time.

We compare LRV-SGD with SGD, RMSProp, bias-corrected Adam (Kingma & Ba, 2015), and the fixed population LRV rule. The population rule uses the true LRVs and is applied from the first update. The two online variants, LRV-SGD-c1 and LRV-SGD-c8, use $c = 1$ and $c = 8$, respectively. All methods use the same noise stream.

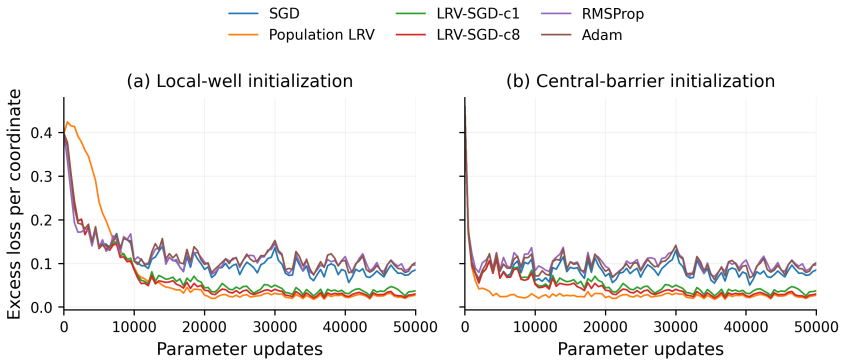


Figure 2: Ten-dimensional double-well optimization under two initializations. The vertical axis shows excess loss per coordinate. The initialization affects the early optimization trajectories, while each method reaches similar loss levels under both initializations later in training. Both LRV-SGD variants attain lower excess loss than SGD. LRV-SGD-c8 attains a lower loss than LRV-SGD-c1 and performs close to the fixed population LRV rule.

6 CONCLUSION

We establish last-iterate convergence results for online SGD with dependent data under local gradient domination and loss-dependent noise. Our analysis explains how subsampling and mini-batch sampling control conditional bias and fluctuations, and how these controls help keep the iterates in the region where convergence is guaranteed. All proofs can be found in the appendix. For observable additive noise, we study a separated approach based on long-run variance. The experiments illustrate the benefits of incorporating sampling and temporal covariance information in online SGD for training dependent data. For future work, one could develop a rigorous convergence analysis of LRV-SGD with observable noise.

AI USE STATEMENT

In this work, we used generative AI tools for improve the paper's wording and organization, check references, and suggest corrections to mathematical statements and notation. We have not used

generative AI tools for proposing research ideas and generating mathematical proofs, neither for the numerical experiments. We have reviewed all AI-assisted work, and we take responsibility for the final text, mathematical claims, references, and conclusions.

Reproducibility Statement

The code, experimental configurations, and scripts used to reproduce the numerical results are available at `https://anonymous.4open.science/r/sgd-dependent-data-code-E4B4/`. Additional implementation details and experimental settings are provided in the corresponding sections and appendices.

References

Alekh Agarwal and John C. Duchi. The generalization ability of online algorithms for dependent data. *IEEE Transactions on Information Theory*, 59(1):573–587, 2013. URL `https://arxiv.org/abs/1110.2529`.

Ahmet Alacaoglu and Hanbaek Lyu. Convergence of first-order methods for constrained nonconvex optimization with dependent data. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 458–489. PMLR, 2023. URL `https://proceedings.mlr.press/v202/alacaoglu23a.html`.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 242–252. PMLR, 2019. URL `https://proceedings.mlr.press/v97/allen-zhu19a.html`.

Jing An and Jianfeng Lu. Convergence of stochastic gradient descent under a local Łojasiewicz condition for deep neural networks. *Journal of Machine Learning*, 4(2):89–107, 2025. doi: 10.4208/jml.240724. URL `https://www.global-sci.com/jml/article/view/13210`.

Donald W. K. Andrews. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59(3):817–858, 1991. doi: 10.2307/2938229.

Vivek S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*, volume 48 of *Texts and Readings in Mathematics*. Hindustan Book Agency, 2008. doi: 10.1007/978-93-86279-38-5.

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

Richard C. Bradley. Basic properties of strong mixing conditions. a survey and some open questions. *Probability Surveys*, 2:107–144, 2005. doi: 10.1214/154957805100000104. URL `https://arxiv.org/abs/math/0511078`.

Sourav Chatterjee. Convergence of gradient descent for deep neural networks, 2022. URL `https://arxiv.org/abs/2203.16462`. arXiv preprint.

Steffen Dereich and Sebastian Kassing. Convergence of stochastic gradient descent schemes for Łojasiewicz-landscapes. *Journal of Machine Learning*, 3(3):245–281, 2024. doi: 10.4208/jml.240109. URL `https://www.global-sci.com/jml/article/view/13201`.

Thinh T. Doan, Lam M. Nguyen, Nhan H. Pham, and Justin Romberg. Finite-time analysis of stochastic gradient descent under markov randomness. *SIAM Journal on Optimization*, 30(3):2684–2718, 2020.

Ron Dorfman and Kfir Yehuda Levy. Adapting to mixing time in stochastic optimization with Markovian data. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 5429–5446. PMLR, 2022. URL `https://proceedings.mlr.press/v162/dorfman22a.html`.

Paul Doukhan. *Mixing: Properties and Examples*, volume 85 of *Lecture Notes in Statistics*. Springer, 1994. doi: 10.1007/978-1-4612-2642-0.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1675–1685. PMLR, 2019. URL https://proceedings.mlr.press/v97/du19c.html.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011. URL https://jmlr.org/papers/v12/duchi11a.html.

John C. Duchi, Alekh Agarwal, Mikael Johansson, and Michael I. Jordan. Ergodic mirror descent. *SIAM Journal on Optimization*, 22(4):1549–1578, 2012. doi: 10.1137/110836043. URL https://arxiv.org/abs/1105.4681.

Mathieu Even. Stochastic gradient descent under Markovian sampling schemes. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 9412–9439. PMLR, 2023. URL https://proceedings.mlr.press/v202/even23a.html.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3354–3361, 2012. doi: 10.1109/CVPR.2012.6248074. URL https://www.cvlibs.net/publications/Geiger2012CVPR.pdf.

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

Robert M. Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. SGD: General analysis and improved rates. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5200–5209. PMLR, 2019. URL https://proceedings.mlr.press/v97/qian19b.html.

Avik Kar, Siddharth Chandak, Rahul Singh, Eric Moulines, Shalabh Bhatnagar, and Nicholas Bambos. High-probability bounds for SGD under the Polyak–Łojasiewicz condition with Markovian noise, 2026. URL https://arxiv.org/abs/2603.14514.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak–Łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases*, volume 9851 of *Lecture Notes in Computer Science*, pp. 795–811. Springer, 2016. doi: 10.1007/978-3-319-46128-1_50. URL https://arxiv.org/abs/1608.04636.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Harold J. Kushner and G. George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2 edition, 2003. doi: 10.1007/b97441.

Hunter Lang, Lin Xiao, and Pengchuan Zhang. Using statistics to automate stochastic optimization. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/e1054bf2d703bca1e8fe101d3ac5efcd-Abstract.html.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022. doi: 10.1016/j.acha.2021.12.009. URL https://doi.org/10.1016/j.acha.2021.12.009.

Kangqiao Liu, Liu Ziyin, and Masahito Ueda. Noise and fluctuation of finite learning rate stochastic gradient descent. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7045–7056. PMLR, 2021. URL https://proceedings.mlr.press/v139/liu21ad.html.

Shaocong Ma, Ziyi Chen, Yi Zhou, Kaiyi Ji, and Yingbin Liang. Data sampling affects the complexity of online SGD over dependent data. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 1296–1305. PMLR, 2022. URL https://proceedings.mlr.press/v180/ma22a.html.

Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

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. doi: 10.1137/070704277.

Whitney K. Newey and Kenneth D. West. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708, 1987. doi: 10.2307/1913610.

Boris T. Polyak. Gradient methods for minimizing functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963. doi: 10.1016/0041-5553(63)90382-3.

Boris T. Polyak and Anatoli B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992. doi: 10.1137/0330046.

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

Herbert Robbins and David Siegmund. A convergence theorem for nonnegative almost supermartingales and some applications. In *Optimizing Methods in Statistics*, pp. 233–257. Academic Press, 1971.

Dhruv Sarkar, Aprameyo Chakrabartty, and Vaneet Aggarwal. High-probability PL-SGD with Markovian noise: Optimal mixing and tail dependence, 2026. URL https://arxiv.org/abs/2606.26316. arXiv preprint.

Naichen Shi, Dawei Li, Mingyi Hong, and Ruoyu Sun. RMSprop converges with proper hyperparameter. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=3UDSdyIcBDA.

Tao Sun, Yuejiao Sun, and Wotao Yin. On Markov chain gradient descent. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://papers.neurips.cc/paper/8195-on-markov-chain-gradient-descent.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2 edition, 2018. ISBN 9780262039246. URL https://mitpress.mit.edu/9780262039246/reinforcement-learning/.

Vladislav B. Tadić. Convergence and convergence rate of stochastic gradient search in the case of multiple and non-isolated extrema. *Stochastic Processes and their Applications*, 125(5):1715–1755, 2015. doi: 10.1016/j.spa.2014.11.001. URL https://arxiv.org/abs/0907.1020.

Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1195–1204. PMLR, 2019a. URL https://proceedings.mlr.press/v89/vaswani19a.html.

Sharan Vaswani, Aaron Mishkin, Issam Laradji, Mark Schmidt, Gauthier Gidel, and Simon Lacoste-Julien. Painless stochastic gradient: Interpolation, line-search, and convergence rates. In *Advances in Neural Information Processing Systems*, volume 32, 2019b. URL https://arxiv.org/abs/1905.09997.

Simon Weissmann, Sara Klein, Waïss Azizian, and Leif Döring. Almost sure convergence of stochastic gradient methods under gradient domination. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=OTwnNBxZFB.

Stephan Wojtowytsch. Stochastic gradient descent with noise of machine learning type part I: Discrete time analysis. *Journal of Nonlinear Science*, 33(3):45, 2023. doi: 10.1007/s00332-023-09903-3. URL https://doi.org/10.1007/s00332-023-09903-3.

Wei Biao Wu. Recursive estimation of time-average variance constants. *The Annals of Applied Probability*, 19(4):1529–1552, 2009. doi: 10.1214/08-AAP587. URL https://arxiv.org/abs/0908.4540.

Zhiming Zhou, Qingru Zhang, Guansong Lu, Hongwei Wang, Weinan Zhang, and Yong Yu. AdaShift: Decorrelation and convergence of adaptive learning rate methods. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HkgTkhRcKQ.