Stochastic Learning Rate With Memory: Optimization in the Stochastic Approximation and Online Learning Settings

Theodoros Mamalis[1], Dušan Stipanović[2], Petros Voulgaris[3]

***Abstract*—In this work, multiplicative stochasticity is applied to the learning rate of stochastic optimization algorithms, giving rise to stochastic learning-rate schemes. In-expectation theoretical convergence results of Stochastic Gradient Descent equipped with this novel learning rate scheme under the stochastic setting, as well as convergence results under the online optimization settings are provided. Empirical results consider the case of an adaptively uniformly distributed multiplicative stochasticity and include not only Stochastic Gradient Descent, but also other popular algorithms equipped with a stochastic learning rate. They demonstrate noticeable optimization performance gains with respect to their deterministic-learning-rate versions.**

***Index Terms*—Optimization algorithms, Stochastic systems, Convergence, Machine learning.**

I. Introduction

MACHINE LEARNING models today range from simple linear regression to deep neural networks. These models try to capture the underlying behavior of systems which produce outputs y when given inputs x. The performance of these models is usually measured via a loss function $\ell(h(x;\theta),y)$ where $h(x;\theta)$ is the predicted output of the model for an input x for given model parameters $\theta \in \mathbb{R}^d$. Then, if the joint distribution P of x and y is known, the estimates θ are the minimizers of the expected loss $\mathbb{E}_{x,y\sim P}[\ell(h(x;\theta),y)]$. Nonetheless, complete knowledge of P is absent in most practical scenarios, and a limited set of training data is available instead. In that case, an estimate of the expected loss is minimized, known as the empirical loss function, giving rise to the empirical loss minimization (ERM) problem $\min\limits_{\theta\in R^d} \frac{1}{n}\sum_{i=1}^{n}\ell(h(x_i;\theta),y_i)$ where the sequence $\{(x_i,y_i)\}_{i=1}^{n}$ denotes the available training data.

Then, if ξ represents a random sample from the the available training data, let the loss function as a composition of $\ell(.,.)$ and $h(.;.)$ be denoted by $L(\theta;\xi)$. Also let $f(\theta) := \mathbb{E}_{\xi}[L(\theta,\xi)]$. This yields a stochastic approximation setting where the computed gradient $g(\theta,\xi)$ is a function of the true gradient $\nabla f(\theta)$. If the former is assumed to be an unbiased estimator of the latter as in this work, then:

$$\mathbb{E}_{\xi_t}[g(\theta_t,\xi_t)] = \nabla f(\theta_t), \tag{1}$$

where ξ_t and θ_t are independent. This a common assumption made in the stochastic approximation setting [1], [2], that also holds for ERM. Due to its simplicity, cost-effectiveness and well studied properties, one of the most commonly used algorithms for ERM is Stochastic Gradient Descent (SGD) preceded by the stochastic gradient method developed in [3] (see [2] for more details), which iteratively gives an estimate of θ in the ERM formulation via the rule:

$$\theta_{t+1} = \theta_t - a_t g_t, \tag{2}$$

for $t = 1, 2, \cdots$ and where g_t is short for $g(\theta_t,\xi_t)$. This rule will be referred to as original SGD or simply SGD.

The learning rate a_t is usually assumed to be deterministic. This work introduces the novel setting where the learning rate in SGD becomes stochastic. That is, in this work $a_t = \eta_t u_t$ where η_t will be referred to as step size and $u_t = \psi(v_b,...,v_t)$ is a function of past random variables $\{v_i\}_{i=b}^{t}$, referred to as stochastic factors (SF) in the rest of the paper. A Multiplicative-Stochastic-Learning-Rate is referred to as MSLR, and, in specific, MSLR-with-memory if $b < t$.

An instance of MSLR-with-memory is for $\psi(v_1,...,v_t) = \prod_{i=1}^{t} v_i$ which will be the theoretical focus of this work. Its efficient implementation when applied on SGD is shown in Algorithm 1. The extra storage and time requirements of MSLR-with-memory SGD compared to the deterministic-learning-rate SGD is, at each t, storing η_{t-1} and a_{t-1}, and generating the SF v_t, which are minimal. Algorithm 1 avoids storing all of the SFs from previous timesteps, but still efficiently uses them at each step yielding minimal extra storage requirements to the deterministic-learning-rate SGD.

Minimization performance can be noticeably improved under appropriate learning schemes e.g., under deterministic adaptive learning rate schemes with memory (such as memory of past gradients, updates etc). Examples include ADAM [4] and some variants (e.g., AMSGrad [5] or ADAMW [6]) or precursors (e.g. [7], [8]). The memory of the stochastic learning rate in this work considers past SFs to update current parameters, and shows significantly improved performance.

Convergence analysis in expectation but for deterministic learning rates has been studied in [9], [1], [10], [11], [2]. This work provides convergence results in-expectation for SGD using stochastic learning rate with memory as well as results under the online learning framework (introduced in [12]). It shows that stochastic learning rates with memory significantly improve optimization, while providing similar technical results to the deterministic-learning-rate case. Convergence analysis in the memoryless setting is made in [13]. Convergence in the almost-surely setting for algorithms including SGD has been done in [14].

This work is supported by the grant from the National Robotics Initiative grant titled “NRI: FND:COLLAB: Multi-Vehicle Systems for Collecting Shadow-Free Imagery in Precision Agriculture” (grant no.2019-04791/project accession no. 1020285) from the USDA National Institute of Food and Agriculture.

[1] Theodoros Mamalis is with the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, 306 N Wright St, Urbana, IL 61801, USA mamalis2@illinois.edu

[2] Dušan Stipanović is with the Coordinated Science Laboratory, University of Illinos at Urbana-Champaign, 1308 W Main St, Urbana, IL 61801, USA. dusan@illinois.edu

[3]Petros Voulgaris is with the Department of Mechanical Engineering, University of Nevada, Reno, NV 89557, USA pvoulgaris@unr.edu

The analysis will allow for resetting SF distributions. which yield resetting learning rate schemes, shown to improve optimization performance ([15], [16]). A memoryful Resetting-MSLR scheme will be referred to as RMSLR with memory, and RUMSLR with memory if its SFs are uniformly distributed (e.g. as shown in Fig. 1).

The remainder of the paper is organized as follows. Section II provides the stochastic approximation setting and the in-expectation convergence results. The convergence results of the online learning setting are given in Section III. Section IV provides a discussion on the RUMSLR scheme used to obtain the empirical results. Section V presents the experimental results for various optimization algorithms using this scheme. Section VI concludes the paper with paths for future work.

II. Stochastic Optimization Setting

This section will provide convergence results for stochastic-learning-rate SGD under the stochastic optimization setting, briefly discussed in Section I. These results hold for both MSLR and RMSLR schemes. The following common ([2]) assumptions will be used in the theorems. Firstly, function $f(\theta)$ is L-smooth, i.e.:

$$f(\theta) \leq f(\theta') + \nabla f(\theta')^T(\theta - \theta') + \frac{L}{2}\|\theta - \theta'\|^2, \quad (3)$$

for all $\theta', \theta \in \mathbb{R}^d$. Furthermore, for all $\theta \in \mathbb{R}^d$, it satisfies:

$$\|\nabla f(\theta)\|^2 \geq 2c(f(\theta) - f_{min}), \quad (4)$$

where f_{min} is the global minimum of $f(\theta)$, and $0 < c \leq L$. It is noted that this assumption only requires the function to attain a global minimum (more detailed discussion in [17]). Also:

$$\mathbb{E}_{\xi_t}[\|g(\theta_t, \xi_t)\|^2] \leq M + M_G\|\nabla f(\theta_t)\|^2 \quad \text{for} \quad t \in \mathbb{N}, \quad (5)$$

for some nonnegative scalars M and $M_G \geq 1$. Cases where Assumptions 3-5 hold include the negative log-likelihood loss of a binomial and multinomial logistic regression (used e.g. in classification tasks), as well as least-squares linear regression (used e.g. in financial analyses), under an ERM setting. Moreover, for the stochastic learning rates it is assumed that the following holds:

$$\sum_{j=1}^{\infty} \mathbb{E}_a[a_j] = \sum_{j=1}^{\infty} \eta_j \mathbb{E}_u[u_j] = \infty$$
$$\sum_{j=1}^{\infty} \mathbb{V}_a[a_j] = \sum_{j=1}^{\infty} \eta_j^2 \mathbb{V}_u[u_j] < \infty \quad (6)$$

These assumptions are the stochastic-learning-rate equivalents to the typical assumptions made in the deterministic-learning-rate case–commonly referred to as Robbins-Monro conditions ([3])–namely, $\sum_j \eta_j = \infty$ and $\sum_j \eta_j^2 < \infty$.

The theorem that follows provides a bound on the optimality gap of the SGD algorithm when using a stochastic-learning-rate with a constant step size:

Theorem 1. *Assume (3-6) are satisfied. Furthermore, assume a positive learning rate of:*

$$a_t = \eta_t u_t, \quad (7)$$

where $\eta_t \in \mathbb{R}$ is either constant or decreasing, and $u_t := \prod_{i=1}^{t} v_i$ where v_i are i.i.d. with $\mathbb{E}_v[v_t] \leq 1$, and $1 > \lambda \geq \frac{\mathbb{V}_u[u_t]}{\mathbb{E}_u^2[u_t]}$, for some positive $\lambda \in \mathbb{R}$. If, moreover:

$$\eta_t < \frac{1}{\mathbb{E}_u[u_t]LM_G}, \quad (8)$$

then the iterates of the SGD algorithm in (2) satisfy:

$$\lim_{t\to\infty} \mathbb{E}[f(\theta_{t+1}) - f_{min}] = 0. \quad (9)$$

Remark 1. *The guarantee in (9) is the same as when SGD uses a decreasing deterministic-learning-rate. However, the stepsize's upper bound $\frac{1}{\mathbb{E}_u[u_t]LM_G}$ is larger than the $\frac{1}{LM_G}$ in the deterministic-learning-rate case. Moreover, the guarantee in (9) holds for arbitrary decreasing step sizes and does not depend on the initial point unlike decreasing stepsizes in previous works (see, e.g., [2] for all of the above). In addition, the optimality gap for the stochastic-learning-rate with constant stepsize is zero, whereas for the deterministic-learning-rate case is nonzero, i.e., $\frac{\eta LM}{2c}$. This is also true for a stochastic learning rate scheme without memory ([13]) which yields a nonzero optimality gap for a constant stepsize, whereas Theorem 1 is able to obtain a zero optimality-gap asymptotically.*

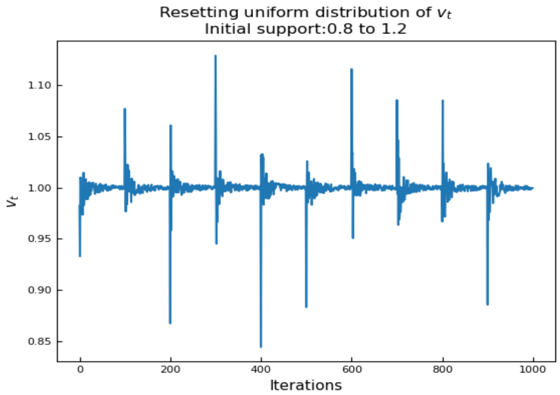

Fig. 1: Plot that illustrates a resetting SF v_t, used in the RMSLR-with-memory scheme considered in the experiments, with $v_t \sim \mathcal{U}(\sqrt[\varepsilon_t]{c_1}, \sqrt[\varepsilon_t]{c_2})$ where $\varepsilon_t = t \pmod{\beta} + 1$, and $c_1 = 0.8, c_2 = 1.2$. For the case depicted in the plot, the value of the resetting parameter is $\beta = 100$, i.e. the distribution of v_t resets every 100 iterations to $\mathcal{U}(0.8, 1.2)$.

Algorithm 1 MSLR-with-memory SGD

```
Require: initial point $\theta_1$, stepsize $\eta_t$, adaptive probability
    distribution $\mathcal{P}_t$.
    Set $\eta_1 = 1, a_1 = 1$.
    while $\theta_t$ not converged do
        $t \leftarrow t + 1$
        $v_t \sim \mathcal{P}_t$
        $a_t \leftarrow \frac{\eta_t}{\eta_{t-1}} a_{t-1} v_t$
        $\theta_t \leftarrow \theta_{t-1} - a_t g_{t-1}$
    end while
    return $\theta_t$
```

Proof. Taking expectations with respect to ξ_t in (3), and from (2) substituting θ_{t+1} for θ and θ_t for θ':

$$\begin{aligned}
&\mathbb{E}_{\xi_t}[f(\theta_{t+1})] \leq f(\theta_t) - a_t \nabla^T f(\theta_t)\mathbb{E}_{\xi_t}[g(\theta_t,\xi_t)] \\
&+ a_t^2 \frac{L}{2}\mathbb{E}_{\xi_t}\left[\|g(\theta_t,\xi_t)\|^2\right] \\
&= f(\theta_t) - a_t\|\nabla f(\theta_t)\|^2 + a_t^2\frac{L}{2}\mathbb{E}_{\xi_t}\left[\|g(\theta_t,\xi_t)\|^2\right] \\
&\leq f(\theta_t) + a_t(a_t\frac{M_G L}{2} - 1)\|\nabla f(\theta_t)\|^2 + a_t^2\frac{L}{2}M \\
&= f(\theta_t) + \frac{1}{2}a_t(a_t M_G L - 2)\|\nabla f(\theta_t)\|^2 + a_t^2\frac{L}{2}M \\
&= f(\theta_t) + \frac{1}{2}(a_t^2 M_G L - 2a_t)\|\nabla f(\theta_t)\|^2 + a_t^2\frac{L}{2}M. \qquad (10)
\end{aligned}$$

The second inequality followed from (1), and the third from (5). Then, taking expectations with respect to a_t, ξ_t for all t, denoting these expectations as $\mathbb{E}_\xi[.]$ and $\mathbb{E}_a[.]$ respectively, and the result of $\mathbb{E}_\xi[\mathbb{E}_a[.]]$ as $\mathbb{E}[.]$:

$$\begin{aligned}
&\mathbb{E}[f(\theta_{t+1})]f(\theta_t) + \mathbb{E}_a[a_t^2]\frac{L}{2}M \\
&+ \frac{1}{2}(\mathbb{E}_a^2[a_t]M_G L + \mathbb{V}_a[a_t]M_G L - 2\mathbb{E}_a[a_t])\|\nabla f(\theta_t)\|^2,
\end{aligned} \qquad (11)$$

where:

$$\mathbb{E}_a[a_t^2] = \mathbb{V}_a[a_t] + \mathbb{E}_a^2[a_t], \qquad (12)$$

was used to form the last term. Then, factoring out $\mathbb{E}_a[a_t]$ from the last term, and using (7) followed by (8) on its remaining quantities:

$$\begin{aligned}
&\mathbb{E}[f(\theta_{t+1})] \leq f(\theta_t) + \mathbb{E}_a[a_t^2]\frac{L}{2}M - \frac{1}{2}\mathbb{E}_a[a_t]\varphi_t\|\nabla f(\theta_t)\|^2 \\
&\leq f(\theta_t) + \mathbb{E}_a[a_t^2]\frac{L}{2}M - \mathbb{E}_a[a_t]c\varphi_t(f(\theta_t) - f_{min}).
\end{aligned} \qquad (13)$$

where the inequality followed from (4), and $\varphi_t := 1 - \frac{\mathbb{V}_u[u_t]}{\mathbb{E}_u^2[u_t]}$. Also, $\tilde{c} \leq \varphi_t < 1$ where $\tilde{c} = 1 - \lambda$. Then, using (12), adding and subtracting $\mathbb{E}_a[a_t]\frac{LM}{2c\varphi_t}, \mathbb{E}_a[a_{t+1}]\frac{LM}{2c\varphi_{t+1}}, f_{min}$ and rearranging:

$$\begin{aligned}
&\mathbb{E}[f(\theta_{t+1}) - f_{min}] - \mathbb{E}_a[a_{t+1}]\frac{LM}{2c\varphi_{t+1}} \\
&\leq (1 - c\varphi_t\mathbb{E}_a[a_t])(\mathbb{E}[f(\theta_t) - f_{min}] - \mathbb{E}_a[a_t]\frac{LM}{2c\varphi_t}) \\
&+ (\frac{\mathbb{E}_a[a_t]}{\varphi_t} - \frac{\mathbb{E}_a[a_{t+1}]}{\varphi_{t+1}})\frac{LM}{2c} + \mathbb{V}_a[a_t]\frac{LM}{2}. \qquad (14)
\end{aligned}$$

Next, a lemma proving monotonicity of φ_t is presented:

Lemma 1. *Let* $\varphi_t := 1 - \frac{\mathbb{V}_u[u_t]}{\mathbb{E}_u^2[u_t]}$. *Then:*

$$-(\varphi_t)^{-1} > -(\varphi_{t+1})^{-1}. \qquad (15)$$

Proof. Using the definition $u_t := \prod_{i=1}^t v_i$, it is that:

$$\begin{aligned}
\mathbb{V}[u_t] &= \mathbb{V}[\prod_{i=1}^t v_i] = \mathbb{E}[\prod_{i=1}^t v_i^2] - \mathbb{E}^2[\prod_{i=1}^t v_i] \\
&= \prod_{i=1}^t(\mathbb{V}[v_i] + \mathbb{E}[v_i]^2) - \prod_{i=1}^t\mathbb{E}^2[v_i], \qquad (16)
\end{aligned}$$

since the v_i's are independent. This means that:

$$\frac{\mathbb{V}[u_t]}{E^2[u_t]} = \frac{\mathbb{V}[\prod_{i=1}^t v_i]}{\mathbb{E}^2[\prod_{i=1}^t v_i]} = \frac{\prod_{i=1}^t(\mathbb{V}[v_i] + \mathbb{E}[v_i]^2)}{\prod_{i=1}^t\mathbb{E}^2[v_i]} - 1. \qquad (17)$$

Using (17), the inequality $\frac{\mathbb{V}[u_t]}{\mathbb{E}^2[u_t]} < \frac{\mathbb{V}[u_{t+1}]}{\mathbb{E}^2[u_{t+1}]}$ is written as:

$$\begin{aligned}
&\frac{\prod_{i=1}^t(\mathbb{V}[v_i] + \mathbb{E}[v_i]^2)}{\prod_{i=1}^t\mathbb{E}^2[v_i]} \\
&< \frac{\prod_{i=1}^t(\mathbb{V}[v_i] + \mathbb{E}[v_i]^2)(\mathbb{V}[v_{t+1}] + \mathbb{E}[v_{t+1}]^2)}{\prod_{i=1}^t\mathbb{E}^2[v_i]\mathbb{E}^2[v_{t+1}]}. \qquad (18)
\end{aligned}$$

This yields $\frac{\mathbb{V}[v_{t+1}] + \mathbb{E}^2[v_{t+1}]}{\mathbb{E}^2[v_{t+1}]} > 0$ which holds for all t, proving $\frac{\mathbb{V}[u_t]}{\mathbb{E}^2[u_t]} < \frac{\mathbb{V}[u_{t+1}]}{\mathbb{E}^2[u_{t+1}]}$, which means $\varphi_{t+1} < \varphi_t$, and thus $-\frac{1}{\varphi_{t+1}} < -\frac{1}{\varphi_t}$. *Q.E.D.*

Then, by unrolling the sequence, using Lemma 1 and $\varphi_t \geq \tilde{c}$:

$$\begin{aligned}
&\mathbb{E}[f(\theta_{t+1}) - f_{min}] - \mathbb{E}_a[a_{t+1}]\frac{LM}{2c\varphi_{t+1}} \\
&\leq \prod_{j=1}^t(1 - c\varphi_j\mathbb{E}_a[a_j])(\mathbb{E}[f(\theta_1) - f_{min}] - \mathbb{E}_a[a_1]\frac{LM}{2c\varphi_1}) \\
&+ \frac{LM}{2c}Q_{1,t}(t) + \frac{LM}{2}S_{1,t}(t).
\end{aligned} \qquad (19)$$

where:

$$\begin{aligned}
Q_{1,t}(t) &:= \sum_{j=1}^t \frac{1}{\tilde{c}}(\mathbb{E}_a[a_j] - \mathbb{E}_a[a_{j+1}])p_j(t) \\
S_{1,t}(t) &:= \sum_{j=1}^t \mathbb{V}_a[a_j]p_j(t), \qquad (20)
\end{aligned}$$

and $p_j(t) := \prod_{n=j+1}^t(1 - c\mathbb{E}_a[a_n])$. The subscripts in $Q_{1,t}(t)$, $S_{1,t}(t)$ denote the lower and upper bounds for their sums, and the argument refers to their dependency on $p_j(t)$. Then, from (8), $M_G \geq 1$, and $c \leq L$:

$$c\varphi_t\mathbb{E}_a[a_t] \leq c\mathbb{E}_a[a_t] = c\eta_t\mathbb{E}_u[u_t] < \frac{c}{M_G L} \leq \frac{c}{L} \leq 1, \quad (21)$$

which yields $0 \leq 1 - c\varphi_t\mathbb{E}_a[a_t] < 1$, since $c\mathbb{E}[a_t]$ is positive, and $0 < \tilde{c} \leq \varphi_t < 1$. Moreover, $p_j(t) < 1$ for all t.

Then, for the second term in (19) this means that:

$$\begin{aligned}
Q_{1,t}(t) &= \sum_{j=1}^t \frac{1}{\tilde{c}}(\mathbb{E}_a[a_j] - \mathbb{E}_a[a_{j+1}])p_j(t) \qquad (22) \\
&\leq \sum_{j=1}^t \frac{1}{\tilde{c}}(\mathbb{E}_a[a_j] - \mathbb{E}_a[a_{j+1}]) = \frac{1}{\tilde{c}}(\mathbb{E}_a[a_1] - \mathbb{E}_a[a_{t+1}]),
\end{aligned}$$

which converges to $\frac{1}{\tilde{c}}\mathbb{E}_a[a_1]$ as $t \to \infty$ since $\lim_{t\to\infty}\mathbb{E}_a[a_{t+1}] = 0$. For the third term in (19):

$$S_{1,t}(t) = \sum_{j=1}^t \mathbb{V}_a[a_j]p_j(t) \leq \sum_{j=1}^t \mathbb{V}_a[a_j] < \infty, \qquad (23)$$

from the assumptions of the theorem.

Then, it is that $\mathbb{E}_a[a_j] \geq \mathbb{E}_a[a_{j+1}] = \mathbb{E}_a[a_j]\mathbb{E}_v[v_{j+1}]$ since $\mathbb{E}_v[v_{j+1}] \leq 1$, therefore $Q_{1,t}(t)$ and $S_{1,t}(t)$ are summations with nonnegative summands. Thus, since $Q_{1,t}(t)$ and $S_{1,t}(t)$ are also convergent there exists K such that $\lim_{t\to\infty} Q_{K+1,t}(t) =$

0 and $\lim_{t\to\infty} S_{K+1,t}(t) = 0$. Next, partitioning the last two terms in (23) at K:

$$\mathbb{E}[f(\theta_{t+1}) - f_{min}] - \mathbb{E}_a[a_{t+1}]\frac{LM}{2c\varphi_{t+1}}$$
$$\leq \prod_{j=1}^{t}(1 - c\varphi_j\mathbb{E}_a[a_j])(\mathbb{E}[f(\theta_1) - f_{min}] - \mathbb{E}_a[a_1]\frac{LM}{2c\varphi_1})$$
$$+ S_{1,K}(t) + S_{K+1,t}(t) + Q_{1,K}(t) + Q_{K+1,t}(t). \tag{24}$$

Moreover:

$$\lim_{t\to\infty} Q_{1,K}(t) = \sum_{j=1}^{K}\frac{1}{\tilde{c}}(\mathbb{E}_a[a_j] - \mathbb{E}_a[a_{j+1}])\lim_{t\to\infty} p_j(t) = 0$$
$$\lim_{t\to\infty} S_{1,K}(t) = \sum_{j=1}^{K}\mathbb{V}_a[a_j]\lim_{t\to\infty} p_j(t) = 0. \tag{25}$$

Finally, from (24), from the first equation of (6) and from $\mathbb{E}_a[a_{t+1}] = \eta_{t+1}\mathbb{E}[u_{t+1}]$ it follows that:

$$\lim_{t\to\infty}\mathbb{E}[f(\theta_{t+1}) - f_{min}] = \lim_{t\to\infty} = \eta\frac{LM}{2c}, \tag{26}$$

which means $\lim_{t\to\infty}\mathbb{E}[f(\theta_{t+1}) - f_{min}] = \eta\frac{LM}{2c} \neq 0$ since $\lim_{t\to\infty}\eta_{t+1}\mathbb{E}[u_{t+1}]\frac{LM}{2c} = 0$ from $\lim_{t\to\infty}\mathbb{E}[u_{t+1}] = 0$ and $\eta_{t+1} = \eta$ being constant. since $\eta_{t+1} = \eta$ is constant. *Q.E.D.*

Concluding this section, from Theorem 1 it is thus observed that for the stochastic approximation setting the optimality-gap bounds for SGD using a stochastic-learning-rate are improved from the ones when using a deterministic-learning-rate. That is, the optimality gap is zero for the former case regardless of whether a fixed or a decreasing stepsize is used, and nonzero for the latter. This improvement is due to the memoryful property of MSLR-with-memory, since for a memoryless MSLR (including RMSLR), the optimality gap for fixed stepsizes would remain nonzero [13].

The reason is that, in the latter case, the presence of noise ($M \neq 0$) causes the optimality gap to be nonzero, since the expected objective values do not converge to the minimizer but to a noise-dependent neighborhood of it. However, a stochastic learning rate injects its SF to the optimality-gap expression and directly controls the gap. Therefore, if appropriately chosen, the SF is able to cancel the effects of the noise and guide the iterates through this neighborhood towards the minimizer. Mathematically, this is observed in equation (26) where the SF expectation causes the optimality-gap to convergence to zero. For a deterministic learning rate, this added degree of freedom of controlling the limiting value of the optimality gap through the SF is absent. This shows the importance of being able to freely choose the properties of the learning-rate randomness, similarly to the stochastic-learning rate schemes introduced in this work.

III. Online Learning Setting

Now, an alternative analysis for stochastic-learning-rate SGD which holds for both MSLR and RMSLR is given by considering the following optimization problem, referred to as online optimization:

$$R_T := \sum_{t=1}^{T} f_t(\theta_t) - \min_{\theta\in X}\sum_{t=1}^{T} f_t(\theta), \tag{27}$$

where $\theta^* = \arg\min_{\theta\in X}\sum_{t=1}^{T} f_t(\theta)$, functions f_1 through f_T are convex, $g_t := \nabla f_t(\theta_t)$, and $X \subseteq \mathbb{R}^d$ is a convex and closed set. The function R_T is called the regret up to timestep T. This optimization framework was introduced in [12]. Then, consider the online algorithm:

$$\nu_t = \theta_t - a_t g_t$$
$$\theta_{t+1} = \Pi_X(\nu_t), \tag{28}$$

where $\Pi_X(y) = \arg\min_{x\in X}\|x - y\|$ is the projection of y on X. The projection step ([5], [12]) ensures that θ_{t+1} is inside the feasible parameter set X, which is not $\mathbb{R}^d$ as in the setting of Section II. This, along with the definition of R_T in equation (27), makes different assumptions necessary for the two settings. Nonetheless, studying an online algorithm with vanishing average regrets is related to studying a stochastic algorithm for ERM through a one-way reduction from the former to the latter, as discussed in [18]. Next, assume that $f_t(\theta)$ are convex for all t:

$$f_t(\theta) \geq f_t(\theta') + \nabla f_t(\theta')^T(\theta - \theta') \quad \text{for all} \quad \theta', \theta \in \mathbb{R}^d. \tag{29}$$

Furthermore, define X as a set with bounded diameter D if:

$$\|\theta' - \theta\| \leq D, \tag{30}$$

for all $\theta', \theta \in X$. Moreover, if $X \subseteq R^d$ is a convex and closed set, then for any $\theta \in X$ and $y \in \mathbb{R}^d$, it is that:

$$\|y - \theta\| \geq \|\theta' - \theta\|, \tag{31}$$

for $\theta' = \Pi_X(y) = \arg\min_{\theta\in X}\|y - \theta\|$. This property proves useful in the subsequent theorem, that assumes a convex and closed set $X \subseteq \mathbb{R}^d$ as the feasible parameter set. It is noted that the regret bounds of the deterministic-learning-rate version of SGD are obtained from the proof of the next Theorem 2 but using $a_t = \frac{a}{\sqrt{t}}$ as the learning rate.

Now it will be shown that stochastic-learning-rate SGD has vanishing average regret for diminishing step size $\eta_t = \frac{a}{\sqrt{t}}$. The theorem for the online learning setting follows:

Theorem 2. *Assume that X is convex, closed, and has bounded diameter D, and that the gradient of f_t is bounded, i.e. $\|\nabla f_t(\theta)\|_\infty \leq G$. Moreover, assume a positive learning rate of $a_t = \frac{a}{\sqrt{t}}u_t$ with $a = \frac{D}{2G}$. Furthermore, assume $\mathbb{E}_u[u_t] \leq 1$ and $\frac{\mathbb{V}_u[u_t]}{\mathbb{E}_u^2[u_t]} \leq 1$ for all t. Then for all $T \geq 1$:*

$$R_{a,T} \leq 2DG\sqrt{T}, \tag{32}$$

where the subscript a denotes the expectation of R_T with respect to the stochastic learning rate sequence $\{a_t\}_{t=1}^{T}$. Therefore, $R_{a,T}/T \to 0$ for $T \to \infty$.

Proof. Firstly, it is noted that:

$$\|\theta_{t+1} - \theta^*\|^2 = \|\Pi_X(\theta_t - a_t g_t) - \theta^*\|^2$$
$$\leq \|\theta_t - a_t g_t - \theta^*\|^2$$
$$= \|\theta_t - \theta^*\|^2 - 2a_t g_t^T(\theta_t - \theta^*) + a_t^2\|g_t\|^2, \tag{33}$$

where the first inequality followed from (31) replacing θ by θ^* and y by $\theta_t - a_t g_t$. Taking expectations with respect to the learning rates $a_1 \cdots a_t$ and denoting the result as $\mathbb{E}_a[.]$:

$$\begin{aligned} &\mathbb{E}_a[\|\theta_{t+1} - \theta^*\|^2] \leq \mathbb{E}_a[\|\theta_t - \theta^*\|^2] \\ &- 2\mathbb{E}_a[a_t]\mathbb{E}_a[g_t^T(\theta_t - \theta^*)] + \mathbb{E}_a[a_t^2]\mathbb{E}_a[\|g_t\|^2], \end{aligned} \tag{34}$$

Rearranging with respect to $g_t^T(\theta_t - \theta^*)$ and using (29):

$$\begin{aligned} &\mathbb{E}_a[f_t(\theta_t) - f_t(\theta^*)] \leq \mathbb{E}_a[g_t^T(\theta_t - \theta^*)] \\ &\leq \frac{\mathbb{E}_a[\|\theta_t - \theta^*\|^2] - \mathbb{E}_a[\|\theta_{t+1} - \theta^*\|^2]}{2\mathbb{E}_a[a_t]} + \frac{\mathbb{E}_a[a_t^2]}{2\mathbb{E}_a[a_t]}\mathbb{E}_a[\|g_t\|^2] \\ &\leq \frac{\mathbb{E}_a[\|\theta_t - \theta^*\|^2] - \mathbb{E}_a[\|\theta_{t+1} - \theta^*\|^2]}{2\mathbb{E}_a[a_t]} + \frac{\mathbb{E}_a[a_t]}{2}\mathbb{E}_a[\|g_t\|^2] \\ &+ \frac{\mathbb{V}_a[a_t]}{2\mathbb{E}_a[a_t]}\mathbb{E}_a[\|g_t\|^2]. \end{aligned} \tag{35}$$

Then, summing over all steps and rearranging:

$$\begin{aligned} &\sum\nolimits_{t=1}^{T} \mathbb{E}_a[[f_t(\theta) - f_t(\theta^*)]] \leq \frac{1}{2\mathbb{E}_a[a_1]}\mathbb{E}_a[\|\theta_1 - \theta^*\|^2] \\ &+ \sum\nolimits_{t=2}^{T} \mathbb{E}_a\|\theta_t - \theta^*\|^2 \left(\frac{1}{2\mathbb{E}_a[a_t]} - \frac{1}{2\mathbb{E}_a[a_{t-1}]}\right) \\ &+ \frac{1}{2}\sum\nolimits_{t=1}^{T} \mathbb{E}_a[a_t]\mathbb{E}_a[\|g_t\|^2] + \frac{1}{2}\sum\nolimits_{t=1}^{T} \frac{\mathbb{V}_a[a_t]}{\mathbb{E}_a[a_t]}\mathbb{E}_a[\|g_t\|^2], \end{aligned} \tag{36}$$

Using the diameter of the convex set D, then expanding the telescoping sum:

$$\begin{aligned} &\sum\nolimits_{t=1}^{T} \mathbb{E}_a[[f_t(\theta) - f_t(\theta^*)]] \leq \frac{1}{2\mathbb{E}_a[a_T]}D^2 \\ &+ \frac{1}{2}\sum\nolimits_{t=1}^{T} \mathbb{E}_a[a_t]\mathbb{E}_a[\|g_t\|^2] + \frac{1}{2}\sum\nolimits_{t=1}^{T} \frac{\mathbb{V}_a[a_t]}{\mathbb{E}_a[a_t]}\mathbb{E}_a[\|g_t\|^2], \end{aligned} \tag{37}$$

where the third line followed by using the diameter D of the convex set X, and the fourth line by using the telescoping sum of the middle term. Using $a_t = \frac{a}{\sqrt{t}}u_t$ it is:

$$\begin{aligned} &\sum\nolimits_{t=1}^{T} \mathbb{E}_a[[f_t(\theta) - f_t(\theta^*)]] \leq \frac{1}{2\mathbb{E}_a[a_T]}D^2 \\ &+ \frac{1}{2}\sum_{t=1}^{T} \frac{a\mathbb{E}_u[u_t]}{\sqrt{t}}\mathbb{E}_a[\|g_t\|^2] + \frac{1}{2}\sum_{t=1}^{T} \frac{a\mathbb{V}_u[u_t]}{\sqrt{t}\mathbb{E}_u[u_t]^2}\mathbb{E}_a[\|g_t\|^2]. \end{aligned} \tag{38}$$

Then, using $\mathbb{E}_u[u_t] \leq 1$, $\frac{\mathbb{V}_u[u_t]}{\mathbb{E}_u^2[u_t]} \leq 1$ and $\sum_{i=1}^{T} \frac{1}{\sqrt{t}} \leq 2\sqrt{T}$:

$$\begin{aligned} &\sum\nolimits_{t=1}^{T} \mathbb{E}_a[[f_t(\theta) - f_t(\theta^*)]] \\ &\leq \frac{1}{2\mathbb{E}_a[a_T]}D^2 + \frac{1}{2}\sum_{t=1}^{T} \frac{a}{\sqrt{t}}\mathbb{E}_a[\|g_t\|^2] + \frac{1}{2}\sum_{t=1}^{T} \frac{a}{\sqrt{t}}\mathbb{E}_a[\|g_t\|^2] \\ &\leq \frac{D^2\sqrt{T}}{2a} + aG^2\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq \frac{D^2\sqrt{T}}{2a} + aG^2 2\sqrt{T}, \end{aligned} \tag{39}$$

Then, minimizing the right hand side of the third inequality yields $a = D/(2G)$ for which $R_{a,T} \leq 2DG\sqrt{T}$ thus concluding the proof. *Q.E.D.*

In conclusion, from Theorem 2 it is thus observed that for the online optimization setting the regret bounds yield the same expected regret rate as the deterministic-learning-rate SGD algorithm, that is $O(\frac{1}{\sqrt{T}})$ in specific (e.g., in [8]).

IV. Discussion on the Stochasticity Factor

In this work, the stochastic learning rate scheme considered for the empirical results is a RMSLR-with-memory scheme and specifically $u_t = \prod_{i=1}^{t} v_t$ with $v_t \sim \mathcal{U}(\sqrt[\varepsilon_t]{c_1}, \sqrt[\varepsilon_t]{c_2})$ where $\varepsilon_t = t \pmod{\beta} + 1$, and $\beta \geq 1$, shown in Fig. 1. After experimenting, this was one scheme that was found to provide the necessary performance gains. In this case the distribution of v_t resets every β number of iterations to $\mathcal{U}(c_1, c_2)$, named $\beta-$RUMSLR-with-memory from the uniformly distributed SF. Moreover, c_1 and c_2 should be kept adequately apart so that the bursts are significant enough, however, low values for c_1 or large values for c_2 may decelerate the algorithm or destabilize it.

The distribution of v_t is chosen to be reset after it is adequately close to unity and enough, β timesteps have passed, e.g., as shown in Fig. 1, where $\beta = 100$, indicated by the bursts that occur every 100 iterations. Furthermore, it can be shown that the condition of Theorem 1, $\frac{\mathbb{V}_u[u_t]}{\mathbb{E}_u^2[u_t]} \leq \lambda$ holds with $\lambda = \frac{1}{3}$ for all c_1, c_2. This is because $(\sqrt[\varepsilon_t]{c_2} - \sqrt[\varepsilon_t]{c_1})^2 < (\sqrt[\varepsilon_t]{c_2} + \sqrt[\varepsilon_t]{c_1})^2$ meaning $\frac{V_v[v_t]}{\mathbb{E}_v^2[v_t]} = \frac{4(\sqrt[\varepsilon_t]{c_2} - \sqrt[\varepsilon_t]{c_1})^2}{12(\sqrt[\varepsilon_t]{c_2} + \sqrt[\varepsilon_t]{c_1})^2} \leq \frac{1}{3}$, since v_t is uniformly distributed. Therefore:

$$\begin{aligned} \frac{\mathbb{V}_u[u_t]}{\mathbb{E}_u^2[u_t]} &= \frac{\prod_{i=1}^{t}(\mathbb{V}_u[u_t] + \mathbb{E}_u[u_t])}{\prod_{i=1}^{t}\mathbb{E}_u[u_t]} - 1 \\ &= \prod\nolimits_{i=1}^{t}\left(\frac{\mathbb{V}_v[v_t]}{\mathbb{E}_v[v_t]} + 1\right) - 1 \leq \left(\frac{1}{3} + 1\right)^t - 1, \end{aligned} \tag{40}$$

which yields $\frac{\mathbb{V}_u[u_t]}{\mathbb{E}_u^2[u_t]} \leq \frac{1}{3}$ for $t = 1$, which also holds for all t since $\left(\frac{1}{3} + 1\right)^t - 1$ is increasing. Additionally, it can be checked numerically that the condition $\mathbb{E}_v[v_t] \leq 1$ as well as that the stochastic equivalent to the Robins-Monro conditions in Assumption 6 are satisfied for the RUMSLR scheme used to derive the empirical results in Section V. It is noted that, since the aforementioned conditions are sufficient and not necessary, stochastic learning rates that do not satisfy these conditions but still improve performance might exist.

Experimentally, it is demonstrated that $\beta-$RUMSLR-with-memory SGD yields noticeable improvements in minimization performance for the loss of a neural network training to fit datasets such as CIFAR-10, and CIFAR-100 [19] when compared to SGD without stochastic learning rates. This scheme was equipped on learning rates of other algorithms as well such as ADAM [4], AMSGrad [5], ADAMW [6], and two SGD with momentum algorithms in [20] and [21] (subsequently referred to as SGD with momentum first and second versions respectively), all of which, even though not theoretically studied in this work, yielded similarly improved performance when using $\beta-$RUMSLR with memory.

V. Results

The experiments concern the minimization of the training cross-entropy loss of neural networks that attempt to fit the MNIST, CIFAR-10 and CIFAR-100 datasets, using stochastic-learning-rate optimization algorithms. In specific, results for the $\beta-$RUMSLR-with-memory scheme and the algorithms in IV are presented in this section. The batch size is 128 and $\beta = 100$ in all experiments. The rest of the hyperparameters,

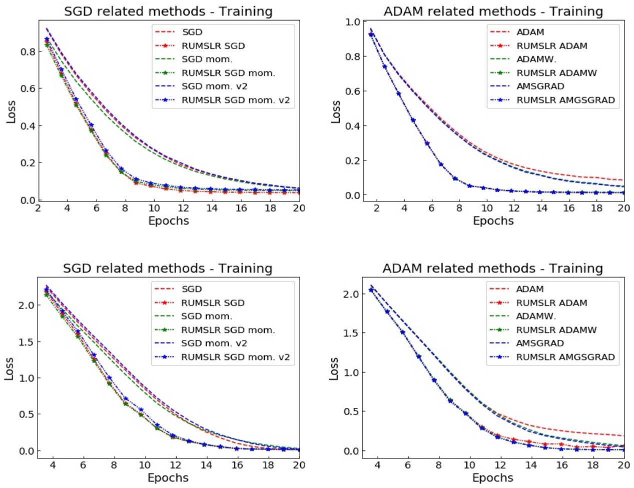

Fig. 3: Plots illustrating β-RUMSLR-with-memory (starred) and deterministic learning rate (dashed) for CIFAR-10 (top row) and CIFAR-100 (bottom row). Algorithms with β-RUMSLR yield noticeable minimization performance gains.

i.e., stepsizes, momentums and SF constants, are chosen via grid search for each algorithm. The stepsizes are 0.1 for SGD, 0.01 for the two SGD-with-momentum versions, and 0.001 for all ADAM-related algorithms for both deterministic- and stochastic-learning-rates. The values of c_1, c_2 are $0.7, 1.3$ for MNIST, and $0.8, 1.2$ for CIFAR-10, and $0.9, 1.1$ for CIFAR-100 where AMSGRAD uses $0.85, 1.15$. For CIFAR-10 and CIFAR-100, a fixed stepsize is used. The network architecture is Pytorch's ResNet18. For MNIST, logistic regression is used, and a diminishing stepsize.

The MNIST dataset is used to investigate convergence of the proposed algorithm when on a convex loss function, and in specific logistic regression, as aforementioned, a well-studied convex loss function. Indeed, the results in Fig. 2 verify the convergence of the $\beta-$RUMSLR-with-memory scheme on MNIST in a convex setting. Moreover, for CIFAR-10 and CIFAR-100, the $\beta-$RUMSLR-with-memory scheme demonstrated significantly faster convergence than when using a deterministic learning rate as shown in Fig. 3.

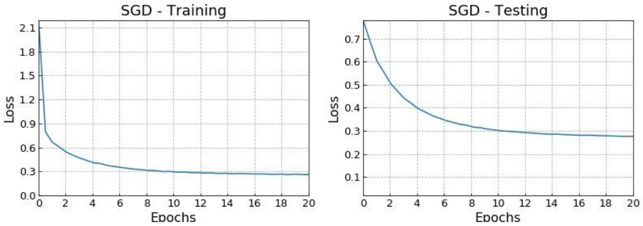

Fig. 2: Plots that illustrate convergence for β-RUMSLR-with-memory SGD in a convex setting.

VI. Conclusion and Future Work

This work introduced learning-rate schemes that make the learning rate stochastic by multiplicatively equipping it with a function of random variables, which are coined stochastic factors. Convergence of the SGD algorithm employing stochastic learning rate schemes with memory of past stochastic factors was theoretically analyzed and compared with known results for the algorithm's deterministic-learning-rate version. Empirical results on popular algorithms demonstrated noticeable increase in optimization performance, presenting stochastic-learning-rate schemes as a viable option for enhancing performance. In-depth generalization performance studies, convergence analysis of algorithms besides SGD, and investigating the effect of various stochastic factor distributions and hyperparameters on algorithm performance are some paths for future work.

References

[1] A. Nemirovski *et al.*, "Robust stochastic approximation approach to stochastic programming," *SIAM J. on Optim.*, vol. 19, no. 4, p. 1574–1609, Jan. 2009.

[2] L. Bottou *et al.*, "Optimization methods for large-scale machine learning," *SIAM Review*, vol. 60, no. 2, pp. 223–311, jan 2018.

[3] H. Robbins and S. Monro, "A stochastic approximation method," *Ann. Math. Statist.*, vol. 22, no. 3, pp. 400–407, 09 1951.

[4] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR (Poster)*, 2015.

[5] S. J. Reddi *et al.*, "On the convergence of adam and beyond," in *ICLR*, 2018.

[6] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*, 2019.

[7] H. B. McMahan and M. J. Streeter, "Adaptive bound optimization for online convex optimization," *CoRR*, vol. abs/1002.4908, 2010.

[8] J. Duchi *et al.*, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, p. 2121–2159, Jul. 2011.

[9] D. Bertsekas and J. Tsitsiklis, "Gradient convergence in gradient methods with errors," *SIAM J. on Optim.*, vol. 10, no. 3, pp. 627–642, 2000.

[10] E. Moulines and F. Bach, "Non-asymptotic analysis of stochastic approximation algorithms for machine learning," in *NeurIPS*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, Eds., vol. 24. Curran Associates, Inc., 2011.

[11] S. Ghadimi and G. Lan, "Stochastic first- and zeroth-order methods for nonconvex stochastic programming," *SIAM J. Optim.*, vol. 23, pp. 2341–2368, 2013.

[12] M. Zinkevich, "Online convex programming and generalized infinitesimal gradient ascent," in *In ICML*, 2003, pp. 928–936.

[13] T. Mamalis *et al.*, "Optimization in the stochastic approximation and online learning settings using stochastic learning rates," in *ACC*, 2022, pp. 4286–4291.

[14] ——, "Accelerated almost-sure convergence rates for nonconvex stochastic gradient descent using stochastic learning rates," *arXiv:0706.1234 [math.FA]*, 2021. [Online]. Available: https://arxiv.org/abs/2110.12634

[15] B. O'Donoghue and E. Candes, "Adaptive restart for accelerated gradient schemes," *Found. Math. Comp.*, vol. 15, pp. 1615–3383, 2015.

[16] I. Loshchilov and F. Hutter, "SGDR: stochastic gradient descent with warm restarts," in *ICLR*, 2017.

[17] H. Karimi *et al.*, "Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition," in *ECML PKDD*, P. Frasconi, N. Landwehr, G. Manco, and J. Vreeken, Eds. Cham: Springer International Publishing, 2016, pp. 795–811.

[18] N. Cesa-Bianchi *et al.*, "On the generalization ability of on-line learning algorithms," *IEEE Trans. Inf. Theory*, vol. 50, no. 9, pp. 2050–2057, 2004.

[19] A. Krizhevsky, "Learning multiple layers of features from tiny images," Tech. Rep., 2009.

[20] B. Polyak, "Some methods of speeding up the convergence of iteration methods," *USSR Comput. Math. Math. Phys.*, vol. 4, pp. 1–17, 1964.

[21] Y. Nesterov, "A method for solving the convex programming problem with convergence rate o($1/k^2$)," *Proc. USSR Acad. Sci.*, vol. 269, pp. 543–547, 1983.