Stochastic Gradient Descent and Adam in Non-Stationary Environments

Sander de Haan
Eliot Walt
Department of Computer Science, EPF Lausanne, Switzerland
Tingting Ni
Department of Mathematics, EPF Lausanne, Switzerland

***Abstract*—Many machine learning applications operate in non-stationary environments. Assessing the performance of Adam and stochastic gradient descent under corresponding circumstances will be the subject of two experiments with differing non-stationarities. Discerning that Adam performs well in the classification of non-stationary Gaussian processes and stochastic gradient descent thrives in the minimization of a curvature-modulated parabola.**

I. Introduction

Machine learning methods are usually based on the assumption that the data generation mechanism does not change over time. However, this assumption is always violated in real world machine leaning applications, such as reinforcement learning, natural language processing, and robot control. Therefore, it is meaningful to figure out how machine learning algorithms work in non-stationary environments.

In this paper, two simple experiments to study the behavior of two optimization algorithms, namely Adam[1] and stochastic gradient descent (SGD)[2], in a synthetic non-stationary stochastic process are put forward. These two cases where the one with fixed minimum and the other with moving minimum but measured with binary cross-entropy.

Subsequently, the experiments yield the observation that SGD can achieve a faster convergence rate than Adam due to non-stationary data generation processes. However, the momentum estimation performed by Adam itself is useful in a non-stationary stochastic process.

II. Experiments

Two experiments will be performed, where the non-stationarity is embedded in a loss function and a sampling distribution respectively.

A. Minimization of curvature-modulated parabola

The first experiment aims to study the behavior of the Adam algorithm compared to stochastic gradient descent, in a simple one dimensional non-stationary environment. The experimental setting consists of a function $f(x,t)$ which is a parabola in x and whose curvature is modulated by a sinusoidal.

$$f(x,t) = \alpha(t)x^2 \tag{1}$$

with,

$$\alpha(t) = \begin{cases} c & \text{if } t \leq T \\ \frac{A}{2}[1 + sin(\omega t)] + \epsilon & \text{if } t > T \end{cases} \tag{2}$$

where c is the initial amplitude, T is the duration of the initial phase after which the environment dynamics starts, A is the maximal amplitude of the sinusoidal, ω is the angular frequency and ϵ is a small value ensuring the amplitude is never zero. The goal is to compare the convergence of the two optimization algorithm towards the true global minimum which is at $x^\star = 0$.

Two sub-experiments will be conducted. First, the influence of the frequency ω is studied. Here, constant values for the amplitudes $c = 1$ and $A = 2$ are used with different values of ω, ranging from 0.4 to 3.2 in a doubling fashion. Second, the impact of the sinusoidal amplitude A is analyzed. Now, the frequency ω and the initial frequency c are fixed to 0.4 and 1 respectively, with values for A ranging from 1 to 8 in a doubling fashion. In both cases, the distance between the current estimate x and the global minimum $x^\star$ is measured and the algorithms are run for 55 iterations.

B. Classification of non-stationary Gaussian processes

The second experiment consists of a binary classification problem modeled as follows. The labels are drawn from a Bernoulli distribution

$$Y \sim Bernoulli(0.5) \tag{3}$$

and the data points two Gaussian distributions based on their label

$$X_{y\sim Y}(t) \sim \mathcal{N}(\mu_{y\sim Y}(t), v_{y\sim Y}(t)I_d),\ \mu_{y\sim Y}, v_{y\sim Y} \in \mathbb{R}^d \tag{4}$$

where y is the label associated to a given sample, d is the dimension of the environment, $\mu = \{\mu_i\}_{i=1}^d$ is the mean along each dimension and $v = \{v_i\}_{i=1}^d$ is the variance along each dimension. A logistic regression model

$$\hat{y} = f(x; w) = \sigma(w^\top x) \tag{5}$$

with binary cross-entropy loss

$$\mathcal{L}_{BCE}(y, \hat{y}) = y \ln(\hat{y}) + (1 - y) \ln(1 - \hat{y}). \tag{6}$$

is considered, where σ is the sigmoid function. Finally, a last random variable describes the lifespan of the data distributions, characterized by an exponential distribution

$$\tau \sim expo(0.01) \tag{7}$$

When the experiment time step reaches the value sampled according to τ, the means and variance of the Gaussian distributions are reset according to, for $i = 1...d$

$$\{\mu_{y\sim Y}\}_i \sim \mathcal{N}(0,1) \tag{8}$$

and

$$\{v_{y\sim Y}\}_i \sim \max(1, \mathcal{N}(0,1)) \tag{9}$$

Then, a new τ is sampled from the exponential distribution.

With this setting, comparing the convergence towards a global minimum is not feasible, however, the evolution of the loss can be analyzed. Additionally, the evolution of the momentum estimates of Adam by computing the decaying averages of past gradients and past squared gradients $\hat{m}_t$ and $\hat{v}_t$ [1]

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t, \tag{10}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2)(g_t \odot g_t), \tag{11}$$

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}, \quad \hat{v}_t = \frac{v_t}{1-\beta_2^t} \tag{12}$$

are studied, where g_t is the gradient on the current mini-batch. Two distinct sets of experiments will be conducted in this environment.

First, the impact of the environment dimension is analyzed. Here, keeping all parameters unchanged, 4000 steps of each algorithms will be performed. During this period, the evolution of the loss, the gradient, and squared gradient estimates maintained by Adam will be recorded for later analysis.

Second, the impact of the exponential decay parameters β_1 and β_2, used by Adam to update its internal estimates[1], are scrutinized. Keeping the dimension constant, 4000 steps of SGD and Adam will be performed for various values of β_1 and β_2. Again, the loss and the momentum estimates will be recorded at each time step.

For both experiments, a time series of the recordings for time steps $t_1, t_2, \ldots, t_{4000}$ is generated, and whenever $t_i = \tau$, the data is regenerated according to equations 8 and 9 and a new τ is sampled according to equation 7. A batch size of 128 is used.

III. Results

A. Minimization of curvature-modulated parabola

The results of the first experiment can be observed in Figure 1 and Figure 2, which we will discuss in order.

First, modulating the frequency has a periodic effect on the convergence of SGD. The frequencies have local changes in convergence, for better and for worse, while on the long term having the same rate of convergence. This is additionally illustrated by the conjunction of the lines in the bottom right of Figure 1. Increasing frequencies appear to inch closer to a straight line analogous to the line at the start of the experiment.

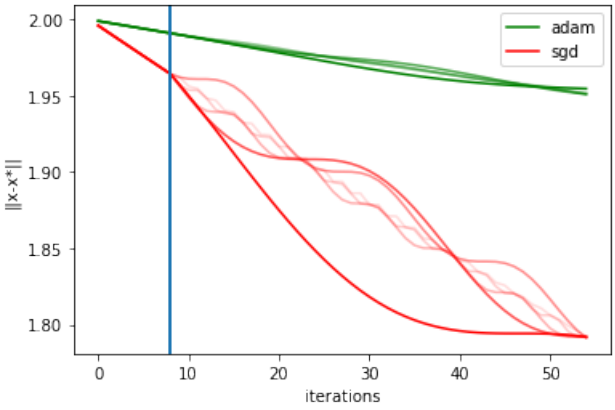

Fig. 1. **Influence of frequency:** the convergence of SGD and Adam on the minimization of a curvature-modulated parabola. The frequencies range from 0.4 (the lowest thick red line) to 3.2 (the most opaque line) with doubling intervals, using a fixed amplitude of 2.

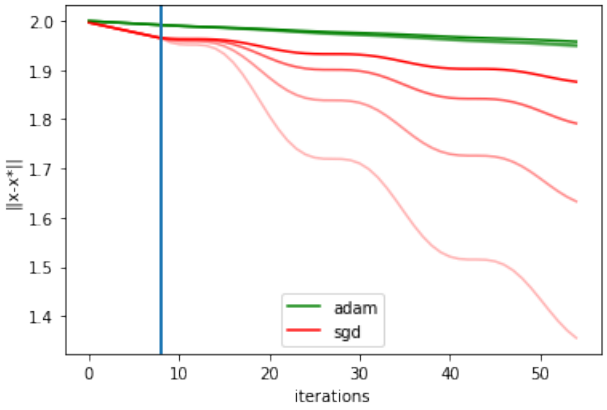

Fig. 2. **Influence of amplitude:** the convergence of SGD and Adam on the minimization of a curvature-modulated parabola. The amplitudes range from 1 (the top thick red line) to 8 (the most opaque line) with doubling intervals, using a fixed frequency of 0.4.

The effect of different frequencies on Adam seems to be negligible, however, it can be noted that the Adam optimizer does not perform well under said circumstances in comparison to SGD. Again, periodicity is observed, now in shorter periods.

Second, modulating the amplitude has a positive effect on the convergence of SGD. Higher amplitude waves on the parabola appear to aid the convergence of SGD with an increasingly positive effect. No significant effect is observed for Adam when increasing the amplitude.

B. Classification of non-stationary Gaussian processes

Figures displaying the results of this experiment can be found in appendix A.

First, the analysis of the impact of dimensions with fixed parameters shows a similar evolution of the binary cross-entropy loss of Adam and SGD for all dimensions in figure A-A. The value of binary cross-entropy loss reaches a peak when the data distribution is changed and later converges

towards zero for both Adam and SGD. In all cases, Adam performs better than SGD.

Second, the analysis of the impact of β_1 and β_2 shows similar results in figure A-B. The loss consistency converges towards zero with spikes corresponding to regeneration of the data. The plots of the momentum also display a similar overall behavior. However, the magnitude of the spikes increases when the values of β_1 and β_2 decrease. This behavior is expected when observing equations 10 and 11 which tells us that the estimates at time $t+1$ are a weighted average of themselves at time t, weighted by β_1 or β_2, and the gradient value or gradient squared, weighted by $(1-\beta_1)$ or $(1-\beta_2)$, at time $t+1$. The value of the gradient contains more noise due to the changes in the data distribution and reducing the values of β_1 and β_2 gives the gradient value more importance relative to the previous estimates. Hence the larger spikes.

Third, concerning the momentum estimates in figures A-A and A-B, it is observed that the peaks of the momentum parameters correspond to the peak of binary cross-entropy loss. Additionally, the absolute value of exponential average trend is equal to the changes of binary cross-entropy loss associated with Adam. If the absolute value of exponential average decreases, then binary cross-entropy loss decreases correspondingly.

IV. Conclusion

In this paper, a comparison has been put forward between stochastic gradient descent and Adam in non-stationary environments through two experiments. A simple minimization task on a curvature-modulated parabola showed that stochastic gradient descent benefits more from vibrations than Adam, which displays robust behavior in a changing environment. A more complex binary classification task, where the non-stationarity is woven through the sampling distributions, reveals nimble convergence after changes in sampling in both algorithms, particularly in Adam. Additionally, tinkering with the inner dials of Adam shows a degeneration when the parameters are decreased leading to large irregularities in the binary cross-entropy loss.

V. Discussion

In general, the impact of adding supplementary optimization algorithms for a broader perspective on non-stationarity should be considered in the future. Specifically, for the minimization of the curvature-modulated parabola the impact of increasing the dimensionality can be considered. Here, the problem would remain convex and would allow for a greater variety of oscillations and possibly combinations of oscillations to be considered. Further, in the classification of non-stationary Gaussian processes, differing distributions can be examined together with more class labels.

APPENDIX A
FIGURES OF EXPERIMENT B

A. *Impact of dimension*

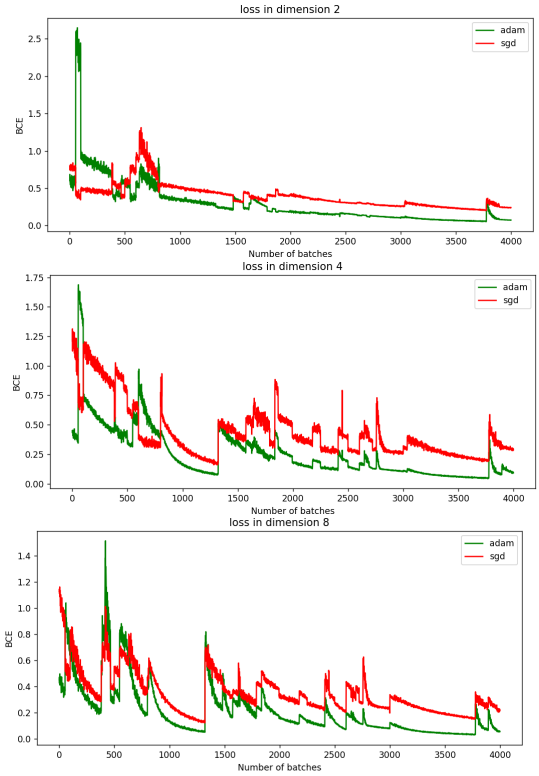

Fig. 3. **Influence of dimension:** the convergence of SGD and Adam on binary cross-entropy loss with dimension 2, 4 and 8.

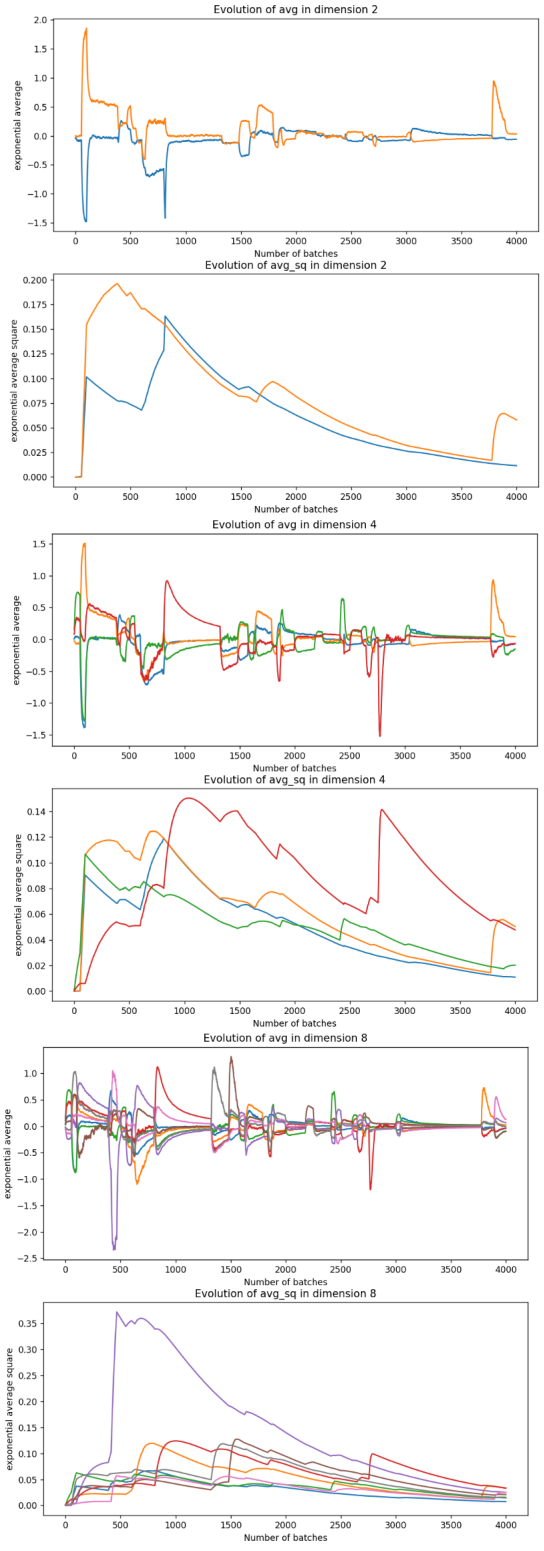

Fig. 4. **Influence of dimension:** first (avg) and second (avg_{sq}) momentum estimates in dimension 2, 4, 8 of Adam. Each curve correspond to a single dimension of the estimates.

B. *Impact of* β_1, β_2

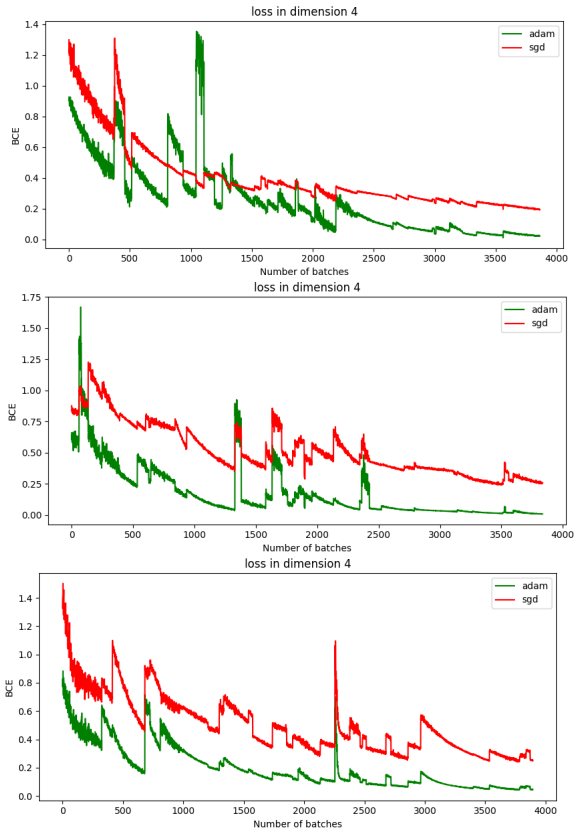

Fig. 5. **Influence of β_1, β_2:** the convergence of SGD and Adam on binary cross-entropy loss with $(\beta_1, \beta_2) = (0.3, 0.399)$ (top), $(\beta_1, \beta_2) = (0.6, 0.699)$ (middle) and $(\beta_1, \beta_2) = (0.9, 0.999)$ (bottom).

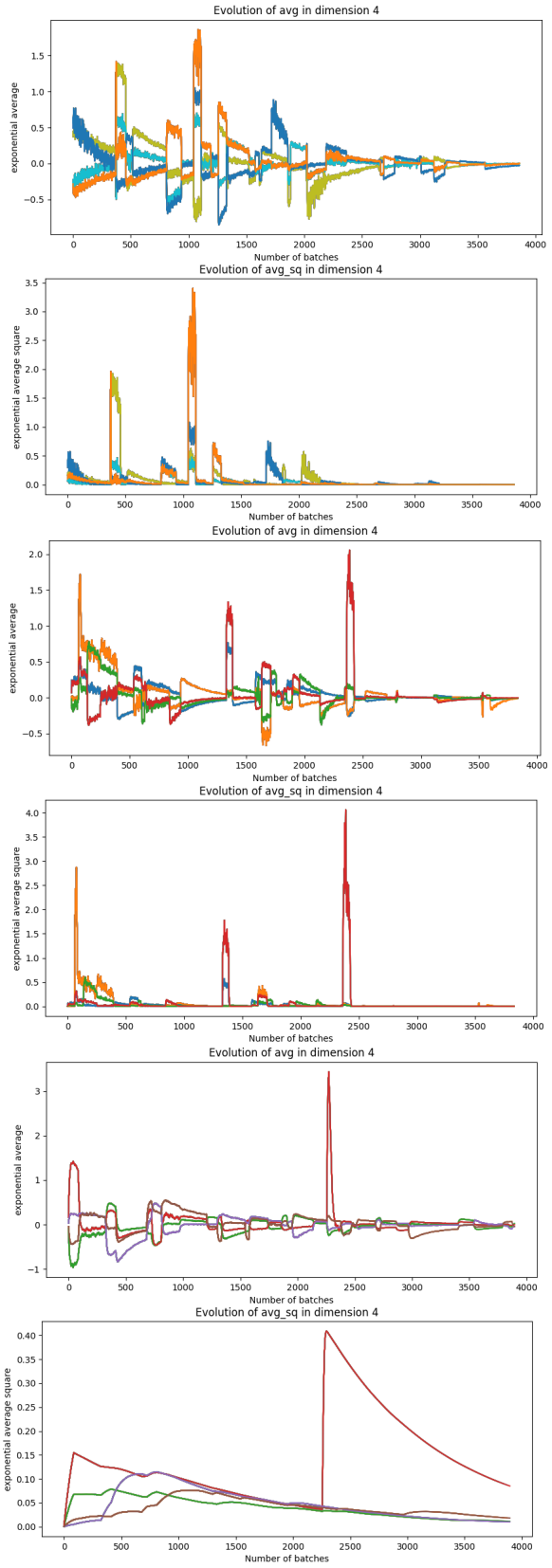

Fig. 6. **Influence of β_1, β_2:** first (avg) and second (avg_{sq}) momentum estimates with $(\beta_1, \beta_2) = (0.3, 0.399)$ (first two figures), $(\beta_1, \beta_2) = (0.6, 0.699)$ (middle two figures) and $(\beta_1, \beta_2) = (0.9, 0.999)$ (last two figures). Each curve correspond to a single dimension of the estimates.

APPENDIX B
PSEUDOCODES

A. SGD algorithm [2]

Algorithm 1 SGD

Require: Differentiable objective function $f(\theta)$, initial parameters θ_0, stepsize parameter α and number of iterations T

for $t = 1, 2, \ldots, T$ **do**
    $g_t \leftarrow \nabla_\theta f(\theta_t)$
    $\theta_t \leftarrow \theta_{t-1} - \alpha \cdot g_t$
end for
Return θ_t (resulting parameters)

B. Adam algorithm [1]

Algorithm 2 Adam

Require: Differentiable objective function $f(\theta)$, initial parameters θ_0, stepsize parameter α, exponential decay rates β_1, β_2 for moment estimates, tolerance parameter $\epsilon > 0$ and number of iterations T

$m_0 \leftarrow 0$
$v_0 \leftarrow 0$
for $t = 1, 2, \ldots, T$ **do**
    $g_t \leftarrow \nabla_\theta f(\theta_t)$
    $m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$
    $v_t \leftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot (g_t \odot g_t)$
    $\hat{m}_t \leftarrow m_t / (1 - \beta_1^t)$
    $\hat{v}_t \leftarrow v_t / (1 - \beta_2^t)$
    $\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \hat{m}_t / (\sqrt{\hat{v}_t} + \epsilon)$
end for
Return θ_t (resulting parameters)

REFERENCES

[1] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG].

[2] Herbert Robbins and Sutton Monro. "A Stochastic Approximation Method". In: The Annals of Mathematical Statistics, Vol.22, No.3, 1951, pp. 400–407.