Stochastic Optimization

Methods for efficient stochastic optimization: Stochastic Gradient Descent & Adam

[1]

BADIS JAOUANI

M2 Mathematics & Applications: Statistics, Sorbonne University

Contents

1. Introduction

Many of Machine Learning algorithms boil down to be a minimization of a certain loss function. The growing need for fast algorithms to solve such problems gave rise to multiple optimization techniques.

In this course, we focus on two stochastic first order methods, the Stochastic Gradient Descent (SGD), and the state of the art ADAM algortihm, that computes adaptive learning, accelerating the convergence. The convergence analysis for the two techniques is done using the Online Learning framework, in which a fundamental quantity, the regret, is defined and bounded. Numerical illustrations are given to compare both algorithms given a convex loss function.

2. Online Convex Optimization framework

The convergence is analyzed through the online setting, given an unknown sequence of convex loss functions $f_1(\theta), ..., f_T(\theta$. At each iteration $t \in \{1, ..., T\}$, we predict θ_t and evaluate it on a previously unknown cost function f_t. Since the sequence is unknown in advance, this is why the following quantity is introduced:

Definition 2.1 (Regret). *Let $\mathcal{X}$ be the set of feasible parameters θ. The regret, at the iteration T is defined as*

$$R(T) = \sum_{t=1}^{T} f_t(\theta_t) - f_t(\theta^*)$$

where $\theta^* = argmin_{\theta \in \mathcal{X}} \sum_{t=1}^{T} f_t(\theta)$

Intuitively, an algorithm performs well if its regret is sublinear as a function of T, SGD and ADAM will be compared based on this quantity. Before proceeding with the convergence proofs, we will need some simple elements from Convex Analysis, that we remind here.

Definition 2.2 (Convex function). *A funsction $f : R^d \to R$ is convex if for all $x, y \in R^d$, for all $\lambda \in [0, 1]$,*

$$\lambda f(x) + (1 - \lambda)f(y) \geq f(\lambda x + (1 - \lambda)y)$$

The following lemma, that we don't prove, is called the gradient trick and will be used as a starting point for the proof of convergence.

Lemma 2.1 (Gradient Trick). *If a function $f : R^d \to R$ is convex, then for all $x, y \in R^d$,*

$$f(y) \geq f(x) + \nabla f(x)^T(y - x)$$

This lemma tells that any convex function can be lower bounded by a hyperplane at its tangent, the idea behind the forthcoming convergence proofs is to take advantage of the update rules of the respective alogirithms to get better upper bounds fo rthe regret.

3. Stochastic Gradient Descent

3.1. Online Gradient Descent regret bound

Stochastic Gradient Descent is an application of the more general Online Gradient Descent (OGD) algorithm, introduced by (Zinkevich, 2003). At each iteration, the OGD algorithm takes a step from the previous point in the direction of the gradient of the previous cost, the pseudo-code is given below, note that the following algorithm is the projected version of the standard OGD, in which we require that the solution lies into a convex set $\mathcal{K} \subset R^d$.

Algorithm 1 Online Gradient Descent

```
Require: Convex set $\mathcal{K}$, number of iterations $T$, step sizes $\alpha_t$
  Initialization initial prediction $\theta_1 \in \mathcal{K}$
  For each iteration $t \geq 1$
  Predict $x_t$ and incur $f_t(x_t)$
  Observe $\nabla f_t(x_t)$
  Update and project
```

$$y_{t+1} = \theta_t - \alpha_t \nabla f_t(\theta_t) \tag{1}$$

$$\theta_{t+1} = \Pi_{\mathcal{K}}(y_{t+1}) \tag{2}$$

In all what follows, the subset $\mathcal{K}$ is assumed to be closed and bounded with diameter D i.e. $||\theta_1 - \theta_2|| \leq D \ \forall \theta_1, \theta_2 \in \mathcal{K}$. The next theorem gives the regret bound of the above procedure, obtained in the case of a general convex loss function.

Theorem 3.1 (OGD regret bound)**.** *Assume that the function f_t has bounded gradients $||\nabla f_t(\theta)|| \leq G$. OGD with step sizes $\alpha_t = \frac{d}{G\sqrt{t}}$ satisfies*

$$R(T) \leq \frac{3}{2} GD\sqrt{T}$$

Proof. Starting with the gradient trick, one gets $f_t(\theta_t) - f_t(\theta^*) \leq \nabla f_t(\theta_t)(\theta_t - \theta^*)$. Thus, we will estimate the linear regret

$$\sum_{t=1}^{T} \nabla f_t(\theta_t)(\theta_t - \theta^*)$$

Using the updates (1) and (2), we have

$$\begin{aligned} ||\theta_t - \theta^*||^2 &\leq ||\Pi_{\mathcal{K}}(\theta_t - \alpha_t \nabla f_t(\theta_t)) - \theta^*||^2 \\ &\leq ||\theta_t - \alpha_t \nabla f_t(\theta_t) - \theta^*||^2 \\ &\leq ||\theta_t - \theta^*||^2 + \alpha_t^2 ||\nabla f_t(\theta_t)||^2 - 2\alpha_t \nabla f_t(\theta_t)^T(\theta_t - \theta^*) \end{aligned}$$

The second inequality comes from the fact that the distance between points inside $\mathcal{K}$ is always lower thant the distance betweend a point inside $\mathcal{K}$ and a point outside $\mathcal{K}$. We then get

$$2\alpha_t \nabla f_t(\theta_t)^T(\theta_t - \theta^*) \leq ||\theta_t - \theta^*||^2 - ||\theta_{t+1} - \theta^*||^2 + \alpha_t^2 G^2. \tag{3}$$

One deduces that (with the convention $1/\alpha_0 = 0$)

$$\begin{aligned} 2\sum_{t=1}^{T} \nabla f_t(\theta_t)^T(\theta_t - \theta^*) &\leq \sum_{t=1}^{T} \frac{||\theta_t - \theta^*||^2 - ||\theta_{t+1} - \theta^*||^2}{\alpha_t} + G^2 \sum_{t=1}^{T} \alpha_t \\ &\leq \sum_{t=1}^{T} ||\theta_t - \theta^*||^2 \left(\frac{1}{\alpha_t} - \frac{1}{\alpha_{t+1}} \right) + G^2 \sum_{t=1}^{T} \alpha_t \\ &\leq D^2 \sum_{t=1}^{T} \left(\frac{1}{\alpha_t} - \frac{1}{\alpha_{t+1}} \right) + G^2 \sum_{t=1}^{T} \alpha_t \\ &\leq \frac{D^2}{\alpha_T} + G^2 \sum_{t=1}^{T} \alpha_t \\ &\leq 3DG\sqrt{T} \end{aligned}$$

using the expression of the step size and that $G^2 \sum_{t=1}^{T} \alpha_t \leq 2\sqrt{T}$. □

One gets a regret bound $R(T) = O(\sqrt{T})$. Better $O(\log(T)$ bounds can be obtained in the case of a strongly convex loss functions, see (Wintenberger, 2020) for a detailed proof.

3.2. Application: Accuracy of Stochastic Gradient Descent

Consider the Convex Optimization problem $(f, \mathcal{K})$. Instead of accurately computing ∇f, wxe use a noisy version of the gradient $\hat{\nabla} f$ that satisfies $E[\hat{\nabla} f(\theta)] = \nabla f(\theta)$ and $E[||\hat{\nabla} f(\theta)||^2] \leq G^2$. The approximation $\hat{\nabla} f$ is then unbiased with bounded variance, this setting is called Stochastic Optimization. The following proposition states the equivalence between Stochastic Optimization and Stochastic Online Convex Optimization if one condition is satisfied.

Proposition 3.1 (Equivalence between SO and OCO)**.** *Any SO problem reduces to Stochastic OCO problem by considering the approximation* $\nabla f_t(\theta_t) = \hat{\nabla} f_t(\theta_t)$.

Based on this equivalence, the SGD algorithm computes a noisy gradient at consecutive points in the decision set and returns the average of all points. The algorithm writes

```
Algorithm 2 Stochastic Gradient Descent
Require: Convex set $\mathcal{K}$, Epochs $T$, step sizes $\alpha_t$
  Initialization initial prediction $\theta_1 \in \mathcal{K}$
  For each iteration $t = 1, ..., T$
  Sample $\hat{\nabla} f_t(x_t)$
  Update and project
                    $y_{t+1} = \theta_t - \alpha_t \hat{\nabla} f_t(\theta_t)$        (4)
                    $\theta_{t+1} = \Pi_{\mathcal{K}}(y_{t+1})$        (5)
  Return $\bar{\theta}_{T+1} = \frac{1}{T+1} \sum_{t=1}^{T+1} \theta_t$
```

The averaging step comes from the use of the expectation in the approximation. The next theorem gives the accuracy of the SGD procedure, linked to the regret bound studied in Theorem 3.1.

Theorem 3.2 (SGD accuracy)**.** *Algorithm 2 with step sizes* $\alpha_t = \frac{D}{G\sqrt{T}}$ *achieves*

$$E[f(\bar{\theta}_t)] \leq \mathit{min}_{\theta^*} f(\theta^*) + \frac{3GD}{2\sqrt{T}}$$

Proof. The regret bound for OGD will be used to derive the accuracy of the SGD procedure. We have

$$
\begin{aligned}
E[f(\bar{\theta}_t)] - f(\theta^*) &\leq \mathbf{E}\frac{1}{T}\sum_t f(\theta_t) - f(\theta^*) && \text{by convexity of } f \text{ (Jensen's inequality)} \\
&\leq \frac{1}{T}\mathbf{E}\sum_t < \nabla f(\theta_t), \theta_t - \theta^* > && \text{by convexity again (gradient trick)} \\
&= \frac{1}{T}\mathbf{E}\sum_t < \hat{\nabla} f(\theta_t), \theta_t - \theta^* > && \text{noisy gradient estimator} \\
&= \frac{1}{T}\mathbf{E}\sum_t f_t(\theta_t) - f_t(\theta^*) && \text{by Algorithm 2, line 4} \\
&= \frac{R(T)}{T} && \text{by definition of the regret} \\
&\leq \frac{3GD}{2\sqrt{T}} && \text{by Theorem 3.1}
\end{aligned}
$$

□

Here also, better $O(\frac{1}{T})$ convergence rates can be achieved in the case of strongly convex loss function, see (Hazan, 2019) for more details.

In the next section, we focus on an improvement of the SGD algorithm, which uses adaptive learning rates and momentum.

4. ADAM

This algorithm was introduced by (Kingma & Ba, 2015). The main focus was to solve optimization problems with high-dimensional parameter spaces. SGD has proven to be very effective in solving such problems and was successful in many machine learning applications. What ADAM brings is the use of momentum and adaptive learning rates which help the algorithim being more stable. The idea behind momentum comes from the RMSProp algorithm (Tieleman & Hinton, 2012) while the use of adaptive learning rates originates from the AdaGrad algorithm intriduced by (Duchi et al., 2011). Consequently, ADAM is a combination of two successful optimization techniques.

4.1. The algorithm

The same Stochastic Optimization setting is used here. We are interested in minimizing the expected value of a noisy objective function $\mathbf{E}[f(\theta)]$ with respect to its parameters θ. We denote $g_t = \nabla_\theta f_t(\theta)$ the gradient wrt. θ evaluated at the step t, the algorithm is presented below.

In the algorithm below, g_t^2 indicates the elementwise square. Empirically, testing the algorithm on machine learning problems results in a good choice of parameters, which are $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. The notations β_1^t and β_2^t indicate β_1 and β_2 to the power t, respectively.

m_t and v_t can be respectively seen as exponential moving averages of the gradient, and a squared gradient, the parameters β_1 and β_2 control the magnitude of the exponential decays of thes averages. Note that the bias correction is necessary since we initialize β_1

Algorithm 3 Adam

```
Require: Stepsize $\alpha$, exponential decay rates for the moment $\beta_1, \beta_2 \in [0, 1)$
Require: Stochastic objective function $f(\theta)$
  Initializations:
  Initial parameter vector $\theta_0$
  $m_0 = 0$ (first moment vector)
  $v_0 = 0$ (second moment vector)
  $t = 0$ (time step)
  While $\theta_t$ not converged do
  $t = t + 1$
  $g_t = \nabla_\theta f_t(\theta_{t-1})$ (Compute the gradient at time step $t$)
  $m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$ (update biased first moment)
  $v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$ (update biased second raw moment)
  $\hat{m}_t = m_t/(1 - \beta_1^t)$ (Compute bias-corrected first moment)
  $\hat{v}_t = v_t/(1 - \beta_2^t)$ (Compute bias-corrected second raw moment)
  $\theta_t = \theta_{t-1} - \alpha \hat{m}_t/(\sqrt{\hat{v}_t} + \epsilon)$
  End While
  Return $\theta_t$
```

and β_2 by vectors of zeros, leading to biased estimates towards 0. We correct the resulting bias by introducing $\hat{m}_t$ and $\hat{v}_t$, which are the bias-corrected estimates.

Let us focus on the update rule of Algorithm 3. The choice of the stepsizes is very careful. Indeed, the effective step taken at iteration t (assuming $\epsilon = 0$) is $\delta_t = \alpha.\hat{m}_t/\sqrt{\hat{v}_t}$. This stepsize is bounded by α in the case $(1 - \beta_1) \leq \sqrt{1 - \beta_2}$ and by $\alpha.(1 - \beta_1)\sqrt{1 - \beta_2}$ otherwise. In the latter case, where the gradients have been zeros across all time steps except the current one, the stepsize can be bigger than α but never too big. In the most common cases, the stepsize is always lower than α.

In what follows, we derive a resulting upper bound of the regret when using Algorithm 7. The same Online Learning framework will be used for the derivation.

4.2. Convergence analysis

We show that Adam has $O(\sqrt{T})$ regret bound, which is the best known regret bound for OCO problems. We denote $g_t = \nabla f_t(\theta_t)$ the gradient at iteration t and $g_{t,i}$ its i^{th} element. Define $g_{1:t,i}$ the vector that contains the i^{th} dimension of the gradients across the time steps until t. To simplify the notations, we define $\gamma = \frac{\beta_1^2}{\sqrt{\beta_2}}$.

To derive the convergence rates, we will make use of two lemmas. The general idea is to substitute the bound we get by the gradient trick with the Adam's update rules.

Lemma 4.1. *Let $g_t = \nabla f_t(\theta_t)$ and $g_{1:t,i}$ as defined above. Assume g_t is bounded i.e. $||g_t||_2 \leq G$ and $||g_t||_\infty \leq G_\infty$. Then,*

$$\sum_{t=1}^{T} \sqrt{\frac{g_{t,i}^2}{t}} \leq 2G_\infty ||g_{1:T,i}||_2 \tag{6}$$

Proof. To prove the inequality, we will conduct an induction over T.

For $T = 1$, the result is obvious since $\sqrt{g_{1,i}^2} \leq 2G_\infty ||g_{1,i}||_2$.

Induction over T:

$$\begin{aligned}
\sum_{t=1}^{T} \sqrt{\frac{g_{t,i}^2}{t}} &= \sum_{t=1}^{T-1} \sqrt{\frac{g_{t,i}^2}{t}} + \sqrt{\frac{g_{T,i}^2}{T}} \\
&\leq 2G_\infty ||g_{1:T-1,i}||_2 + \sqrt{\frac{g_{T,i}^2}{T}} && \text{by the induction hypothesis} \\
&= 2G_\infty \sqrt{||g_{1:T,i}||_2^2 - g_{T,i}^2} + \sqrt{\frac{g_{T,i}^2}{T}} && \text{by adding and substracting } g_T^2
\end{aligned}$$

We then use the following inequality:

$$\left(||g_{1:T,i}||_2 - \frac{g_{T,i}^2}{2||g_{1:T,i}||_2}\right)^2 = ||g_{1:T,i}||_2^2 - g_{T,i}^2 + \frac{g_{T,i}^4}{2||g_{1:T,i}||_2^2} \geq ||g_{1:T,i}||_2^2 - g_{T,i}^2$$

Taking the square root of both sides, we obtain

$$\begin{aligned}
\sqrt{||g_{1:T,i}||_2^2 - g_T^2} &\leq ||g_{1:T,i}|| - \frac{g_{T,i}^2}{2||g_{1:T,i}||_2} \\
&\leq ||g_{1:T,i}|| - \frac{g_{T,i}^2}{2\sqrt{TG_\infty^2}}
\end{aligned}$$

since $||g_{1:T,i}||_2 = \sqrt{\sum_{t=1}^{T} g_{t,i}^2} \leq \sqrt{Tg_{1,i}^2} \leq \sqrt{TG_\infty^2}$.

We now substitute the $\sqrt{||g_{1:T,i}||_2^2 - g_T^2}$ term to get

$$\sum_{t=1}^{T} \sqrt{\frac{g_{t,i}^2}{t}} \leq 2G_\infty \sqrt{||g_{1:T,i}||_2^2 - g_T^2} + \sqrt{\frac{g_{T,i}^2}{T}} \leq 2G_\infty ||g_{1:T,i}||_2$$

□

The next lemma uses the update rule to derive an upper bound of the sum of a fraction between first order and second order moments estimates.

Lemma 4.2. *Let* $\gamma = \frac{\beta_1^2}{\sqrt{\beta_2}}$. *For* $\beta_1, \beta_2 \in [0, 1)$ *satisfying* $\gamma < 1$, *and an objective function with bounded gradients i.e.* $||g_t||_2 \leq G$ *and* $||g_t||_\infty \leq G_\infty$, *the following inequality holds*

$$\sum_{t=1}^{T} \frac{\hat{m}_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} \leq \frac{2G_\infty}{(1-\gamma)^2\sqrt{1-\beta_2}} ||g_{1:T,i}||_2 \tag{7}$$

Proof. Since $\sqrt{1-\beta_2^t} \leq 1$, one can write $\frac{\sqrt{1-\beta_2^t}}{(1-\beta_1)^2} \leq \frac{1}{(1-\beta_1)^2}$. We now expand the summation term and use the Adam's update rule to derive the upper bound (7).

$$\sum_{t=1}^{T} \frac{\hat{m}_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} = \sum_{t=1}^{T-1} \frac{\hat{m}_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} + \frac{\sqrt{1-\beta_2^T}}{(1-\beta_1^T)^2} \frac{(\sum_{k=1}^{T}(1-\beta_1)\beta_1^{T-k} g_{k,i})^2}{\sqrt{T\sum_{j=1}^{T}(1-\beta_2)\beta_2^{T-j} g_{j,i}^2}}$$

This inequality comes from the Adam's recursive update rules, written as a sum of all previous terms:

$$v_t = (1-\beta_2)\sum_{i=1}^{t} \beta_2^{t-i} g_i^2 \qquad \hat{v}_t = v_t/(1-\beta_2^t) \tag{8}$$

$$m_t = (1-\beta_1)\sum_{i=1}^{t} \beta_1^{t-i} g_i \qquad \hat{m}_t = m_t/(1-\beta_1^t) \tag{9}$$

We the get, by Cauchy-Schwarz,

$$\begin{aligned}
\sum_{t=1}^{T} \frac{\hat{m}_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} &\leq \sum_{t=1}^{T-1} \frac{\hat{m}_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} + \frac{\sqrt{1-\beta_2^T}}{(1-\beta_1^T)^2} \sum_{k=1}^{T} \frac{T((1-\beta_1)\beta_1^{T-k} g_{k,i})^2}{\sqrt{T\sum_{j=1}^{T}(1-\beta_2)\beta_2^{T-j} g_{j,i}^2}} \\
&\leq \sum_{t=1}^{T-1} \frac{\hat{m}_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} + \frac{\sqrt{1-\beta_2^T}}{(1-\beta_1^T)^2} \sum_{k=1}^{T} \frac{T((1-\beta_1)\beta_1^{T-k} g_{k,i})^2}{\sqrt{T(1-\beta_2)\beta_2^{T-k} g_{k,i}^2}} \\
&\leq \sum_{t=1}^{T-1} \frac{\hat{m}_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} + \frac{\sqrt{1-\beta_2^T}}{(1-\beta_1^T)^2} \frac{(1-\beta_1)^2}{\sqrt{T(1-\beta_2)}} \sum_{k=1}^{T} T\left(\frac{\beta_1^2}{\sqrt{\beta_2}}\right)^{T-k} \sqrt{g_{k,i}^2} \\
&= \sum_{t=1}^{T-1} \frac{\hat{m}_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} + \frac{\sqrt{1-\beta_2^T}}{(1-\beta_1^T)^2} \frac{(1-\beta_1)^2}{\sqrt{T(1-\beta_2)}} \sum_{k=1}^{T} T\left(\frac{\beta_1^2}{\sqrt{\beta_2}}\right)^{T-k} ||g_{k,i}||_2 \\
&\leq \sum_{t=1}^{T-1} \frac{\hat{m}_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} + \frac{T}{\sqrt{T(1-\beta_2)}} \sum_{k=1}^{T} \gamma^{T-k} ||g_{k,i}||_2
\end{aligned}$$

The second inequality follows from keeping only one term of the positive sum in the denominator. The third and fourth lines (equalities) are just term rearrangements. The last inequality follow from the assumption$\frac{\sqrt{1-\beta_2^t}}{(1-\beta_1)^2} \leq \frac{1}{(1-\beta_1)^2}$.

In a similar way, by using the Adam's update rules, one can get un upper bound all the terms in the summation.

$$\begin{aligned}
\sum_{t=1}^{T} \frac{\hat{m}_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} &\leq \sum_{t=1}^{T} \frac{||g_{t,i}||_2}{\sqrt{t(1-\beta_2)}} \sum_{j=0}^{T-t} t\gamma^j \\
&\leq \sum_{t=1}^{T} \frac{||g_{t,i}||_2}{\sqrt{t(1-\beta_2)}} \sum_{j=0}^{T} t\gamma^j
\end{aligned}$$

Knowing that $\gamma < 1$, and using the inequality $\sum_t t\gamma^t < \frac{1}{(1-\gamma)^2}$ which follows from the sum of arithmetic-geometric series, one gets

$$\sum_{t=1}^{T} \frac{||g_{t,i}||_2}{\sqrt{t(1-\beta_2)}} \sum_{j=0}^{T} t\gamma^j \leq \frac{1}{(1-\gamma)^2\sqrt{1-\beta_2}} \sum_{t=1}^{T} \frac{||g_{t,i}||_2}{\sqrt{t}}$$

The result follows from applying Lemma 4.1 to get

$$\sum_{t=1}^{T} \frac{\hat{m}_{t,i}^2}{\sqrt{t\hat{v}_{t,i}}} \leq \frac{2G_\infty}{(1-\gamma)^2\sqrt{1-\beta_2}} ||g_{1:T,i}||_2$$

□

Now that we have the two lemmas, we can state the theorem that gives the upper bound of the regret. This theorem holds for a learning rate $\alpha_t$decaying at a rate of $t^{-1/2}$.

Theorem 4.1. *Assume that the function f_t has bounded gradients i.e. $||g_t||_2 \leq G$ and $||g_t||_\infty \leq G_\infty$for all $\theta \in R^d$. Assume the distance between any θ_t generated by Adam is bounded, $||\theta_n - \theta_m||_2 \leq D$, $|\theta_n - \theta_m||_\infty \leq D_\infty$for any $m, n \in \{1, ...T\}$ and $\beta_1, \beta_2 \in [0, 1)$ satisfying $\frac{\beta_1^2}{\sqrt{\beta_2}} < 1$. Let $\alpha_t = \frac{\alpha}{\sqrt{t}}$ and $\beta_{1,t} = \beta_1 \lambda^{t-1}$, $\lambda \in (0, 1)$. Adam achieves the following regret bound, for all $T \geq 1$.*

$$R(T) \leq \frac{D^2}{2\alpha(1-\beta_1)} \sum_{i=1}^{d} \sqrt{T\hat{v}_{T,i}} + \frac{\alpha(\beta_1+1)G_\infty}{(1-\beta_1)\sqrt{1-\beta_2}(1-\gamma)^2} \sum_{i=1}^{d} ||g_{1:T,i}||_2 + \sum_{i=1}^{d} \frac{D_\infty^2 G_\infty \sqrt{1-\beta_2}}{2\alpha(1-\beta_1)(1-\lambda)^2}$$

We can clearly see the $\sqrt{T}$ dependence of the regret's upper bound. There is also a dependence on the dimension of the parameters space, on the upper bounds of the loss function's gradients, and on the sequence of parameters returned by Adam.

Proof. We start by applying the gradient trick (Lemma 2.1) to get

$$f_t(\theta_t) - f_t(\theta^*) \leq g_t^T())\theta_t - \theta^*) = \sum_{i=1}^{d} g_{t,i}(\theta_{t,i} - \theta_{,i}^*)$$

we now use the update rules of Algorithm 3 to write the recursion between the different terms

$$\begin{aligned}\theta_{t+1} &= \theta_t - \alpha_t \hat{m}_t / \sqrt{\hat{v}_t} \\ &= \theta_t - \frac{\alpha_t}{1-\beta_1^t} \left(\frac{\beta_{1,t}}{\sqrt{\hat{v}_t}} m_{t-1} + \frac{1-\beta_{1,t}}{\sqrt{\hat{v}_t}} g_t \right)\end{aligned}$$

Let us focus on the i^{th} dimension of the parameter vector θ_t at each iteration. Substracting $\theta_{,i}^*$ and squaring both sides, we get

$$(\theta_{t+1,i} - \theta_{,i}^*)^2 = (\theta_{t,i} - \theta_{,i}^*)^2 - \frac{2\alpha_t}{1-\beta_1^t} \left(\frac{\beta_{1,t}}{\sqrt{\hat{v}_t}} m_{t-1} + \frac{1-\beta_{1,t}}{\sqrt{\hat{v}_t}} g_t \right) (\theta_{t,i} - \theta_{,i}^*) + \alpha_t^2 \left(\frac{\hat{m}_{t,i}}{\sqrt{\hat{v}_{t,i}}} \right)^2$$

We want do dervive an upper bound for $g_{t,i}(\theta_{t,i} - \theta_{,i}^*)$, we use the update rule of the second raw moment to get

$$
\begin{aligned}
g_{t,i}(\theta_{t,i} - \theta^*_{,i}) &= \frac{(1-\beta_1^t)\sqrt{\hat{v}_t}}{2\alpha_t(1-\beta_{1,t})}\left((\theta_{t,i}-\theta^*_{,i})^2 - (\theta_{t+1,i}-\theta^*_{,i})^2\right) \\
&\quad + \frac{\beta_{1,t}}{1-\beta_{1,t}}\frac{\hat{v}^{1/4}_{t-1,i}}{\sqrt{\alpha_{t-1}}}(\theta^*_{,i}-\theta_{t,i})\sqrt{\alpha_{t-1}}\frac{m_{t-1,i}}{\hat{v}^{1/4}_{t-1,i}} + \frac{\alpha_t(1-\beta_1^t)\sqrt{\hat{v}_{t,i}}}{2(1-\beta_{1,t})}\left(\frac{\hat{m}_{t,i}}{\sqrt{\hat{v}_{t,i}}}\right)^2 \\
&\leq \frac{1}{2\alpha_t(1-\beta_{1,t})}\left((\theta_{t,i}-\theta^*_{,i})^2 - (\theta_{t+1,i}-\theta^*_{,i})^2\right)\sqrt{\hat{v}_{t,i}} \\
&\quad + \frac{\beta_{1,t}}{2\alpha_{t-1}(1-\beta_{1,t})}(\theta^*_{,i}-\theta_{t,i})^2\sqrt{\hat{v}_{t-1,i}} + \frac{\beta_1\alpha_{t-1}}{2(1-\beta_1)}\frac{m^2_{t-1,i}}{\sqrt{\hat{v}_{t-1,i}}} + \frac{\alpha_t}{2(1-\beta_1)}\frac{\hat{m}^2_{t,i}}{\sqrt{\hat{v}_{t,i}}}
\end{aligned}
$$

The inequality follows from using Young's inequality $ab \leq a^2/2 + b^2/2$ with

$$
\begin{aligned}
a &= \frac{\hat{v}^{1/4}_{t-1,i}}{\sqrt{\alpha_{t-1}}}(\theta^*_{,i}-\theta_{t,i}) \\
b &= \sqrt{\alpha_{t-1}}m_{t-1,i}/\hat{v}^{1/4}_{t-1,i}
\end{aligned}
$$

and the fact that $\beta_{1,t} \leq \beta_1$, meaning $1/(1-\beta_{1,t}) \leq 1/(1-\beta_1)$.

Summing over the dimension and the iterations, and applying Lemma 4.2 to the last inequality, one gets

$$
\begin{aligned}
R(T) \leq{}& \sum_{i=1}^{d}\frac{1}{2\alpha_1(1-\beta_1)}(\theta_{1,i}-\theta^*_{,i})^2\sqrt{\hat{v}_{1,i}} + \sum_{i=1}^{d}\sum_{t=2}^{T}\frac{1}{2(1-\beta_1)}(\theta_{t,i}-\theta^*_{,i})^2\left(\frac{\sqrt{\hat{v}_{t,i}}}{\alpha_t} - \frac{\sqrt{\hat{v}_{t-1,i}}}{\alpha_{t-1}}\right) \\
&+ \frac{\alpha G_\infty}{(1-\beta_1)(1-\gamma)^2\sqrt{1-\beta_2}}\sum_{i=1}^{d}||g_{1:T,i}||_2 + \frac{\beta_1\alpha G_\infty}{(1-\beta_1)(1-\gamma)^2\sqrt{1-\beta_2}}\sum_{i=1}^{d}||g_{1:T,i}||_2 \\
&+ \sum_{i=1}^{d}\sum_{t=1}^{T}\frac{\beta_{1,t}}{2\alpha_t(1-\beta_{1,t})}(\theta_{t,i}-\theta^*_{,i})^2\sqrt{\hat{v}_{t,i}}
\end{aligned}
$$

We notice the appearance of a telescoping in the second term, which is characteristic to the derivations of regret bounds in Online Optimization. We now use the assumption $||\theta_n - \theta_m||_2 \leq D$, $|\theta_n - \theta_m||_\infty \leq D_\infty$ for all $m, n \in \{1, ..., T\}$, and compute the resulting term of the telescoping sum. We get

$$
\begin{aligned}
R(T) &\leq \frac{D^2}{2\alpha(1-\beta_1)}\sum_{i=1}^{d}\sqrt{T\hat{v}_{T,i}} + \frac{\alpha(1+\beta_1)G_\infty}{(1-\beta_1)(1-\gamma)^2\sqrt{1-\beta_2}}\sum_{i=1}^{d}||g_{1:T,i}||_2 \\
&\quad + \frac{D_\infty^2}{2\alpha}\sum_{i=1}^{d}\sum_{t=1}^{T}\frac{\beta_{1,t}}{1-\beta_{1,t}}\sqrt{t\hat{v}_{t,i}} \\
&\leq \frac{D^2}{2\alpha(1-\beta_1)}\sum_{i=1}^{d}\sqrt{T\hat{v}_{T,i}} + \frac{\alpha(1+\beta_1)G_\infty}{(1-\beta_1)(1-\gamma)^2\sqrt{1-\beta_2}}\sum_{i=1}^{d}||g_{1:T,i}||_2 \\
&\quad + \frac{D_\infty^2 G_\infty\sqrt{1-\beta_2}}{2\alpha}\sum_{i=1}^{d}\sum_{t=1}^{T}\frac{\beta_{1,t}}{1-\beta_{1,t}}\sqrt{t}
\end{aligned}
$$

Using $\beta_{1,t} \leq \beta_1 < 1$ and the arithmetic-geometric sum's upper bound $\sum_t t\lambda^t < \frac{1}{(1-\lambda)^2}$ we compute the following bound

$$\begin{aligned}\sum_{t=1}^{T} \frac{\beta_{1,t}}{1-\beta_{1,t}}\sqrt{t} &\leq \sum_{t=1}^{T} \frac{1}{1-\beta_1}\lambda^{t-1}\sqrt{t} \\ &\leq \sum_{t=1}^{T} \frac{1}{1-\beta_1}\lambda^{t-1}t \\ &\leq \frac{1}{(1-\beta_1)(1-\lambda)^2}\end{aligned}$$

Replacing the last term in the regret's upper bound, we obtain the stated result. □

To conclude on the convergence of the average regret of Adam, we observe that

$$\sum_{i=1}^{d} ||g_{1:T,i}||_2 = \sum_{i=1}^{d} \sqrt{\sum_{t=1}^{T} g_{t,i^2}} \leq \sum_{i=1}^{d} \sqrt{TG_\infty^2} = dG_\infty\sqrt{T}$$

which leads to the following corollary, by simply showing that $\lim_{T\to\infty}\frac{R(T)}{T} = 0$, using the above upper bound.

Corollary 4.1. *Assume that the function f_t has bounded gradients i.e. $||g_t||_2 \leq G$ and $||g_t||_\infty \leq G_\infty$ for all $\theta \in R^d$. Assume the distance between any θ_t generated by Adam is bounded, $||\theta_n - \theta_m||_2 \leq D$, $|\theta_n - \theta_m||_\infty \leq D_\infty$ for any $m, n \in \{1, ...T\}$ and $\beta_1, \beta_2 \in [0, 1)$ satisfying $\frac{\beta_1^2}{\sqrt{\beta_2}} < 1$. Let $\alpha_t = \frac{\alpha}{\sqrt{t}}$ and $\beta_{1,t} = \beta_1\lambda^{t-1}$, $\lambda \in (0, 1)$. Adam achieves the following guarantee, for all $T \geq 1$.*

$$\frac{R(T)}{T} = O\left(\frac{1}{\sqrt{T}}\right)$$

Consequently, Adam achieves the best known regret bound for Online Convex Optimization. Some variants of the algorithm exist, namely the AdaMax variant (Kingma & Ba, 2015), in which the update rule is slightly modified and the computation of the second momet v_t is no longer necessary. In what follows, the base version of Adam is used for the numerical simulations.

5. Numerical illustrations

In this section, numerical comparison is given between SGD, SGD with Nesterov momzntum and Adam. We show that Adam outperforms SGD in the case of training a linear SVM, and outperforms both algorithms in a Logistic Regression setting. The data sets are the following:

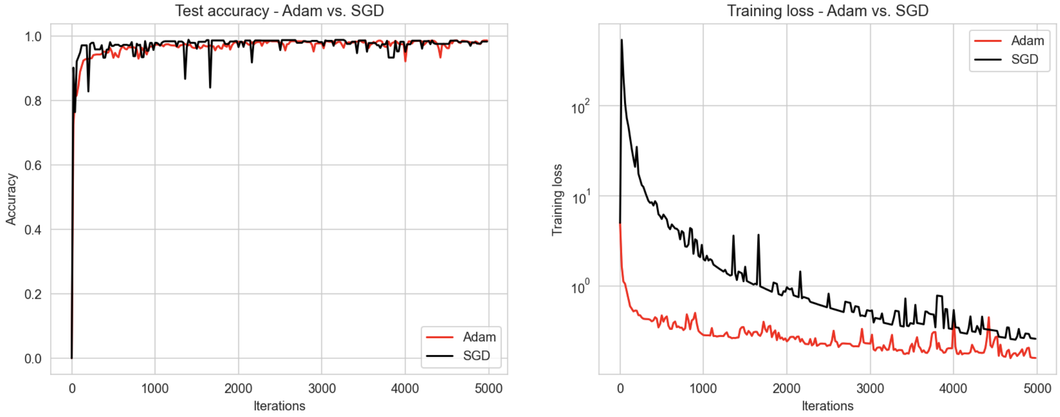
- MNIST data set, for which we consider the binary classification problem 0 vs. other digits for which training will be performed using SGD and Adam.
- A simulated data set according to a logistic model with a fixed set of parameters θ^*, the objective is to train a logitic regression classifier and use the Optimization algorithms (SGD and Adam) to approximate θ^* with the sequence $(\theta_t)_{t\in\{1,...,T\}}$.

5.1. Linear SVM training

We evaluate the two algorithms on the minimization of the L^2 regularized hinge loss, we empirically show that Adam can efficiently solve this problem, which is a Convex Optimization problem. The used parameters are the following:

- $\beta_1 = 0.9, \beta_2 = 0.999, \alpha = 0.01$ for Adam
- $\alpha = 0.01$ for SGD

The regularization parameter is $\lambda = 0.01$, we keep it as low as possible, even though it helps turning the problem into a strongly convex optimization problem, the idea is to compare the algorithms but also to achieve high accuracy on the testing set.


Test accuracy - Adam vs. SGD

Training loss - Adam vs. SGD

Figure 1. Test accuracy (left) and training loss function (right) - Adam vs. SGD

As it can be seen on the left figure, both algorithms achieve very high accuracy on the testing set (almost 1), Adam seems to be less oscillating and thus more stable. In addition, looking to the right figure we notice that Adam with well chosen parameters outperforms SGD as it achieves much lower values of the cost function in fewer iterations. Adam converges faster in this high dimensional setting , which matches the empirical results stated by (Kingma & Ba, 2015).

5.2. Logistic Regression setting

Let us npw focus on the Logistic Regression setting. The first step is to simulate the data set. We fix the number of samples to $n = 10000$ and the number of parameters (dimensionality of the problem to $p = 5$. The parameter vector to be approximated is

$$\theta^* = (-2, -1, 0, 1, 2, 5, 8, 9, 5, 0.2, 0.5, 0.8, 1.2, 5.9, 6.2)$$

The dimensionality of this problem is $p = 15$. The algorithms to be compared here are Adam, SGD as well as SGD with Nesterov momentum.

The parameters are the following:

- Adam: $\alpha = 0.05$, the other parameters are kept unchanged.
- SGD: $\alpha = 0.2$
- SGD Nesterov: similar to SGD

The function to minimize is the logistic loss, for which the gradient is the following

$$\nabla \mathcal{L}(x, y, \theta) = \big(Sigmoid(x^T\theta) - y\big)\, x$$

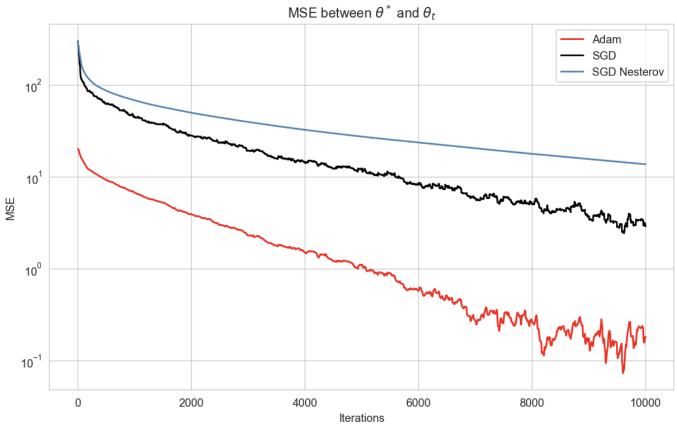


Figure 2. Mean Sqared Error between θ^* and the sequenceθ_t obtained by the three algorithms

The obtained results are shown in figure 2. Clearly, Adam outperforms SGD and SGD with Nesterov moments, which confirms Adam is a very efficient technique. More experiments with Neural Networks can be found in (Kingma & Ba, 2015)'s original paper, the results also show the efficiency of Adam's algorithm.

This document, as well as all the material for numerical simulations, can be found on `https://github.com/Badisj/Stochastic-Optimization`.

References

Diederik P. Kingma, Jimmy Lei Ba (2015): *ADAM: A Method for Stochastic Optimization.* , SarXiv:1412.6980v9 [cs.LG] 30 Jan 2017.

Duchi, John / Hazan, Elad / Singer, Yoram (2011): *Adaptive Subgradient Methods for Online Learning and Stochastic Optimization.* , Journal of Machine Learning Research 12 (2011) 2121-2159.

Hazan, Elad (2019): *Introduction to Online Convex Optimization.* , arXiv:1909.05207 [cs.LG], 2019.

Tijmen Tieleman, Geoffrey Hinton (2012): *Lecture 6.5 - RMSProp, COURSERA: Neural Networks for Machine Learning..* , Technical report (2012).

Wintenberger, Olivier (2020): *Online Convex Optimization.* , Sorbonne Université, 2020.

Zinkevic, Martin (2003): *Online Convex Programming and Generalized Infinitesimal Gradient Ascent.* , Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.