Convergence of Gradient Descent and Its Variants

Group 10

Anup Roy
puna20@iitk.ac.in
20111403

Jaivardhan Kapoor
jkapoor@iitk.ac.in
150300

Abstract

In this text, we survey prominent Gradient Descent techniques for optimization. Both, deterministic and stochastic methods are reviewed, such as SGD, Momentum, AdaGrad, ADAM and NAG. Convergence analyses of these algorithms are given, for objectives with various constraints on convexity, strong smoothness and strong convexity. Particularly for Adam, we review a recent work showing that the algorithm does not always converge, and restate the rigorous proof of the counterexample. Finally, the text aims to act as a reference for the reader to refer to convergence analyses of the above-mentioned methods, along with certain comments on the performance of these methods.

1. Introduction

Gradient descent is an optimization algorithm used to minimize a function by iteratively moving in the direction of steepest descent as defined by the negative of the gradient. (Ruder, 2016). Gradient Descent was formulated by Cauchy (1847) centuries ago. Many variants of this method have arrived, since, and are used in various fields.

Gradient Descent is predominantly used in training Deep Networks. More specifically, variants of Gradient Descent with Stochastic Update rules are used.

This survey aims to look at various variants of Gradient Descent and analyze the convergence of each variant in simple settings.

2. Preliminaries

2.1. Notation

We follow the general notation, where $\mathbf{x}^*$ is an optimal point to be learned, *i.e.* a local minima w.r.t. to a function, say $f : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ is the intersection of the domain set of f and the feasible set of points. The point $\mathbf{x}^t$ represents our approximation of the optimal point at a time step t, and the point $\hat{\mathbf{x}}$ represents the optimal point as estimated by the algorithm.

With an abuse of notation, we assume $\frac{\mathbf{a}}{\mathbf{M}}$ to be the same as $\mathbf{M}^{-1}\mathbf{a}$, where $\mathbf{a} \in \mathbb{R}^d$ and $\mathbf{M} \in \mathbb{R}^{d\times d}$ and $\sqrt{(.)}$ or $(.)^{1/2}$ to be element-wise square root operators.

2.2. Convex Functions

Convexity of a function simplifies the complexity of optimization by inducing inequalities that are helpful for convergence. Below, we define the condition for a function to be convex.

Definition 1.1 (Convex Fucntion). A function $f : \mathcal{X} \to \mathbb{R}$ is said to be convex, iff $\forall\, \mathbf{x}, \mathbf{y} \in \mathcal{X}$,

$$f(\mathbf{y}) \quad \geq \quad f(\mathbf{x}) + \langle \nabla f(\mathbf{x}),\ \mathbf{y} - \mathbf{x} \rangle \tag{1}$$

If a function $f : \mathcal{X} \to \mathbb{R}$ is convex, then all the following inequalities are equivalent,

1. $\forall\, \mathbf{x}, \mathbf{y} \in \mathcal{X}$
$$f(\mathbf{y}) \quad \geq \quad f(\mathbf{x}) + \langle \nabla f(\mathbf{x}),\ \mathbf{y} - \mathbf{x} \rangle$$
2. $\forall\, \mathbf{x}, \mathbf{y} \in \mathcal{X}$ and $\forall\, \alpha \in [\,0, 1\,]$
$$f(\alpha \cdot \mathbf{x} + (1-\alpha)\cdot \mathbf{y}) \quad \leq \quad \alpha \cdot f(\mathbf{x}) + (1-\alpha)\cdot f(\mathbf{y}) \tag{2}$$
3. If f is twice differentiable, then $\forall\, \mathbf{x} \in \mathcal{X}$
$$\nabla^2 f(\mathbf{x}) \quad \succeq \quad 0 \tag{3}$$

Below, we define two other inequalities, Strong Convexity and Strong Smoothness, that if a function satisfies, we can prove stronger convergence bounds for that function.

Definition 1.2 (Strong Convexity). A function $f : \mathcal{X} \to \mathbb{R}$ is said to be α-SC [1] if $\forall\, \mathbf{x}, \mathbf{y} \in \mathcal{X}$,

$$f(\mathbf{y}) \quad \geq \quad f(\mathbf{x}) + \langle \nabla f(\mathbf{x}),\ \mathbf{y} - \mathbf{x} \rangle + \frac{\alpha}{2}\,\|\,\mathbf{y} - \mathbf{x}\,\|_2 \tag{4}$$

Definition 1.3 (Strong Smoothness). A function $f : \mathcal{X} \to \mathbb{R}$ is said to be α-SS if $\forall\, \mathbf{x}, \mathbf{y} \in \mathcal{X}$,

$$f(\mathbf{y}) \quad \leq \quad f(\mathbf{x}) + \langle \nabla f(\mathbf{x}),\ \mathbf{y} - \mathbf{x} \rangle + \frac{\beta}{2}\,\|\,\mathbf{y} - \mathbf{x}\,\|_2 \tag{5}$$

We are now equipped with the basic tools sufficient to tackle the analysis of gradient based optimization methods. We discuss some of the deterministic methods in the next section, followed by Stochastic Methods of optimization in Section 4. We then conclude with some comments on the different techniques of optimization.

3. Deterministic Methods

Deterministic Methods of optimization use the actual value of the function f to compute the optimization step. We will discuss this in more detail when we discuss stochastic methods of optimization. We discuss three such methods of optimization, Vanilla Gradient Descent, Momentum and Nesterov's Accelerated Gradient Method and discuss their convergence under certain conditions.

[1] α-SC (α-SS) denotes that the function is α-Strongly Convex (α-Strongly Smooth)

The convergence of an algorithm is measured using the regret of the algorithm, which is defined, for a data point $\mathbf{x}$ as

$$\mathcal{R}\left[\,\mathbf{x}\,\right] \quad \triangleq \quad f(\mathbf{x}) - f(\mathbf{x}^*)$$

Therefore, our objective, when we are showing the convergence of an method is bound the regret of that function.

The algorithm for Vanilla Gradient Descent is given in Algorithm 1. Most descent algorithms follow the same rules, with minor additions and improvements to the optimization (update) step. The function h determines the form of the output, for example, h can be an average function, *i.e.* $h(\mathbf{x}^1 \ldots \mathbf{x}^T) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{x}^t$, or we can simply set h to return the last time step's estimate, *i.e.* $h(\mathbf{x}^1 \ldots \mathbf{x}^T) = \mathbf{x}^T$.

```
Algorithm 1: Deterministic Gradient Descent

Input:  Step sizes $\{\eta_t > 0\}_{t=1}^{T}$ and a function $h : \mathcal{S} \mapsto \mathcal{X}$ where $\mathcal{S}$ is a sequence of data points.
Output: $\hat{\mathbf{x}} \in \mathcal{H}$, where $\hat{\mathbf{x}} = h\left(\mathbf{x}^{(1)} \ldots \mathbf{x}^{(T)}\right)$
Steps:
  1. Initialize $\mathbf{x}^0 \in \mathcal{H}$
  2. For $t = 1 \ldots T$, do
        $\mathbf{g}_t = \nabla f\left(\mathbf{x}^t\right)$                                (Gradient Step)
        $\mathbf{x}^{t+1} = \mathbf{x}^t - \alpha_t \cdot \mathbf{g}_t$                   (Optimization Step)
  3. Return $\hat{\mathbf{x}} = h\left(\mathbf{x}^1 \ldots \mathbf{x}^T\right)$
```

3.1. Gradient Descent

Vanilla Gradient Descent iteratively solves the optimization problem, using the gradient of the function f at a time step. The idea is to update the parameter $\mathbf{x}$ in the opposite direction of the gradient of the optimization objective.

The projection step ensures that the predictions remain within the feasible set of points, *i.e.* $\mathbf{X}$.

In case of Vanilla Gradient Descent, the values of $\{\alpha_t\}_{t=1}^{T}$ are kept to be equal to the step sizes. Therefore, the update step can be written as

$$\mathbf{x}^{t+1} \quad = \quad \mathbf{x}^t - \eta_t \cdot \nabla f\left(\mathbf{x}^t\right) \tag{\textsc{Vanilla GD}}$$

In the follow subsections, we discuss the convergence and the necessary conditions required for this convergence for different settings for the optimizer function f.

Note. For the rest of the article, we use Φ_t to denote the difference between the t^{th} estimate of the optimal point and the real optimal value, *i.e.* $f\left(\mathbf{x}^t\right) - f\left(\mathbf{x}^*\right)$ and D_t to denote the difference between the current point estimate and the optimal point, *i.e.* $\left\|\,\mathbf{x}^t - \mathbf{x}^*\,\right\|_2$. $\mathbf{\Phi}_t$ is known as the Lyapunov function which basically defines the potential at a time step.

The convergence analysis were borrowed from the analysis discussed in the class of CS777, IITK Alam (2018).

3.1.1. When f is Convex with Bounded Gradients

First, we state the result, and later we give the derivation for the result.

Theorem 1.1. If $f : \mathbf{X} \rightarrow \mathbb{R}$ is convex and $\forall \mathbf{x} \in \mathcal{X}, \nabla f(\mathbf{x})$ exists, then for bounded gradients, we say

$$\frac{1}{T}\sum_{t=0}^{T} \mathbf{\Phi}_t \quad \leq \quad \frac{1}{2\sqrt{T}}\mathrm{D}_0 \cdot \mathrm{G} \tag{6}$$

Proof. From the convexity (equation 1) of the function f, we have

$$\begin{aligned} \Phi_t \quad &\leq \quad \left\langle \nabla f\left(\mathbf{x}^t\right), \, \mathbf{x}^t - \mathbf{x}^* \right\rangle \\ &= \quad \frac{1}{\eta}\left\langle \eta \cdot \nabla f\left(\mathbf{x}^t\right), \, \mathbf{x}^t - \mathbf{x}^* \right\rangle \end{aligned}$$

Here, we mention two properties, which will be used here, as well as a few times in later proofs

Property 1.1. For any two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^m$,

$$\|\, \mathbf{a} + \mathbf{b} \,\|_2^2 - \|\, \mathbf{a} \,\|_2^2 - \|\, b \,\|_2^2 \quad \overset{a}{=} \quad 2\langle a, \; b\rangle \quad \overset{b}{=} \quad \|\, \mathbf{a} \,\|_2^2 + \|\, b \,\|_2^2 - \|\, \mathbf{a} - \mathbf{b} \,\|_2^2 \tag{7}$$

Using property 7a, we can write the above inequality as

$$\begin{aligned} \Phi_t \quad &\leq \quad \frac{1}{2\eta}\left(\|\, \mathbf{x}^t - \mathbf{x}^* \,\|_2^2 + \eta^2 \,\|\, \nabla f\left(\mathbf{x}^t\right) \|_2^2 - \|\, \mathbf{x}^t - \eta \cdot \nabla f\left(\mathbf{x}^t\right) - \mathbf{x}^* \,\|\right) \\ &\leq \quad \frac{1}{2\eta}\left(\mathrm{D}_t^2 + \eta^2 \mathrm{G}^2 - \mathrm{D}_{t+1}^2\right) \\ &= \quad \frac{1}{2\eta}\left(\mathrm{D}_t^2 - \mathrm{D}_{t+1}^2\right) + \frac{\eta}{2}\mathrm{G}^2 \end{aligned}$$

Adding for $t = 0 \ldots T$, we get

$$\begin{aligned} \sum_{t=0}^{T} \Phi_t \quad &\leq \quad \frac{1}{2\eta}\left(\mathrm{D}_0^2 - \mathrm{D}_{T+1}^2\right) + \frac{\eta T}{2} \cdot \mathrm{G}^2 \\ \Longrightarrow \frac{1}{T}\sum_{t=0}^{T} \Phi_t \quad &\leq \quad \frac{1}{2\eta T}\mathrm{D}_0^2 + \frac{\eta}{2} \cdot \mathrm{G}^2 \end{aligned}$$

Since this inequality is true for any choice of η, we can minimize the RHS with respect to η to get an even stronger bound.

Therefore, we get

$$\frac{1}{T}\sum_{t=0}^{T} \mathbf{\Phi}_t \quad \leq \quad \frac{1}{2\sqrt{T}}\mathrm{D}_0 \cdot \mathrm{G}$$

This proves theorem 1.1 □

However, how does the above inequality ensure that gradient descent actually gives us a good estimate of the optimal point $\mathbf{x}^*$? This can, in fact, be seen as another result of convexity in the function, since, using the convexity properties of f, we can claim

$$f\left(\hat{\mathbf{x}}\right) \quad = \quad f\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}^t\right) \quad \leq \quad \frac{1}{2\sqrt{T}}\mathrm{D}_0^2\mathrm{G}^2 + f\left(\mathbf{x}^*\right)$$

Therefore, substituting this in equation 6, we can write

$$\mathcal{R}\left[\,\hat{\mathbf{x}}\,\right] \quad = \quad f\left(\hat{\mathbf{x}}\right) - f\left(\mathbf{x}^*\right) \quad \leq \quad \frac{1}{2\sqrt{T}}\mathrm{D}_0^2\mathrm{G}^2 \tag{8}$$

Hence for a case when the function f is convex and has bounded gradients, we can say that our regret is bounded with an order of $\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$, given the return function, *i.e.* h is an averaging function.

3.1.2. When f is Convex and β-Strongly Smooth

We now look at a more restrictive setting, in the sense that this setting allows us to have a much stronger bound than the bound given in equation 8. Again, we state the result first, then give a convergence proof for the same.

Theorem 1.2. If $f : \mathbf{X} \to \mathbb{R}$ is convex, β-smooth and $\forall\, x$, $\nabla f\left(\mathbf{x}^t\right)$ exists, we can say

$$\frac{1}{T}\sum_{t=0}^{T}\Phi_t \quad \leq \quad \frac{1}{2\eta}\cdot\frac{\mathrm{D}_0^2}{T} \tag{9}$$

Proof. From the convexity and smoothness of the function f, we have, respectively

$$\begin{aligned} f\left(\mathbf{x}^*\right) \quad &\geq \quad f\left(\mathbf{x}^t\right) - \left\langle \nabla f\left(\mathbf{x}^t\right),\ \mathbf{x}^t - \mathbf{x}^* \right\rangle \\ f\left(\mathbf{x}^{t+1}\right) \quad &\leq \quad f\left(\mathbf{x}^t\right) + \left\langle \nabla f\left(\mathbf{x}^t\right),\ \mathbf{x}^{t+1} - \mathbf{x}^t \right\rangle + \frac{\beta}{2}\left\|\,\mathbf{x}^{t+1} - \mathbf{x}^t\,\right\|_2^2 \end{aligned} \tag{10}$$

From the update equation of Gradient Descent, we can replace $\mathbf{x}^{t+1}$ with $\mathbf{x}^t - \eta_t \cdot \nabla f\left(\mathbf{x}^t\right)$. Therefore, we get

$$f\left(\mathbf{x}^{t+1}\right) \quad \leq \quad f\left(\mathbf{x}^t\right) + \left(\frac{\beta}{2} - \frac{1}{\eta_t}\right)\left\|\,\eta_t\cdot\nabla f\left(\mathbf{x}^t\right)\right\|_2^2 \tag{11}$$

Subtracting equation 10 from 11, we get

$$\begin{aligned} \Phi_{t+1} \quad &\leq \quad \left(\frac{\beta}{2} - \frac{1}{\eta_t}\right)\left\|\,\eta_t\cdot\nabla f\left(\mathbf{x}^t\right)\right\|_2^2 - \left\langle \nabla f\left(\mathbf{x}^t\right),\ \mathbf{x}^t - \mathbf{x}^* \right\rangle \\ &= \quad \left(\frac{\beta}{2} - \frac{1}{\eta_t}\right)\left\|\,\eta_t\cdot\nabla f\left(\mathbf{x}^t\right)\right\|_2^2 - \frac{1}{\eta_t}\left\langle \eta_t\cdot\nabla f\left(\mathbf{x}^t\right),\ \mathbf{x}^t - \mathbf{x}^* \right\rangle \end{aligned}$$

Using property 7a, we can write this, similarly to the previous case, as

$$\begin{aligned} \Phi_{t+1} \quad &\leq \quad \left(\frac{\beta}{2} - \frac{1}{\eta_t}\right)\left\|\,\eta_t\cdot\nabla f\left(\mathbf{x}^t\right)\right\|_2^2 + \frac{1}{2\eta_t}\left(\mathrm{D}_t^2 + \left\|\,\eta_t\cdot\nabla f\left(\mathbf{x}^t\right)\right\|_2^2 - \mathrm{D}_{t+1}^2\right) \\ \Phi_{t+1} \quad &\leq \quad \frac{1}{2\eta_t}\left(\mathrm{D}_t^2 - \mathrm{D}_{t+1}^2\right) + \left(\frac{\beta}{2} - \frac{1}{2\eta_t}\right)\left\|\,\eta_t\cdot\nabla f\left(\mathbf{x}^t\right)\right\|_2^2 \end{aligned}$$

Suppose if we set $\eta_t \leq \frac{1}{\beta}$, then the second term is always positive. Hence, we can write

$$\Phi_t \quad \leq \quad \frac{1}{2\eta_t}\left(\mathrm{D}_t^2 - \mathrm{D}_{t+1}^2\right)$$

Adding for $t = 0 \ldots T$, we get

$$\begin{aligned} \sum_{t=0}^{T} \Phi_t &\leq \frac{1}{2\eta_t}\left(\mathrm{D}_0^2 - \mathrm{D}_{T+1}^2\right) \\ \implies \frac{1}{T}\sum_{t=0}^{T} \Phi_t &\leq \frac{1}{2\eta}\frac{\mathrm{D}_0^2}{T} \end{aligned}$$

This completes the proof. □

Therefore, we can see that this bound offers much more than the bound in the previous case, as we can see that for large values of T, the bound will tend towards 0, and hence we can be sure our estimate of $f(\mathbf{x}^*)$ is good.

Also, similar to the previous case, we can write, using the properties of convexity,

$$\begin{aligned} f(\hat{\mathbf{x}}) \quad = \quad f\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}^t\right) &\leq \frac{1}{2\eta}\cdot\frac{\mathrm{D}_0^2}{T} + f(\mathbf{x}^*) \\ \implies \mathcal{R}[\hat{\mathbf{x}}] &\leq \frac{1}{2\eta}\cdot\frac{\mathrm{D}_0^2}{T} \qquad (12) \end{aligned}$$

Remark. In this case, setting $\eta = \frac{1}{\beta}$ would result in the most optimal bound.

Remark. The bound in this case is $\mathcal{O}\left(\frac{1}{T}\right)$ and therefore the convergence will be much faster and better in this case as opposed to simply convex case with bounded gradients.

3.2. Momentum

The gradient descent algorithm described above exhibits good convergence properties for "nice" functions (having smoothness and convexity/strong convexity properties). However, for poorly scaled objectives, for which in each dimension, the objective changes very differently, Gradient Descent may converge very slowly. This is because sometimes the updates get trapped in narrow valleys, where the direction of steepest descent for a given rate parameter causes the updates to oscillate between two sides of the smaller axis of the valley, and correspondingly move very slowly through the larger axis of the valley. Ruder (2016)

A solution devised for this problem was inspired from a physical analogy of a ball rolling down a hill. The ball has some momentum associated with it, and thus on the basis of this virtue, it can move past small bumps (local minima) and narrow valleys (poorly scaled regions). It takes into account a convex combination of the current and previous updates, ain to giving the updates a short-term memory. An additional parameter β is added to the GD updates to account for this momentum term, and correspondingly the updates become

$$\begin{aligned} \mathbf{z}^{t+1} &= \beta\cdot\mathbf{x}^t + \nabla f(\mathbf{x}^t) \\ \mathbf{x}^{t+1} &= \mathbf{x}^t - \alpha\cdot\mathbf{z}^{t+1} \end{aligned}$$

which amounts to

$$\mathbf{x}^{t+1} \quad = \quad \mathbf{x}^t - \alpha\nabla f(\mathbf{x}^t) + \beta(\mathbf{x}^t - \mathbf{x}^{t-1}) \qquad \text{(Momentum)}$$

The convergence rate of this method compared with the above method shows that the *condition number*, denoted by $\kappa = \frac{\beta}{\alpha}$, plays a large role in the speedup obtained using momentum. The gradient descent method, α-s.c. and β-s.s convex function gives the following result

$$\| \mathbf{x}^t - \mathbf{x}^* \| \quad = \quad \frac{\kappa - 1}{\kappa + 1} \left\| \mathbf{x}^0 - \mathbf{x}^* \right\|$$

However, using momentum yields us the following result:

$$\| \mathbf{x}^t - \mathbf{x}^* \| \quad = \quad \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \left\| \mathbf{x}^0 - \mathbf{x}^* \right\|$$

This means for a sufficient large $\kappa = 100$ the optimum is reached 10 times faster.

We consider a convex f which is β-strongly smooth, and derive the convergence rate for the Lyapunov function, $\Phi_t = f(\mathbf{x}^t) - f(\mathbf{x}^*)$.

Theorem 1.3. If f is convex and L-SS, then the above updates satisfy, for learning rate α an momentum parameter β,

$$\mathcal{R}\left[\, \hat{\mathbf{x}} \,\right] \quad = \quad f\left(\hat{\mathbf{x}}\right) - f\left(\mathbf{x}^*\right) \quad \leq \quad \frac{1}{T}\left(2\beta F + \frac{1-\beta}{2\alpha}\mathrm{D}_0^2\right) \tag{13}$$

That is, the error scale down inversely with the number of time steps.

Proof. We assume $\beta \in []0, 1)$. Also, define:

$$\mathbf{p}_t \quad = \quad \frac{\beta}{1-\beta}(\mathbf{x}^t - \mathbf{x}^{t-1})$$

from which we have

$$\mathbf{x}^{t+1} + \mathbf{p}t + 1 \quad = \quad \mathbf{x}^t + \mathbf{p}^t - \frac{\alpha}{1-\beta}\nabla f(\mathbf{x}^t)$$

Consider

$$\begin{aligned}
\left\| \mathbf{x}^{t+1} + \mathbf{p}^{t+1} - \mathbf{x}^* \right\|^2 \quad &= \quad \| \mathbf{x}^t + \mathbf{p}^t - \mathbf{x}^* \|^2 - \frac{2\alpha}{1-\beta} \cdot \langle \mathbf{x}^t + \mathbf{p}^t - \mathbf{x}^*, \nabla f(\mathbf{x}^t) \rangle + \left(\frac{\alpha}{1-\beta}\right)^2 \| \nabla f(\mathbf{x}^t) \|^2 \\
&= \quad \| \mathbf{x}^t + \mathbf{p}^t - \mathbf{x}^* \|^2 - \frac{2\alpha}{1-\beta} \cdot \langle \mathbf{x}^t - \mathbf{x}^*, \nabla f(\mathbf{x}^t) \rangle - \frac{2\alpha\beta}{(1-\beta)^2} \langle \mathbf{x}^t - \mathbf{x}^{t-1}, \nabla f(\mathbf{x}^t) \rangle \\
&+ \quad \left(\frac{\alpha}{1-\beta}\right)^2 \| \nabla f(\mathbf{x}^t) \|^2
\end{aligned}$$

From the strongly smooth property, it follows that:

$$\| \nabla f(\mathbf{x}^t) \|^2 \quad \leq \quad L\langle \mathbf{x}^t - \mathbf{x}^{t-1}, \nabla f(\mathbf{x}^t) \rangle f(\mathbf{x}^t) - f(\mathbf{x}^*) + \frac{1}{2L} \| \nabla f(\mathbf{x}^t) \|^2 \quad \leq \quad \langle \mathbf{x}^t - \mathbf{x}^{t-1}, \nabla f(\mathbf{x}^t) \rangle \tag{14}$$

Using these inequalities, we get

$$\begin{aligned}\left\| \mathbf{x}^{t+1} + \mathbf{p}^{t+1} - \mathbf{x}^* \right\|^2 &= \| \mathbf{x}^t + \mathbf{p}^t - \mathbf{x}^* \|^2 - \frac{2\alpha}{1-\beta}(f(\mathbf{x}^t) - f(\mathbf{x}^*) + \frac{1}{2L} \| \nabla f(\mathbf{x}^t) \|^2) \\ &- \frac{2\alpha}{1-\beta}\langle \mathbf{x}^t - \mathbf{x}^*, \nabla f(\mathbf{x}^t)\rangle + \left(\frac{\alpha}{1-\beta}\right)^2 \| \nabla f(\mathbf{x}^t) \|^2\end{aligned}$$

Applying convexity to the inner product, we obtain

$$\begin{aligned}\left\| \mathbf{x}^{t+1} + \mathbf{p}^{t+1} - \mathbf{x}^* \right\|^2 &= \| \mathbf{x}^t + \mathbf{p}^t - \mathbf{x}^* \|^2 - \frac{2\alpha}{1-\beta}(f(\mathbf{x}^t) - f(\mathbf{x}^*) + \frac{1}{2L} \| \nabla f(\mathbf{x}^t) \|^2) \\ &- \frac{2\alpha}{1-\beta}(f(\mathbf{x}^t) - f(\mathbf{x}^{t-1})) + \left(\frac{\alpha}{1-\beta}\right)^2 \| \nabla f(\mathbf{x}^t) \|^2\end{aligned}$$

Now subtract $\frac{2\alpha\beta}{(1-\beta)^2} f(\mathbf{x}^*)$ from both sides and collect the terms to obtain

$$\begin{aligned}\left(\frac{2\alpha}{1-\beta} + \frac{\alpha\beta}{(1-\beta)^2}\right)(f(\mathbf{x}^t) - f(\mathbf{x}^*)) + \left\| \mathbf{x}^{t+1} + \mathbf{p}^{t+1} - \mathbf{x}^* \right\|^2 &\leq \frac{2\alpha\beta}{(1-\beta)^2}(f(\mathbf{x}^{t-1}) - f(\mathbf{x}^*)) \\ &+ \| \mathbf{x}^t + \mathbf{p}^t - \mathbf{x}^* \|^2 \\ &+ \left(\frac{\alpha}{1-\beta}\right)\left(\frac{\alpha}{1-\beta} - \frac{1}{L}\right) \| \nabla f(\mathbf{x}^t) \|^2\end{aligned}$$

For $\alpha \in (0, (1-\beta)/L]$, we take $\alpha = (1-\beta)/L$, so that the last term vanishes to give

$$\begin{aligned}\left(\frac{2\alpha}{1-\beta} + \frac{\alpha\beta}{(1-\beta)^2}\right)(f(\mathbf{x}^t) - f(\mathbf{x}^*)) + \left\| \mathbf{x}^{t+1} + \mathbf{p}^{t+1} - \mathbf{x}^* \right\|^2 &\leq \frac{2\alpha\beta}{(1-\beta)^2}(f(\mathbf{x}^{t-1}) - f(\mathbf{x}^*)) \\ &+ \| \mathbf{x}^t + \mathbf{p}^t - \mathbf{x}^* \|^2\end{aligned}$$

Sum over both sides, to get

$$\left(\frac{2\alpha}{1-\beta}\right)\sum_{t=1}^{T} \Phi_t \leq \left(\frac{2\alpha\beta}{(1-\beta)^2}\right)\Phi_1 + \left\| \mathbf{x}^1 - \mathbf{x}^* \right\|^2$$

This gives us our required inequality. □

From the bound above, we can see that the function h will be an averaging function in this case as well, and we can find the regret bound for this case similarly. Also, in this case, the regret bound will be $\mathcal{O}\left(\frac{1}{T}\right)$.

3.3. Nesterov's Accelerated Gradient

NAG also takes advantage of the momentum term in the update equations, however the updates are made smarter to slow down when close to an optima. NAG obtains this by adding a lookahead term in the update equations. Figure 1 below roughly describes the difference between the updates. Notice

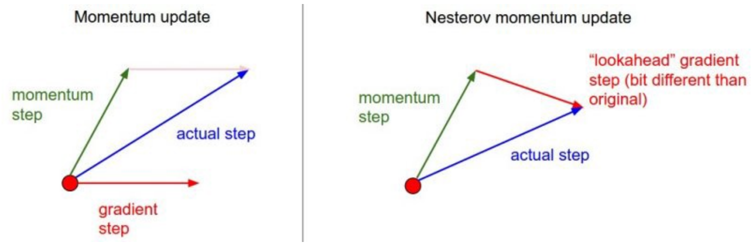

Figure 1: A look on updates for Momentum and NAG (*Source: Stanford CS231n Class*)

that the update in NAG first applies the momentum term, and then at the new position, applies the corrective gradient update.

For NAG, we consider the case where f is convex and β-SS.

Theorem 1.4. If f is convex and β-SS, then Nesterov Accelerated Gradient satisfies

$$f\left(\mathbf{x}^T\right) - f\left(\mathbf{x}^*\right) \quad \leq \quad \frac{2\beta \mathrm{D}_0^2}{(T-1)^2} \tag{15}$$

Proof. From the β-Smoothness of the function f, we have

$$\begin{aligned} f\left(\mathbf{x}^{t+1}\right) &\leq f(\mathbf{z}^{(t)}) + \left\langle \nabla f(\mathbf{z}^{(t)}),\ \mathbf{x}^{t+1} - \mathbf{z}^{(t)} \right\rangle + \frac{\beta}{2} \left\| \mathbf{x}^{t+1} - \mathbf{z}^{(t)} \right\|_2^2 \\ \implies f\left(\mathbf{x}^{t+1}\right) &\leq f(\mathbf{z}^{(t)}) + \left(\frac{\beta}{2} - \frac{1}{\eta}\right) \left\| \mathbf{x}^{t+1} - \mathbf{z}^{(t)} \right\|_2^2 \end{aligned} \tag{16}$$

Using the convexity of function f, we can write

$$\begin{aligned} f(\mathbf{x}) &\geq f(\mathbf{z}^{(t)}) + \left\langle \nabla f(\mathbf{z}^{(t)}),\ \mathbf{x} - \mathbf{z}^{(t)} \right\rangle \\ \implies f(\mathbf{x}) &\geq f(\mathbf{z}^{(t)}) + \frac{1}{\eta} \left\langle \mathbf{x}^{t+1} - \mathbf{z}^{(t)},\ \mathbf{x} - \mathbf{z}^{(t)} \right\rangle \end{aligned} \tag{17}$$

using equations 16 and 17, we can say

$$f\left(\mathbf{x}^{t+1}\right) - f(\mathbf{x}) \quad \leq \quad \left(\frac{\beta}{2} - \frac{1}{\eta}\right) \left\| \mathbf{x}^{t+1} - \mathbf{z}^{(t)} \right\|_2^2 - \frac{1}{\eta} \left\langle \mathbf{x}^{t+1} - \mathbf{z}^{(t)},\ \mathbf{x} - \mathbf{z}^{(t)} \right\rangle \tag{18}$$

We can now move on to find a bound similar to given in the theorem. First, note

$$\begin{aligned} \lambda_t^2 \cdot \Phi_{t+1} - \lambda_{t-1}^2 \cdot \Phi_t &= \lambda_t \cdot (\lambda_t \cdot \Phi_{t+1} - (\lambda_t - 1) \cdot \Phi_t) \\ &= \lambda_t \cdot \left(f\left(\mathbf{x}^{t+1}\right) - f\left(\mathbf{x}^*\right) \right) + \lambda_t \cdot (\lambda_t - 1) \left(f\left(\mathbf{x}^{t+1}\right) - f\left(\mathbf{x}^t\right) \right) \end{aligned}$$

Since $\forall\, t > 0$, $\lambda_t > 1$, using equation 18, we can derive the following inequality

$$\begin{aligned} \lambda_t^2 \cdot \Phi_{t+1} - \lambda_{t-1}^2 \cdot \Phi_t \quad \leq \quad & \lambda_t^2 \cdot \left(\frac{\beta}{2} - \frac{1}{\eta}\right) \left\| \mathbf{x}^{t+1} - \mathbf{z}^{(t)} \right\|_2^2 + \\ & + \ \frac{\lambda_t}{\eta} \cdot \left\langle \mathbf{x}^{t+1} - \mathbf{z}^{(t)},\ \mathbf{x}^* + (\lambda_t - 1) \cdot \mathbf{x}^t - \lambda_t \cdot \mathbf{z}^{(t)} \right\rangle \end{aligned}$$

Now, suppose we fix $\eta = 1/\beta$. Therefore, we can rewrite the above inequality as

$$\begin{aligned}\lambda_t^2 \cdot \Phi_{t+1} - \lambda_{t-1}^2 \cdot \Phi_t \quad &\leq \quad -\frac{\beta}{2} \cdot \Bigg(\left\| \lambda_t \cdot \left(\mathbf{x}^{t+1} - \mathbf{z}^{(t)} \right) \right\|_2^2 - \\ &\qquad - \; 2 \left\langle \lambda_t \cdot \left(\mathbf{x}^{t+1} - \mathbf{z}^{(t)} \right) , \; \mathbf{x}^* + (\lambda_t - 1) \cdot \mathbf{x}^t - \lambda_t \cdot \mathbf{z}^{(t)} \right\rangle \Bigg)\end{aligned}$$

From property 7b, where $\mathbf{a} = \lambda_t \cdot \left(\mathbf{x}^{t+1} - \mathbf{z}^{(t)} \right)$ and $b = \mathbf{x}^* + (\lambda_t - 1) \cdot \mathbf{x}^t - \lambda_t \cdot \mathbf{z}^{(t)}$, we can write this as

$$\begin{aligned}\lambda_t^2 \cdot \Phi_{t+1} - \lambda_{t-1}^2 \cdot \Phi_t \quad &\leq \quad -\frac{\beta}{2} \cdot \Bigg(\left\| \lambda_t \cdot \mathbf{x}^{t+1} - (\lambda_t - 1) \cdot \mathbf{x}^t - \mathbf{x}^* \right\|_2^2 - \\ &\qquad - \left\| \lambda_t \cdot \mathbf{z}^{(t)} - (\lambda_t - 1) \cdot \mathbf{x}^t - \mathbf{x}^* \right\| \Bigg)_2^2 \qquad (19)\end{aligned}$$

Now, from the definition of $\mathbf{z}^{(t+1)}$, we can write

$$\begin{aligned}\mathbf{z}^{(t+1)} \quad &= \quad \mathbf{x}^{t+1} + \gamma_t \cdot \left(\mathbf{x}^t - \mathbf{x}^{t+1} \right) \\ \implies \lambda_{t+1} \cdot \mathbf{z}^{(t+1)} \quad &= \quad \lambda_{t+1} \cdot \mathbf{x}^{t+1} + (1 - \lambda_t) \cdot \left(\mathbf{x}^t - \mathbf{x}^{t+1} \right) \\ \implies \lambda_{t+1} \cdot \mathbf{z}^{(t+1)} - (\lambda_{t+1} - 1) \cdot \mathbf{x}^{t+1} \quad &= \quad \lambda_t \cdot \mathbf{x}^{t+1} - (\lambda_t - 1) \cdot \mathbf{x}^t \qquad (20)\end{aligned}$$

We can substitute the term in equation 19 using equation 20. Now, define $\mathbf{u}_t = \lambda_t \cdot \mathbf{z}^{(t)} - (\lambda_t - 1) \cdot \mathbf{x}^t - \mathbf{x}^*$. Therefore, we can write

$$\lambda_t^2 \cdot \Phi_{t+1} - \lambda_{t-1}^2 \cdot \Phi_t \quad \leq \quad \frac{\beta}{2} \cdot \left(\| \mathbf{u}_t \|_2^2 - \| \mathbf{u}_{t+1} \|_2^2 \right)$$

Adding this from $t = 0, 1 \ldots (T-1)$

$$\lambda_{T-1}^2 \Phi_T \quad \leq \quad \frac{\beta}{2} \cdot \left(\| \mathbf{u}_0 \|_2^2 - \| \mathbf{u}_T \|_2^2 \right) + \lambda_0^2 \cdot \Phi_0 \quad \leq \quad \frac{\beta}{2} \cdot \| \mathbf{u}_0 \|_2^2 + \lambda_0^2 \cdot \Phi_0$$

We know $\lambda_0 = 0$. Also, using induction, it is easy to see that $\forall\, t \geq 2,\ \lambda_t > t/2$. therefore, we have

$$\Phi_T \quad \leq \quad \frac{2\beta \cdot \| \mathbf{u}_0 \|_2^2}{(T-1)^2} \quad = \quad \frac{2\beta \cdot \mathrm{D}_0^2}{(T-1)^2}$$

□

From the bound given above, we can see that the function h simply returns the value from the last iteration. Also, the bound in this case is much stronger, with $\mathcal{O}\left(\frac{1}{T^2}\right)$.

4. Stochastic Optimization

Deterministic methods, although guaranteeing convergence, can be slow and each time step can be expensive to perform. However, most convex functions can be written as a sum of simpler convex optimization objectives. In such a case, optimization can be made more effective by taking gradient steps with respect to each subfunction. Such an optimization technique is termed as Stochastic Optimization.

Logistic Regression with mini-batch learning can be seen as a simple example for Stochastic Optimization. The generic adaptive method for Stochastic Optimization is given in Algorithm 2, however we still do not give the forms of the functions $\{\phi_t, \psi_t\}_{t=1}^T$, which vary depending on different algorithms.

Stochastic Optimization has a natural annealing property which makes then suitable for even non-convex optimization, however a theoretical convergence guarantee cannot be formulate in that case.

```
Algorithm 2: Generic Adaptive Method

Input: step sizes $\left\{\eta_t \in \mathbb{R}^+\right\}_{t=1}^T$, a sequence of functions $\{\phi_t, \psi_t\}_{t=1}^T$, and a sequence of convex sub-objectives $\{f_t\}_{t=1}^T$
Output: $\hat{\mathbf{x}} \in \mathcal{H}$, where $\hat{\mathbf{x}} = h\left(\mathbf{x}^{(1)} \dots \mathbf{x}^{(T)}\right)$
Steps:
  1. Initialize $\mathbf{x}^{(0)} \in \mathcal{H}$
  2. For $t = 1 \dots T$, do
         $\mathbf{g}_t = \nabla f_t\left(\mathbf{x}^t\right)$
         $\mathbf{m}_t = \phi_t\left(\mathbf{g}_1 \dots \mathbf{g}_t\right)$
         $V_t = \psi_t\left(\mathbf{g}_1 \dots \mathbf{g}_t\right)$
         $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta_t V_t^{-1/2} \mathbf{m}_t$
  3. Return $\hat{\mathbf{x}} = h\left(\mathbf{x}^{(1)} \dots \mathbf{x}^{(T)}\right)$
```

4.1. Stochastic Gradient Descent

Parts of this section is borrowed from Kapoor (2018). In cases of large amount of data, the objective function that has to be calculated and correspondingly its gradient for the update steps in optimization procedures scale linearly with the data. Thus, sometimes the subset of data is chosen and gradients are calculated according to that subset of data. SGD is restricted to optimizations of these types, where the data parameters control the value of the function, and the data is assumed to be sampled from an unknown data distribution, $\mathcal{D}$.

Let us denote the parameter of choosing as $\boldsymbol{\theta} \sim \mathcal{D}$. Then the function is of the form $f(\mathbf{x}; \boldsymbol{\theta})$, such that

$$f(\mathbf{x}) = \mathop{\mathbb{E}}_{\theta \sim \mathcal{D}} \left[f(\mathbf{x}; \boldsymbol{\theta}) \right]$$

Therefore, we can choose our sub-functions to be of the form

$$f_t(\mathbf{x}) \quad \triangleq \quad f(\mathbf{x}; \boldsymbol{\theta}_t)$$

where $\boldsymbol{\theta}_t \sim \mathcal{D}$ is an arbitrary value sampled from the data distribution i.i.d.

In the update equations for SGD, we assume ϕ_t to return the last gradient, and ψ_t to return $\mathbf{I}$. More formally, SGD follows the update equations from Algorithm 2.

$$\phi_t\left(\mathbf{g}_1 \ldots \mathbf{g}_t\right) \quad = \quad \mathbf{g}_t \quad \text{and} \quad \psi_t\left(\mathbf{g}_1 \ldots \mathbf{g}_t\right) \quad = \quad \mathbf{I} \qquad \text{(SGD)}$$

Also, since the data points ($\boldsymbol{\theta}_t$) are assumed to be i.i.d, we can say the following about the gradients $\mathbf{g}_t$

$$\mathbb{E}\left[\mathbf{g}_t \,\middle|\, \mathcal{H}_t\right] \quad = \quad \nabla f(\mathbf{x})$$

where $\mathcal{H}_t$ is the history of $\mathbf{g}_t$, $\mathcal{H}_t = \{\mathbf{g}_1, \mathbf{g}_2 \ldots \mathbf{g}_{t-1}\}$. The above equation is satisfied as

$$\mathbb{E}\left[\,\mathbf{g}_t(\mathbf{x})\,\right] \quad = \quad \mathbb{E}_{\theta}\left[\,\nabla f(\mathbf{x},\theta)\,\right] \quad = \quad \nabla\mathbb{E}_{\theta}\left[\,(\mathbf{x},\theta)\,\right] \quad = \quad \nabla f(\mathbf{x})$$

The convergence analysis of the above algorithm is presented for convex f with bounded gradients. It is also assumed that $\mathcal{C}$ is convex. Another assumption we make is that $f(\mathbf{x},\theta)$ is convex for all θ. Also, $\|\, \mathbf{g}_t \,\|_2 < G$.

Using convexity condition between $\mathbf{x}^t$ and $\mathbf{x}^*$, where the latter is the optimum, we get

$$f(\mathbf{x}^t) - f(\mathbf{x}^*) \quad \leq \quad \langle \nabla f(\mathbf{x}^t), \mathbf{x}^t - \mathbf{x}^* \rangle$$

This is not particularly useful to us, since the algorithm deals with $\mathbf{g}_t$ instead of $\nabla f(\mathbf{x}^t)$. Therefore we work with $f(\mathbf{x},\theta)$ and $\mathbf{g}_t$, and then take expectation over δ on both sides. This will get rid of θ.

$$f(\mathbf{x}^t,\theta^t) - f(\mathbf{x}^*,\theta^t) \quad \leq \quad \langle \mathbf{g}_t, \mathbf{x}^t - \mathbf{x}^* \rangle$$

We now take expectation ($\mathbb{E}\left[\,\cdot|\mathcal{H}_t\,\right]$) on both sides:

$$\begin{aligned} f(\mathbf{x}^t) - f(\mathbf{x}^*) \quad &\leq \quad \langle \mathbf{g}_t, \mathbf{x}^t - \mathbf{x}^* \rangle \\ &\leq \quad \eta\frac{G^2}{2} + \frac{D_t^2}{2\eta} - \frac{\left\|\, \mathbf{z}^{t+1} - \mathbf{x}^* \,\right\|_2^2}{2\eta} \\ &\leq \quad \eta\frac{G^2}{2} + \frac{D_t^2}{2\eta} - \frac{D_{t_1}^2}{2\eta} \end{aligned}$$

We now take the expectation w.r.t $\mathcal{H}_t$ on both the sides, and sum over all t. The term $D_t^2 - D_{t+1}^2$ telescopes, giving the following result for $\eta = \frac{D_0}{G\sqrt{T}}$:

$$\frac{1}{T}\mathop{\mathbb{E}}_{\mathcal{H}_t}\left[\,f(\mathbf{x}^t)\,\right] \quad \leq \quad f(\mathbf{x}^*) + \frac{2G^2D_0^2}{\sqrt{T}}$$

Since the procedure is stochastic, we need a Chernoff-like bound for the convergence of $f(\mathbf{x}^t)$- $f(\mathbf{x}^*)$. For this, we propose 3 claims, for which the proof can be referred to in the cited article:

Claim 1.1. With high probability$(1-\delta)$,

$$\sum(f(\mathbf{x}^t) - f(\mathbf{x}^t,\theta^t)) \quad \leq \quad \sqrt{T}log(\frac{1}{\delta})$$

Claim 1.2. With high probability$(1-\delta)$,

$$\sum(f(\mathbf{x}^*,\theta^t) - f(\mathbf{x}^*)) \quad \leq \quad \sqrt{T}log(\frac{1}{\delta})$$

Claim 1.3. With high probability$(1-\delta)$,

$$\sum(f(\mathbf{x}^t,\theta^t) - f(\mathbf{x}^*,\theta^t)) \quad \leq \quad \sqrt{T}log(\frac{1}{\delta})$$

Combining the above 3 inequalities together, we get, w.h.p$(1-\delta)$

$$\frac{1}{T}\sum(f(\mathbf{x}^t)-f(\mathbf{x}^*)) \quad \leq \quad \frac{1}{\sqrt{T}}log(\frac{3}{\delta})$$

which gives us a confidence bound on the convergence of the function to its optimum.

4.2. Adam

In case of Adam (and AdaGrad, later), we define the regret to be with respect to each subfunction, and therefore the total regret is written as

$$R(T) \quad \triangleq \quad \sum_{t=1}^{T} f_t(\hat{\mathbf{x}^t}) - f_t(\mathbf{x}^*) \tag{21}$$

This is because all operations are done pointwise, and therefore the expected values of the gradients is difficult to compute.

Adam, short for Adaptive Moment Estimation, given by Kingma and Ba (2014), is an Adaptive Method and a variant of AdaGrad, which gives Exponentially Moving Averages. The key idea is to use exponentially moving average as function ψ_t instead of the simple averaging function we have seen until now.

Although other popular Adaptive Methods based on Exponentially Moving Averages exist, such as RMSprop, Nadam and AdaDelta, we only discuss Adam and its convergence in this article.

The update step for Adam is given as below

$$\begin{aligned} \phi_t\left(\mathbf{g}_1 \ldots \mathbf{g}_t\right) &= \beta_1 \cdot \phi_{t-1}\left(\mathbf{g}_1 \ldots \mathbf{g}_{t-1}\right) + (1-\beta_1) \cdot \mathbf{g}_t \\ \psi_t\left(\mathbf{g}_1 \ldots \mathbf{g}_t\right) &= \beta_2 \cdot \psi_{t-1}\left(\mathbf{g}_1 \ldots \mathbf{g}_{t-1}\right) + (1-\beta_2) \cdot \text{diag}\left(\mathbf{g}_t^2\right) \end{aligned} \tag{Adam}$$

There is an additional initial bias correction step (see Kingma and Ba, 2014), where we write

$$\widehat{\mathbf{m}}_t \quad = \quad \frac{\mathbf{m}_t}{1-\beta_1^t} \quad \text{and} \quad \widehat{\mathbf{V}}_t \quad = \quad \frac{\mathbf{V}_t}{1-\beta_2^t}$$

and use these instead of $\mathbf{m}_t$ and $\mathbf{V}_t$ in the update step.

4.2.1. Convergence Analysis of Adam

We give the same convergence analysis as given by Kingma and Ba (2014), but there are a few mistakes in the convergence analysis provided by them. We tackle these mistakes and try to give a better proof. As was previously done, we first state the regret bound and the assumptions required to obtain the bound.

Note. $\mathbf{g}_{1:t,i}$ denotes the vector $\left[g_{1,i}, g_{2,i} \ldots g_{t,i}\right]^{\mathrm{T}}$ for the following discussion

Theorem 1.5. For a series of convex sub-functions $\{f_t\}_{t=1}^T$ which have bounded gradients, *i.e.* $\forall\, \mathbf{x} \in \mathbb{R}$, $\|\nabla f_t(\mathbf{x})\|_2 \leq \mathrm{G}$ and $\|\nabla f_t(\mathbf{x})\|_\infty \leq \mathrm{G}_\infty$ and distance between any points $\mathbf{x}^t$ generated by Adam is bounded, *i.e.* $\forall\, i,j \in [T]$, $\left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2 \leq \mathrm{D}$ and $\left\|\mathbf{x}_i - \mathbf{x}_j\right\|_\infty \leq \mathrm{D}_\infty$, then Adam achieves the following guarantee, for all $T \geq 1$,

$$R(T) \quad \leq \quad \frac{1}{1-\beta_1}\left(\frac{\mathrm{D}^2\sqrt{T}}{2\eta}\mathrm{Tr}\left(\widehat{\mathbf{V}}^{1/2}\right) + \frac{\eta(1+\beta_1)}{\sqrt{1-\beta_2}(1-\gamma)^2}\sum_{i=1}^{d}\|\mathbf{g}_{1:T,i}\|_2 + \frac{d\,\mathrm{D}_\infty^2\mathrm{G}_\infty\sqrt{1-\beta_2}}{2\eta(1-\lambda)^2}\right)$$

where $\beta_1, \beta_2 \in [0,1)$ such that $\gamma \triangleq \frac{\beta_1^2}{\sqrt{\beta_2}} < 1$, and we set $\eta_t = \frac{\eta}{\sqrt{t}}$ and $\beta_{1,t} = \beta_1 \lambda^{t-1}$ where $\lambda \in (0,1)$

Before giving the proof, we first give two lemmas which will help us in deriving the final regret bound for Adam.

Lemma 1.5.1. If gradients $\mathbf{g}_t = \nabla f_t\left(\mathbf{x}^t\right)$ are bounded, *i.e.* $\|\mathbf{g}_t\|_2 \leq \mathrm{G}$ and $\|\mathbf{g}_t\|_\infty \leq \mathrm{G}_\infty$, then $\forall\, i \in [\,d\,]$

$$\sum_{t=1}^{T} \sqrt{\frac{g_{t,i}^2}{t}} \quad \leq \quad 2G_\infty \left\| \mathbf{g}_{1:T,i} \right\|_2$$

Proof. The proof is straightforward using induction over T. The base case, with $T = 1$ clearly satisfies the hypothesis. Now, suppose the hypothesis is true for all $T-1$, then we can write

$$\begin{aligned}
\sum_{t=1}^{T} \sqrt{\frac{g_{t,i}^2}{t}} \quad &\leq \quad 2G_\infty \left\| \mathbf{g}_{1:T-1,i} \right\|_2 + \sqrt{\frac{g_{T,i}^2}{T}} \\
&= \quad 2\mathrm{G}_\infty \sqrt{\left\| \mathbf{g}_{1:T,i} \right\|_2^2 - g_{T,i}^2} + \sqrt{\frac{g_{T,i}^2}{T}} \\
&\leq \quad 2\mathrm{G}_\infty \sqrt{\left(\left\| \mathbf{g}_{1:T,i} \right\|_2 - \frac{g_{T,i}^2}{2\left\| \mathbf{g}_{1:T,i} \right\|_2} \right)^2} + \sqrt{\frac{g_{T,i}^2}{T}} \\
&\leq \quad 2\mathrm{G}_\infty \left(\left\| \mathbf{g}_{1:T,i} \right\|_2^2 - \frac{g_{T,i}^2}{2\sqrt{T\mathrm{G}_\infty^2}} \right) + \sqrt{\frac{g_{T,i}^2}{T}}
\end{aligned}$$

where the last inequality comes from the fact that $2\left\| \mathbf{g}_{1:T,i} \right\| > g_{T,i}^2$ and the boundedness of the gradients. From this, we can directly write

$$\sum_{t=1}^{T} \sqrt{\frac{g_{t,i}^2}{t}} \quad \leq \quad 2G_\infty \left\| \mathbf{g}_{1:T,i} \right\|_2$$

□

We now give the second lemma that will help us prove the regret bound for Adam.

Lemma 1.5.2. For $\beta_1, \beta_2 \in [0,1)$ that satisfy $\gamma = \frac{\beta_1^2}{\sqrt{\beta_2}} < 1$, and bounded gradients, *i.e.* $\|\mathbf{g}_t\|_2 \leq \mathrm{G}$ and $\|\mathbf{g}_t\|_\infty \leq \mathrm{G}_\infty$, the following holds

$$\sum_{t=1}^{T} \frac{\widehat{m}_{t,i}^2}{\sqrt{t\,\widehat{v}_{t,i}}} \quad \leq \quad \frac{2\mathrm{G}_\infty}{(1-\gamma)\sqrt{1-\beta_2}} \left\| \mathbf{g}_{1:T,i} \right\|_2$$

Proof. Expanding the last term in the summation on the LHS, we get

$$
\begin{aligned}
\sum_{t=1}^{T} \frac{\widehat{m}_{t,i}^{2}}{\sqrt{t\,\widehat{v}_{t,i}}} &= \sum_{t=1}^{T-1} \frac{\widehat{m}_{t,i}^{2}}{\sqrt{t\,\widehat{v}_{t,i}}} + \frac{\sqrt{1-\beta_2^T}}{\left(1-\beta_1^T\right)^2} \frac{\left(\sum_{t=1}^{T}(1-\beta_1)\,\beta_1^{T-t} g_{t,i}\right)^2}{\sqrt{T\sum_{t=1}^{T}(1-\beta_2)\,\beta_2^{T-t} g_{t,i}^2}} \\
&\overset{(a)}{\leq} \sum_{t=1}^{T-1} \frac{\widehat{m}_{t,i}^{2}}{\sqrt{t\,\widehat{v}_{t,i}}} + \frac{1}{\sqrt{T(1-\beta_2)}} \frac{\left(\sum_{t=1}^{T}\beta_1^{T-t} g_{t,i}\right)^2}{\sqrt{\sum_{t=1}^{T}\beta_2^{T-t} g_{t,i}^2}} \\
&\overset{(b)}{=} \sum_{t=1}^{T-1} \frac{\widehat{m}_{t,i}^{2}}{\sqrt{t\,\widehat{v}_{t,i}}} + \frac{1}{\sqrt{T(1-\beta_2)}} \sum_{t=1}^{T} \frac{\left(\beta_1^{T-t} g_{t,i}\right)^2}{\sqrt{\sum_{t'=1}^{T}\beta_2^{T-t'} g_{t',i}^2}} \\
&\overset{(c)}{\leq} \sum_{t=1}^{T-1} \frac{\widehat{m}_{t,i}^{2}}{\sqrt{t\,\widehat{v}_{t,i}}} + \frac{1}{\sqrt{T(1-\beta_2)}} \sum_{t=1}^{T} \frac{\left(\beta_1^{T-t} g_{t,i}\right)^2}{\sqrt{\beta_2^{T-t} g_{t,i}^2}} \\
&\overset{(d)}{\leq} \sum_{t=1}^{T-1} \frac{\widehat{m}_{t,i}^{2}}{\sqrt{t\,\widehat{v}_{t,i}}} + \frac{1}{\sqrt{T(1-\beta_2)}} \sum_{t=1}^{T} \left(\frac{\beta_1^2}{\sqrt{\beta_2}}\right)^{T-t} g_{t,i} \\
&\overset{(e)}{\leq} \sum_{t=1}^{T-1} \frac{\widehat{m}_{t,i}^{2}}{\sqrt{t\,\widehat{v}_{t,i}}} + \frac{1}{\sqrt{T(1-\beta_2)}} \sum_{t=1}^{T} \gamma^{T-t} g_{t,i}
\end{aligned}
$$

The first inequality (a) is from the assumption $\forall\, t \in \mathbb{N},\ \frac{\sqrt{1-\beta_2^t}}{\left(1-\beta_1^t\right)^2} \leq \frac{1}{(1-\beta_1)^2}$. The inequality (c) is from the fact that $\forall\, t \in [\,T\,],\ \beta_2^{T-t} g_{t,i}^2 \leq \sum_{t'=1}^{T} \beta_2^{T-t'} g_{t',i}^2$.

Using the same arguments, we can upper bound the summation. Therefore, we get

$$
\begin{aligned}
\sum_{t=1}^{T} \frac{\widehat{m}_{t,i}^{2}}{\sqrt{t\,\widehat{v}_{t,i}}} &= \sum_{t=1}^{T} \left\{ \frac{g_{t,i}}{\sqrt{t(1-\beta_2)}} \sum_{t'=0}^{T-t} \gamma^{t'} \right\} \\
&\leq \sum_{t=1}^{T} \left\{ \frac{g_{t,i}}{\sqrt{t(1-\beta_2)}} \sum_{t'=0}^{\infty} \gamma^{t'} \right\} \\
&\leq \frac{1}{(1-\gamma)\sqrt{1-\beta_2}} \sum_{t=1}^{T} \frac{g_{t,i}}{\sqrt{t}}
\end{aligned}
$$

Applying Lemma 1.5.1, we can write

$$
\sum_{t=1}^{T} \frac{\widehat{m}_{t,i}^{2}}{\sqrt{t\,\widehat{v}_{t,i}}} \leq \frac{2\mathrm{G}_{\infty}}{(1-\gamma)\sqrt{1-\beta_2}} \sum_{t=1}^{T} \left\| \mathbf{g}_{1:T,i} \right\|
$$

□

We can now move towards a proof for Theorem 1.5. As suggested by Kingma and Ba (2014), the update using β_1 at a time step t is replaced by $\beta_{1,t} = \beta_1 \lambda^{t-1}$ where $\lambda < 1$ but very close to 1.

Proof. From the convexity of the sub-functions, we have for all $t = 1, 2 \ldots T$,

$$
f_t\left(\mathbf{x}^t\right) - f_t\left(\mathbf{x}^*\right) \leq \left\langle \mathbf{g}_t,\ \mathbf{x}^t - \mathbf{x}^* \right\rangle = \sum_{i=1}^{d} g_{t,i}\left(\mathbf{x}_{t,i} - \mathbf{x}_i^*\right)
$$

From the Adam update rules, we have

$$\begin{aligned}
\mathbf{x}^{t+1} &= \mathbf{x}^t - \eta_t \widehat{\mathbf{V}}_t^{-1/2} \widehat{\mathbf{m}}_t \\
&= \mathbf{x}^t - \frac{\eta_t}{\beta_1^t} \widehat{\mathbf{V}}_t^{-1/2} \left(\beta_{1,t} \cdot \mathbf{m}_{t-1} + \left(1-\beta_{1,t}\right) \cdot \mathbf{g}_t\right) \\
\implies \left(\mathbf{x}^{t+1} - \mathbf{x}^*\right) &= \left(\mathbf{x}^t - \mathbf{x}^*\right) - \frac{\eta_t}{1-\beta_1^t} \widehat{\mathbf{V}}_t^{-1/2} \left(\beta_{1,t} \cdot \mathbf{m}_{t-1} + \left(1-\beta_{1,t}\right) \cdot \mathbf{g}_t\right)
\end{aligned}$$

If we consider only the i^{th} coordinate, we get for all $i \in [\,d\,]$

$$\begin{aligned}
\left(x_{t+1,i} - x_i^*\right) &= \left(x_{t,i} - x_i^*\right) - \frac{\eta_t}{1-\beta_1^t} \left(\beta_{1,t} \cdot \frac{m_{t-1,i}}{\sqrt{\widehat{v}_{t,i}}} + \left(1-\beta_{1,t}\right) \cdot \frac{g_{t,i}}{\sqrt{\widehat{v}_{t,i}}}\right) \\
\left(x_{t+1,i} - x_i^*\right)^2 &= \left(x_{t,i} - x_i^*\right)^2 + \eta_t^2 \left(\frac{\widehat{m}_{t,i}}{\sqrt{\widehat{v}_{t,i}}}\right)^2 \\
&- \frac{2\eta_t}{1-\beta_1^t} \left(\beta_{1,t} \cdot \frac{m_{t-1,i}}{\sqrt{\widehat{v}_{t,t}}} + \left(1-\beta_{1,t}\right) \cdot \frac{g_{t,i}}{\sqrt{\widehat{v}_{t,i}}}\right) \left(x_{t,i} - x_i^*\right) \\
\implies g_{t,i}\left(x_{t,i} - x_i^*\right) &= \frac{\sqrt{\widehat{v}_{t,i}}\left(1-\beta_1^t\right)}{2\eta_t\left(1-\beta_{1,t}\right)} \left(\left(x_{t,i} - x_i^*\right)^2 - \left(x_{t+1,i} - x_i^*\right)^2\right) \\
&+ \frac{\eta_t\left(1-\beta_1^t\right)}{2\left(1-\beta_{1,t}\right)} \frac{\widehat{m}_{t,i}^2}{\sqrt{\widehat{v}_{t,i}}} + \frac{\beta_{1,t}}{1-\beta_{1,t}} \cdot m_{t-1,i}\left(x_i^* - x_{t,i}\right)
\end{aligned}$$

Consider the last term on the RHS in the above expression.

$$\begin{aligned}
\frac{\beta_{1,t}}{1-\beta_{1,t}} \cdot m_{t-1,i}\left(x_i^* - x_{t,i}\right) &= \sqrt{\frac{\beta_{1,t}}{1-\beta_{1,t}}} \sqrt{\frac{\widehat{v}_{t-1,i}^{1/2}}{\eta_{t-1}}} \left(x_i^* - x_{t,i}\right) \cdot \sqrt{\frac{\beta_{1,t}}{1-\beta_{1,t}}} \sqrt{\frac{\eta_{t-1}}{\widehat{v}_{t-1,i}^{1/2}}} m_{t-1,i} \\
&\leq \frac{\beta_{1,t}\sqrt{\widehat{v}_{t-1,i}}}{2\eta_{t-1}\left(1-\beta_{1,t}\right)} \left(x_{t,i} - x_i^*\right)^2 + \frac{\eta_{t-1}\beta_{1,t}}{2\left(1-\beta_{1,t}\right)} \frac{m_{t-1,i}^2}{\sqrt{\widehat{v}_{t-1,i}}}
\end{aligned}$$

Using this, and the fact that $\beta_{1,t} \leq \beta_1$, we can write

$$\begin{aligned}
g_{t,i}\left(x_{t,i} - x_i^*\right) &\leq \frac{\sqrt{\widehat{v}_{t,i}}}{2\eta_t\left(1-\beta_1\right)} \left(\left(x_{t,i} - x_i^*\right)^2 - \left(x_{t+1,i} - x_i^*\right)^2\right) \\
&+ \frac{\eta_t}{2\left(1-\beta_1\right)} \frac{\widehat{m}_{t,i}^2}{\sqrt{\widehat{v}_{t,i}}} + \frac{\beta_{1,t}\sqrt{\widehat{v}_{t-1,i}}}{2\eta_{t-1}\left(1-\beta_1\right)} \left(x_{t,i} - x_i^*\right)^2 \\
&+ \frac{\beta_1\eta_{t-1}}{2\left(1-\beta_1\right)} \frac{m_{t-1,i}^2}{\sqrt{\widehat{v}_{t-1,i}}}
\end{aligned}$$

Summing this over $t = 1 \ldots T$ and rearranging, we can write

$$\begin{aligned}
\sum_{t=1}^{T} g_{t,i}\left(x_{t,i} - x_i^*\right) \quad \leq \quad & \sum_{t=1}^{T} \frac{1 + \beta_{1,t} - \beta_{1,t}}{2\eta_t\left(1 - \beta_1\right)} \left(x_{t,i} - x_i^*\right)^2 \sqrt{\widehat{v}_{t,i}} \\
& - \sum_{t=2}^{T} \frac{1}{2\eta_{t-1}\left(1 - \beta_1\right)} \left(x_{t,i} - x_i^*\right)^2 \sqrt{\widehat{v}_{t-1,i}} \\
& + \sum_{t=2}^{T} \frac{\beta_{1,t}}{2\eta_{t-1}\left(1 - \beta_1\right)} \left(x_{t,i} - x_i^*\right)^2 \sqrt{\widehat{v}_{t-1,i}} \\
& + \sum_{t=1}^{T} \frac{\eta_t}{2\left(1 - \beta_1\right)} \frac{\widehat{m}_{t,i}^2}{\sqrt{\widehat{v}_{t,i}}} + \sum_{t=2}^{T} \frac{\beta_1 \eta_{t-1}}{2\left(1 - \beta_1\right)} \frac{m_{t-1,i}^2}{\sqrt{\widehat{v}_{t-1,i}}} \\[1em]
\leq \quad & \sum_{t=1}^{T} \frac{\beta_{1,t}}{2\eta_t\left(1 - \beta_1\right)} \left(x_{t,i} - x_i^*\right)^2 \sqrt{\widehat{v}_{t,i}} \\
& + \sum_{t=2}^{T} \frac{1}{2\left(1 - \beta_1\right)} \left(x_{t,i} - x_i^*\right)^2 \left(\frac{\sqrt{\widehat{v}_{t,i}}}{\eta_t} - \frac{\sqrt{\widehat{v}_{t-1,i}}}{\eta_{t-1}}\right) \\
& + \sum_{t=1}^{T} \frac{\eta_t}{2\left(1 - \beta_1\right)} \frac{\widehat{m}_{t,i}^2}{\sqrt{\widehat{v}_{t,i}}} + \sum_{t=2}^{T} \frac{\beta_1 \eta_{t-1}}{2\left(1 - \beta_1\right)} \frac{m_{t-1,i}^2}{\sqrt{\widehat{v}_{t-1,i}}} \\
& + \frac{1}{2\eta_1\left(1 - \beta_1\right)} \left(x_{1,i} - x_i^*\right)^2 \sqrt{\widehat{v}_{1,i}}
\end{aligned}$$

Using the assumptions $\left\| \mathbf{x}^t - \mathbf{x}^* \right\|_2 \leq \mathrm{D}$ and $\left\| \mathbf{x}^t - \mathbf{x}^* \right\|_\infty \leq \mathrm{D}_\infty$ and applying Lemma 1.5.2, we can rewrite this as

$$\begin{aligned}
\sum_{t=1}^{T} g_{t,i}\left(x_{t,i} - x_i^*\right) \quad = \quad & \frac{\mathrm{D}^2}{2\eta\left(1 - \beta_1\right)} \sqrt{T\widehat{v}_{t,i}} + \frac{\eta\left(1 + \beta_1\right) G_\infty}{\left(1 - \beta_1\right)\sqrt{1 - \beta_2}\left(1 - \gamma\right)} \left\| \mathbf{g}_{1:T,i} \right\|_2 \\
& + \frac{\mathrm{D}_\infty^2}{2\eta} \sum_{t=1}^{T} \frac{\beta_{1,t}}{1 - \beta_{1,t}} \sqrt{t\widehat{v}_{t,i}} \\
\implies \sum_{i=1}^{d} \sum_{t=1}^{T} g_{t,i}\left(x_{t,i} - x_i^*\right) \quad = \quad & \frac{\mathrm{D}^2}{2\eta\left(1 - \beta_1\right)} \sum_{i=1}^{d} \sqrt{T\widehat{v}_{t,i}} + \frac{\eta\left(1 + \beta_1\right) G_\infty}{\left(1 - \beta_1\right)\sqrt{1 - \beta_2}\left(1 - \gamma\right)} \sum_{i=i}^{d} \left\| \mathbf{g}_{1:T,i} \right\|_2 \\
& + \frac{\mathrm{D}_\infty^2}{2\eta} \sum_{i=1}^{d} \sum_{t=1}^{T} \frac{\beta_{1,t}}{1 - \beta_{1,t}} \sqrt{t\widehat{v}_{t,i}}
\end{aligned}$$

Note, the LHS is simply the regret term. Therefore, we have

$$\begin{aligned}
R(T) \quad = \quad & \frac{\mathrm{D}^2}{2\eta\left(1 - \beta_1\right)} \sum_{i=1}^{d} \sqrt{T\widehat{v}_{t,i}} + \frac{\eta\left(1 + \beta_1\right) G_\infty}{\left(1 - \beta_1\right)\sqrt{1 - \beta_2}\left(1 - \gamma\right)} \sum_{i=i}^{d} \left\| \mathbf{g}_{1:T,i} \right\|_2 \\
& + \frac{\mathrm{D}_\infty^2}{2\eta} \sum_{i=1}^{d} \sum_{t=1}^{T} \frac{\beta_{1,t}}{1 - \beta_{1,t}} \sqrt{t\widehat{v}_{t,i}}
\end{aligned}$$

We can observe that $\sqrt{\widehat{v}_{t,i}} \leq \| \mathbf{g}_{1:t,i} \|_2$, and

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

Therefore, we can write the final regret bound as

$$R(T) \quad \leq \quad \frac{1}{1-\beta_1} \left(\frac{\mathrm{D}^2 \sqrt{T}}{2\eta} \mathrm{Tr}\left(\widehat{\mathbf{V}}^{1/2}\right) + \frac{\eta(1+\beta_1) G_\infty}{\sqrt{1-\beta_2}(1-\gamma)} \sum_{i=1}^{d} \| \mathbf{g}_{1:T,i} \|_2 + \frac{d\,\mathrm{D}_\infty^2 \mathrm{G}_\infty \sqrt{1-\beta_2}}{2\eta(1-\lambda)^2} \right)$$

□

The above analysis can be convincing that Adam is a good optimizer, and provides very good convergence. However, Reddi et al. (2018) have shown that Adam, in fact, does not provably converge, even for simple convex problems. We discuss this non-convergence in brief in the following section.

4.2.2. Non-Convergence of Adam

Reddi et al. (2018) proved the non-convergence of Adam by an example. We discuss the same proof by example here.

Consider a setting where each subfunction f_t is a linear function (and therefore convex) with the domain $\mathcal{X} = [-1, 1]$. We write the form of each subfunction as follows

$$f_t(\mathbf{x}) \quad = \quad \begin{cases} Cx, & \text{for } t \bmod 3 = 1, \\ -x, & \text{otherwise} \end{cases}$$

where C is constant such that $C \geq 2$. Clearly, the regret (Equation 21) is minimized for the point $x^* = -1$. Suppose we start the algorithm at $x_0 = 1$ (without violating any assumptions regarding the algorithm).

Note. This is a constrained optimization problem, and therefore we need to introduce a projection step. The projection in this case is simple, since the vector space is 1-D, in which case the projection will be a Euclidean projection (Reddi et al., 2018). We denote the update before taking the projection with a hat symbol above the data point. This however does not disturb the analysis provided in the prior section.

Consider the execution of Adam algorithm with

$$\beta_1 = 0, \quad \beta_2 = \frac{1}{1+C^2} \quad \text{and} \quad \eta_t = \frac{\eta}{\sqrt{t}}$$

where $\eta \leq \sqrt{1-\beta_2}$. All the conditions of Adam are satisfied, and therefore, Adam should converge for this objective.

The proof is given using induction, where the outline is that every third update step sets the value of $\mathbf{x}_t$ to be equal to 1. More formally, $\forall\, t \in \mathbb{N} \bigcup \{0\}$, we have $\mathbf{x}_{3t} = 1$. This is satisfied for $\mathbf{x} = 0$ from our choice of the starting point. Then, we need to prove that if vx_{3t} is 1, then $\mathbf{x}_{3t+3}$ is also equal to 1.

Firstly, observe that we can write the gradients as

$$\nabla f_t(x) \quad = \quad \begin{cases} C & \text{for } t \bmod 3 = 1 \\ -1 & \text{otherwise} \end{cases}$$

Then from the (3t + 1)$^{\text{th}}$ update, we can write

$$\begin{aligned} \hat{x}_{3t+1} &= x_{3t} - \frac{\eta C}{\sqrt{(3t+1)\left(\beta_2 v_{3t} + (1-\beta_2)C^2\right)}} \\ &= 1 - \frac{\eta C}{\sqrt{(3t+1)\left(\beta_2 v_{3t} + (1-\beta_2)C^2\right)}} \\ &\geq 1 - \frac{\eta C}{\sqrt{(3t+1)(1-\beta_2)C^2}} \end{aligned}$$

Since we have $0 < \eta < \sqrt{1-\beta_2}$, we can say $0 < x_{3t+1} < 1$.

The next two updates are given as

$$\begin{aligned} \hat{x}_{3t+2} &= x_{3t+1} + \frac{\eta}{\sqrt{(3t+2)\left(\beta_2 v_{3t+1} + (1-\beta_2)\right)}} \\ \hat{x}_{3t+3} &= x_{3t+2} + \frac{\eta}{\sqrt{(3t+3)\left(\beta_2 v_{3t+2} + (1-\beta_2)\right)}} \end{aligned}$$

Therefore, $x_{3t+3} \geq x_{3t+2} > x_{3t+1} > 0$. Hence if x_{3t+2} is equal to 1, then from the projection step, x_{3t+3} has to be equal to 1. Otherwise, we can write the value of $\hat{x}_{3t+3}$ as

$$\begin{aligned} \hat{x}_{3t+3} &= \hat{x}_{3t+2} + \frac{\eta}{\sqrt{(3t+3)\left(\beta_2 v_{3t+2} + (1-\beta_2)\right)}} \\ &= x_{3t+1} + \frac{\eta}{\sqrt{(3t+2)\left(\beta_2 v_{3t+1} + (1-\beta_2)\right)}} + \frac{\eta}{\sqrt{(3t+3)\left(\beta_2 v_{3t+2} + (1-\beta_2)\right)}} \\ &\geq 1 - \frac{\eta C}{\sqrt{(3t+1)(1-\beta_2)C^2}} + \frac{\eta}{\sqrt{(3t+2)\left(\beta_2 v_{3t+1} + (1-\beta_2)\right)}} \\ &\qquad + \frac{\eta}{\sqrt{(3t+3)\left(\beta_2 v_{3t+2} + (1-\beta_2)\right)}} \end{aligned}$$

Let $T = \frac{\eta}{\sqrt{(3t+2)\left(\beta_2 v_{3t+1} + (1-\beta_2)\right)}} + \frac{\eta}{\sqrt{(3t+3)\left(\beta_2 v_{3t+2} + (1-\beta_2)\right)}}$, then

$$\begin{aligned} T &\geq \frac{\eta}{\sqrt{\beta_2 C^2 + (1-\beta_2)}} \left(\frac{1}{\sqrt{3t+2}} - \frac{1}{\sqrt{3t+3}} \right) \\ &\geq \frac{\eta}{\sqrt{\beta_2 C^2 + (1-\beta_2)}} \left(\frac{1}{\sqrt{2(3t+1)}} - \frac{1}{\sqrt{2(3t+1)}} \right) \\ &\geq \frac{\sqrt{2}\eta}{(3t+1)\sqrt{\beta_2 C^2 + (1-\beta_2)}} \end{aligned}$$

Using the fact that $\beta_2 = 1/(1+C^2)$, the last expression is equal to $\frac{\eta}{(3t+1)(1-\beta_2)}$

Therefore,

$$T \quad \geq \quad \frac{\eta}{(3t+1)(1-\beta_2)}$$

Puttin this back, we get

$$\hat{x}_{3t+3} \quad \geq \quad 1 - \frac{\eta}{(3t+1)(1-\beta_2)} + T \quad \geq \quad 1$$

Hence, our claim is true.

Reddi et al. (2018) proved an extension to this as well for more general settings, which allowed the analysis to extend to other methods such as RmsProp, Nadam, etc. This simple one dimensional convex example shows that Adam in fact does not provably converge.

Reddi et al. (2018) claim that this non-convergence is attributed to the fact the quantity

$$\Gamma_{t+1} = \left(\frac{\sqrt{\mathbf{V}_{t+1}}}{\eta_{t+1}} - \frac{\sqrt{\mathbf{V}_t}}{\eta_t} \right)$$

is not positive-definite, which is the case for other stochastic methods such as SGD and AdaGrad.

Remark. We believe that the proof is invalid because of the inconsistencies in the update equations and the analysis, where an alternate to β_1 is used, *i.e.* $\beta_{1,t}$. The updates are assumed to be using this while the main theorem is being proved, however, this creates two problems (a) the bias is not corrected anymore, (b) Lemma 1.5.2 does not directly hold when β_1 is replaced by $\beta_{1,t}$ due to the added λ^r term (for some r) in the numerator. This suggests that there are inconsistencies in the proof, and the proof is not reliable.

4.3. AdaGrad

AdaGrad (Duchi et al., 2010) is a special case of Adam, and uses the same update rules, with the values of $\beta_1 = 0$ and $\beta_2 = 1$. From Algorithm 2, we can write the functions for the update rules as

$$\phi_t\left(\mathbf{g}_1 \ldots \mathbf{g}_t\right) \quad = \quad \mathbf{g}_t \quad \text{and} \quad \psi_t\left(\mathbf{g}_1 \ldots \mathbf{g}_t\right) \quad = \quad \frac{\text{diag}\left(\sum_{i=1}^{t} \mathbf{g}_i^2\right)}{t} \qquad (\text{AdaGrad})$$

Since in this case $\beta_1 = 0$, we can set the value of $lambda = 0$ as it would not affect the updates at all. Since in this case, the inconsistencies discussed in the previous section are removed, we can use the proof to get a convergence bound for AdaGrad. In this case, we directly state the results and skip the intermediate calculations.

Theorem 1.6. For a series of convex sub-functions $\{f_t\}_{t=1}^{T}$ which have bounded gradients, *i.e.* $\forall \mathbf{x} \in \mathbb{R}$, $\|\nabla f_t(\mathbf{x})\|_2 \leq \mathrm{G}$ and $\|\nabla f_t(\mathbf{x})\|_\infty \leq \mathrm{G}_\infty$ and distance between any points $\mathbf{x}^t$ generated by AdaGrad is bounded, *i.e.* $\forall i, j \in [T]$, $\left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2 \leq \mathrm{D}$ and $\left\|\mathbf{x}_i - \mathbf{x}_j\right\|_\infty \leq \mathrm{D}_\infty$, then AdaGrad achieves the following guarantee, for all $T \geq 1$,

$$R(T) \quad \leq \quad \frac{1}{2\eta} \sum_{i=1}^{d} \sqrt{T \mathbf{g}_{1:T,i}^2} + \sum_{i=1}^{d} \sum_{t=1}^{T} \frac{\eta_t}{2} \frac{g_{t,i}^2}{\sqrt{\sum_{j=1}^{t} g_{j,i}^2}}$$

where $\beta_1, \beta_2 \in [0,1)$ such that $\gamma \triangleq \frac{\beta_1^2}{\sqrt{\beta_2}} < 1$, and we set $\eta_t = \frac{\eta}{\sqrt{t}}$ and $\beta_{1,t} = \beta_1 \lambda^{t-1}$ where $\lambda \in (0,1)$

Remark. Duchi et al. (2010) have presented great details in their paper, as well as proved convergence guarantees for various cases with AdaGrad.

5. Conclusion

We surveyed different methods of convex optimization, both deterministic and stochastic. We also looked at the non-convergence of Adam, which is a powerful and widely used optimizer, yet it does not guarantees convergence, even for simple objectives. We point some inconsistencies in the proof of convergece for Adam, and present an example that shows the non-convergence of Adam even in the simplest of settings. However, still Adam remains a powerful optimizer, due to its fast convergence, if present, and it converges for most objectives considering random initialization.

While the analyses of these algorithms is certainly a useful tool for selecting the appropriate optimization method for a given problem, in practice many of the objectives are non-convex. In that case, rigorous proofs may not be available for the convergence of these algorithms, or even if they converge. In that case, we are forced to look at empirical evidence of performance of these methods on non-convex objectives and particular cases such as saddle points, shallow local minima, and so on.

Empirical results in most cases, along with non-convex have shown that adaptive gradient methods are better at performing than their counterparts. Momentum is in particular a very useful method to traverse shallow local minima and poorly scaled objectives, common in neural network training. As the theory for neural neworks develops further, we may find newer problems to overcome and develop algorithms suitable for those tasks. Many new variants for convex optimization are arriving, however only very few of the stick in the Machine Learning Community.

References

Sebastian Ruder. An overview of gradient descent optimization algorithms. *CoRR*, abs/1609.04747, 2016. URL http://arxiv.org/abs/1609.04747.

M. Augustine Cauchy. Méthode générale pour la résolution des syst'emes d'équations simultanées. 1847.

Afroz Alam. Lecture 15 - optimization techniques ii. *Topics in Learning Theory, CS777, Indian Institute of Technology Kanpur*, 2018.

Jaivardhan Kapoor. Lecture 18 - sparse recovery ii. *Topics in Learning Theory, CS777, Indian Institute of Technology Kanpur*, 2018.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL http://arxiv.org/abs/1412.6980.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. 2018. URL https://openreview.net/forum?id=ryQu7f-RZ.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. (UCB/EECS-2010-24), Mar 2010. URL http://www2.eecs.berkeley.edu/Pubs/TechRpts/2010/EECS-2010-24.html.