Systematic Trading Strategies with Machine Learning Algorithms

The Adam Optimizer
Optional Lecture Notes

22 May 2025

Contents

1 Position of the problem

In this section, we study the optimization problem, which can be written as follows:

$$\min_{\theta \in \mathbb{R}^d} f(\theta) \quad \text{where} \quad f(\theta) := \mathbb{E}_{s \sim \mathbb{P}}[\mathcal{L}(\theta, s)], \tag{1}$$

where:

- The function f is called **the objective function**.
- The function $\mathcal{L}$ is the loss function.
- $\mathbb{P}$ is the unknown data distribution on the domain $\mathcal{S}$
- θ is the set of parameters we wish to optimize.

In the following sections, we are going to focus on the Adam optimizer. Section 2 is a brief introduction to the history of the Adam algorithm. Section 3 is a description of the algorithm. In section 4, we analyze the convergence behavior of the algorithm in the nonconvex setting.

2 Brief history of the Adam optimizer

The Adam algorithm was first introduced in 2015 [1]. The authors proposed a proof of convergence which was found to have problems. In 2018, [3] clarified the inconsistency of the previous paper and fixed the proof in the convex setting. In [2], the authors conducted the proof for the non convex case under some useful parameter settings.

In the following sections, we provide a detailed version of the proof provided in the original paper [2].

3 Description of the algorithm

The Adam algorithm defined in [2] is summarized in Algorithm 1

Algorithm 1 The Adam Optimizer

Require:

Initial parameter value: $\theta_1 \in \mathbb{R}^d$

Learning rates: $\{\eta_t\}_{t=1}^T$

Decay parameters: $0 \leq \beta_1, \beta_2 \leq 1$

Stability parameter: $\epsilon > 0$

Ensure: ϵ-First Order Stationary Point θ_{T+1}

Set $\boldsymbol{m}_0 = 0, \boldsymbol{v}_0 = 0$

for $t = 1$ **to** T **do**

Draw a batch $(s_t^i)_{i \in \mathcal{B}_t}$ from $\mathbb{P}$

Compute $\boldsymbol{g}_t = \frac{1}{|\mathcal{B}_t|} \sum\limits_{s \in \mathcal{B}_t} \nabla \mathcal{L}(\theta_t, s)$

Update $\boldsymbol{m}_t = \beta_1 \boldsymbol{m}_{t-1} + (1 - \beta_1)\, \boldsymbol{g}_t$

Update $\boldsymbol{v}_t = \boldsymbol{v}_{t-1} - (1 - \beta_2)\, (\boldsymbol{v}_{t-1} - \boldsymbol{g}_t \circ \boldsymbol{g}_t)$

Update $\theta_{t+1} = \theta_t - \eta_t \frac{\boldsymbol{m}_t}{\sqrt{\boldsymbol{v}_t} + \epsilon}$

4 The convergence behavior of the Adam optimizer

4.1 Preliminaries

We are fetching for First Order Stationary Points. We would like to prove that under some assumptions on the loss function (not necessarily convex), we can have:

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[\left\| \nabla f\left(\boldsymbol{\theta}_t\right) \right\|_2^2 \right] \leq h(T) \quad \text{with} \quad \lim_{T \to +\infty} h(T) = 0$$

Where T is the number of batches

4.2 Assumptions

- $(\mathcal{A}_1)$: We assume the loss function $\mathcal{L}$ to be L-smooth, which means that:

$$\forall \theta_1, \theta_2 \in \mathbb{R}^d \ \forall s \in \mathcal{S} \quad \|\nabla \mathcal{L}(\theta_2, s) - \nabla \mathcal{L}(\theta_1, s)\|_2 \leq L\|\theta_2 - \theta_1\|_2 \tag{2}$$

- $(\mathcal{A}_2)$: We assume the loss function $\mathcal{L}$ to have bounded gradient: i.e,

$$\exists G \in \mathbb{R}_+ \ \forall \theta \in \mathbb{R}^d \ \forall s \in \mathcal{S} \quad \|\nabla \mathcal{L}(\theta, s)\|_2 \leq G \tag{3}$$

- $(\mathcal{A}_3)$: We assume the variance of the loss function $\mathcal{L}$ to be bounded: i.e,

$$\forall \theta \in \mathbb{R}^d \quad \mathbb{E}\left[\|\nabla \mathcal{L}(\boldsymbol{\theta}; \xi) - \nabla \mathcal{L}(\boldsymbol{\theta})\|_2^2 | \mathcal{F}_t\right] \leq \sigma^2 \tag{4}$$

  where the sigma-algebra $\mathcal{F}_t$ represents *the information known at time t*

4.3 The convergence theorem

Theorem 4.3.1 (Convergence of the Adam Algorithm)**.** *Let $\eta_t = \eta$ for all $t \in [T]$. Furthermore, assume that ϵ, β_2 and η are chosen such that the following conditions are satisfied:*

$$\eta \leq \frac{2G\sqrt{1-\beta_2}}{L} \tag{5}$$

$$1 - \beta_2 \leq \frac{\epsilon^4}{16G^2(G+\epsilon)^2} \tag{6}$$

Then, for $(\theta_t)_t$ generated using ADAM (Algorithm 1), we have the following inequality:

1. *If the batch size b_t is fixed (i.e, $b_t = b_0$ for all t). Then,*

$$\exists c_1, c_2 \in \mathbb{R}_+ \quad \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla f(\theta_t)\|_2^2\right] \leq \frac{c_1}{T} + c_2 \tag{7}$$

2. *If the batch size $b_t = b_0 T$ for all t. Then,*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla f(\boldsymbol{\theta}_t)\|_2^2\right] = O\left(\frac{1}{T}\right) \tag{8}$$

3. *If the batch size in linear in time (i.e, $b_t = b_0 t$ for all t). Then,*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla f(\boldsymbol{\theta}_t)\|_2^2\right] = O\left(\frac{\ln(T)}{T}\right) \tag{9}$$

4. *If the batch size is of the form $b_t = \lceil b_0 t^\gamma \rceil$ for all t (with $0 < \gamma < 1$). Then,*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla f\left(\boldsymbol{\theta}_t\right)\right\|_2^2\right] = O\left(\frac{1}{T^\gamma}\right) \tag{10}$$

Proof. We would like to understand the change in function value between two successive iterations of the algorithm 1. In the whole proof, we will consider $\beta_1 = 0$

1. **Showing that the objective function f is L-smooth**

   For all $\theta_1, \theta_2 \in \mathbb{R}^d$

$$\begin{aligned}
\left\|\nabla f\left(\boldsymbol{\theta}_1\right) - \nabla f\left(\boldsymbol{\theta}_2\right)\right\|_2 &= \left\|\nabla \mathbb{E}_{s\sim\mathbb{P}}\left[\mathcal{L}\left(\boldsymbol{\theta}_1; s\right)\right] - \nabla \mathbb{E}_{s\sim\mathbb{P}}\left[\mathcal{L}\left(\boldsymbol{\theta}_2; s\right)\right]\right\|_2 \quad \text{(from the definition 1)}\\
&= \left\|\mathbb{E}_{s\sim\mathbb{P}}\left[\nabla\mathcal{L}\left(\boldsymbol{\theta}_1; s\right)\right] - \mathbb{E}_{s\sim\mathbb{P}}\left[\nabla\mathcal{L}\left(\boldsymbol{\theta}_2; s\right)\right]\right\|_2\\
&= \left\|\mathbb{E}_{s\sim\mathbb{P}}\left[\nabla\mathcal{L}\left(\boldsymbol{\theta}_1; s\right) - \nabla\mathcal{L}\left(\boldsymbol{\theta}_2; s\right)\right]\right\|_2\\
&\le \mathbb{E}_{s\sim\mathbb{P}}\left[\left\|\nabla\mathcal{L}\left(\boldsymbol{\theta}_1; s\right) - \nabla\mathcal{L}\left(\boldsymbol{\theta}_2; s\right)\right\|_2\right]\\
&\le \mathbb{E}_{s\sim\mathbb{P}}\left[L\left\|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\right\|_2\right] \quad \text{(from the assumption 2)}\\
&= L\|\theta_2 - \theta_1\|_2
\end{aligned}$$

   Therefore, f is L-smooth

2. **Deducing the change in the objective value between two successive iterations**.

   Let us consider $t \in [T]$. As f is L-smooth, we can deduce that:

$$f\left(\boldsymbol{\theta}_{t+1}\right) \le f\left(\boldsymbol{\theta}_t\right) + \nabla^\top f\left(\boldsymbol{\theta}_t\right)\left(\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t\right) + \frac{L}{2}\left\|\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t\right\|_2^2 \tag{11}$$

   From the update equations in algorithm 1 we have:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t\frac{\mathbf{g}_t}{\left(\sqrt{\mathbf{v}_t} + \epsilon\right)}$$

   Which can be expressed component-wise as follows:

$$\forall i \in [d] \quad \boldsymbol{\theta}_{i,t+1} = \boldsymbol{\theta}_{i,t} - \eta_t\frac{\mathbf{g}_{i,t}}{\left(\sqrt{v_{i,t}} + \epsilon\right)} \tag{12}$$

   From 11 and 12, we deduce the following inequality:

$$f\left(\boldsymbol{\theta}_{t+1}\right) \le f\left(\boldsymbol{\theta}_t\right) - \eta_t\sum_{i=1}^{d}\left(\left[\nabla f\left(\boldsymbol{\theta}_t\right)\right]_i \times \frac{\mathbf{g}_{i,t}}{\sqrt{\mathbf{v}_{i,t}} + \epsilon}\right) + \frac{L\eta_t^2}{2}\sum_{i=1}^{d}\frac{\mathbf{g}_{i,t}^2}{\left(\sqrt{\mathbf{v}_{i,t}} + \epsilon\right)^2}$$

   Let us introduce the following notations for each time t':

b_t' be the size of $\mathcal{B}_{t'}$.

The sigma-algebra $\mathcal{F}_{t'}$ represents *the information known at time t'*.

Consequently:

$$\boxed{\begin{aligned}\mathbb{E}\left[f\left(\boldsymbol{\theta}_{t+1}\right) \mid \mathcal{F}_t\right] \leq f\left(\boldsymbol{\theta}_t\right) & \underbrace{-\eta_t \sum_{i=1}^{d}\left(\left[\nabla f\left(\boldsymbol{\theta}_t\right)\right]_i \times \mathbb{E}\left[\frac{\mathbf{g}_{i,t}}{\sqrt{\mathbf{v}_{i,t}}+\epsilon} \mid \mathcal{F}_t\right]\right)}_{\text{(a)}} \\ & + \underbrace{\frac{L\eta_t^2}{2}\sum_{i=1}^{d}\mathbb{E}\left[\frac{\mathbf{g}_{i,t}^2}{\left(\sqrt{\mathbf{v}_{i,t}}+\epsilon\right)^2} \mid \mathcal{F}_t\right]}_{(b)}\end{aligned}} \tag{13}$$

3. **Bounding the first term (a) in 13**

We have:

$$\begin{aligned}\mathbb{E}\left[\frac{\mathbf{g}_{i,t}}{\sqrt{\mathbf{v}_{i,t}}+\epsilon} \mid \mathcal{F}_t\right] &= \mathbb{E}\left[\frac{\boldsymbol{g}_{i,t}}{\sqrt{\boldsymbol{v}_{i,t}}+\epsilon} - \frac{\boldsymbol{g}_{i,t}}{\sqrt{\beta_2\boldsymbol{v}_{i,t-1}}+\epsilon} + \frac{\boldsymbol{g}_{i,t}}{\sqrt{\beta_2\boldsymbol{v}_{i,t-1}}+\epsilon} \mid \mathcal{F}_t\right] \\ &= \mathbb{E}\left[\frac{\mathbf{g}_{i,t}}{\sqrt{\mathbf{v}_{i,t}}+\epsilon} - \frac{\mathbf{g}_{i,t}}{\sqrt{\beta_2\mathbf{v}_{i,t-1}}+\epsilon} \mid \mathcal{F}_t\right] + \underbrace{\mathbb{E}\left[\frac{\mathbf{g}_{i,t}}{\sqrt{\beta_2\mathbf{v}_{i,t-1}}+\epsilon} \mid \mathcal{F}_t\right]} \\ &= \mathbb{E}\left[\frac{\mathbf{g}_{i,t}}{\sqrt{\mathbf{v}_{i,t}}+\epsilon} - \frac{\mathbf{g}_{i,t}}{\sqrt{\beta_2\mathbf{v}_{i,t-1}}+\epsilon} \mid \mathcal{F}_t\right] + \frac{[\nabla f(\boldsymbol{\theta})]_i}{\sqrt{\beta_2\mathbf{v}_{i,t-1}}+\epsilon}\end{aligned}$$

Which enables us to rewrite (a) defined in 13 as follows:

$$\begin{aligned}(a) &= -\eta_t \sum_{i=1}^{d}\left(\left[\nabla f\left(\boldsymbol{\theta}_t\right)\right]_i \times \left[\frac{\left[\nabla f\left(\boldsymbol{\theta}_t\right)\right]_i}{\sqrt{\beta_2\boldsymbol{v}_{i,t-1}}+\epsilon} + \mathbb{E}\left[\frac{\boldsymbol{g}_{i,t}}{\sqrt{\boldsymbol{v}_{i,t}}+\epsilon} - \frac{\boldsymbol{g}_{i,t}}{\sqrt{\beta_2\boldsymbol{v}_{i,t-1}}+\epsilon} \mid \mathcal{F}_t\right]\right]\right) \\ &= -\eta_t \sum_{i=1}^{d}\frac{\left[\nabla f\left(\boldsymbol{\theta}_t\right)\right]_i^2}{\sqrt{\beta_2\boldsymbol{v}_{i,t-1}}+\epsilon} \underbrace{-\eta_t \sum_{i=1}^{d}\left[\nabla f\left(\boldsymbol{\theta}_t\right)\right]_i \times \mathbb{E}\left[\frac{\boldsymbol{g}_{i,t}}{\sqrt{\boldsymbol{v}_{i,t}}+\epsilon} - \frac{\boldsymbol{g}_{i,t}}{\sqrt{\beta_2\boldsymbol{v}_{i,t-1}}+\epsilon} \mid \mathcal{F}_t\right]}_{(a_1)}\end{aligned} \tag{14}$$

Let us bound the term (a_1) in 14:

$$\begin{aligned}
-\eta_t \sum_{i=1}^{d} \left[\nabla f\left(\boldsymbol{\theta}_t\right)\right]_i \times \mathbb{E}\left[\frac{\boldsymbol{g}_{i,t}}{\sqrt{\boldsymbol{v}_{i,t}}+\epsilon} - \frac{\boldsymbol{g}_{i,t}}{\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}}+\epsilon} \mid \mathcal{F}_t\right] \\
\leq \left|\eta_t \sum_{i=1}^{d} \left[\nabla f\left(\boldsymbol{\theta}_t\right)\right]_i \times \mathbb{E}\left[\frac{\boldsymbol{g}_{i,t}}{\sqrt{\boldsymbol{v}_{i,t}}+\epsilon} - \frac{\boldsymbol{g}_{i,t}}{\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}}+\epsilon} \mid \mathcal{F}_t\right]\right| \\
\leq \eta_t \sum_{i=1}^{d} \left|\left[\nabla f\left(\boldsymbol{\theta}_t\right)\right]_i\right| \left|\times \mathbb{E}\left[\frac{\boldsymbol{g}_{i,t}}{\sqrt{\boldsymbol{v}_{i,t}}+\epsilon} - \frac{\boldsymbol{g}_{i,t}}{\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}}+\epsilon} \mid \mathcal{F}_t\right]\right| \\
\leq \eta_t \sum_{i=1}^{d} \left|\left[\nabla f\left(\boldsymbol{\theta}_t\right)\right]_i\right| \times \mathbb{E}\left[\underbrace{\left|\frac{\boldsymbol{g}_{i,t}}{\sqrt{\boldsymbol{v}_{i,t}}+\epsilon} - \frac{\boldsymbol{g}_{i,t}}{\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}}+\epsilon}\right|}_{(a_2)} \mid \mathcal{F}_t\right]
\end{aligned} \tag{15}$$

By using the update rule $\mathbf{v}_{i,t} = \beta_2 \mathbf{v}_{i,t-1} + (1-\beta_2)\,\mathbf{g}_{i,t}^2$ from algorithm 1, we can bound the term (a_2) in 15:

$$\begin{aligned}
(a_2) &= \left|\boldsymbol{g}_{i,t}\right| \left|\frac{1}{\sqrt{\boldsymbol{v}_{i,t}}+\epsilon} - \frac{1}{\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}}+\epsilon}\right| \\
&= \frac{\left|\boldsymbol{g}_{i,t}\right|}{\left(\sqrt{\boldsymbol{v}_{i,t}}+\epsilon\right)\left(\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}}+\epsilon\right)} \left|\sqrt{\boldsymbol{v}_{i,t}} - \sqrt{\beta_2 \boldsymbol{v}_{i,t-1}}\right| \\
&= \frac{\left|\boldsymbol{g}_{i,t}\right|}{\left(\sqrt{\boldsymbol{v}_{i,t}}+\epsilon\right)\left(\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}}+\epsilon\right)} \frac{\left|\boldsymbol{v}_{i,t} - \beta_2 \boldsymbol{v}_{i,t-1}\right|}{\sqrt{\boldsymbol{v}_{i,t}} + \sqrt{\beta_2 \boldsymbol{v}_{i,t-1}}} \\
&= \frac{\left|\boldsymbol{g}_{i,t}\right|}{\left(\sqrt{\boldsymbol{v}_{i,t}}+\epsilon\right)\left(\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}}+\epsilon\right)} \frac{(1-\beta_2)\,\mathbf{g}_{i,t}^2}{\sqrt{\beta_2 \mathbf{v}_{i,t-1} + (1-\beta_2)\,\mathbf{g}_{i,t}^2} + \sqrt{\beta_2 \mathbf{v}_{i,t-1}}} \quad \text{(by using the update rule)} \\
&\leq \frac{\left|\boldsymbol{g}_{i,t}\right|}{\left(\sqrt{\boldsymbol{v}_{i,t}}+\epsilon\right)\left(\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}}+\epsilon\right)} \frac{(1-\beta_2)\,\mathbf{g}_{i,t}^2}{\sqrt{\beta_2 \mathbf{v}_{i,t-1} + (1-\beta_2)\,\mathbf{g}_{i,t}^2}} \quad \left(\text{since } \sqrt{\beta_2 \mathbf{v}_{i,t-1}} \geq 0\right) \\
&\leq \frac{\left|\boldsymbol{g}_{i,t}\right|}{\epsilon\left(\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}}+\epsilon\right)} \frac{(1-\beta_2)\,\mathbf{g}_{i,t}^2}{\sqrt{(1-\beta_2)\,\mathbf{g}_{i,t}^2}} \quad \left(\text{since } \sqrt{v_{i,t}} \geq 0 \text{ and } \beta_2 \mathbf{v}_{i,t-1} \geq 0\right) \\
&= \frac{\sqrt{1-\beta_2}\mathbf{g}_{i,t}^2}{\epsilon\left(\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}}+\epsilon\right)}
\end{aligned} \tag{16}$$

From 14, 15 and 16, we deduce the following inequality:

$$(a_1) \leq \eta_t \sum_{i=1}^{d} \left(\mid [\nabla f(\boldsymbol{\theta}_t)]_i \mid \frac{\sqrt{1-\beta_2}}{\epsilon} \mathbb{E}\left[\frac{\mathbf{g}_{i,t}^2}{\sqrt{\beta_2 \mathbf{v}_{i,t-1}} + \epsilon} \mid \mathcal{F}_t \right] \right) \tag{17}$$

Therefore, we deduce a bound for (a) from 14 and 17:

$$(a) \leq -\eta_t \sum_{i=1}^{d} \frac{[\nabla f(\boldsymbol{\theta}_t)]_i^2}{\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}} + \epsilon} + \eta_t \sum_{i=1}^{d} \left(\mid [\nabla f(\boldsymbol{\theta}_t)]_i \mid \frac{\sqrt{1-\beta_2}}{\epsilon} \mathbb{E}\left[\frac{\mathbf{g}_{i,t}^2}{\sqrt{\beta_2 \mathbf{v}_{i,t-1}} + \epsilon} \mid \mathcal{F}_t \right] \right) \tag{18}$$

By using the assumption 3, we can also bound the term $\mid [\nabla f(\boldsymbol{\theta}_t)]_i$ for all $i \in [d]$.

Indeed,

$$\begin{aligned} \forall i \in [d] \quad |[\nabla f(\boldsymbol{\theta}_t)]_i| &\leq \|\nabla f(\boldsymbol{\theta_t})\|_2 \\ &:= \|\mathbb{E}_{s\sim\mathbb{P}}[\mathcal{L}(\theta_t, s)]\|_2 \\ &\leq \mathbb{E}_{s\sim\mathbb{P}}[\|\nabla \mathcal{L}(\boldsymbol{\theta_t}; s)\|] \\ &\leq G \quad \text{(from assumption3)} \end{aligned}$$

So,

$$\forall i \in [d] \quad |[\nabla f(\boldsymbol{\theta}_t)]_i| \leq G \tag{19}$$

From 18 and 19 we deduce:

$$\boxed{(a) \leq -\eta_t \sum_{i=1}^{d} \frac{[\nabla f(\boldsymbol{\theta}_t)]_i^2}{\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}} + \epsilon} + \frac{\eta_t G \sqrt{1-\beta_2}}{\epsilon} \sum_{i=1}^{d} \mathbb{E}\left[\frac{\mathbf{g}_{i,t}^2}{\sqrt{\beta_2 \mathbf{v}_{i,t-1}} + \epsilon} \mid \mathcal{F}_t \right]} \tag{20}$$

4. **Bounding the second term (b) in 13**

   By using the update rule $\mathbf{v}_{i,t} = \beta_2 \mathbf{v}_{i,t-1} + (1-\beta_2)\, \mathbf{g}_{i,t}^2$ from algorithm 1 in the expression (b), we get:

$$
\begin{aligned}
(b) &:= \frac{L\eta_t^2}{2} \sum_{i=1}^{d} \mathbb{E}\left[\frac{\mathbf{g}_{i,t}^2}{\left(\sqrt{\mathbf{v}_{i,t}} + \epsilon\right)^2} \mid \mathcal{F}_t\right] \\
&= \frac{L\eta_t^2}{2} \sum_{i=1}^{d} \mathbb{E}\left[\frac{\boldsymbol{g}_{i,t}^2}{\left(\sqrt{\beta_2 \boldsymbol{v}_{i,t-1} + (1-\beta_2)\, \boldsymbol{g}_{i,t}^2} + \epsilon\right)^2} \mid \mathcal{F}_t\right] \\
&\le \frac{L\eta_t^2}{2} \sum_{i=1}^{d} \mathbb{E}\left[\frac{\boldsymbol{g}_{i,t}^2}{\left(\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}} + \epsilon\right)^2} \mid \mathcal{F}_t\right] \quad (\text{since } (1-\beta_2)\, \boldsymbol{g}_{i,t}^2 \ge 0) \\
&\le \frac{L\eta_t^2}{2\epsilon} \sum_{i=1}^{d} \mathbb{E}\left[\frac{g_{i,t}^2}{\sqrt{\beta_2 \mathbf{v}_{i,t-1}} + \epsilon} \mid \mathcal{F}_t\right] \quad (\text{since} \sqrt{\beta_2 \mathbf{v}_{i,t-1}} \ge 0)
\end{aligned}
$$

So,

$$
\boxed{(b) \le \frac{L\eta_t^2}{2\epsilon} \sum_{i=1}^{d} \mathbb{E}\left[\frac{g_{i,t}^2}{\sqrt{\beta_2 \mathbf{v}_{i,t-1}} + \epsilon} \mid \mathcal{F}_t\right]} \tag{21}
$$

5. **Combining the upper bounds on (a) and (b) defined in 13**

   By combining the upper bounds 20 and 21, we get the following inequality:

$$
\begin{aligned}
\mathbb{E}\left[f(\boldsymbol{\theta}_{t+1}) \mid \mathcal{F}_t\right] \le f(\boldsymbol{\theta}_t) & \underbrace{-\eta_t \sum_{i=1}^{d} \frac{\left[\nabla f(\boldsymbol{\theta}_t)\right]_i^2}{\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}} + \epsilon}}_{(c)} + \underbrace{\frac{\eta_t G\sqrt{1-\beta_2}}{\epsilon} \sum_{i=1}^{d} \mathbb{E}\left[\frac{\mathbf{g}_{i,t}^2}{\sqrt{\beta_2 \mathbf{v}_{i,t-1}} + \epsilon} \mid \mathcal{F}_t\right]}_{(d)} \\
& + \underbrace{\frac{L\eta_t^2}{2\epsilon} \sum_{i=1}^{d} \mathbb{E}\left[\frac{g_{i,t}^2}{\sqrt{\beta_2 \mathbf{v}_{i,t-1}} + \epsilon} \mid \mathcal{F}_t\right]}_{(e)}
\end{aligned} \tag{22}
$$

- Bounding the sum of (d) and (e) defined in 22:

  We have:

$$
\begin{aligned}
(d) + (e) &= \left(\frac{\eta_t G\sqrt{1-\beta_2}}{\epsilon} + \frac{L\eta_t^2}{2\epsilon}\right) \sum_{i=1}^{d} \mathbb{E}\left[\frac{\mathbf{g}_{i,t}^2}{\sqrt{\beta_2 \mathbf{v}_{i,t-1}} + \epsilon} \mid \mathcal{F}_t\right] \\
&\le \frac{1}{\epsilon}\left(\frac{\eta_t G\sqrt{1-\beta_2}}{\epsilon} + \frac{L\eta_t^2}{2\epsilon}\right) \sum_{i=1}^{d} \mathbb{E}\left[\mathbf{g}_{i,t}^2 \mid \mathcal{F}_t\right] \quad (\text{since } \sqrt{\beta_2 \mathbf{v}_{i,t-1}} \ge 0)
\end{aligned} \tag{23}
$$

- Bounding the expression (c) defined in 22:

  To that end, we first need to prove that $\forall i \in [d]\ \forall t' \in [T]\ v_{i,t'} \leq G^2$. We can do it by induction on t'.

  - It's true for $t' = 0$
  - Let's consider $t' \in [T]$ such that $\forall i \in [d]\ v_{i,t'-1} \leq G^2$. We have:

$$
\begin{aligned}
\forall i \in [d]\ v_{i,t'} &= \beta_2 \mathbf{v}_{i,t-1} + (1-\beta_2)\, \mathbf{g}_{i,t}^2 \\
&\leq \beta_2 G^2 + (1-\beta_2)\, \mathbf{g}_{i,t}^2 \quad \text{(by induction hypothesis)} \\
&\leq \beta_2 G^2 + (1-\beta_2)\, \|\mathbf{g}_t\|_2^2 \\
&= \beta_2 G^2 + (1-\beta_2) \left\| \frac{1}{b_t} \sum_{s \in \mathcal{B}_t} \nabla \mathcal{L}(\theta_t, s) \right\|_2^2 \quad \text{(by definition of } g_t\text{)} \\
&\leq \beta_2 G^2 + (1-\beta_2)\, \frac{1}{b_t^2} \sum_{s \in \mathcal{B}_t} ||\nabla \mathcal{L}(\theta_t, s)||_2^2 \\
&\leq \beta_2 G^2 + (1-\beta_2)\, \frac{1}{b_t^2} \sum_{s \in \mathcal{B}_t} G^2 \quad \text{(by assumption 3)} \\
&= \beta_2 G^2 + (1-\beta_2)\, \frac{1}{b_t^2} b_t G^2 \\
&= \beta_2 G^2 + \frac{(1-\beta_2)}{b_t} G^2 \\
&\leq \beta_2 G^2 + (1-\beta_2)\, G^2 \quad (\text{since} b_t \geq 1) \\
&= G^2
\end{aligned}
$$

  We conclude by induction that,

$$
\forall i \in [d]\ \forall t' \in [T]\ v_{i,t'} \leq G^2 \tag{24}
$$

  Consequently,

$$
\begin{aligned}
(c) := -\eta_t \sum_{i=1}^{d} \frac{\left[\nabla f\left(\boldsymbol{\theta}_t\right)\right]_i^2}{\sqrt{\beta_2 \boldsymbol{v}_{i,t-1}} + \epsilon} \\
&\leq -\frac{\eta_t}{\sqrt{\beta_2} G + \epsilon} \sum_{i=1}^{d} \left[\nabla f\left(\boldsymbol{\theta}_t\right)\right]_i^2 \quad \text{(by using 24)} \\
&= -\frac{\eta_t}{\sqrt{\beta_2} G + \epsilon} \left\|\nabla f\left(\boldsymbol{\theta}_t\right)\right\|_2^2
\end{aligned} \tag{25}
$$

- Combining the results:

  By using the inequalities 23 and 25, the upper bound in 22 becomes:

$$\boxed{\begin{aligned}\mathbb{E}\left[f\left(\boldsymbol{\theta}_{t+1}\right) \mid \mathcal{F}_t\right] \leq f\left(\boldsymbol{\theta}_t\right) &- \frac{\eta_t}{\sqrt{\beta_2}G+\epsilon}\left\|\nabla f\left(\boldsymbol{\theta}_t\right)\right\|_2^2 \\ &+ \frac{1}{\epsilon}\left(\frac{\eta_t G\sqrt{1-\beta_2}}{\epsilon}+\frac{L\eta_t^2}{2\epsilon}\right)\underbrace{\mathbb{E}\left[\left\|\mathbf{g}_t\right\|_2^2 \mid \mathcal{F}_t\right]}_{(f)}\end{aligned}} \tag{26}$$

6. **Bounding the last term (f)**

Let us introduce the following notations:

$$\xi_t := \frac{1}{b_t}\sum_{s\in\mathcal{B}_t}\left(\nabla\mathcal{L}(\theta_t, s) - \nabla f(\theta_t)\right) \tag{27}$$

$$\forall s \in \mathcal{B}_t \quad Y_s := \nabla\mathcal{L}(\theta_t, s) - \nabla f(\theta_t) \tag{28}$$

$$Y := \sum_{s\in\mathcal{B}_t} Y_s \tag{29}$$

Then,

$$\xi_t := \frac{1}{b_t}Y \tag{30}$$

Consequently,

$$
\begin{aligned}
(f) &:= \mathbb{E}\left[\|\mathbf{g}_t\|_2^2 \mid \mathcal{F}_t\right] \\
&= \mathbb{E}\left[\left\|\frac{1}{b_t}\sum_{s\in\mathcal{B}_t}\nabla\mathcal{L}(\theta_t, s)\right\|_2^2 \mid \mathcal{F}_t\right] \quad \text{(by definition)} \\
&= \mathbb{E}\left[\left\|\frac{1}{b_t}\sum_{s\in\mathcal{B}_t}\left(\nabla\mathcal{L}(\theta_t, s) - \nabla f(\theta_t) + \nabla f(\theta_t)\right)\right\|_2^2 \mid \mathcal{F}_t\right] \\
&= \mathbb{E}\left[\left\|\left(\frac{1}{b_t}\sum_{s\in\mathcal{B}_t}\left(\nabla\mathcal{L}(\theta_t, s) - \nabla f(\theta_t)\right)\right) + \nabla f(\theta_t)\right\|_2^2 \mid \mathcal{F}_t\right] \\
&= \mathbb{E}\left[\|\xi_t + \nabla f(\theta_t)\|_2^2 \mid \mathcal{F}_t\right] \quad \text{(by definition 27)} \\
&= \mathbb{E}\left[\left(\xi_t + \nabla f(\theta_t)\right)^\top \left(\xi_t + \nabla f(\theta_t)\right) \mid \mathcal{F}_t\right] \\
&= \mathbb{E}\left[\|\xi_t\|_2^2 \mid \mathcal{F}_t\right] + \underbrace{\mathbb{E}\left[\xi_t \mid \mathcal{F}_t\right]^\top}_{=0} \nabla f(\theta_t) + \nabla f(\theta_t)^\top \underbrace{\mathbb{E}\left[\xi_t \mid \mathcal{F}_t\right]}_{=0} + \|\nabla f(\theta_t)\|_2^2 \\
&= \frac{1}{b_t^2}\mathbb{E}\left[\|Y\|_2^2 \mid \mathcal{F}_t\right] + \|\nabla f(\theta_t)\|_2^2 \quad \text{(using 30)} \\
&= \frac{1}{b_t^2}\mathbb{E}\left[\left\|\left(\sum_{s\in\mathcal{B}_t} Y_s\right)^\top \left(\sum_{s'\in\mathcal{B}_t} Y_{s'}\right)\right\|_2^2 \mid \mathcal{F}_t\right] + \|\nabla f(\theta_t)\|_2^2 \quad \text{(using 29)} \\
&= \frac{1}{b_t^2}\sum_{s,s'\in\mathcal{B}_t}\mathbb{E}\left[Y_s^\top Y_{s'} \mid \mathcal{F}_t\right] + \|\nabla f(\theta_t)\|_2^2 \\
&= \frac{1}{b_t^2}\sum_{\substack{s,s'\in\mathcal{B}_t\\ s\neq s'}}\mathbb{E}\left[Y_s^\top Y_{s'} \mid \mathcal{F}_t\right] + \frac{1}{b_t^2}\sum_{\substack{s,s'\in\mathcal{B}_t\\ s=s'}}\mathbb{E}\left[Y_s^\top Y_{s'} \mid \mathcal{F}_t\right] + \|\nabla f(\theta_t)\|_2^2 \\
&= \frac{1}{b_t^2}\sum_{\substack{s,s'\in\mathcal{B}_t\\ s\neq s'}}\mathbb{E}\left[Y_s \mid \mathcal{F}_t\right]^\top \underbrace{\mathbb{E}\left[Y_{s'} \mid \mathcal{F}_t\right]}_{=0} + \frac{1}{b_t^2}\sum_{s\in\mathcal{B}_t}\mathbb{E}\left[\|Y_s\|_2^2 \mid \mathcal{F}_t\right] + \|\nabla f(\theta_t)\|_2^2 \quad (\text{since} Y_s Y_{s'} \mid \mathcal{F}_t) \\
&= \frac{1}{b_t^2}\sum_{s\in\mathcal{B}_t}\mathbb{E}\left[\|\nabla\mathcal{L}(\theta_t, s) - \nabla f(\theta_t)\|_2^2 \mid \mathcal{F}_t\right] + \|\nabla f(\theta_t)\|_2^2 \quad \text{(by definition 28)} \\
&\leq \frac{1}{b_t^2}\sum_{s\in\mathcal{B}_t}\sigma^2 + \|\nabla f(\theta_t)\|_2^2 \quad \text{(using the assumption 4)} \\
&\leq \frac{\sigma^2}{b_t} + \|\nabla f(\theta_t)\|_2^2
\end{aligned}
$$

We conclude the following bound on the (f) term defined in 26:

$$(f) \leq \frac{\sigma^2}{b_t} + \|\nabla f(\theta_t)\|_2^2 \tag{31}$$

By using the bound 31, the inequality 26 becomes:

$$\mathbb{E}\left[f\left(\boldsymbol{\theta}_{t+1}\right) \mid \mathcal{F}_t\right] \leq f\left(\boldsymbol{\theta}_t\right) - \frac{\eta_t}{\sqrt{\beta_2} G+\epsilon}\left\|\nabla f\left(\boldsymbol{\theta}_t\right)\right\|_2^2 + \frac{1}{\epsilon}\left(\frac{\eta_t G \sqrt{1-\beta_2}}{\epsilon} + \frac{L \eta_t^2}{2 \epsilon}\right)\left(\frac{\sigma^2}{b_t} + \|\nabla f(\theta_t)\|_2^2\right)$$

Which results in the following inequality:

$$\boxed{\begin{aligned}\mathbb{E}\left[f\left(\boldsymbol{\theta}_{t+1}\right) \mid \mathcal{F}_t\right] \leq f\left(\boldsymbol{\theta}_t\right) + \left\|\nabla f\left(\boldsymbol{\theta}_t\right)\right\|_2^2 & \underbrace{\left(\frac{-\eta_t}{\sqrt{\beta_2} G+\epsilon} + \frac{1}{\epsilon}\left(\frac{\eta_t G \sqrt{1-\beta_2}}{\epsilon} + \frac{L \eta_t^2}{2 \epsilon}\right)\right)}_{(i)} \\ & + \frac{\eta_t \sigma^2}{\epsilon b_t} \underbrace{\left(\frac{G \sqrt{1-\beta_2}}{\epsilon} + \frac{L \eta_t}{2 \epsilon}\right)}_{(ii)}\end{aligned}} \tag{32}$$

7. **Incorporating the assumptions on the hyperparameters**

   The final step of the proof is to use the conditions on the hyperparameters to bound (i) and (ii).

   - Using the choice of η

     Based on the condition 6, the learning rate is chosen to be fixed such that:

     $$\eta \leq \frac{2 G \sqrt{1-\beta_2}}{L}$$

     Therefore,

     $$\frac{L \eta}{2 \epsilon} \leq \frac{G \sqrt{1-\beta_2}}{\epsilon} \tag{33}$$

     We can then bound the first term (i) as follows:

     $$\begin{aligned}(i) &:= -\eta\left(\frac{1}{\sqrt{\beta_2} G+\epsilon} - \frac{1}{\epsilon}\left(\frac{G \sqrt{1-\beta_2}}{\epsilon} + \frac{L \eta_t}{2 \epsilon}\right)\right) \\ & \leq -\eta\left(\frac{1}{\sqrt{\beta_2} G+\epsilon} - \frac{1}{\epsilon}\left(\frac{G \sqrt{1-\beta_2}}{\epsilon} + \frac{G \sqrt{1-\beta_2}}{\epsilon}\right)\right) \quad \text{(using 33)} \\ & = -\eta\left(\frac{1}{\sqrt{\beta_2} G+\epsilon} - \underbrace{\frac{2 G \sqrt{1-\beta_2}}{\epsilon^2}}_{(iii)}\right)\end{aligned} \tag{34}$$

We can use the inequality 33 to bound the second term (ii) as follows:

$$
\begin{aligned}
(ii) &:= \frac{\eta_t \sigma^2}{\epsilon b_t}\left(\frac{G\sqrt{1-\beta_2}}{\epsilon} + \frac{L\eta_t}{2\epsilon}\right) \\
&\leq \frac{\eta_t \sigma^2}{\epsilon b_t}\left(\frac{G\sqrt{1-\beta_2}}{\epsilon} + \frac{G\sqrt{1-\beta_2}}{\epsilon}\right) \quad \text{(using 33)} \\
&= \frac{2\eta\sigma^2 G\sqrt{1-\beta_2}}{\epsilon^2 b_t}
\end{aligned} \tag{35}
$$

- Using the choice of β_2 to bound (iii)
  By using condition 6, we can bound (iii):

$$
\begin{aligned}
\frac{2G\sqrt{1-\beta_2}}{\epsilon^2} &= \sqrt{\frac{4G^2(1-\beta_2)}{\epsilon^4}} \\
&\leq \sqrt{\frac{4G^2}{\epsilon^4}\frac{\epsilon^4}{16G^2(G+\epsilon)^2}} \quad \text{(using 6)} \\
&= \frac{1}{2}\frac{1}{G+\epsilon} \\
&\leq \frac{1}{2}\frac{1}{\sqrt{\beta_2}G+\epsilon} \quad (\text{since } \beta_2 \leq 1)
\end{aligned} \tag{36}
$$

- Deducing a new bound for (i)
  From 36, we conclude that:

$$
\frac{1}{\sqrt{\beta_2}G+\epsilon} - \frac{2G\sqrt{1-\beta_2}}{\epsilon^2} \geq \frac{1}{2}\frac{1}{\sqrt{\beta_2}G+\epsilon}
$$

  The inequality 34 becomes

$$
(i) \leq -\frac{\eta}{2(\sqrt{\beta_2}G+\epsilon)} \tag{37}
$$

Finally, by using 37 and 35, we get the following update to the inequality 32:

$$
\boxed{\mathbb{E}\left[f\left(\boldsymbol{\theta}_{t+1}\right) \mid \mathcal{F}_t\right] \leq f\left(\boldsymbol{\theta}_t\right) - \frac{\eta}{2(\sqrt{\beta_2}G+\epsilon)}\left\|\nabla \mathcal{L}\left(\boldsymbol{\theta}_t\right)\right\|_2^2 + \frac{2\eta\sigma^2 G\sqrt{1-\beta_2}}{\epsilon^2 b_t}} \tag{38}
$$

8. **Concluding according to the batch size**

   Notations:

Let us define the following constants:

$$\Delta = \frac{\eta}{2(\sqrt{\beta_2}G + \epsilon)}$$
$$\alpha = \frac{2\eta\sigma^2 G\sqrt{1-\beta_2}}{\epsilon^2}$$

The inequality 38 can then be written as follows:

$$\mathbb{E}\left[f\left(\boldsymbol{\theta}_{t+1}\right) \mid \mathcal{F}_t\right] \leq f\left(\boldsymbol{\theta}_t\right) - \Delta \left\|\nabla\mathcal{L}\left(\boldsymbol{\theta}_t\right)\right\|_2^2 + \frac{\alpha}{b_t} \tag{39}$$

By taking the expected value of the inequality 39,

$$\mathbb{E}\left[f\left(\boldsymbol{\theta}_{t+1}\right)\right] \leq \mathbb{E}\left[f\left(\boldsymbol{\theta}_t\right)\right] - \Delta\mathbb{E}\left[\left\|\nabla\mathcal{L}\left(\boldsymbol{\theta}_t\right)\right\|_2^2\right] + \frac{\alpha}{b_t}$$

Which can be rearranged as follows:

$$\mathbb{E}\left[\left\|\nabla\mathcal{L}\left(\boldsymbol{\theta}_t\right)\right\|_2^2\right] \leq \frac{\mathbb{E}\left[f\left(\boldsymbol{\theta}_t\right)\right] - \mathbb{E}\left[f\left(\boldsymbol{\theta}_{t+1}\right)\right]}{\Delta} + \frac{\alpha}{\Delta b_t} \tag{40}$$

By summing 40 for all $t \in [T]$, we get:

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla\mathcal{L}\left(\boldsymbol{\theta}_t\right)\right\|_2^2\right] &\leq \frac{1}{T\Delta}\sum_{t=1}^{T}\left(\mathbb{E}\left[f\left(\boldsymbol{\theta}_t\right)\right] - \mathbb{E}\left[f\left(\boldsymbol{\theta}_{t+1}\right)\right]\right) + \frac{\alpha}{T\Delta}\sum_{t=1}^{T}\frac{1}{b_t} \\
&= \frac{1}{T\Delta}\left(f\left(\boldsymbol{\theta}_1\right) - \mathbb{E}\left[f\left(\boldsymbol{\theta}_{T+1}\right)\right]\right) + \frac{\alpha}{T\Delta}\sum_{t=1}^{T}\frac{1}{b_t} \quad \text{(using telescoping sum)} \\
&\leq \frac{1}{T\Delta}\left(f\left(\boldsymbol{\theta}_1\right) - f\left(\boldsymbol{\theta}^*\right)\right) + \frac{\alpha}{T\Delta}\sum_{t=1}^{T}\frac{1}{b_t} \quad \left(\text{where } \theta^* := \arg\min_{\theta} f(\theta)\right)
\end{aligned}$$

We conclude that:

$$\boxed{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla\mathcal{L}\left(\boldsymbol{\theta}_t\right)\right\|_2^2\right] \leq \frac{f\left(\boldsymbol{\theta}_1\right) - f\left(\boldsymbol{\theta}^*\right)}{T\Delta} + \frac{\alpha}{T\Delta}\sum_{t=1}^{T}\frac{1}{b_t}} \tag{41}$$

- If the batch size is fixed: $b_t = b_0$ for all t:
  Then, the inequality 41 becomes:

  $$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\nabla\mathcal{L}\left(\boldsymbol{\theta}_t\right)\right\|_2^2\right] \leq \frac{f\left(\boldsymbol{\theta}_1\right) - f\left(\boldsymbol{\theta}^*\right)}{T\Delta} + \frac{\alpha}{\Delta b_0}$$

Let us denote $c_1 = \frac{f(\boldsymbol{\theta}_1)-f(\boldsymbol{\theta}^*)}{\Delta}$ and $c_2 = \frac{\alpha}{\Delta b_0}$, we conclude the first part 7 of the theorem :

$$\boxed{\exists c_1, c_2 \in \mathbb{R}_+ \quad \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla \mathcal{L}\left(\boldsymbol{\theta}_t\right)\|_2^2\right] \leq \frac{c_1}{T} + c_2}$$

- If the batch size $b_t = b_0 T$ for all t:
  Then, the inequality 41 becomes:

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla \mathcal{L}\left(\boldsymbol{\theta}_t\right)\|_2^2\right] &\leq \frac{c_1}{T} + \frac{c_2}{T}\sum_{t=1}^{T}\frac{1}{T} \\ &= \frac{c_1 + c_2}{T} \end{aligned}$$

  We conclude the part 8 of the theorem, i.e:

$$\boxed{\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla \mathcal{L}\left(\boldsymbol{\theta}_t\right)\|_2^2\right] = O\left(\frac{1}{T}\right)}$$

- If the batch size in linear in time (i.e, $b_t = b_0 t$ for all t):
  Then, the inequality 41 becomes:

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla \mathcal{L}\left(\boldsymbol{\theta}_t\right)\|_2^2\right] \leq \underbrace{\frac{c_1}{T} + \frac{c_2}{T}\sum_{t=1}^{T}\frac{1}{t}}_{(\mathcal{E}_1)} \tag{42}$$

  We would like to find an equivalent to $(\mathcal{E}_1)$.
  By applying the mean value theorem to the function $\Phi : t \mapsto \ln(t)$ between t and $t+1$ (for a fixed $t \in [T]$), there exists $c_t \underset{t\to+\infty}{\sim} t$ such that:

$$\Phi(t+1) - \Phi(t) = \frac{1}{c_t} \sim \frac{1}{t}$$

  As, $\sum(\frac{1}{t})_t$ diverges, we conclude that:

$$\sum_{t=1}^{T}\left(\Phi(t+1) - \Phi(t)\right) \underset{t\to+\infty}{\sim} \sum_{t=1}^{T}\frac{1}{t}$$

  By telescoping sum, it implies:

$$\ln(T) \underset{T\to+\infty}{\sim} \sum_{t=1}^{T}\frac{1}{t}$$

Which gives, by deviding by T:

$$\frac{1}{T}\sum_{t=1}^{T}\frac{1}{t} \underset{T\to+\infty}{\sim} \frac{\ln(T)}{T}$$

On the other hand,

$$\frac{1}{T} = o\left(\frac{\ln(T)}{T}\right)$$

Which gives an equivalent to the bound $(\mathcal{E}_1)$ in 42:

$$(\mathcal{E}_1) \underset{T\to+\infty}{\sim} c_2\frac{\ln(T)}{T} \tag{43}$$

From 41 and 43 we conclude the part 9 of the theorem:

$$\boxed{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla f\left(\boldsymbol{\theta}_t\right)\|_2^2\right] = O\left(\frac{\ln(T)}{T}\right)}$$

- <u>If the batch size is of the form $b_t = \lceil b_0 t^{\gamma}\rceil$ for all t (with $0 < \gamma < 1$):</u>
Then, the inequality 41 becomes:

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla \mathcal{L}\left(\boldsymbol{\theta}_t\right)\|_2^2\right] \leq \underbrace{\frac{c_1}{T} + \frac{c_2}{T}\sum_{t=1}^{T}\frac{1}{t^{\gamma}}}_{(\mathcal{E}_2)} \tag{44}$$

We would like to find an equivalent to $(\mathcal{E}_2)$:
By applying the mean value theorem to the function $\Psi : t \mapsto \frac{1}{1-\gamma}t^{1-\gamma}$ between $t-1$ and t, there exists $c_t \sim t$ such that :

$$\Psi(t) - \Psi(t-1) = \frac{1}{c_t^{\gamma}} \underset{t\to+\infty}{\sim} \frac{1}{t^{\gamma}}$$

And the Riemann series $\sum_t \frac{1}{t^{\gamma}}$ diverges (since $\gamma < 1$), so we have:

$$\sum_{t=1}^{T}(\Psi(t) - \Psi(t-1)) \underset{t\to+\infty}{\sim} \sum_{t=1}^{T}\frac{1}{t^{\gamma}}$$

Hence, by telescoping sum:

$$\frac{T^{1-\gamma}}{1-\gamma} \underset{T\to+\infty}{\sim} \sum_{t=1}^{T}\frac{1}{t^{\gamma}}$$

Consequently:

$$\frac{c_2}{T}\sum_{t=1}^{T}\frac{1}{t^\gamma} \underset{T\to+\infty}{\sim} \frac{c_2}{(1-\gamma)T^\gamma}$$

And since:

$$\frac{1}{T} = o\left(\frac{1}{T^\gamma}\right)$$

We conclude the following equivalent to the bound $(\mathcal{E}_2)$:

$$(\mathcal{E}_2) \underset{T\to+\infty}{\sim} \frac{c_2}{(1-\gamma)T^\gamma} \tag{45}$$

From 41 and 45 we conclude the second part 9 of the theorem:

$$\boxed{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla f(\boldsymbol{\theta}_t)\|_2^2\right] = O\left(\frac{1}{T^\gamma}\right)}$$

□

References

[1] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[2] S Reddi, Manzil Zaheer, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. In *Proceeding of 32nd Conference on Neural Information Processing Systems (NIPS 2018)*, 2018.

[3] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.