Convergence of Gradient Descent and Its Variants

Gurpreet Singh
guggu@iitk.ac.in
150259

Jaivardhan Kapoor
jkapoor@iitk.ac.in
150300

Abstract

In this survey, we review a set of methods and techniques used to analyze the convergence of methods based around convex optimization using Gradient Descent and Stochastic Gradient Descent Variants.

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 articles 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.

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.

We define a function $\Pi : \mathbb{R}^d \to \mathcal{X}$ as the projection operation with respect to a positive definite matrix $\mathbf{A}$, such that

$$\Pi_{\mathcal{X},\mathbf{A}}(\mathbf{x}) = \arg\min_{\mathbf{z} \in \mathcal{X}} \left\| \mathbf{A}^{1/2}(\mathbf{z} - \mathbf{x}) \right\|_2 \tag{1}$$

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 $\forall\, \mathbf{x}, \mathbf{y} \in \mathcal{X}$,

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

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 $\forall\, \alpha \in [\,0, 1\,]$
$$f(\alpha \cdot \mathbf{x} + (1 - \alpha) \cdot \mathbf{y}) \quad \leq \quad \alpha \cdot f(\mathbf{x}) + (1 - \alpha) \cdot f(\mathbf{y}) \tag{3}$$
3. If f is twice differentiable, then $\forall\, \mathbf{x} \in \mathcal{X}$
$$\nabla^2 f(\mathbf{x}) \quad \succeq \quad 0 \tag{4}$$

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 $\forall\, \mathbf{x}, \mathbf{y} \in \mathcal{X}$,

$$f(\mathbf{y}) \quad \geq \quad f(\mathbf{x}) + \langle \nabla f(\mathbf{x}),\ \mathbf{y} - \mathbf{x} \rangle + \frac{\alpha}{2} \left\| \mathbf{y} - \mathbf{x} \right\|_2 \tag{5}$$

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

$$f(\mathbf{y}) \quad \leq \quad f(\mathbf{x}) + \langle \nabla f(\mathbf{x}),\ \mathbf{y} - \mathbf{x} \rangle + \frac{\beta}{2} \left\| \mathbf{y} - \mathbf{x} \right\|_2 \tag{6}$$

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. The later sections deal with the noise variants of Gradient Descent methods, along with some comments on non-convex 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.

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.* $h(\mathbf{x}^{(1)} \dots \mathbf{x}^{(T)}) = \frac{1}{T} \sum_{t=1}^{T} \mathbf{x}_t$, or we can simply set h to return the last time step's estimate, *i.e.* $h(\mathbf{x}^{(1)} \dots \mathbf{x}^{(T)}) = \mathbf{x}^{(T)}$.

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

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 $\hat{\mathbf{x}} = h\left(\mathbf{x}^{(1)} \dots \mathbf{x}^{(T)}\right)$

Steps:

1. Initialize $\mathbf{x}^{(0)} \in \mathcal{H}$
2. For $t = 1 \dots T$, do

$$\begin{aligned} \mathbf{g}_t &= \nabla f(\mathbf{x}^{(t)}) \\ \mathbf{z}^{(t+1)} &= \mathbf{x}^{(t)} - \alpha_t \cdot \mathbf{g}_t && \text{(Optimization Step)} \\ \mathbf{x}^{(t+1)} &= \Pi_{\mathcal{X}}\left(\mathbf{z}^{(t+1)}\right) && \text{(Projection Step)} \end{aligned}$$

3. Return $\hat{\mathbf{x}} = h\left(\mathbf{x}^{(1)} \dots \mathbf{x}^{(T)}\right)$

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

$$\mathbf{x}^{(t+1)} = \Pi_{\mathcal{X}}\left(\mathbf{x}^{(t)} - \eta_t \cdot \nabla f(\mathbf{x}^{(t)})\right) \qquad \text{(Vanilla GD)}$$

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.* $f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*)$ and D_t to denote the difference between the current point estimate and the optimal point, *i.e.* $\left\|\mathbf{x}^{(t)} - \mathbf{x}^*\right\|_2$

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

$$\frac{1}{T}\sum_{t=1}^{T} \Phi_t \leq \frac{1}{\sqrt{T}} \mathrm{D}_0^2 \mathrm{G}^2 \tag{7}$$

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

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

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 $\mathbf{a}, \mathbf{b} \in \mathbb{R}^m$,

$$\|\mathbf{a}+\mathbf{b}\|_2^2 - \|\mathbf{a}\|_2^2 - \|b\|_2^2 \overset{(a)}{=} 2\langle a,\ b\rangle \overset{(b)}{=} \|\mathbf{a}\|_2^2 + \|b\|_2^2 - \|\mathbf{a}-\mathbf{b}\|_2^2 \tag{8}$$

Using property 8a, we can write the above inequality as

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

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

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

This proves theorem 1.2 □

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

$$f\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}^{(t)}\right) \leq \sum_{t=1}^{T} f(\mathbf{x}^{(t)})$$

Therefore, substituting this in equation 10, we can write

$$f\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}^{(t)}\right) - f(\mathbf{x}^*) \leq \frac{1}{2\eta T}\mathrm{D}_0^2 + \frac{\eta}{2}\cdot\mathrm{G}^2 \tag{9}$$

Hence for a case when the function f is convex and has bounded gradients, we can say that This allows us to bound the value of Φ_T as T increases, 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 9. 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 $\forall x$, $\nabla f(\mathbf{x}^{(t)})$ exists, we can say

$$\Phi_t \quad \leq \quad \frac{1}{2\eta} \cdot \frac{\mathrm{D}_0^2}{T} \tag{10}$$

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

$$\begin{aligned} f(\mathbf{x}^*) \quad &\geq \quad f(\mathbf{x}^{(t)}) - \left\langle \nabla f(\mathbf{x}^{(t)}),\ \mathbf{x}^{(t)} - \mathbf{x}^* \right\rangle \\ f(\mathbf{x}^{(t+1)}) \quad &\leq \quad f(\mathbf{x}^{(t)}) + \left\langle \nabla f(\mathbf{x}^{(t)}),\ \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \right\rangle + \frac{\beta}{2} \left\| \mathbf{x}^{(t+1)} - \mathbf{x}^{(t)} \right\|_2^2 \end{aligned} \tag{11}$$

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

$$f(\mathbf{x}^{(t+1)}) \quad \leq \quad f(\mathbf{x}^{(t)}) + \left(\frac{\beta}{2} - \frac{1}{\eta_t} \right) \left\| \eta_t \cdot \nabla f(\mathbf{x}^{(t)}) \right\|_2^2 \tag{12}$$

Subtracting equation 11 from 12, we get

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

Using property 8a, we can write this, similarly to the previous case, as

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

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

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

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,

$$f\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}^{(t)}\right) \quad \leq \quad \frac{1}{2\eta}\cdot\frac{\mathrm{D}_0^2}{T} \tag{13}$$

3.2. Momentum

<++>

3.3. Nesterov's Accelerated Gradient

<++>

<++>

4. Adaptive Methods based on Exponentially Moving Averages

4.1. Stochastic Gradient Descent

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

4.2. AdaGrad

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

<++>

References

M. Augustine Cauchy. M´ethode g´en´erale pour la r´esolution des syst'emes d'´equations simultan´ees. 1847.

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

Algorithm 2: Generic Adaptive Method

Input: step size $\{\alpha > 0\}_{t=1}^{T}$, and a sequence of functions $\{\phi_t, \psi_t\}_{t=1}^{T}$

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

$$\begin{aligned} \mathbf{g}_t &= \nabla f_t\left(x_t\right) \\ \mathbf{m}_t &= \phi_t\left(\mathbf{g}_1 \dots \mathbf{g}_t\right) \\ V_t &= \phi_t\left(\mathbf{g}_1 \dots \mathbf{g}_t\right) \\ \mathbf{z}^{(t+1)} &= \mathbf{x}^{(t)} - \alpha_t V_t^{-1/2} \mathbf{m}_t \\ \mathbf{x}^{(t)} &= \Pi_{\mathcal{H}, V^{1/2}}\left(\mathbf{z}^{(t+1)}\right) \end{aligned}$$

3. Return $\hat{\mathbf{x}} = h\left(\mathbf{x}^{(1)} \dots \mathbf{x}^{(T)}\right)$