Gradient Descent

Zachary Ross

March 2, 2023

In the following writeup, we explore gradient descent (GD) methods and their convergence. These are adaptations to the general method

$$x_{t+1} = x_t - \eta \nabla f(x_t), \tag{1}$$

where η is the learning rate and $f : \mathbb{R}^n \to \mathbb{R}$ is a *convex* objective function with a unique minimum, which use previous descent directions as part of current iterates computation.

Contents

1 Convergence Analysis **2**
1.1 Smoothness 2
1.2 Strong Convexity 6
1.3 Stochastic Gradient Descent 8

2 Accelerated Methods **10**
2.1 Polyak's Momentum 10
2.2 Nesterov's Momentum 11

3 First Order Methods **11**

1 Convergence Analysis

In general, convergence proofs follow the form of analyzing a convergent sequence $\{x_i\}_{i=1}^t$ which optimizes a function f. We begin by exploring the general convergence of a convex function and prove Equation 1 when different aspects of the curvature of f are known, such as an upper bound, lower bound, and stochastic properties under SGD. This requires a few definitions prior to analysis, mainly be defining convexity from a theoretical point of view, then adding upper and lower bounds to this definition which will aid in these proofs.

Where convenient, we use the following notations: $x^* = \inf_x f(x)$ is the optimal point of f (which exists due to convexity), $f^* \triangleq f(x^*)$ is the value of f at the optimal point, $f_t \triangleq f(x_t)$ is the value of f at x_t, $\nabla f_t \triangleq \nabla f(x_t)$ is the gradient evaluation of f at x_t, and $\delta_t \triangleq f_t - f^*$ is the error at x_t.

Definition 1. A function $f : \mathbb{R}^n \to \mathbb{R}$ is *convex* if for all $x, y \in \mathbb{R}^n$, for all $\lambda \in [0, 1]$,

$$\lambda f(x) + (1 - \lambda) f(y) \geq f(\lambda x + (1 - \lambda) y). \tag{2}$$

This definition states that any interpolation between two points evaluated on f is greater than evaluating f on the interpolated points. This provides an upper bound for a convex function over any two points and is demonstrated by the red line in Figure 1. Likewise, we have that a convex function can be lower-bounded at every point by a hyperplane which runs tangential to the function. This concept is demonstrated by the blue line in Figure 1.

Lemma 1. *If a function f is convex and differentiable, then for all $x, y \in \mathbb{R}^n$,*

$$f(y) - f(x) \geq \nabla f(x)^\intercal (y - x) \tag{3}$$

Proof. For $x, y \in \mathbb{R}^n$, Equation 2 can be restated as

$$f(y) - f(x) \geq \frac{f(x + \lambda(y - x)) - f(x)}{\lambda} \quad \text{for} \quad \lambda \in (0, 1]. \tag{4}$$

Taking the limit w.r.t. λ gives us

$$\lim_{\lambda \to 0^+} \frac{f(x + \lambda(y - x)) - f(x)}{\lambda} = \nabla f(x)^\intercal (y - x) \tag{5}$$

□

1.1 Smoothness

Smoothness is often an assumption when analyzing the convergence of gradient descent algorithms, as it characterizes the curvature of a function quite nicely. Most proofs benefit from the assumption that the gradient will tend to *decrease* as elements in the sequence move more towards an optimal point, which is the property that smoothness describes.

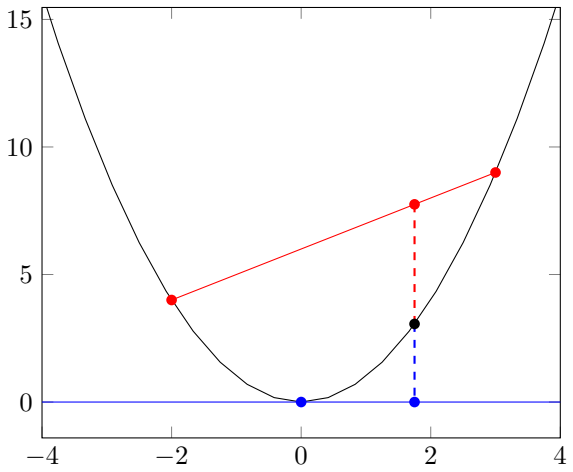

Figure 1: Graph of the convex function $f(x) = x^2$. The solid red line indicates an interpolation between points $(-2, 4)$ and $(3, 9)$ on the graph while the dashed red line demonstrates the distance between the theoretical upper bound and the evaluation for $x = 1.75$. Likewise, the solid blue line indicates the hyperplane which runs tangential to the point $(0, 0)$ while the dashed blue line demonstrates the distance between the theoretical lower bound and the evaluation for the same x.

Example 1. Consider the functions $f(x) = x^2$ and $g(x) = |x|$. The function f is much more desirable when dealing with gradient descent as ∇f scales with changes in x as it moves closer and closer 0. The function g yields greater difficult, since ∇g gives almost no information of the direction of the gradient, only yielding whether to move in the positive or negative direction, which will cause infinite oscillation if the learning rate is constant.

Definition 2. A continuously differentiable function f is β-smooth if its gradient is β-Lipschitz

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\| . \tag{6}$$

Using smoothness, one can derive both upper and lower quadratic bounds to characterize f.

Lemma 2 (Quadratic Bounds)**.** *Let f be β-smooth on $\mathbb{R}^n$. Then for any $x, y \in \mathbb{R}^n$ we have*

$$|f(y) - f(x) - \nabla f(x)^\intercal (y - x)| \leq \frac{\beta}{2} \|y - x\|^2 . \tag{7}$$

Lemma 2 can be equivalently stated as providing upper and lower bounds on a continuously differentiable, β-smooth function f at $y \in \mathbb{R}^n$

$$-\frac{\beta}{2} \|y - x\|^2 \leq f(y) - f(x) - \nabla f(x)^\intercal (y - x) \leq \frac{\beta}{2} \|y - x\|^2 \tag{8}$$

where for a convex f, Lemma 1 implies that

$$0 \leq f(y) - f(x) - \nabla f(x)^\intercal (y - x) \leq \frac{\beta}{2} \|y - x\|^2 . \tag{9}$$

In fact, when f indeed is convex *and* β-smooth, we can apply a much tighter lower bound using the following lemma.

Lemma 3. *Let f be convex and β-smooth on $\mathbb{R}^n$. Then for any $x, y \in \mathbb{R}^n$ we have and*

$$\frac{1}{2\beta} \|\nabla f(y) - \nabla f(x)\|^2 \leq f(y) - f(x) - \nabla f(x)^\intercal (y - x). \tag{10}$$

Lemma 4. *Let f be convex and β-smooth. Then the gradient descent update listed in Equation 1 with learning rates $\eta_t \leq \frac{1}{\beta}$ satisfies*

$$\|x_{t+1} - x^*\|^2 \leq \|x_t - x^*\|^2 . \tag{11}$$

Proof. By Lemma 3, we can rearrange to bound the difference

$$f_t - f^* \leq -\nabla f_t^\intercal (x^* - x_t) - \frac{1}{2\beta} \|\nabla f_t\|^2 . \tag{12}$$

Since x^* was chosen s.t. $x^* = \inf_x f(x)$, we have $f(x_t) - f(x^*) \geq 0$ so that

$$0 \leq f_t - f^* \leq -\nabla f_t^\intercal (x^* - x_t) - \frac{1}{2\beta} \|\nabla f_t\|^2 , \tag{13}$$

and by rearranging and excluding the middle,

$$\nabla f_t^\intercal (x^* - x_t) \leq -\frac{1}{2\beta} \|\nabla f_t\|^2 . \tag{14}$$

Note that the expansion of parameter updates based on gradient descent is given by

$$\|x_{t+1} - x^*\|^2 = \|x_t - x^*\|^2 - 2\eta_t \nabla f_t^\intercal (x_t - x^*) + \eta_t^2 \|\nabla f_t\|^2 . \tag{15}$$

whereby substituting Equation 14 gives us

$$\|x_{t+1} - x^*\|^2 \leq \|x_t - x^*\|^2 + \eta_t \left(\eta_t - \frac{1}{\beta} \right) \|\nabla f_t\|^2 \tag{16}$$

Choice of $\eta_t \leq \frac{1}{\beta}$ implies that $\eta_t - \frac{1}{\beta} \leq 0$ so that

$$\|x_{t+1} - x^*\|^2 \leq \|x_t - x^*\|^2 . \tag{17}$$

□

Theorem 1. *Let f and η_t be defined as in Lemma 4. Then the objective difference satisfies*

$$f_{t+1} - f^* \leq \frac{\|x_1 - x^*\|^2}{\sum_{i=1}^t 1 - \frac{\beta \eta_i}{2}} \tag{18}$$

Proof. By Lemma 3, the progress in objective function satisfies

$$f_{t+1} - f_t \leq -\eta_t \left(1 - \eta_t \frac{\beta}{2}\right) \|\nabla f_t\|^2 . \tag{19}$$

We can add and subtract f^* to determine a forumlation relative to f^*

$$(f_{t+1} - f^*) - (f_t - f^*) = \delta_{t+1} - \delta_t \leq -\eta_t \left(1 - \eta_t \frac{\beta}{2}\right) \|\nabla f_t\|^2 . \tag{20}$$

By Lemma 1 we have that $\delta_t \leq -\nabla f_t^\intercal (x^* - x_t)$, which by the Cauchy-Schwartz inequality gives us

$$\delta_t^2 \leq \left(\nabla f_t^\intercal (x^* - x_t)\right)^2 \leq \|\nabla f_t\|^2 \|x_t - x^*\|^2 , \tag{21}$$

which can be plugged back into Equation **??** to get

$$\delta_{t+1} - \delta_t \leq -\eta_t \left(1 - \eta_t \frac{\beta}{2}\right) \frac{\delta_t^2}{\|x_t - x^*\|^2}. \tag{22}$$

The difference in parameter updates given by Lemma 4 tells us

$$\delta_{t+1} - \delta_t \leq -\eta_t \left(1 - \eta_t \frac{\beta}{2}\right) \frac{\delta_t^2}{\|x_1 - x^*\|^2}. \tag{23}$$

By dividing both sides by $\delta_{t+1}\delta_t$, we can make use of the fact that since $f_{t+1} \leq f_t \implies \delta_{t+1} \leq \delta_t$, then the left side of Equation 23 can be reduced to

$$\frac{\delta_{t+1} - \delta_t}{\delta_{t+1}\delta_t} = \frac{1}{\delta_t} - \frac{1}{\delta_{t+1}} \tag{24}$$

and the right side as

$$\begin{aligned} -\eta_t \left(1 - \eta_t \frac{\beta}{2}\right) \frac{\delta_t^2}{\|x_1 - x^*\|^2} \frac{1}{\delta_{t+1}\delta_t} &= -\eta_t \left(1 - \eta_t \frac{\beta}{2}\right) \frac{1}{\|x_1 - x^*\|^2} \frac{\delta_t}{\delta_{t+1}} \\ &\leq -\eta_t \left(1 - \eta_t \frac{\beta}{2}\right) \frac{1}{\|x_1 - x^*\|^2}. \end{aligned} \tag{25}$$

Putting these two together yields the equation

$$\frac{1}{\delta_t} - \frac{1}{\delta_{t+1}} \leq -\eta_t \left(1 - \eta_t \frac{\beta}{2}\right) \frac{1}{\|x_1 - x^*\|^2}. \tag{26}$$

which forms the telescopic series

$$\sum_{i=1}^{t} \frac{1}{\delta_i} - \frac{1}{\delta_{i+1}} \leq -\sum_{i=1}^{t} \eta_i \left(1 - \eta_i \frac{\beta}{2}\right) \frac{1}{\|x_1 - x^*\|^2}, \tag{27}$$

or put more simply

$$\frac{1}{\delta_1} - \frac{1}{\delta_{t+1}} \leq -\sum_{i=1}^{t} \eta_i \left(1 - \eta_i \frac{\beta}{2}\right) \frac{1}{\|x_1 - x^*\|^2}. \tag{28}$$

Since $\delta_1 \geq \delta_{t+1} \geq 0$,

$$-\frac{1}{\delta_{t+1}} \leq \frac{1}{\delta_1} - \frac{1}{\delta_{t+1}}. \tag{29}$$

By using this, taking the negative reciprocal, and plugging back in the definition of δ_t, we get the desired equation

$$f_{t+1} - f^* \leq \frac{\|x_1 - x^*\|^2}{\sum_{i=1}^{t} \eta_i \left(1 - \eta_i \frac{\beta}{2}\right)}. \tag{30}$$

□

Corollary 1. *With constant learning rate $\eta_i = \frac{1}{\beta}$, residual error in parameter updates are bounded by*

$$f_{t+1} - f^* \leq \frac{2\beta}{t} \|x_0 - x^*\|^2 \tag{31}$$

Proof. Substitute the learning update into Theorem 1 to get

$$\begin{aligned} f_{t+1} - f^* &\leq \frac{\|x_1 - x^*\|^2}{\sum_{i=1}^{t} \frac{1}{2\beta}} \\ &= \frac{\|x_1 - x^*\|^2}{\frac{t}{2\beta}} \\ &= \frac{2\beta}{t} \|x_1 - x^*\|^2 \end{aligned} \tag{32}$$

□

1.2 Strong Convexity

Definition 3. A function f is α-strongly convex if for $\alpha > 0$ and $x \in \mathbb{R}^n$

$$f(x) - \frac{\alpha}{2} \|x\|^2 \tag{33}$$

is convex.

Strong convexity allows us to apply a tighter lower bound to that derived in Lemmas 1 and 3 using the following Lemma.

Lemma 5. *Let f be α-strong convex. Then for all $x, y \in \mathbb{R}^n$ we have*

$$\frac{\alpha}{2} \|y - x\|^2 \leq f(y) - f(x) - \nabla f(x)^\intercal (y - x) \tag{34}$$

Lemma 6 (Coercivity of the Gradient)**.** *Let f be β-smooth and α-strongly convex. Then for all x and y we have*

$$(\nabla f(x) - \nabla f(y))^\intercal (x - y) \geq \frac{\alpha\beta}{\alpha + \beta} \|x - y\|^2 + \frac{1}{\alpha + \beta} \|\nabla f(x) - \nabla f(y)\|^2. \tag{35}$$

Theorem 2 (General Convergence)**.** *Let f be β-smooth and α-strongly convex. Then the gradient descent update listed in Equation 1 with learning rates η_i satisfies*

$$f_{t+1} - f^* \leq \frac{\beta}{2} \|x_1 - x^*\|^2 \prod_{i=1}^{t} (1 - \eta_i \beta)^2. \tag{36}$$

Proof. Since f is β-smooth, convex, and $\nabla f^* = 0$ by definition, Lemma 3 yields

$$f_t - f^* \leq \frac{\beta}{2} \|x_t - x^*\|^2 . \tag{37}$$

Additionally, we get from definition of β-smooth that

$$\|\nabla f_t\| \leq \beta \|x_t - x^*\| . \tag{38}$$

We can then bound Equation 37 by bounding $\|x_t - x^*\|^2$.

$$\begin{aligned}
\|x_{t+1} - x^*\|^2 &= (x_t - \eta_t \nabla f_t - x^*)^\intercal (x_t - \eta_t \nabla f_t - x^*) \\
&= ((x_t - x^*) - \eta_t \nabla f_t)^\intercal ((x_t - x^*) - \eta_t \nabla f_t) \\
&= \|x_t - x^*\|^2 - 2\eta_t \nabla f_t{}^\intercal (x_t - x^*) + \eta_t^2 \|\nabla f_t\|^2 \\
&= \|x_t - x^*\|^2 - 2\eta_t [\nabla f_t - \nabla f(x^\star)]^\intercal (x_t - x^*) + \eta_t^2 \|\nabla f_t\|^2 \\
&\leq \left[1 - 2\eta_t \frac{\alpha\beta}{\alpha + \beta}\right] \|x_t - x^*\|^2 + \left[\eta_t^2 - 2\eta_t \frac{1}{\alpha + \beta}\right] \|\nabla f_t\|^2 \\
&= \left[1 - 2\eta_t \frac{\alpha\beta}{\alpha + \beta} + \eta_t^2 \beta^2 - 2\eta_t \frac{\beta^2}{\alpha + \beta}\right] \|x_t - x^*\|^2 \\
&= \left[1 - 2\eta_t \beta + \eta_t^2 \beta^2\right] \|x_t - x^*\|^2 \\
&= (1 - \eta_t \beta)^2 \|x_t - x^*\|^2 \\
&\leq \frac{\beta}{2} \|x_1 - x^*\|^2 \prod_{i=1}^{t} (1 - \eta_i \beta)^2
\end{aligned} \tag{39}$$

□

Although the previous equation derives a setting for the general convergence of such a function, we can get much stricter bounds under specified learning rates.

Theorem 3 (Convergence for $\eta = 2/(\alpha + \beta)$)**.** *Let f be β-smooth and α-strongly convex. Then the gradient descent update listed in Equation 1 with constant learning rate $\eta = \frac{2}{\alpha + \beta}$ satisfies*

$$f_{t+1} - f^* \leq \frac{\beta}{2} \exp\left(-\frac{4t}{\kappa + 1}\right) \|x_1 - x^*\|^2 . \tag{40}$$

Proof. We use the same steps leading up to the previous proof, and begin modifi-

cations where the two differ. Let $\kappa = \frac{\beta}{\alpha}$ be the *condition number*.

$$\begin{aligned}
\|x_{t+1} - x^*\|^2 &\leq \left[1 - 2\eta\frac{\alpha\beta}{\alpha+\beta}\right]\|x_t - x^*\|^2 + \left[\eta^2 - 2\eta\frac{1}{\alpha+\beta}\right]\|\nabla f_t\|^2 \\
&= \left[1 - 2\eta\frac{\alpha\beta}{\alpha+\beta}\right]\|x_t - x^*\|^2 \\
&= \left(\frac{\kappa-1}{\kappa+1}\right)^2\|x_t - x^*\|^2 \\
&\leq \left(\frac{\kappa-1}{\kappa+1}\right)^{2t}\|x_1 - x^*\|^2 \\
&\leq \exp\left(-\sum_{i=1}^{2t}\frac{2}{\kappa+1}\right)\|x_1 - x^*\|^2 \\
&= \exp\left(-\frac{4t}{\kappa+1}\right)\|x_1 - x^*\|^2
\end{aligned} \tag{41}$$

□

1.3 Stochastic Gradient Descent

Stochastic gradient descent slightly differs from Equation 1 in that instead of a deterministic valued ∇f, we consider the case where the optimization task is

$$\inf_x f(x) = \mathbb{E}[F(x,\xi)] \tag{42}$$

where $F : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable, ξ is a random variable supported on the distribution $\Xi \subseteq \mathbb{R}^d$, and we have generally that either $F(\cdot,\xi)$ is not given explicitly and/or the distribution of Ξ is unknown. This is the case for most machine learning problems and hence is important to study alongside general gradient descent methods. In these cases, we may only rely upon the known properties of f (e.g. β-smooth, convexity, etc) and the noisy information in ∇f to optimize parameters in this form. We assume that the noisy gradients of f can be given to us by a Stochastic First-Order Oracle (SFO) in the form of $g(x,\xi)$. From hereafter, we use the notation $g_t \triangleq g(x_t,\xi_t)$. This random variable g is assumed to have the properties that it is

1. an unbiased estimate
$$\mathbb{E}[g_t] = \nabla f_t, \tag{43}$$
2. and has an upper bound on the noise distribution, i.e. for some $\sigma > 0$, the max noise level of the gradient estimation is
$$\mathbb{E}\left[\|g_t - \nabla f_t\|^2\right] \leq \sigma^2. \tag{44}$$

Theorem 4. *Let f be convex and β-smooth. Then the stochastic gradient descent update using the update equation listed in Equation 1 with learning rates $\eta_t \leq \frac{1}{\beta}$ satisfies*

$$\min_{i=2,3,\ldots,t}\mathbb{E}[\|\nabla f_i\|^2 | x_{i-1}] \leq 2\frac{f_1 - f^*}{\sum_{i=2}^t \eta_i} + \frac{\sigma^2}{\beta}\frac{t-1}{\sum_{i=2}^t \eta_i}. \tag{45}$$

Proof. Let $\Delta_t = g_t - \nabla f_t$. Lemma 2 gives us the inequality

$$\begin{aligned} f_{t+1} - f_t &\leq -\eta_t \nabla f_t{}^\intercal g_t + \eta_t^2 \frac{\beta}{2} \left\| g_t \right\|^2 \\ &= -\eta_t \left\| \nabla f_t \right\|^2 - \eta_t \nabla f_t{}^\intercal \Delta_t + \eta_t^2 \frac{\beta}{2} \left\| g_t \right\|^2 \end{aligned} \tag{46}$$

By taking the expectation of both sides and noting that $\mathbb{E}[\Delta_t | x_t] = \nabla f_t - \nabla f_t = 0$, we get

$$\mathbb{E}[f_{t+1}|x_t] - f_t \leq -\eta_t \left\| \nabla f_t \right\|^2 + \eta_t^2 \frac{\beta}{2} \mathbb{E}[\left\| g_t \right\|^2 | x_t] \tag{47}$$

where the last term expands out to

$$\begin{aligned} \mathbb{E}[\left\| g_t \right\|^2 | x_t] &= \mathbb{E}[\left\| \Delta_t + \nabla f_t \right\|^2 | x_t] \\ &= \mathbb{E}[\left\| \Delta_t \right\|^2 | x_t] + 2\mathbb{E}[\nabla f_t{}^\intercal \Delta_t | x_t] + \mathbb{E}[\left\| \nabla f_t \right\|^2 | x_t] \\ &= \left\| \nabla f_t \right\|^2 + \sigma^2 \end{aligned} \tag{48}$$

yielding the equation

$$\mathbb{E}[f_{t+1}|x_t] - f_t \leq -\eta_t \left(1 - \eta_t \frac{\beta}{2}\right) \left\| \nabla f_t \right\|^2 + \eta_t^2 \sigma^2 \frac{\beta}{2} \tag{49}$$

or

$$\eta_t \left(1 - \eta_t \frac{\beta}{2}\right) \left\| \nabla f_t \right\|^2 \leq f_t - \mathbb{E}[f_{t+1}|x_t] + \eta_t^2 \sigma^2 \frac{\beta}{2}. \tag{50}$$

We can then add these together and take the iterated expectation to get

$$\sum_{i=2}^{t} \eta_i \left(1 - \eta_i \frac{\beta}{2}\right) \mathbb{E}[\left\| \nabla f_i \right\|^2 | x_{i-1}] \leq \sum_{i=2}^{t} \mathbb{E}[f_i | x_{i-1}] - \mathbb{E}[f_{i+1}|x_i] + \eta_i^2 \sigma^2 \frac{\beta}{2}. \tag{51}$$

Bounding the LHS by min and taking into account the telescoping yields

$$\begin{aligned} \min_{i=2,3,\ldots,t} \mathbb{E}[\left\| \nabla f_i \right\|^2 | x_{i-1}] \sum_{i=2}^{t} \eta_i \left(1 - \eta_i \frac{\beta}{2}\right) &\leq \mathbb{E}[f_2|x_1] - \mathbb{E}[f_{t+1}|x_t] + \sigma^2 \frac{\beta}{2} \sum_{i=2}^{t} \eta_i^2 \\ &\leq f_1 - f^* + \sigma^2 \frac{\beta}{2} \sum_{i=2}^{t} \eta_i^2 \end{aligned} \tag{52}$$

and by dividing both sides by the coefficient of the LHS,

$$\min_{i=2,3,\ldots,t} \mathbb{E}[\left\| \nabla f_i \right\|^2 | x_{i-1}] \leq \frac{f_1 - f^*}{\sum_{i=2}^{t} \eta_i \left(1 - \eta_i \frac{\beta}{2}\right)} + \sigma^2 \frac{\beta}{2} \frac{\sum_{i=2}^{t} \eta_i^2}{\sum_{i=2}^{t} \eta_i \left(1 - \eta_i \frac{\beta}{2}.\right)} \tag{53}$$

Since $\eta_i \leq \frac{1}{\beta}$, $\sum_{i=2}^{t} \eta_i^2 \leq (t-1) \frac{1}{\beta}^2$ and $\sum_{i=2}^{t} \eta_i (1 - \eta_i \beta / 2) \geq \frac{1}{2} \sum_{i=2}^{t} \eta_i$

$$\begin{aligned} \min_{i=2,3,\ldots,t} \mathbb{E}[\left\| \nabla f_i \right\|^2 | x_{i-1}] &\leq \frac{f_1 - f^*}{\frac{1}{2} \sum_{i=2}^{t} \eta_i} + \sigma^2 \frac{\beta}{2} \frac{(t-1) \frac{1}{\beta}^2}{\frac{1}{2} \sum_{i=2}^{t} \eta_i} \\ &= 2 \frac{f_1 - f^*}{\sum_{i=2}^{t} \eta_i} + \frac{\sigma^2}{\beta} \frac{t-1}{\sum_{i=2}^{t} \eta_i} \end{aligned} \tag{54}$$

□

2 Accelerated Methods

The intuition behind using accelerated methods is that it introduces a physics-inspired component to GD which mimics a dampened oscillator ordinary differential equation (ODE). This effect minimizes the oscillations that tend to occur in gradient descent by using previous gradient computations to penalize rapid changes in direction and reward movement towards an optimal point.

2.1 Polyak's Momentum

We reconstruct this via a method similar to as was done by Gadat et al.: the ODE limit for gradient descent is given by the equation

$$\frac{dx}{dt} = -\nabla f(x). \tag{55}$$

This can be modified to mimic a dampened oscillator ODE by including acceleration and dampening term $\gamma \geq 0$

$$\frac{d^2x}{dt^2} + \gamma \frac{dx}{dt} = -\nabla f(x) \tag{56}$$

and then rearranged to interpret in terms of acceleration, along with a simplification to velocity v of x, i.e. $v = \frac{dx}{dt}$,

$$\frac{dv}{dt} = -\gamma v - \nabla f(x). \tag{57}$$

Conversion to a discrete time step $\sqrt{\eta}$ with a forward-difference yields the equations

$$\frac{dv_t}{dt} \approx \frac{v_{t+1} - v_t}{\sqrt{\eta}}, \quad \text{and} \quad v_t = \frac{dx_t}{dt} \approx \frac{x_{t+1} - x_t}{\sqrt{\eta}} \tag{58}$$

which can be used to derive a formula for the dampened oscillator GD. Let $\beta = 1 - \gamma\sqrt{\eta}$. Then at iteration $t-1$

$$\begin{aligned}
\frac{v_t - v_{t-1}}{\sqrt{\eta}} &= -\gamma v_{t-1} - \nabla f(x_{t-1}) \\
v_t - v_{t-1} &= -\gamma\sqrt{\eta} v_{t-1} - \sqrt{\eta}\nabla f(x_{t-1}) \\
v_t &= \beta v_{t-1} - \sqrt{\eta}\nabla f(x_{t-1}) \\
\frac{x_{t+1} - x_t}{\sqrt{\eta}} &= \beta \frac{x_t - x_{t-1}}{\sqrt{\eta}} - \sqrt{\eta}\nabla f(x_{t-1}) \\
x_{t+1} &= x_t + \beta(x_t - x_{t-1}) - \eta\nabla f(x_{t-1})
\end{aligned} \tag{59}$$

By modifying this equation to evaluate the gradient at x_t rather than x_{t-1}, this yields the *heavyball* method as was discovered by Polyak

$$x_{t+1} = x_t + \beta(x_t - x_{t-1}) - \eta\nabla f(x_t). \tag{60}$$

2.2 Nesterov's Momentum

Despite the effectiveness of this method, Polyak's still has the potential to oscillate infinitely under specific conditions, as observed by Lessard et al.. This is due to the gradient being evaluated *before* momentum is applied and x_t sometimes serving as a poor approximate for x_{t-1}. Nesterov altered this method slightly by evaluating the gradient *after* momentum is applied. This derivation can be made by reinterpreting Equation 61 as a backwards difference with time step $\sqrt{\delta}$ for some $\delta \geq 0$

$$\frac{dv_t}{dt} \to \frac{v_t - v_{t-1}}{\sqrt{\delta}}, \quad \text{and} \quad v_t = \frac{dx_t}{dt} \to \frac{x_t - x_{t-1}}{\sqrt{\delta}} \tag{61}$$

and re-deriving Equation 62 under this condition. Let $\beta = \left(1 + \gamma\sqrt{\delta}\right)^{-1}$ and $\eta = \beta\delta$. Then at iteration $t+1$

$$\begin{aligned}
\frac{v_{t+1} - v_t}{\sqrt{\delta}} &= -\gamma v_{t+1} - \nabla f(x_{t+1}) \\
v_{t+1} - v_t &= -\gamma\sqrt{\delta} v_{t+1} - \sqrt{\delta}\nabla f(x_{t+1}) \\
\beta^{-1} v_{t+1} - v_t &= -\sqrt{\delta}\nabla f(x_{t+1}) \\
\frac{x_{t+1} - x_t}{\beta\sqrt{\delta}} - \frac{x_t - x_{t-1}}{\sqrt{\delta}} &= -\sqrt{\delta}\nabla f(x_{t+1}) \\
\frac{x_{t+1} - x_t}{\beta\sqrt{\delta}} &= \frac{x_t - x_{t-1}}{\sqrt{\delta}} - \sqrt{\delta}\nabla f(x_{t+1}) \\
x_{t+1} - x_t &= \beta(x_t - x_{t-1}) - \eta\nabla f(x_{t+1}) \\
x_{t+1} &= x_t + \beta(x_t - x_{t-1}) - \eta\nabla f(x_{t+1}).
\end{aligned} \tag{62}$$

By evaluating the gradient at $x_t + \beta(x_t - x_{t-1})$ rather than x_{t+1}, this then yields Nesterov's accelerated method.

$$x_{t+1} = x_t + \beta(x_t - x_{t-1}) - \eta\nabla f(x_t + \beta(x_t - x_{t-1})). \tag{63}$$

The difference between these methods is shown in Figure 2.

Research Question 1. Heavyball and Nesterov's momentum use either forward or backwards differences, which both yield a $\mathcal{O}(\epsilon)$ derivative approximation. Center differences, defined at time step t by

$$\frac{dx_t}{dt} \approx \frac{x_{t+1} - x_{t-1}}{\delta} \tag{64}$$

have been shown to yield a $\mathcal{O}(\epsilon^2)$ approximation [citation needed]. Would there be any performance benefits to using this derivative approximation over the other momentum methods?

3 First Order Methods

This section discusses methods of the form

$$x_{t+1} = x_t - \eta H_t^{-1}\nabla f(x_t) \tag{65}$$

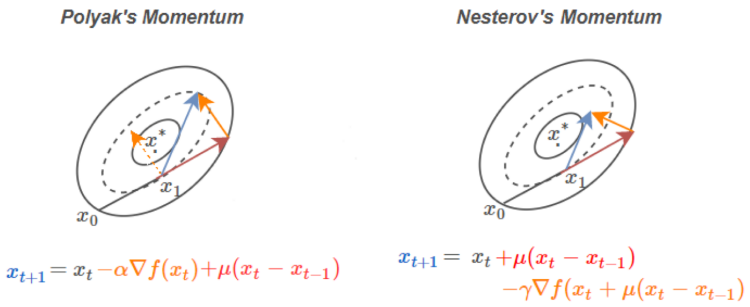

Figure 2: Comparison between Polyak's and Nesterov's momentum. The gradient descent step (orange arrow) is perpendicular to the level set before applying momentum to x_1 (red arrow) in Polyak's algorithm, whereas it is perpendicular to the level set after applying momentum to x_1 in Nesterov's algorithm. Graphic provided by Mitliagka.

where H_t is a first-order hessian approximation. Hessian approximation methods make use of the geometry of the data to approximate the Fisher information matrix, which is itself a hessian approximation, given by

$$I(x) = \mathbb{E}_x \left[\nabla f(x) \nabla f(x)^\intercal \right] \tag{66}$$

to find a solution to the root-finding problem

$$f(x^*) \approx f(x_t) + \nabla f(x_t)^\intercal (x^* - x_t) + \frac{1}{2}(x^* - x_t)^\intercal H_t^{-1} (x^* - x_t). \tag{67}$$

First order methods approximate the diagonal of the matrix in Equation 66 via Hadamard product rather than outerproduct, i.e.

$$I(x) \approx \operatorname{diag} \left\{ \mathbb{E}_x \left[\nabla f(x) \odot \nabla f(x) \right] \right\}. \tag{68}$$

rather than computing the full matrix.

Kingma and Ba provide a method which combines earlier methods AdaGrad and RMSProp to build *Adam*, which is theoretically defined for some approximate g_t s.t. $\mathbb{E}[g_t] = \nabla f(x_t)$ using

$$x_{t+1} = x_t - \eta \frac{\mathbb{E}\left[g_t\right]}{\sqrt{\mathbb{E}\left[{g_t}^2\right]}}. \tag{69}$$

This method can be seen as minimizing a *Signal-to-Noise ratio*, as it uses the gradient's first moment (mean) divided by the gradient's square root of the second moment (un-adjusted standard deviation) as a descent direction. As this value reaches an optima, the mean will tend to decline and become overcome by the noise.

Both $\mathbb{E}\left[g_t\right]$ and $\mathbb{E}\left[{g_t}^2\right]$ are approximated using exponential moving averages, where if $g_t \sim \rho(g_t)$ is the gradient distribution and g_t is selected s.t. $\mathbb{E}[g_t] = \nabla f(x_t)$ then

$$\begin{aligned} m_t &= \beta_1 m_{t-1} + (1 - \beta_1) g_t, \text{ and} \\ v_t &= \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \end{aligned} \tag{70}$$

where m_0 and v_0 are zero-value initialized. Due to this initialization, both terms require bias correction, as we can see by taking the expectation of the closed form solution to v_t (the same applies to m_t)

$$\begin{aligned}\mathbb{E}[v_t] &= \mathbb{E}\left[(1-\beta_2)\sum_{i=1}^{t}\beta_2^{t-i}g_i^2\right] \\ &= \mathbb{E}\left[g_t^2\right](1-\beta_2)\sum_{i=1}^{t}\beta_2^{t-i} + \xi \\ &= \mathbb{E}\left[g_t^2\right](1-\beta_2^t) + \xi.\end{aligned} \tag{71}$$

The bias correction is then employed by including the bias adjustment term to each equation

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}, \quad \text{and} \quad \hat{v}_t = \frac{v_t}{1-\beta_2^t} \tag{72}$$

which is substituted in the gradient descent equation to get

$$x_{t+1} = x_t - \eta\frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\epsilon} \tag{73}$$

where $\epsilon \ll 1$ is used to ensure the denominator never reaches 0. Note the use of ξ in Equation 71. Although this means there will never be 0 bias, the used of a β large enough will ensure that this value is sufficiently small in most cases.

References

[1] Sébastien Gadat, Fabien Panloup, and Sofiane Saadane. Stochastic heavy ball. *Electronic Journal of Statistics*, 12(1):461 – 529, 2018. doi: 10.1214/18-EJS1395. URL `https://doi.org/10.1214/18-EJS1395`.

[2] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL `http://arxiv.org/abs/1412.6980`.

[3] Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

[4] Ioannis Mitliagka. Lecture 6 nesterov's momentum, stochastic gradient descent, Winter 2020.

[5] Yurii Evgen'evich Nesterov. A method of solving a convex programming problem with convergence rate o\bigl(k^2\bigr). In *Doklady Akademii Nauk*, volume 269, pages 543–547. Russian Academy of Sciences, 1983.

[6] Boris Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr Computational Mathematics and Mathematical Physics*, 4:1–17, 12 1964. doi: 10.1016/0041-5553(64)90137-5.