UNIVERSITY OF
ILLINOIS
URBANA-CHAMPAIGN

Analysis of Nesterov’s Accelerated Gradient Descent

AUTHORS

Ashwath Karthikeyan

May 27, 2024

Contents

1 Nesterov's Accelerated Gradient Descent Method 1

1.1 Necessary Conditions 1

1.2 Method 1

1.3 Proving Convergence 2

1.3.1 Proving Lemma 1 2

1.3.2 Proving Lemma 2 3

1.4 Proving the Rate of Convergence 3

References 7

1 Nesterov's Accelerated Gradient Descent Method

Nesterov's method, also known as Nesterov's accelerated gradient (NAG), is a pivotal algorithm in optimization, particularly for tackling problems involving smooth convex functions. Developed by Yurii Nesterov in 1983, this method significantly improves the convergence rates of gradient-based optimization techniques. By incorporating a look-ahead feature, where the algorithm anticipates future gradient values, Nesterov's method is able to take larger, more effective steps towards the optimum solution. This foresight allows it to outpace standard gradient descent by reducing the number of iterations needed to achieve a given accuracy, making it especially beneficial in high-dimensional problem spaces where traditional methods falter. This acceleration is crucial for applications in machine learning, data analysis, and beyond, where fast and efficient optimization algorithms are essential.

1.1 Necessary Conditions

For NAG to converge, there are some necessary conditions the program has to satisfy.

The objective functions is defined as $f : R^n \to R$. $f(x)$ is convex and $f \in C_l^{1,1}(R^n)$, which means it is a C^1 function and the Gradient is Lipschitz continuous, with Lipschitz constant l, i.e.,

$$||\nabla f(x) - \nabla f(u)|| \leq l||x-u|| \; \forall x, u \in R^n \; (l > 0)$$

1.2 Method

Let's now define the NAG Method:

$$\text{Let } g(x) = x - \alpha \nabla f(x) \forall x \in R^n.$$

To begin with, we choose $y_1 = x_0 \in R^n, t_1 = 1$ and $\alpha \in (0, \frac{1}{l}]$

For $k \geq 1$, compute :

$$x_k = g(y_k) = y_k - \alpha \nabla f(y_k)$$
$$t_{k+1} = \frac{1 + \sqrt{1 + 4t_k^2}}{2}$$
$$y_{k+1} = x_k + \frac{t_k - 1}{t_{k+1}}(x_k - x_{k-1})$$

In the momentum-based gradient descent method, the update for y_{k+1} considers both the current position x_k and the previous position x_{k-1}. Unlike traditional gradient descent that relies solely on the current gradient, this method uses a "momentum" term. This term helps the method to build up speed along directions that consistently reduce the loss, allowing it to move faster through shallow areas and smooth out updates. Essentially, it uses past information to make smarter, quicker steps towards the solution, leading to potentially faster convergence.

Convergence Proof of Nesterov's Accelerated Gradient Descent

To demonstrate the convergence of NAG method, we establish the roles of two key lemmas in ensuring that the function values $f(x_k)$ generated by the iterations approach the function value at a minimum $f(x^*)$. This is inspired by the method used in the FISTA paper by Beck and Teboulle.[1]

Lemma 1

The first lemma provides an inequality that expresses the progress made in one iteration of gradient descent:

$$f(g(y)) - f(x) \leq -\frac{\alpha}{2}\|\nabla f(y)\|^2 + \langle \nabla f(y), y - x \rangle$$

where $g(y) = y - \alpha \nabla f(y)$, and α is the step size.

Contribution to Convergence:

1. Gradient Descent Iteration Impact: It indicates that the function value decreases unless the gradient $\nabla f(y)$ is zero, with the decrease quantified by $-\frac{\alpha}{2}\|\nabla f(y)\|^2$.
2. Control by Gradient Norm: The decrease in function value is proportional to the squared norm of the gradient, promoting a faster decrease when far from the minimum.

Lemma 2

The second lemma states:

$$f(g(y)) - f(x) \leq \frac{1}{2\alpha}(\|y - x\|^2 - \|g(y) - x\|^2)$$

demonstrating a "squared distance decrease".

Contribution to Convergence:

1. Contraction Property: The lemma implies that the iterative process contracts the distance between the updated point $g(y)$ and any other point x.
2. Guaranteeing Effective Steps: Each gradient step not only reduces the function value but also methodically approaches the minimum by reducing the error in terms of distance.

1.3 Proving Convergence

We will proceed with proving the two lemmas based on the paper by Beck and Teboulle.[1]

1.3.1 Proving Lemma 1

We know that for the method to converge the function must demonstrate Lipshitz continuity of gradient throughout. So, we have:

$$f(z) \leq f(y) + \langle \nabla f(y), z - y \rangle + \frac{l}{2}||z - y||^2 \;\; \forall\, y, z \in R^n$$

As we discussed above, for Nesterov's AGD,

$$g(y) = y - \alpha \nabla f(y)$$

Substituting for $z = g(y)$, we get:

$$f(g(y)) = f(y - \alpha \nabla f(y)) \leq f(y) + \langle \nabla f(y), -\alpha \nabla f(y) \rangle + \frac{l}{2}|| - \alpha \nabla f(y)||^2$$

$$= f(y) - \alpha||\nabla f(y)||^2 + \frac{l\alpha^2}{2}||\nabla f(y)||^2$$

$$= f(y) - \alpha(1 - \frac{l\alpha}{2})||\nabla f(y)||^2$$

We have mentioned before that

$$0 < \alpha \leq \frac{1}{l} \implies (1 - \frac{l\alpha}{2}) \geq \frac{1}{2}$$

Which gives us

$$f(g(y)) \leq f(y) - \frac{\alpha}{2}||\nabla f(y)||^2$$

We also know that

$$\langle \nabla f(y), x - y \rangle \leq f(x) - f(y) \implies f(y) \leq f(x) + \langle \nabla f(y), y - x \rangle$$

Which makes the inequality:

$$f(g(y)) \leq f(x) + \langle \nabla f(y), y - x \rangle - \frac{\alpha}{2}||\nabla f(y)||^2$$

And finally, rearranging the terms we get:

$$f(g(y)) - f(x) \leq -\frac{\alpha}{2}||\nabla f(y)||^2 + \langle \nabla f(y), y - x \rangle$$

Which is exactly what is stated in Lemma 1.

1.3.2 Proving Lemma 2

We have:

$$\frac{\alpha}{2}||\nabla f(y)||^2 - \langle \nabla f(y), y - x \rangle = \frac{\alpha}{2}(||\nabla f(y)||^2 - \frac{2}{\alpha}\langle \nabla f(y), y - x \rangle)$$

$$= \frac{\alpha}{2}(||\nabla f(y) - \frac{\alpha}{2}(y - x)||^2 - \frac{1}{\alpha^2}||y - x||^2$$

$$= \frac{1}{2\alpha}(||g(y) = x||^2 = ||y - x||^2)$$

$$f(g(y)) - f(x) \leq \frac{1}{2\alpha}(\|y - x\|^2 - \|g(y) - x\|^2)$$

This is exactly Lemma 2.

1.4 Proving the Rate of Convergence

Vanilla Gradient Descent has the rate $O(\frac{1}{k})$. This is fine, but for large flat or only slightly curved surfaces, it takes forever to reach the solution [2].

Nesterov, in his paper[3], states that the rate of convergence of his Accelerated Gradient Descent Method is of the order of $O(\frac{1}{k^2})$.

We will now attempt to prove this. Remember the Lemmas we proved before:

Lemma 1: $f(g(y)) - f(x) \leq -\frac{\alpha}{2}||\nabla f(y)||^2 + \langle \nabla f(y), y - x\rangle$

Lemma 2: $f(g(y)) - f(x) \leq \frac{1}{2\alpha}(||y-x||^2 - ||g(y)-x||^2)$

$$\forall\, x, y \in R^n$$

We know that, by definition:

$$f(x_{k+1}) - f(x_k) = f(g(y_{k+1})) - f(x_k)$$

Plugging in Lemma 1, we get:

$$f(x_{k+1}) - f(x_k) \leq -\frac{\alpha}{2}||\nabla f(y_{k+1})||^2 + \langle \nabla f(y_{k+1}), y_{k+1} - x_k\rangle$$

Also we know:

$$x_{k+1} = y_{k+1} - \alpha\nabla f(y_{k+1}) \implies \nabla f(y_{k+1}) = \frac{y_{k+1} - x_{k+1}}{\alpha}$$

Therefore:

$$f(x_{k+1}) - f(x_k) \leq -\frac{\alpha}{2}||\frac{y_{k+1} - x_{k+1}}{\alpha}||^2 + \langle \frac{y_{k+1} - x_{k+1}}{\alpha}, y_{k+1} - x_k\rangle$$

$$f(x_{k+1}) - f(x_k) \leq -\frac{1}{2\alpha}||y_{k+1} - x_{k+1}||^2 + \frac{1}{\alpha}\langle y_{k+1} - x_{k+1}, y_{k+1} - x_k\rangle \tag{1}$$

Similarly,

$$f(x_{k+1}) - f(x^*) \leq -\frac{1}{2\alpha}||y_{k+1} - x_{k+1}||^2 + \frac{1}{\alpha}\langle y_{k+1} - x_{k+1}, y_{k+1} - x^*\rangle \tag{2}$$

Let us now define $v_k = f(x_k) - f(x^*)$

Putting together (1) and (2), we get

$$v_{k+1} - v_k \leq -\frac{1}{2\alpha}||y_{k+1} - x_{k+1}||^2 + \frac{1}{\alpha}\langle y_{k+1} - x_{k+1}, y_{k+1} - x_k\rangle \tag{3}$$

$$v_{k+1} \leq -\frac{1}{2\alpha}||y_{k+1} - x_{k+1}||^2 + \frac{1}{\alpha}\langle y_{k+1} - x_{k+1}, y_{k+1} - x^*\rangle \tag{4}$$

Multiplying (3) by $(t_{k+1} - 1)$ and adding to (4) we get:

$$t_{k+1}v_{k+1} - (t_{k+1} - 1)v_k \leq -\frac{1}{2\alpha}t_{k+1}||y_{k+1} - x_{k+1}||^2 + \frac{1}{\alpha}\langle y_{k+1} - x_{k+1}, t_{k+1}y_{k+1} - (t_{k+1} - 1)x_k - x^*\rangle \tag{5}$$

Multiplying both sides by (t_{k+1}), we get:

$$t_{k+1}^2v_{k+1} - (t_{k+1}^2 - 1)v_k \leq -\frac{1}{2\alpha}t_{k+1}^2||y_{k+1} - x_{k+1}||^2 + \frac{t_{k+1}}{\alpha}\langle y_{k+1} - x_{k+1}, t_{k+1}y_{k+1} - (t_{k+1} - 1)x_k - x^*\rangle \tag{6}$$

But we know that $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$ from definition of the NAG method. With this we calculate that $t_k^2 = t_{k+1}^2 - t_{k+1}$. This gives us:

$$t_{k+1}^2v_{k+1} - t_k^2v_k \leq -\frac{1}{2\alpha}||t_{k+1}(y_{k+1} - x_{k+1})||^2 + \frac{t_{k+1}}{\alpha}\langle y_{k+1} - x_{k+1}, t_{k+1}y_{k+1} - (t_{k+1} - 1)x_k - x^*\rangle$$

$$= -\frac{1}{2\alpha}(||t_{k+1}(y_{k+1} - x_{k+1})||^2 - 2t_{k+1}\langle y_{k+1} - x_{k+1}, t_{k+1}y_{k+1} - (t_{k+1} - 1)x_k - x^*\rangle)$$

We know that

$$||a||^2 - 2\langle a, b\rangle = ||a - b||^2 - ||b||^2$$

So that gives us

$$t_{k+1}^2 v_{k+1} - t_k^2 v_k \leq -\frac{1}{2\alpha}(||t_{k+1}x_{k+1} - (t_{k+1-1})x_k - x^*||^2 - ||t_{k+1}y_{k+1} - (t_{k+1} - 1)x_k - x^*||^2)$$

We know:

$$t_{k+1}\ y_{k+1} = t_{k+1}x_k + (t_k - 1)(x_k - x_{k-1})$$

Let $u_k = t_k x_k - (t_k - 1)x_{k-1} - x^*$

Then:

$$t_{k+1}^2 v_{k+1} - t_k^2 v_k \leq -\frac{1}{2\alpha}(||u_{k+1}||^2 - ||u_k||^2)$$

Now, let's make a series with this inequality

$$t_2^2 v_2 - t_1^2 v_1 \leq -\frac{1}{2\alpha}(||u_2||^2 - ||u_1||^2)$$

$$t_3^2 v_2 - t_2^2 v_1 \leq -\frac{1}{2\alpha}(||u_3||^2 - ||u_2||^2)$$

$$\vdots$$

$$t_3^2 v_2 - t_2^2 v_1 \leq -\frac{1}{2\alpha}(||u_3||^2 - ||u_2||^2)$$

Adding all these terms together cancels out like terms and gives:

$$t_k^2 v_k - t_1^2 v_1 \leq -\frac{1}{2\alpha}(||u_k||^2 - ||u_1||^2)$$

So,

$$t_k^2 v_k \leq t_1^2 v_1 + \frac{1}{2\alpha}(||u_1||^2 - ||u_k||^2)$$

$$t_k^2 v_k \leq t_1^2 v_1 + \frac{1}{2\alpha}(||u_1||^2)$$

So from the definition,

$$t_1 = 1, u_1 = f(x_1) - f(x^*), u_1 = x_1 - x^*$$

So,

$$t_k^2 v_k \leq 1[f(x_1) - f(x^*)] + \frac{1}{2\alpha}||x_1 - x^*||^2$$

But we know that:

$$f(x_1) - f(x^*) = f(g(y_1)) - f(x^*) \leq \frac{1}{2\alpha}(||y_1 - x^*||^2 - ||x_1 - x^*||^2)$$

Which makes:

$$t_k^2 v_k \leq \frac{1}{2\alpha}(||y_1 - x^*||^2 - ||x_1 - x^*||^2) + \frac{1}{2\alpha}||x_1 - x^*||^2$$

$$t_k^2 v_k \leq \frac{1}{2\alpha}||x_0 - x^*||^2$$

Simplifying this gives us:

$$v_k \leq \frac{||x_0 - x^*||^2}{2\alpha t_k^2}$$

But, we know that

$$t_k \geq \frac{k+1}{2}$$

$$v_k \leq \frac{||x_0 - x^*||^2}{2\alpha(\frac{k+1}{2})^2}$$

This finally gives us:

$$0 \leq f(x_k) - f(x^*) \leq \frac{2||x_0 - x^*||^2}{\alpha(k+1)^2}$$

$$\forall k \in N$$

And when $\alpha = \frac{1}{l}$, we can see the equation become

$$0 \leq f(x_k) - f(x^*) \leq \frac{2l||x_0 - x^*||^2}{a(k+1)^2}$$

This makes the convergence in the order of $O(\frac{1}{k^2})$ which is a great improvement from regular gradient descent.

References

[1] Beck, A. and Teboulle, M., “A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems,” pp. 183–202, 2009.

[2] A. Beck and M. Teboulle, “Mirror descent and nonlinear projected subgradient methods for convex optimization,” pp. 167–175, 2003.

[3] Nesterov, Y., “Introductory Lectures on Convex Optimization: A Basic Course,” 2004.