Backtracking Line Search Method

Dhruv Malik

April 5th 2018

Overview. In the first section, I'm going to first analyze a method which uses backtracking line search in a random direction chosen to be positively correlated with the gradient. I'm going to give an analysis with the standard Armijo condition, and show that at least with the tools that I have, one cannot show convergence. I'm going to explain why this difficulty arises. Then in the second section, I will then use an algorithm that I just made, inspired by these difficulties. I will explain why I think my method is strictly better than Armijo's work, and show linear convergence in expectation.

1 Armijo's Method

Setup. Consider a function $f(x) : \mathbb{R}^d \mapsto \mathbb{R}$ which has L-Lipschitz gradients:

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2} \|y - x\|_2^2 \tag{1}$$

and also satisfies γ-strong convexity:

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\gamma}{2} \|y - x\|_2^2 \tag{2}$$

Armijo Condition. At timestep t, given a direction w, $\alpha_0 > 0$ and $\beta \in (0, 1)$, set step size $\alpha_t = \alpha_0 \beta^i$ for the smallest integer $i \geq 0$ such that:

$$f(x_t - \alpha_t w) \leq f(x_t) - \frac{\alpha_t}{2} w^\top \nabla f(x_t) \tag{3}$$

Method. Consider the following zero order optimization method: we randomly sample a direction r from a uniform distribution over the hypersphere with unit radius which is centered at the origin. Define w to be the following random variable:

$$w = \begin{cases} r, & \text{if } r^\top \nabla f(x) \geq 0 \\ -r, & \text{if } r^\top \nabla f(x) < 0 \end{cases} \tag{4}$$

At any time step t, given this direction w, determine the value of α_t in the manner described by the Armijo condition. The update step is then:

$$x_{t+1} = x_t - \alpha_t w \tag{5}$$

Lemma 1. *If the unit vector w is selected using the procedure above, then:*

$$\mathbb{E}[w^\top \nabla f(x)] = O\left(\frac{\|\nabla f(x)\|_2}{\sqrt{d}} \right) \tag{6}$$

Proof. See the other document. □

Lemma 2. *If we have:*

$$\alpha > \frac{\|\nabla f(x)\|_2}{2\gamma} \tag{7}$$

then the Armijo condition is not satisfied.

Proof. Consider the following series of implications:

$$\begin{align}
\alpha &> \frac{\|\nabla f(x)\|_2}{2\gamma} \tag{8}\\
\alpha &\geq \frac{w^\top \nabla f(x)}{2\gamma} \tag{9}\\
\gamma\alpha &\geq \frac{w^\top \nabla f(x)}{2} \tag{10}\\
\gamma\alpha^2 &\geq \alpha\frac{w^\top \nabla f(x)}{2} \tag{11}\\
-\alpha w^\top \nabla f(x) + \gamma\alpha^2 &\geq -\alpha\frac{w^\top \nabla f(x)}{2} \tag{12}\\
f(x) - \alpha w^\top \nabla f(x) + \gamma\alpha^2 &\geq f(x) - \alpha\frac{w^\top \nabla f(x)}{2} \tag{13}
\end{align}$$

By strong convexity, we then have that:

$$\begin{align}
f(x - \alpha w) &\geq f(x) + (x - \alpha w - x)^\top \nabla f(x) + \gamma \left\|x - \alpha w - x\right\|_2^2 \tag{14}\\
&= f(x) - \alpha w^\top \nabla f(x) + \gamma\alpha^2 \tag{15}\\
&> f(x) - \alpha\frac{w^\top \nabla f(x)}{2} \tag{16}
\end{align}$$

This clearly shows that the Armijo condition is not satisfied. □

Lemma 3. *If we have:*

$$\alpha \leq \frac{w^\top \nabla f(x)}{L} \tag{17}$$

then the Armijo condition is satisfied.

Proof. Consider the following series of implications:

$$\begin{align}
\alpha &\leq \frac{w^\top \nabla f(x)}{L} \tag{18}\\
\frac{L}{2}\alpha^2 &\leq \frac{\alpha}{2} w^\top \nabla f(x) \tag{19}\\
-\alpha w^\top \nabla f(x) + \frac{L}{2}\alpha^2 &\leq -\frac{\alpha}{2} w^\top \nabla f(x) \tag{20}\\
f(x) - \alpha w^\top \nabla f(x) + \frac{L}{2}\alpha^2 &\leq f(x) - \frac{\alpha}{2} w^\top \nabla f(x) \tag{21}
\end{align}$$

By Lipschitz gradients, we have the following:

$$\begin{aligned} f(x-\alpha w) &\leq f(x) + (x-\alpha w - x)^\top \nabla f(x) + \frac{L}{2}\left\|x-\alpha w - x\right\|_2^2 && (22)\\ &= f(x) - \alpha w^\top \nabla f(x) + \frac{L}{2}\alpha^2 && (23)\\ &\leq f(x) - \frac{\alpha}{2} w^\top \nabla f(x) && (24) \end{aligned}$$

This clearly shows that the Armijo condition is satisfied. □

Lemma 4. *At any time step t, we have that:*

$$\alpha \geq \beta \frac{w^\top \nabla f(x)}{L} \tag{25}$$

Proof. This follows directly from Lemma 3 and the fact that we select α_t in a manner that takes the largest step size such that the Armijo condition is satisfied. □

Before we jump to the real theorem, I'm going to quickly give the result for Armijo's method where we are not picking directions randomly, and instead doing line search in the exact direction of the gradient. This is so I can illustrate the differences between this and the random direction method that I described above. This proof was not given in the resource that Ashwin sent.

Theorem 1. *Pick $\alpha_0 = \frac{1}{L}$ and $\beta > \frac{3}{4}$. If we line search in the exact direction of the gradient and satisfy Armijo's condition with $w = \nabla f(x)$ at all time steps, then we have linear convergence.*

Proof. We begin by using the property of Lipschitz gradients and then apply Lemma 4:

$$\begin{aligned} f(x_{t+1}) &= f(x_t - \alpha w) && (26)\\ &= f(x_t - \alpha \nabla f(x_t)) && (27)\\ &\leq f(x_t) - \alpha_t (\nabla f(x_t))^\top \nabla f(x_t) + \frac{L}{2}\alpha_t^2 \left\|\nabla f(x)\right\|_2^2 && (28)\\ &\leq f(x_t) - \frac{\beta}{L}\left\|\nabla f(x_t)\right\|_2^2 + \frac{1}{2L}\left\|\nabla f(x)\right\|_2^2 && (29)\\ &\leq f(x_t) - \frac{3}{4L}\left\|\nabla f(x_t)\right\|_2^2 + \frac{2}{4L}\left\|\nabla f(x)\right\|_2^2 && (30)\\ &\leq f(x_t) - \frac{1}{4L}\left\|\nabla f(x_t)\right\|_2^2 && (31) \end{aligned}$$

The theorem follows by noting that we have the PL Inequality and then following the standard argument - note that we don't need strong convexity for this theorem. □

Now let's move on to analyzing our random direction method. Observe the key step above of picking $\alpha_0 = \frac{1}{L}$. Obviously this is not useful in our method, since it leaves a constant that cannot be dealt with. Instead I use Lemma 2 (from strong convexity) to get an upper bound on α_t.

Theorem 2. *If we use the above method, then we can only get linear convergence with $d \lessapprox 8$.*

Proof. We begin by using the property of Lipschitz gradients:

$$\begin{align}
f(x_{t+1}) &= f(x_t - \alpha w) \tag{32}\\
&\leq f(x_t) - \alpha_t w^\top \nabla f(x_t) + \frac{L}{2}\alpha_t^2 \tag{33}\\
&\leq f(x_t) - \frac{\beta}{L}(w^\top \nabla f(x_t))^2 + \frac{L}{2}\alpha_t^2 \tag{34}\\
&\leq f(x_t) - \frac{\beta}{L}(w^\top \nabla f(x_t))^2 + \frac{L}{2}\left(\frac{\|\nabla f(x)\|_2}{2\gamma}\right)^2 \tag{35}
\end{align}$$

where the last two inequalities follow from Lemma 2 and Lemma 4. Now, in expectation:

$$\begin{align}
\mathbb{E}[f(x_{t+1})] &\leq \mathbb{E}\left[f(x_t) - \frac{\beta}{L}(w^\top \nabla f(x_t))^2 + \frac{L}{2}\left(\frac{\|\nabla f(x_t)\|_2}{2\gamma}\right)^2\right] \tag{36}\\
&\leq f(x_t) - \frac{\beta}{L}(\mathbb{E}[w^\top \nabla f(x_t)])^2 + \mathbb{E}\left[\frac{L}{2}\left(\frac{\|\nabla f(x_t)\|_2}{2\gamma}\right)^2\right] \tag{37}\\
&= f(x_t) - \frac{\beta}{L}\left(\frac{\|\nabla f(x_t)\|_2}{\sqrt{d}}\right)^2 + \frac{L}{2}\left(\frac{\|\nabla f(x_t)\|_2}{2\gamma}\right)^2 \tag{38}\\
&\leq f(x_t) - \frac{\beta}{L}\left(\frac{\|\nabla f(x_t)\|_2}{\sqrt{d}}\right)^2 + \frac{L}{2}\left(\frac{\|\nabla f(x_t)\|_2}{2\gamma}\right)^2 \tag{39}\\
&= f(x_t) + \frac{L^2 d - 8\beta\gamma^2}{8Ld\gamma^2}\,\|\nabla f(x_t)\|_2^2 \tag{40}
\end{align}$$

The theorem follows because we need the constant term to be negative. Clearly this is a very weak result. It can probably be improved, but the key is that we would need a stricter upper bound on the value of α_t. Since I can't use $\alpha_0 = \frac{1}{L}$, I tried to use strong convexity, although this is also too loose as we can clearly see. So why not pick α_0 to be something related to the inner product or the norm of the gradient itself? This would deal with everything. But at this point, we are basically picking the step size, and it's not really backtracking line search in it's true sense. This leads to my method in the next section. □

2 My Method

Discussion. The reason why the analysis above failed is because my upper bound on the value of α_t was too loose. The issue is that we have no idea when the Armijo condition will actually be satisfied. The best we can do is use strong convexity to see when it is certainly not satisfied, but as we observed this does not give us much. If we had a way to get a tighter upper bound on α_t, then things might be different.

Now, let's go a different route. It seems to me that the main reason we wanted to look in this avenue is so that we get linear convergence in expectation. While trying to get stuff to work earlier, I discovered this new method, which gives us linear convergence in expectation. Here's why I think my method is better than trying to satisfy the Armijo condition given earlier:

- Satisfying the Armijo condition requires us to know the gradient, or at least know the value of $w^\top \nabla f(x)$. In my method, we need to know the same information.

- Satisfying the Armijo condition requires us to use functional evaluations as we try different values of α_t. My method does not require any functional evaluations.

This suggests to me that my method is strictly better.

Setup. Consider a function $f(x) : \mathbb{R}^d \mapsto \mathbb{R}$ which has L-Lipschitz gradients:

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2} \left\| y - x \right\|_2^2 \tag{41}$$

and also satisfies the PL Inequality:

$$\frac{1}{2} \left\| \nabla f(x) \right\|_2^2 \geq \mu(f(x) - f(x^*)) \tag{42}$$

Modified Step Size Condition. At any time step t, select step size α_t in the following manner:

$$\alpha_t = \frac{w^\top \nabla f(x)}{Ld} \tag{43}$$

Method. Consider the following zero order optimization method: we randomly sample a direction r from a uniform distribution over the hypersphere with unit radius which is centered at the origin. Define w to be the following random variable:

$$w = \begin{cases} r, & \text{if } r^\top \nabla f(x) \geq 0 \\ -r, & \text{if } r^\top \nabla f(x) < 0 \end{cases} \tag{44}$$

Then, follow the update rule $x_{t+1} = x_t - \alpha_t w$. We will show linear convergence in expectation.

Theorem 3. *If we follow the above method, then we get that:*

$$\mathbb{E}[f(x_{t+1}) - f(x^*)] \leq \mathbb{E}\left[\left(1 - \frac{u}{Ld^2}\right)(f(x_t) - f(x^*))\right] \tag{45}$$

Proof. Given a fixed x_t:

$$\begin{aligned}
\mathbb{E}[f(x_{t+1})] &\leq \mathbb{E}[f(x_t) - \alpha_t w^\top \nabla f(x_t) + \frac{L}{2}\alpha_t^2] && (46)\\
&\leq \mathbb{E}\left[f(x_t) - \frac{(w^\top \nabla f(x_t))^2}{Ld} + \frac{L}{2}\left(\frac{w^\top \nabla f(x_t)}{Ld}\right)^2\right] && (47)\\
&\leq f(x_t) - \frac{1}{Ld}(\mathbb{E}(w^\top \nabla f(x_t)))^2 + \frac{L}{2}\left(\frac{\left\| \nabla f(x_t) \right\|_2}{Ld}\right)^2 && (48)\\
&= f(x_t) - \frac{1}{Ld}\left(\frac{\left\| \nabla f(x_t) \right\|_2}{\sqrt{d}}\right)^2 + \frac{L}{2}\left(\frac{\left\| \nabla f(x_t) \right\|_2}{Ld}\right)^2 && (49)\\
&= f(x_t) - \frac{1}{Ld^2}\left\| \nabla f(x_t) \right\|_2^2 + \frac{1}{2Ld^2}\left\| \nabla f(x_t) \right\|_2^2 && (50)\\
&= f(x_t) - \frac{1}{2Ld^2}\left\| \nabla f(x_t) \right\|_2^2 && (51)\\
&\leq f(x_t) - \frac{1}{Ld^2}\mu(f(x_t) - f(x^*)) && (52)
\end{aligned}$$

The theorem follows directly from this. □