Analysis of Newton's Method

Raghav Somani

October 12, 2019

Consider a function $f : \mathbb{R}^d \to \mathbb{R}$ which is strongly convex, smooth, twice differentiable and second order smooth. We are interested in minimizing f by steepest descent using its second order derivatives. Newton's method can be viewed as minimizing the second order approximation of the function at every iterate. Because f is twice differentiable and has smooth second derivatives, one can consider its Taylor series approximation till the second order term to directly and minimize the approximation hoping to also minimize the true function f since the second order derivatives are smooth.

$$\begin{aligned}\arg\min_{\Delta\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}+\Delta\mathbf{x}) &\approx \arg\min_{\Delta\mathbf{x}\in\mathbb{R}^d}\left\{f(\mathbf{x}) + \langle\nabla f(\mathbf{x}), \Delta\mathbf{x}\rangle + \frac{1}{2}\left\langle\Delta\mathbf{x}, \nabla^2 f(\mathbf{x})\Delta\mathbf{x}\right\rangle\right\} \\ &= -\nabla^2 f(\mathbf{x})^{-1}\nabla f(\mathbf{x}) \end{aligned} \tag{0.1}$$

The Newton step therefore has a closed form expression that can be computed given access to the second order derivatives of f. $\Delta\mathbf{x}$ is essentially the steepest descent made according the the norm $\|\cdot\|_{\nabla^2 f(\mathbf{x})}$ at $\mathbf{x}$. The decrease in the function value is therefore

$$f(\mathbf{x}) - \min_{\Delta\mathbf{x}\in\mathbb{R}^d}\left\{f(\mathbf{x}) + \langle\nabla f(\mathbf{x}), \Delta\mathbf{x}\rangle + \frac{1}{2}\left\langle\Delta\mathbf{x}, \nabla^2 f(\mathbf{x})\Delta\mathbf{x}\right\rangle\right\} = \frac{1}{2}\left\langle\nabla f(\mathbf{x}), \nabla^2 f(\mathbf{x})^{-1}\nabla f(\mathbf{x})\right\rangle \tag{0.2}$$

Since f is strongly convex, the decrease is strictly positive. The Newton's algorithm is therefore an iterative application of this update.

1 Setup

The function $f : \mathbb{R}^d \to \mathbb{R}$ is assumed to satisfy

- Strong convexity and Smoothness: $\exists\ \mu, L \in \mathbb{R}_{++} \ni \mu \leq L, L\mathbf{I} \succeq \nabla^2 f(\mathbf{x}) \succeq \mu\mathbf{I}\ \forall\ \mathbf{x} \in \mathbb{R}^d$, and
- Second order smoothness: $\exists\ \rho \in \mathbb{R}_{++} \ni \left\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\right\|_2 \leq \rho.\left\|\mathbf{x}-\mathbf{y}\right\|_2\ \forall\ \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

The second order smoothness parameter ρ essentially controls how well can the function f be approximated by its quadratic approximation.

Algorithm 1: Newton's algorithm $\left(\mathbf{x}_1, \epsilon, \{\eta_t\}_{t\in\mathbb{N}}\right)$

```
for t = 1, 2, ... do
    λ_t^2 = ∇f(x_t)^T ∇^2 f(x_t)^{-1} ∇f(x_t)
    if λ_t^2/2 < ε then
        return x_t
    end
    x_{t+1} = x_t − η_t ∇^2 f(x_t)^{-1} ∇f(x_t)
end
```

2 Analysis

The analysis of the Newton's method can be broken down into two phases - damped and quadratically convergent, which we will see why.
For an parameter $\gamma > 0$, which we will choose later, assume that in the damped phase $\|\nabla f(\mathbf{x})\|_2 \geq \gamma$. From strong convexity of f, we have

$$\begin{aligned} f(\mathbf{x}_{t+1}) &\leq f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{L}{2} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|_2^2 \\ &= f(\mathbf{x}_t) - \eta_t \left\langle \nabla f(\mathbf{x}_t), \nabla^2 f(\mathbf{x}_t)^{-1} \nabla f(\mathbf{x}_t) \right\rangle + \frac{L\eta_t^2}{2} \left\| \nabla^2 f(\mathbf{x}_t)^{-1} \nabla f(\mathbf{x}_t) \right\|_2^2 \\ &\leq f(\mathbf{x}_t) - \left(\eta_t - \frac{L\eta_t^2}{2\mu} \right) \lambda_t^2 \qquad \text{(Using strong convexity of } f\text{, and definition of } \lambda_t^2 \text{ in Algorithm 1).} \end{aligned} \tag{2.1}$$

Setting $\eta_t = \frac{\mu}{L}$, Equation (2.1) becomes

$$\begin{aligned} f(\mathbf{x}_{t+1}) &\leq f(\mathbf{x}_t) - \frac{\mu}{2L} \lambda_t^2 \\ &\leq f(\mathbf{x}_t) - \frac{\mu}{2L^2} \left\| \nabla^2 f(\mathbf{x}_t) \right\|_2^2 \qquad \left(\text{Since } L^{-1}\mathbf{I} \preceq \nabla^2 f(\mathbf{x}_t)^{-1}\right) \\ &\leq f(\mathbf{x}_t) - \frac{\mu}{2L^2} \gamma^2. \end{aligned} \tag{2.2}$$

Let $\mathbf{x}^*$ be the minimizer of f. And since the decrease in the function value is at-least a constant, the number of iterations cannot exceed $\frac{2L^2(f(\mathbf{x}_0) - f(\mathbf{x}^*))}{\gamma^2 \mu}$.
Now, let us considering the quadratically convergent phase when $\|\nabla f(\mathbf{x}_t)\|_2 < \gamma$. Let $\Delta_t := -\nabla^2 f(\mathbf{x}_t) \nabla f(\mathbf{x}_t)$, then

$$\begin{aligned} & \left\| \nabla^2 f(\mathbf{x}_t + \eta_t \Delta_t) - \nabla^2 f(\mathbf{x}_t) \right\|_2 = \eta_t \rho \|\Delta_t\|_2^2 \qquad \text{(From second order smoothness of } f\text{)} \\ \implies & \Delta_t^T \left(\nabla^2 f(\mathbf{x}_t + \eta_t \Delta_t) - \nabla^2 f(\mathbf{x}_t) \right) \Delta_t \leq \eta_t \rho \|\Delta_t\|_2^3. \end{aligned} \tag{2.3}$$

Consider f along the direction Δ_t from $\mathbf{x}_t$ as a new function $g(\eta_t) := f(\mathbf{x}_t + \eta_t \Delta_t)$. Clearly, $\nabla^2 g(\eta_t) = \Delta_t^T \nabla^2 f(\mathbf{x}_t + \eta_t \Delta_t) \Delta_t$, $\nabla^2 g(0) = \lambda_t^2$, and $\nabla g(0) = -\lambda_t^2$. Now Equation (2.3) is nothing but

$$\begin{aligned} \nabla^2 g(\eta_t) &\leq \nabla^2 g(0) + \eta_t \rho \|\Delta_t\|_2^3 \\ &= \lambda_t^2 + \eta_t \rho \|\Delta_t\|_2^3 \\ &\leq \lambda_t^2 + \eta_t \rho \frac{\lambda_t^3}{\mu^{3/2}}. \end{aligned} \tag{2.4}$$

Integrating Equation (2.4) twice, we get

$$g(\eta_t) \leq g(0) - \eta_t \lambda_t^2 + \frac{\eta_t^2}{2} \lambda_t^2 + \frac{\eta_t^3 \rho}{6\mu^{3/2}} \lambda_t^3. \tag{2.5}$$

Setting $\eta_t = 1$, Equation (2.5) becomes

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \lambda_t^2 \left(\frac{1}{2} - \frac{\rho}{6\mu^{3/2}} \lambda_t \right). \tag{2.6}$$

From the definition of λ_t, we have

$$\begin{aligned} \lambda_t^2 &= \nabla f(\mathbf{x}_t) \nabla^2 f(\mathbf{x}_t)^{-1} \nabla f(\mathbf{x}_t) \\ &\leq \frac{1}{\mu} \|\nabla f(\mathbf{x}_t)\|_2^2 \\ \implies \lambda_t &\leq \frac{\|\nabla f(\mathbf{x})\|_2}{\mu^{1/2}}. \end{aligned} \tag{2.7}$$

Using Equation (2.5) in Equation (2.6), we get

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \lambda_t^2 \left(\frac{1}{2} - \frac{\rho}{6\mu^2} \|\nabla f(\mathbf{x}_t)\|_2 \right)$$

$$< f(\mathbf{x}_t) - \lambda_t^2 \left(\frac{1}{2} - \frac{\rho}{6\mu^2}\gamma \right). \tag{2.8}$$

Setting $\gamma \leq \frac{3\mu^2}{2\rho}$, Equation (2.8) becomes

$$\begin{aligned} f(\mathbf{x}_{t+1}) &< f(\mathbf{x}_t) - \frac{1}{4}\lambda_t^2 \\ &= f(\mathbf{x}_t) - \frac{1}{4}\nabla f(\mathbf{x}_t)\nabla^2 f(\mathbf{x}_t)^{-1}\nabla f(\mathbf{x}_t). \end{aligned} \tag{2.9}$$

Therefore, setting $\eta_t = 1$ ensures a positive decrease in the function value. Using the second order smoothness condition, we can bound the gradient at the next step as

$$\begin{aligned} \|\nabla f(\mathbf{x}_{t+1})\|_2 &= \left\|\nabla f(\mathbf{x}_{t+1}) - \nabla f(\mathbf{x}_t) - \nabla^2 f(\mathbf{x}_t)\Delta_t\right\|_2 \\ &= \left\| \int_{\eta_t=0}^{\eta_t=1} \left(\nabla^2 f(\mathbf{x}_t + \eta_t\Delta_t) - \nabla f(\mathbf{x}_t)\right)\Delta_t d\eta_t \right\|_2 \\ &\leq \frac{\rho}{2}\|\Delta_t\|_2^2 \\ &= \frac{\rho}{2}\left\|\nabla^2 f(\mathbf{x}_t)^{-1}\nabla f(\mathbf{x}_t)\right\|_2^2 \\ &\leq \frac{\rho}{2\mu^2}\|\nabla f(\mathbf{x}_t)\|_2^2. \end{aligned} \tag{2.10}$$

Therefore if $\gamma = \frac{\mu^2}{\rho}$ ensures the shrinkage of gradient norms in Equation (2.10). We now have the recurrence

$$\frac{\rho}{2\mu^2}\|\nabla f(\mathbf{x}_{t+1})\|_2 \leq \left(\frac{\rho}{2\mu^2}\|\nabla f(\mathbf{x}_t)\|_2\right)^2. \tag{2.11}$$

Applying Equation (2.11) recursively for we have that for $T \geq t$,

$$\frac{\rho}{2\mu^2}\|\nabla f(\mathbf{x}_T)\|_2 \leq \left(\frac{\rho}{2\mu^2}\|\nabla f(\mathbf{x}_t)\|_2\right)^{2^{T-t}} \leq \left(\frac{1}{2}\right)^{2^{T-t}}. \tag{2.12}$$

From the strong convexity of f, we have

$$\begin{aligned} f(\mathbf{x}^*) &\geq f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{x}^* - \mathbf{x}_t\rangle + \frac{\mu}{2}\|\mathbf{x}^* - \mathbf{x}_t\|_2^2 \\ &\geq \min_{\mathbf{y}\in\mathbb{R}^d}\left\{ f(\mathbf{x}_t) + \langle \nabla f(\mathbf{x}_t), \mathbf{y} - \mathbf{x}_t\rangle + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}_t\|_2^2 \right\} \\ &= f(\mathbf{x}_t) - \frac{1}{2\mu}\|\nabla f(\mathbf{x}_t)\|_2^2. \end{aligned} \tag{2.13}$$

Using Equation (2.13) in Equation (2.12) we get

$$f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq \frac{1}{2\mu}\|\nabla f(\mathbf{x}_t)\|_2 \leq \frac{2\mu^3}{\rho^2}\left(\frac{1}{2}\right)^{2^{T-t}}. \tag{2.14}$$

Therefore, for $T = \lg\lg\frac{2\mu^3}{\rho^2\epsilon}$, we have $f(\mathbf{x}_T) - f(\mathbf{x}^*) \leq \epsilon$. Therefore by adding the iterations required in the two phases, we get a total of

$$\frac{2L^2\rho^2}{\mu^5}(f(\mathbf{x}_0) - f(\mathbf{x}^*)) + \lg\lg\frac{2\mu^3}{\rho^2\epsilon} \tag{2.15}$$

iterations. It is to note that if $\|\nabla f(\mathbf{x}_0)\|_2 \leq \frac{\mu^2}{\rho}$, then the algorithm doesn't have to go through the first phase and the rate of convergence is purely quadratic which makes the effective iteration complexity to be

$$\frac{2L^2\rho^2}{\mu^5}(f(\mathbf{x}_0) - f(\mathbf{x}^*))\mathbb{1}\left(\|\nabla f(\mathbf{x}_0)\|_2 > \frac{\mu^2}{\rho}\right) + \lg\lg\frac{2\mu^3}{\rho^2\epsilon}. \tag{2.16}$$

References

[1] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.