Columbia University in the City of New York
Optimization Methods for Machine Learning
Instructors: Satyen Kale
Authors: Chao Qin
Email: cq2199@columbia.edu

SCRIBE

22

Newton's Method

In this lecture, we study the Newton's method over $K = \mathbb{R}^d$.

Algorithm 1: Newton's Method

1. Start with arbitrary $\mathbf{x}_0$
2. For $t = 0, 1, \ldots, T-1$,
   (a) Compute $\lambda(\mathbf{x}_t) = \nabla f(\mathbf{x}_t)^\top \nabla^2 f(\mathbf{x}_t)^{-1} \nabla f(\mathbf{x}_t)$.
   (b) If $\lambda(\mathbf{x}_t) \geq \frac{\alpha^4}{\beta\gamma^2}$, then set $\eta_t = \frac{\alpha}{\beta}$, else set $\eta_t = 1$.
   (c) Set $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \nabla^2 f(\mathbf{x}_t)^{-1} \nabla f(\mathbf{x}_t)$
3. Output $\mathbf{x}_T$

The standard Newton's method uses no step sizes (i.e. $\eta_t = 1$ for all t. However, this can be shown to converge only when $\mathbf{x}_0$ is very close to the optimal point $\mathbf{x}^*$. To fix this issue we add a step size in the above algorithm. This step size is determined based on the value of the so-called *Newton decrement*, i.e. $\lambda(\mathbf{x}_t) := \nabla f(\mathbf{x}_t)^\top \nabla^2 f(\mathbf{x}_t)^{-1} \nabla f(\mathbf{x}_t)$. If $\lambda(\mathbf{x}_t) \geq \frac{\alpha^4}{\beta\gamma^2}$, (definitions of α, β, γ to follow), then we set $\eta_t = \frac{\alpha}{\beta}$, else we set $\eta_t = 1$.

To analyze the algorithm, we need to make the following assumptions.

1. f is α-strongly convex and β-smooth.
2. $\nabla^2 f$ is γ-Lipschitz, i.e.,
   $$\left\| \nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{x}') \right\| \leq \gamma \left\| \mathbf{x} - \mathbf{x}' \right\|$$
   The matrix norm on the LHS is the spectral norm.

1. Analysis

By β-smoothness,

$$\begin{aligned} f(\mathbf{x}_{t+1}) &\le f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top \left(-\eta_t \nabla^2 f(\mathbf{x}_t)^{-1} \nabla f(\mathbf{x}_t)\right) + \frac{\beta}{2}\eta_t^2 \left\| \nabla^2 f(\mathbf{x}_t)^{-1} \nabla f(\mathbf{x}_t) \right\|^2 \\ &= f(\mathbf{x}_t) - \eta_t \nabla f(\mathbf{x}_t)^\top \nabla^2 f(\mathbf{x}_t)^{-1} \nabla f(\mathbf{x}_t) + \frac{\beta}{2}\eta_t^2 \nabla f(\mathbf{x}_t)^\top \nabla^2 f(\mathbf{x}_t)^{-2} \nabla f(\mathbf{x}_t) \\ &\le f(\mathbf{x}_t) - \eta_t \nabla f(\mathbf{x}_t)^\top \nabla^2 f(\mathbf{x}_t)^{-1} \nabla f(\mathbf{x}_t) + \frac{\beta}{2\alpha}\eta_t^2 \nabla f(\mathbf{x}_t)^\top \nabla^2 f(\mathbf{x}_t)^{-1} \nabla f(\mathbf{x}_t) \\ &= f(\mathbf{x}_t) - \eta_t \lambda(\mathbf{x}_t) + \frac{\beta}{2\alpha}\eta_t^2 \lambda(\mathbf{x}_t) \end{aligned}$$

where $\lambda(\mathbf{x}_t) \triangleq \nabla f(\mathbf{x}_t)^\top \nabla^2 f(\mathbf{x}_t)^{-1} \nabla f(\mathbf{x}_t)$ and the last inequality follows from $\nabla^2 f(\mathbf{x}_t)^{-1} \preceq \frac{1}{\alpha} I$ due to α-sc. By choosing $\eta_t = \frac{\alpha}{\beta}$, we have

$$f(\mathbf{x}_{t+1}) \le f(\mathbf{x}_t) - \frac{\alpha}{2\beta}\lambda(\mathbf{x}_t). \tag{1}$$

In addition,

$$\lambda(\mathbf{x}_t) = \nabla f(\mathbf{x}_t)^\top \nabla^2 f(\mathbf{x}_t)^{-1} \nabla f(\mathbf{x}_t) \ge \frac{1}{\beta} \left\| \nabla f(\mathbf{x}_t) \right\|^2 \ge \frac{\alpha^2}{\beta} \left\| \mathbf{x}_t - \mathbf{x}^* \right\|^2 \tag{2}$$

where the first inequality uses $\nabla^2 f(\mathbf{x}_t)^{-1} \succeq \frac{1}{\beta} I$ due to β-smoothness and the other inequality follows from $\left\| \nabla f(\mathbf{x}_t) \right\| = \left\| \nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}^*) \right\| \ge \alpha \left\| \mathbf{x}_t - \mathbf{x}^* \right\|$ due to α-sc.

Depending on the value of $\lambda(\mathbf{x}_t)$, the analysis of the algorithm factors neatly into two cases. In the first case, when the iterates $\mathbf{x}_t$ are far from the optimal point $\mathbf{x}^*$, then $\lambda(\mathbf{x}_t)$ is large (at least $\frac{\alpha 4}{\beta \gamma^2}$) and then we set $\eta_t = \frac{\alpha}{\beta}$. This is called the *damped Newton phase* of the algorithm since the Newton step $\nabla^2 f(\mathbf{x}_t)^{-1} \nabla f(\mathbf{x}_t)$ is damped by a factor of $\frac{\alpha}{\beta}$ before doing the update. We will show in the analysis that the damped Newton phase lasts for only a *constant* number of steps. Then, $\mathbf{x}_t$ becomes close enough to $\mathbf{x}^*$, at which point $\lambda(\mathbf{x}_t)$ becomes small enough so that $\eta_t = 1$. This is called the *quadratically convergent phase* since at this points the algorithm converges *doubly exponentially fast* to the optimum point: i.e. in order to reach ϵ suboptimality, we need only $O(\log(\log(\frac{1}{\epsilon})))$ steps in this phase. The detailed analysis follows.

1. **Damped Newton phase.** If $\lambda(\mathbf{x}_t) \ge \frac{\alpha^4}{\beta \gamma^2}$, we set $\eta_t = \frac{\alpha}{\beta}$.
   By Equation (1),
   $$f(\mathbf{x}_{t+1}) \le f(\mathbf{x}_t) - \frac{\alpha^5}{2\beta^2\gamma^2}$$
   Thus, the function value reduces by a constant amount, $\frac{\alpha^5}{2\beta^2\gamma^2}$, for each iteration in this phase. Thus, the number of iterations in this phase is bounded by $\frac{2\beta^2\gamma^2(f(\mathbf{x}_0) - f(\mathbf{x}^*))}{\alpha^5}$. Typically, we have a finite lower bound on $f(\mathbf{x}^*)$ (generally, this lower bound is just 0) so this bound on the number of iterations in this phase of the algorithm is just a *constant*.

2. **Quadratically convergent phase.** If $\lambda(\mathbf{x}_t) < \frac{\alpha^4}{\beta \gamma^2}$, we set $\eta_t = 1$.
   By Equation (2),
   $$\left\| \mathbf{x}_t - \mathbf{x}^* \right\| < \frac{\alpha}{\gamma}.$$

Notice that

$$\begin{aligned}
\mathbf{x}_{t+1} - \mathbf{x}^* &= \mathbf{x}_t - \mathbf{x}^* - \nabla^2 f(\mathbf{x}_t)^{-1} \nabla f(\mathbf{x}_t) \\
&= \nabla^2 f(\mathbf{x}_t)^{-1} \left[\nabla^2 f(\mathbf{x}_t)(\mathbf{x}_t - \mathbf{x}^*) - \nabla f(\mathbf{x}_t) \right] \\
&= \nabla^2 f(\mathbf{x}_t)^{-1} \left[\nabla^2 f(\mathbf{x}_t)(\mathbf{x}_t - \mathbf{x}^*) - \int_{u=0}^{1} \nabla^2 f(\mathbf{x}^* + u(\mathbf{x}_t - \mathbf{x}^*))(\mathbf{x}_t - \mathbf{x}^*) \mathrm{d}u \right] \\
&= \nabla^2 f(\mathbf{x}_t)^{-1} \int_{u=0}^{1} \left[\nabla^2 f(\mathbf{x}_t) - \nabla^2 f(\mathbf{x}^* + u(\mathbf{x}_t - \mathbf{x}^*)) \right] (\mathbf{x}_t - \mathbf{x}^*) \mathrm{d}u
\end{aligned}$$

The penultimate equality follows using $\nabla f(\mathbf{x}_t) = \int_{u=0}^{1} \nabla^2 f(\mathbf{x}^* + u(\mathbf{x}_t - \mathbf{x}^*))(\mathbf{x}_t - \mathbf{x}^*) du$ by the fundamental theorem of calculus. This in turn is based on the fact that $\frac{d[\nabla f(\mathbf{x}^* + u(\mathbf{x}_t - \mathbf{x}^*))]}{du} = \nabla^2 f(\mathbf{x}^* + u(\mathbf{x}_t - \mathbf{x}^*))(\mathbf{x}_t - \mathbf{x}^*)$ by the chain rule. Now we can upper bound $\| x_{t+1} - \mathbf{x}^* \|$, by using the Cauchy-Schwarz inequality, the sub-multiplicativity of the spectral norm of matrices, and subadditivity of the ℓ_2 norm on the RHS as follows:

$$\begin{aligned}
\| \mathbf{x}_{t+1} - \mathbf{x}^* \| &\leq \left\| \nabla^2 f(\mathbf{x}_t)^{-1} \right\| \| \mathbf{x}_t - \mathbf{x}^* \| \int_{u=0}^{1} \left\| \nabla^2 f(\mathbf{x}_t) - \nabla^2 f(\mathbf{x}^* + u(\mathbf{x}_t - \mathbf{x}^*)) \right\| \mathrm{d}u \\
&\leq \frac{1}{\alpha} \| \mathbf{x}_t - \mathbf{x}^* \| \int_{u=0}^{1} \gamma \| \mathbf{x}_t - (\mathbf{x}^* + u(\mathbf{x}_t - \mathbf{x}^*) \| \mathrm{d}u \\
&= \frac{1}{\alpha} \left(\int_{u=0}^{1} \gamma(1-u) \mathrm{d}u \right) \| \mathbf{x}_t - \mathbf{x}^* \|^2 \\
&= \frac{\gamma}{2\alpha} \| \mathbf{x}_t - \mathbf{x}^* \|^2
\end{aligned}$$

where the second inequality follows from $\nabla^2 f(\mathbf{x}_t)^{-1} \preceq \frac{1}{\alpha}$ and Assumption 2.

Now, let t_0 be the first time step at which $\lambda(\mathbf{x}_{t_0}) < \frac{\alpha^4}{\beta\gamma^2}$. By the analysis of the damped Newton phase, we have $t_0 \leq \frac{2\beta^2\gamma^2(f(\mathbf{x}_0) - f(\mathbf{x}^*))}{\alpha^5}$. We have $\|\mathbf{x}_{t_0} - \mathbf{x}^*\| \leq \frac{\alpha}{\gamma}$. By a simple induction using the analysis above, we can show that for any $s \geq 0$, we have

$$\|\mathbf{x}_{t_0+s} - \mathbf{x}^*\| \leq \frac{\alpha}{\gamma} \cdot \left(\frac{1}{2}\right)^{2^s - 1}.$$

Thus, to reach a point $\mathbf{x}_t$ such that $\|\mathbf{x}_t - \mathbf{x}^*\| \leq \epsilon$, we need $\log_2(\log_2(\frac{2\alpha}{\gamma\epsilon}))$ iterations in this phase

So overall, we need at most

$$\frac{2\beta^2\gamma^2(f(\mathbf{x}_0) - f(\mathbf{x}^*))}{\alpha^5} + \log_2\left(\log_2\left(\frac{2\alpha}{\gamma\epsilon}\right)\right)$$

iterations of Newton's algorithm. This convergence rate is significantly faster than any variant of gradient descent we have studied in class.