ESTR2520: Optimization Methods **2023–24 Second Term**

Handout 8: Analysis of Newton's Method

Instructor: Hoi-To Wai Last updated: March 29, 2024

This note analyzes the Newton method introduced in the FTEC lecture.

1 Recap: Newton's Method

Our quest is the solve the following **unconstrained** problem:

$$\min_{x \in \mathbb{R}^n} \; f(x) \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex and twice differentiable function. Like in the last lecture, we assume:

Assumption 1 *The function $f(x)$ is convex and lower bounded, i.e., $\min_{x \in \mathbb{R}^n} f(x) > -\infty$.*

Assumption 2 *The gradient of $f(x)$ is L Lipschitz continuous, i.e., there exists $L \in \mathbb{R}_+$ such that*

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad x, y \in \mathbb{R}^n.$$

Assumption 3 *The function $f(x)$ is μ strongly convex (with $\mu > 0$) such that*

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x) + \frac{\mu}{2}\|y - x\|^2, \; \forall \; x, y \in \mathbb{R}^n,$$

As $f(x)$ is twice differentiable, the above is equivalent to requiring $\nabla^2 f(x) - \mu I_n$ is PSD. Let us first recall the Newton's method:

```
Newton's Method

Input: $x^0$ - initialization, set the iteration counter as $t = 0$.
Repeat
    Select a step size $\gamma > 0$
    $x^{t+1} = x^t - \gamma(\nabla^2 f(x^t))^{-1} \nabla f(x^t)$
Until convergence / stopping criterion.
```

In the last lecture, we have seen that the GD method method has a linear convergence of

$$\|x^t - x^\star\|^2 \leq (1 - \mu/L)^t \|x^0 - x^\star\|^2 = \mathcal{O}((1 - \mu/L)^t)$$

In this lecture, we shall show that the Newton's method converges at much faster rate, i.e.,

$$\|x^t - x^\star\|^2 \leq C \, \|x^t - x^\star\|^4$$

for some $C < \infty$. It has a *quadratic convergence* rate. To this end, we need one extra assumption:

Assumption 4 *The Hessian of $f(x)$ is L_h Lipschitz continuous, i.e., there exists L_h such that*

$$\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \leq L_h \|x - y\|, \; \forall \; x, y \in \mathbb{R}^n$$

Note that $\|\cdot\|_2$ denotes the **spectral norm** of a matrix (recall $\nabla^2 f(x) \in \mathbb{R}^{n\times n}$), defined as

$$\|A\|_2 = \sup_{x\in\mathbb{R}^n} \frac{\|Ax\|}{\|x\|} \quad \text{where it satisfies} \quad \underbrace{\|Ay\|}_{\text{Euclidean norm}} \leq \underbrace{\|A\|_2}_{\text{spectral norm}} \underbrace{\|y\|}_{\text{Euclidean norm}}.$$

For symmetric matrix, we have $\|A\|_2 = \lambda_{\max}(A)$.

2 Convergence Analysis for the Newton's Method

Theorem 1: Convergence Rate of Newton's Method

Under Assumption 1, 2, 3, 4. Consider the iterates $\{x^t\}_{t\geq 0}$ generated by the Newton's method with $\gamma = 1$. Suppose that $\|x^0 - x^\star\| < \frac{2\mu}{L_h}$. Then for any $t \geq 0$,

$$\|x^{t+1} - x^\star\| \leq \frac{L_h}{2\mu}\|x^t - x^\star\|^2$$

Proof We shall use the following definition for the gradient[1]:

$$\nabla f(x^t) = \int_0^1 \nabla^2 f(x^\star + z(x^t - x^\star))\,\mathrm{d}z\,(x^t - x^\star)$$

This formula is immediately useful as we observe

$$\begin{aligned}
\|x^{t+1} - x^\star\| &= \|x^t - (\nabla^2 f(x^t))^{-1}\nabla f(x^t) - x^\star\| \\
&= \left\|(\nabla^2 f(x^t))^{-1}\left[\nabla^2 f(x^t)(x^t - x^\star) - \nabla f(x^t)\right]\right\| \\
&= \left\|(\nabla^2 f(x^t))^{-1}\left[\nabla^2 f(x^t)(x^t - x^\star) - \int_0^1 \nabla^2 f(x^\star + z(x^t - x^\star))\,\mathrm{d}z\,(x^t - x^\star)\right]\right\| \\
&= \left\|(\nabla^2 f(x^t))^{-1}\left[\nabla^2 f(x^t) - \int_0^1 \nabla^2 f(x^\star + z(x^t - x^\star))\,\mathrm{d}z\right](x^t - x^\star)\right\| \\
&\leq \left\|(\nabla^2 f(x^t))^{-1}\right\|_2 \left\|\nabla^2 f(x^t) - \int_0^1 \nabla^2 f(x^\star + z(x^t - x^\star))\,\mathrm{d}z\right\|_2 \|x^t - x^\star\|.
\end{aligned}$$

Notice that $\left\|(\nabla^2 f(x^t))^{-1}\right\|_2 \leq 1/\mu$ and

$$\begin{aligned}
&\left\|\nabla^2 f(x^t) - \int_0^1 \nabla^2 f(x^\star + z(x^t - x^\star))\,\mathrm{d}z\right\|_2 \\
&= \left\|\int_0^1 (\nabla^2 f(x^t) - \nabla^2 f(x^\star + z(x^t - x^\star)))\,\mathrm{d}z\right\|_2 \\
&\leq \int_0^1 \left\|\nabla^2 f(x^t) - \nabla^2 f(x^\star + z(x^t - x^\star))\right\|_2\,\mathrm{d}z \\
&\leq \int_0^t L_h\|x^t - (x^\star + z(x^t - x^\star))\|\mathrm{d}z = L_h\|x^t - x^\star\|\int_0^1 z\,\mathrm{d}z = \frac{L_h}{2}\|x^t - x^\star\|
\end{aligned}$$

[1] The formula is similar to $f'(x^t) = f'(x^\star) + \int_0^1 f''(x^\star + z(x^t - x^\star))(x^t - x^\star)\,\mathrm{d}z$ and that $f'(x^\star) = 0$.

Substituting this back into the first inequality, we get

$$\|x^{t+1} - x^\star\| \le \frac{L_h}{2\mu} \|x^t - x^\star\|^2.$$

This concluding the proof. □

3 Perspective

Theorem 1 actually only characterizes the **local convergence** for the Newton's method as it requires $\|x^0 - x^\star\| < \frac{2\mu}{L_h}$, i.e., the method is initialized at a point that is sufficiently close to optimal. This contrasts sharply with the GD method we analyzed last week, whose convergence is **global** as we did not impose any restriction on $\|x^0 - x^\star\|$.

Of course, the Newton's method may be initialized with $\|x^0 - x^\star\| \nless \frac{2\mu}{L_h}$. As such, in practice, the method is typically applied with **two phases**.

- In phase one, it will be used with a backtracking line search design for the step size γ. The algorithm will exhibit **linear convergence** that is not slower than the GD method.
- In phase two (basically when $\|x^t - x^\star\| < \frac{2\mu}{L_h}$), the backtracking line search will return $\gamma = 1$ and the **quadratic convergence** proven in Theorem 1 applies.

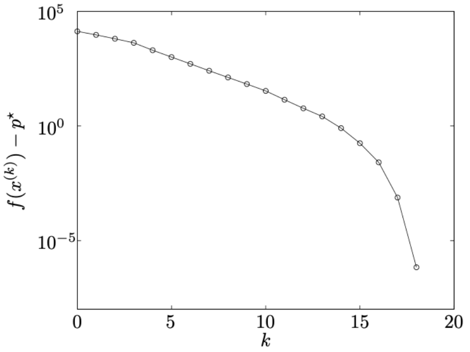

Figure 9.23 Error versus iteration of Newton's method, for a problem in $\mathbf{R}^{10000}$. A backtracking line search with parameters $\alpha = 0.01$, $\beta = 0.5$ is used. Even for this large scale problem, Newton's method requires only 18 iterations to achieve very high accuracy.

(taken from [1])

The proof we presented in the above requires the Lipschitz continuity of Hessian. However, this is not the only method to prove the quadratic convergence of Newton's method, other approach includes using a property called *self-concordance* that is slightly more relaxed; see [1].

Lastly, we should emphasize again that the Newton's method is significantly more complex than the GD method for high-dimensional problem due to its use of matrix inverses. Nevertheless, it remains an active research topic as people explores various ways to accelerate its practical convergence, e.g., via variants such as Quasi Newton method.

References

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