CMPT 409/981: Optimization for Machine Learning

Lecture 8

Sharan Vaswani

October 1, 2024

Newton Method

We have seen that for quadratics, the Newton method converges to the minimizer in one step.

• Let us analyze the convergence of Newton for general L-smooth, μ-strongly convex functions. For this, we will consider two phases for the update:

$$w_{k+1} = w_k - \eta_k \left[\nabla^2 f(w_k)\right]^{-1} \nabla f(w_k) \,,$$

Phase 1 (Damped Newton): For some α to be chosen later, if $\|\nabla f(w_k)\|^2 > \alpha$ ("far" from the solution), use the Newton method with the step-size η_k set according to the Back-tracking Armijo line-search.

Phase 2 (Pure Newton): If $\|\nabla f(w_k)\|^2 \leq \alpha$ ("close" to the solution), use the Newton method with step-size equal to 1.

Newton Method - Phase 2

Let us first analyze the convergence rate for Phase 2. For this, we will need an additional assumption that the Hessian is Lipschitz continuous with constant $M > 0$:

$$\left\|\nabla^2 f(w) - \nabla^2 f(v)\right\| \leq M \left\|w - v\right\|.$$

Claim: In Phase 2 of the Newton method, the iterates satisfy the following inequality,

$$\left\|w_{k+1} - w^*\right\| \leq \frac{M}{2\mu} \left\|w_k - w^*\right\|^2$$

Proof:

$$\begin{aligned}
w_{k+1} - w^* &= w_k - w^* - [\nabla^2 f(w_k)]^{-1} \nabla f(w_k) \qquad \text{(Newton update with step-size 1.)} \\
&= [\nabla^2 f(w_k)]^{-1} \left[[\nabla^2 f(w_k)](w_k - w^*) - \nabla f(w_k)\right] \\
\implies \left\|w_{k+1} - w^*\right\| &= \left\|[\nabla^2 f(w_k)]^{-1} \left[[\nabla^2 f(w_k)](w_k - w^*) - \nabla f(w_k)\right]\right\| \\
\implies \left\|w_{k+1} - w^*\right\| &\leq \left\|[\nabla^2 f(w_k)]^{-1}\right\| \left\|[\nabla^2 f(w_k)](w_k - w^*) - \nabla f(w_k)\right\| \\
&\qquad \text{(By definition of the matrix norm)}
\end{aligned}$$

Newton Method - Phase 2

Recall that $\|w_{k+1} - w^*\| \leq \left\|[\nabla^2 f(w_k)]^{-1}\right\| \left\|[\nabla^2 f(w_k)](w_k - w^*) - \nabla f(w_k)\right\|$.

$$\|w_{k+1} - w^*\| \leq \frac{1}{\mu} \left\|\left[[\nabla^2 f(w_k)](w_k - w^*) - \nabla f(w_k)\right]\right\| \qquad \text{(Since } \nabla^2 f(w) \succeq \mu I_d\text{)}$$

$$\implies \|w_{k+1} - w^*\| \leq \frac{1}{\mu} \left\|[\nabla^2 f(w_k)](w_k - w^*) + \nabla f(w^*) - \nabla f(w_k)\right\| \tag{1}$$

Now let us bound $\nabla f(w^*) - \nabla f(w_k)$. By the fundamental theorem of calculus, for all x, y, $f(y) = f(x) + \int_{t=0}^{1} [\nabla f(t\,y + (1-t)\,x)]\,(y-x)\,dt$. This theorem also holds for the vector-valued gradient function,

$$\nabla f(y) = \nabla f(x) + \int_{t=0}^{1} \left[\nabla^2 f\,(t\,y + (1-t)\,x)\right]\,(y-x)\,dt$$

Using the above statement with $x = w^*$ and $y = w_k$,

$$\implies \nabla f(w_k) - \nabla f(w^*) = \int_{t=0}^{1} \left[\nabla^2 f\,(t\,w_k + (1-t)\,w^*)\right]\,(w_k - w^*)\,dt \tag{2}$$

Newton Method - Phase 2

Combining eqs. (1) and (2),

$$
\begin{aligned}
&\left\| w_{k+1} - w^* \right\| \\
&\le \frac{1}{\mu} \left\| [\nabla^2 f(w_k)](w_k - w^*) + \nabla f(w^*) - \nabla f(w_k) \right\| \\
&= \frac{1}{\mu} \left\| \left[[\nabla^2 f(w_k)](w_k - w^*) - \int_{t=0}^{1} \left[\nabla^2 f\left(t\, w_k + (1-t)\, w^*\right) \right] (w_k - w^*)\, dt \right] \right\| \\
&= \frac{1}{\mu} \left\| \left[\int_{t=0}^{1} [\nabla^2 f(w_k)](w_k - w^*)\, dt - \int_{t=0}^{1} \left[\nabla^2 f\left(t\, w_k + (1-t)\, w^*\right) \right] (w_k - w^*)\, dt \right] \right\| \\
&= \frac{1}{\mu} \left\| \int_{t=0}^{1} \left[\nabla^2 f(w_k) - \nabla^2 f\left(t\, w_k + (1-t)\, w^*\right) \right] (w_k - w^*)\, dt \right\| \\
&\le \frac{1}{\mu} \int_{t=0}^{1} \left\| \left[\nabla^2 f(w_k) - \nabla^2 f\left(t\, w_k + (1-t)\, w^*\right) \right] (w_k - w^*) \right\| \, dt \qquad \text{(Jensen's inequality)} \\
&\le \frac{1}{\mu} \int_{t=0}^{1} \left\| \nabla^2 f(w_k) - \nabla^2 f\left(t\, w_k + (1-t)\, w^*\right) \right\| \left\| w_k - w^* \right\| \, dt \quad \text{(Definition of matrix norm)}
\end{aligned}
$$

Newton Method - Phase 2

From the previous slide,

$$\|w_{k+1} - w^*\| \leq \frac{1}{\mu} \int_{t=0}^{1} \left\|\nabla^2 f(w_k) - \nabla^2 f\left(t\, w_k + (1-t)\, w^*\right)\right\| \, \|w_k - w^*\| \; dt$$

Since the Hessian is M-Lipschitz,

$$\begin{aligned}
&\leq \frac{1}{\mu} \int_{t=0}^{1} M \, \|w_k - t\, w_k - (1-t)\, w^*\| \, \|w_k - w^*\| \; dt \\
&= \frac{M}{\mu} \, \|w_k - w^*\| \int_{t=0}^{1} \|(1-t)(w_k - w^*)\| \, dt \\
&= \frac{M}{\mu} \, \|w_k - w^*\|^2 \int_{t=0}^{1} (1-t)\, dt
\end{aligned}$$

$$\implies \|w_{k+1} - w^*\| \leq \frac{M}{2\mu} \, \|w_k - w^*\|^2$$

Newton Method - Phase 2

Recall that for Phase 2 of the Newton method, $\|w_{k+1} - w^*\| \leq c \, \|w_k - w^*\|^2$ where $c := \frac{M}{2\mu}$.

Claim: If in Phase 2, $\|w_0 - w^*\| \leq \frac{1}{2c} = \frac{\mu}{M}$, then after T iterations of the Pure Newton update, $\|w_T - w^*\| \leq \left(\frac{1}{2}\right)^{2^T} \frac{1}{c} = \left(\frac{1}{2}\right)^{2^T} \frac{2\mu}{M}$.

Proof: Let us prove it by induction.

Base-case: For $T = 0$, $\|w_T - w^*\| \leq \frac{\mu}{M}$ which is true by our assumption.

Inductive hypothesis: If the statement is true for iteration k, then $\|w_k - w^*\| \leq \left(\frac{1}{2}\right)^{2^k} \frac{1}{c}$.

$$\|w_{k+1} - w^*\| \leq c \, \|w_k - w^*\|^2 \leq c \left(\left(\frac{1}{2}\right)^{2^k} \frac{1}{c} \right)^2 = \frac{1}{c} \left(\frac{1}{2}\right)^{2^{k+1}},$$

which completes the induction. Hence, $\|w_T - w^*\| \leq \left(\frac{1}{2}\right)^{2^T} \frac{2\mu}{M}$. For $\|w_T - w^*\| \leq \epsilon$, we need T such that,

$$\left(\frac{1}{2}\right)^{2^T} \frac{2\mu}{M} \leq \epsilon \implies T \geq \frac{1}{\log(2)} \log\left(\frac{\log\left(2\mu/M\epsilon\right)}{\log(2)} \right)$$

Newton Method - Phase 2

- From the previous slide, we can conclude that Phase 2 of the Newton method requires $O\left(\log\left(\log\left(1/\epsilon\right)\right)\right)$ iterations to achieve an ϵ sub-optimality.

- This rate of convergence is often referred to as **quadratic** or **super-linear** convergence. Note that there is no dependence on κ and the dependence on $\frac{\mu}{M}$ is in the $\log\log$.

- But the bound is true only if $\|w_0 - w^*\| \leq \frac{\mu}{M}$ i.e. we enter Phase 2 only when we are "close enough" to the solution. This is referred to as **local convergence**. Hence, the Newton method has super-linear local convergence.

- Algorithmically, since we do not know w^*, we do not know when to start Phase 2 of the algorithm. By strong-convexity,

$$\|\nabla f(x) - \nabla f(y)\| \geq \mu \|x - y\| \implies \|w_0 - w^*\| \leq \frac{1}{\mu} \|\nabla f(w_0)\|$$

Hence, in order to ensure that $\|w_0 - w^*\| \leq \frac{\mu}{M}$, it suffices to guarantee that $\|\nabla f(w_0)\|^2 \leq \alpha := \frac{\mu^4}{M^2}$. This can be checked algorithmically.

Questions?

Newton Method

Theorem: If $\|\nabla f(w)\|^2 \leq \alpha = \frac{\mu^4}{M^2}$, the algorithm switches to Phase 2 for T iterations of the pure Newton step and ensures that $\|w_T - w^*\| \leq \left(\frac{1}{2}\right)^{2^T} \frac{2\mu}{M}$.

• In order to prove global convergence for the Newton method i.e. starting from any initialization, we need to prove that Phase 1 of the Newton step can result in an iterate w such that $\|\nabla f(w)\|^2 \leq \alpha$ and we can switch to Phase 2.

• Recall that for Phase 1, we will use the Backtracking Armijo line-search. For a prospective step-size $\tilde{\eta}_k$, check the (more general) Armijo condition,

$$f(w_k - \tilde{\eta}_k d_k) \leq f(w_k) - c\,\tilde{\eta}_k \underbrace{\langle \nabla f(w_k), d_k \rangle}_{\text{Newton decrement}}$$

where $c \in (0, 1)$ is a hyper-parameter and $d_k = [\nabla^2 f(w_k)]^{-1} \nabla f(w_k)$ is the Newton direction. If $\tilde{\eta}_k$ satisfies the above condition, use the Newton update with $\eta_k = \tilde{\eta}_k$.

Q: Why does the Newton direction make an acute angle with the gradient direction? Ans: Because the Newton decrement is positive since the inverse Hessian is positive definite.

Newton Method - Phase 1

- Using a similar proof as the standard Back-tracking Armijo line-search, we can show that the step-size returned by the back-tracking procedure at iteration k is lower-bounded as: $\eta_k \geq \min\left\{\frac{2\mu\,(1-c)}{L}, \eta_{\max}\right\}$ (Need to prove this in Assignment 2).
- At iteration k, η_k is the step-size returned by the Back-tracking Armijo line-search and satisfies the general Armijo condition. Hence,

$$f(w_k - \eta_k d_k) - f^* \leq [f(w_k) - f^*] - c\,\eta_k\,\langle \nabla f(w_k), d_k \rangle$$

$$\implies f(w_{k+1}) - f^* \leq [f(w_k) - f^*] - c\,\eta_k\,\langle \nabla f(w_k), [\nabla^2 f(w_k)]^{-1} \nabla f(w_k) \rangle$$

Since $\nabla^2 f(w_k)$ is P.S.D, $\langle \nabla f(w_k), [\nabla^2 f(w_k)]^{-1} \nabla f(w_k) \rangle \geq 0$ and we need to lower-bound it,

$$\langle \nabla f(w_k), [\nabla^2 f(w_k)]^{-1} \nabla f(w_k) \rangle \geq \lambda_{\min}[\nabla^2 f(w_k)]^{-1}\,\|\nabla f(w_k)\|^2$$

$$\implies f(w_{k+1}) - f^* \leq [f(w_k) - f^*] - c\,\eta_k\,\lambda_{\min}[\nabla^2 f(w_k)]^{-1}\,\|\nabla f(w_k)\|^2$$

$$f(w_{k+1}) - f^* \leq [f(w_k) - f^*] - \frac{c\,\eta_k}{L}\,\|\nabla f(w_k)\|^2$$

(Since $\lambda_{\min}[\nabla^2 f(w_k)]^{-1} = \frac{1}{\lambda_{\max}[\nabla^2 f(w_k)]} = \frac{1}{L}$)

Newton Method - Phase 1

Recall that $f(w_{k+1}) - f^* \leq [f(w_k) - f^*] - {}^{c\,\eta_k}/_{L} \left\|\nabla f(w_k)\right\|^2$.

$$f(w_{k+1}) - f^* \leq [f(w_k) - f^*] - \frac{c \, \min\left\{\frac{2\mu\,(1-c)}{L}, \eta_{\max}\right\}}{L} \left\|\nabla f(w_k)\right\|^2 \quad \text{(Lower-bound on } \eta_k\text{)}$$

$$\leq [f(w_k) - f^*] - \frac{\min\left\{\frac{\mu}{2L}, \frac{\eta_{\max}}{2}\right\}}{L} \left\|\nabla f(w_k)\right\|^2 \quad \text{(Setting } c = {}^1/_2\text{)}$$

$$\leq \left(1 - \frac{\mu \min\left\{\frac{\mu}{L}, \eta_{\max}\right\}}{L}\right) [f(w_k) - f^*] \quad (\left\|\nabla f(w_k)\right\|^2 \geq 2\mu[f(w_k) - f^*])$$

$$\implies f(w_{k+1}) - f^* \leq \left(1 - \frac{\mu^2 \, \min\{1, \kappa\eta_{\max}\}}{L^2}\right) [f(w_k) - f^*]$$

Recursing from $k = 0$ to $\tau - 1$ and setting $\eta_{\max} = 1$

$$f(w_\tau) - f^* \leq \left(1 - \frac{1}{\kappa^2}\right)^\tau [f(w_0) - f^*] \leq \exp\left(\frac{-\tau}{\kappa^2}\right) [f(w_0) - f^*]$$

Newton Method

Recall that $f(w_\tau) - f^* \le \exp\left(\frac{-\tau}{\kappa^2}\right)[f(w_0) - f^*]$. Phase 1 terminates when $\|\nabla f(w_\tau)\|^2 = \alpha$. Using L-smoothness, $\|\nabla f(w_\tau)\|^2 \le 2L[f(w_\tau) - f^*]$. To terminate Phase 1, we want

$$2L[f(w_\tau) - f^*] = 2L \exp\left(\frac{-\tau}{\kappa^2}\right)[f(w_0) - f^*] = \alpha$$

$$\implies \tau = \kappa^2 \log\left(\frac{2L\,M^2\,[f(w_0) - f^*]}{\mu^4}\right) \qquad \text{(Since } \alpha = \tfrac{\mu^4}{M^2}\text{)}$$

- Hence, iterations required for global convergence to an ϵ sub-optimality is,

$$\underbrace{\kappa^2 \log\left(\frac{2L\,M^2\,[f(w_0) - f^*]}{\mu^4}\right)}_{\text{Phase 1}} + \underbrace{\frac{1}{\log(2)} \log\left(\frac{\log\left({}^{2\mu}/_{M\epsilon}\right)}{\log(2)}\right)}_{\text{Phase 2}} = O\left(\kappa^2 + \log\left(\log\left({}^{1}/_{\epsilon}\right)\right)\right)$$

- Recall that GD requires $O\left(\kappa \log\left({}^{1}/_{\epsilon}\right)\right)$ iterations. If we do a matrix inversion in every iteration, cost of each iteration is $O(d^3)$. Since computing gradients is linear in d, the cost of each GD iteration is $O(d)$. Comparing computational complexity:

Gradient Descent: $O\left(d\kappa \log\left({}^{1}/_{\epsilon}\right)\right)$ Newton Method: $O\left(\left(d^3\kappa^2 + d^3 \log\left(\log\left({}^{1}/_{\epsilon}\right)\right)\right)\right)$

- Newton method is more efficient than GD for small d (low-dimension) and small ϵ (high precision).

Questions?