Optimization Methods (CS1.404), Spring 2024
Lecture 13

Naresh Manwani

Machine Learning Lab, IIIT-H

February 22nd, 2024

Diagonal Scaling to Improve Condition Number

- Consider the problem $\min_{\mathbf{x}} \ \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x} - \mathbf{c}^T\mathbf{x}$, where $\mathbf{H}$ is a symmetric positive definite matrix.
- Condition number of Hessian matrix controls the convergence rate of steepest descent.
- Faster convergence if Hessian is closer to scalar multiple of identity matrix.
- Can we transform the problem into another space in which the condition number of the Hessian becomes Identity?
- Let $\mathbf{H} = \mathbf{L}\mathbf{L}^T$ be the Cholesky decomposition of H.
- Define $\mathbf{y} = \mathbf{L}^T\mathbf{x}$.
- Consider the transformed function $h(\mathbf{y}) = f(\mathbf{L}^{-T}\mathbf{y})$.

Diagonal Scaling to Improve Condition Number

$$
\begin{aligned}
h(\mathbf{y}) = f(\mathbf{L}^{-T}\mathbf{y}) &= \frac{1}{2}\mathbf{y}^T\mathbf{L}^{-1}\mathbf{H}\mathbf{L}^{-T}\mathbf{y} - \mathbf{c}^T(\mathbf{L}^{-T}\mathbf{y}) \\
&= \frac{1}{2}\mathbf{y}^T\mathbf{L}^{-1}\mathbf{L}\mathbf{L}^T\mathbf{L}^{-T}\mathbf{y} - \mathbf{c}^T(\mathbf{L}^{-T}\mathbf{y}) \\
&= \frac{1}{2}\mathbf{y}^T\mathbf{y} - \mathbf{c}^T(\mathbf{L}^{-T}\mathbf{y})
\end{aligned}
$$

- The hessian of $h(\mathbf{y})$ is identity matrix.
- Let us apply steepest descent on $\mathbf{y}$ space.

$$\mathbf{y}^{k+1} = \mathbf{y}^k - \nabla h(\mathbf{y}^k) = \mathbf{y}^k - \mathbf{L}^{-1}\nabla f(\mathbf{L}^{-T}\mathbf{y}^k)$$

- Applying transformation $\mathbf{L}^{-T}$ on both sides

$$
\begin{aligned}
&\mathbf{L}^{-T}\mathbf{y}^{k+1} = \mathbf{L}^{-T}\mathbf{y}^k - \mathbf{L}^{-T}\mathbf{L}^{-1}\nabla f(\mathbf{L}^{-T}\mathbf{y}^k) \\
\Rightarrow \quad &\mathbf{x}^{k+1} = \mathbf{x}^k - \mathbf{H}^{-T}\nabla f(\mathbf{x}^k) = \mathbf{x}^k - \mathbf{H}^{-1}\nabla f(\mathbf{x}^k)
\end{aligned}
$$

- This method is called **Newton Method**.

Newton Method

- Consider $\min_{\mathbf{x} \in \mathbb{R}^n} \; f(\mathbf{x})$, where $f \in \mathbb{C}^2(\mathbb{R}^n)$.
- Newton method used second order information to find out the descent direction.
- At each iteration, it uses second order Taylor series approximation of f at $\mathbf{x}_k$ and finds the minimum of it to get $\mathbf{x}_{k+1}$.

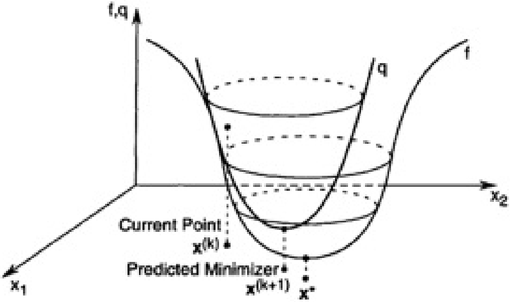
$$\mathbf{x}_{k+1} = \arg\min_{\mathbf{x} \in \mathbb{R}^n} \left\{ f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^T(\mathbf{x} - \mathbf{x}_k) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_k)^T \nabla^2 f(\mathbf{x}_k)(\mathbf{x} - \mathbf{x}_k) \right\}$$

- The above formula is well defined only if we further assume that $\nabla^2 f(\mathbf{x}_k)$ is positive definite. Under this assumption, the unique minimizer is

$$\mathbf{x}_{k+1} = \mathbf{x}_k - (\nabla^2 f(\mathbf{x}_k))^{-1} \nabla f(\mathbf{x}_k)$$

- **Newton Direction:** $\mathbf{d}_N = -(\nabla^2 f(\mathbf{x}_k))^{-1} \nabla f(\mathbf{x}_k)$.

Geometry of Newton Method



Pure Newton Method

- **Input:** $\epsilon > 0$ -tolerance parameter
- **Initialization:** Pick $\mathbf{x}_0 \in \mathbb{R}^n$ arbitrarily
- **General Step :** For any $k = 0, 1, 2, \ldots$ execute the following steps:
  1. Compute the Newton's direction, which is the solution to the linear system: $\nabla^2 f(\mathbf{x}_k)\mathbf{d}_k = -\nabla f(\mathbf{x}_k)$.
  2. Set $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{d}_k$
  3. If $\|\nabla f(\mathbf{x}_{k+1})\| \leq \epsilon$, then STOP and output $\mathbf{x}_{k+1}$.

Convergence of Newton Method for Quadratic Functions

- Newton method requires that $\nabla^2 f(\mathbf{x})$ is positive definite for every $\mathbf{x}$ (strict convexity).
- Consider quadratic function $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x} - \mathbf{c}^T\mathbf{x}$ such that matrix $\mathbf{H}$ is real symmetric and positive definite matrix.
- We know that the unique global minimizer of f is $\mathbf{x}^* = \mathbf{H}^{-1}\mathbf{c}$.
- We see that $\nabla f(\mathbf{x}) = \mathbf{H}\mathbf{x} - \mathbf{c}$ and $\nabla^2 f(\mathbf{x}) = \mathbf{H}$.
- Applying Newton method on this function for $\mathbf{x}_0$ as initial point, we see that

$$\mathbf{x}_1 = \mathbf{x}_0 - \nabla^2 f(\mathbf{x}_0)^{-1}\nabla f(\mathbf{x}_0) = \mathbf{x}_0 - \mathbf{H}^{-1}(\mathbf{H}\mathbf{x}_0 - \mathbf{c}) = \mathbf{H}^{-1}\mathbf{c}$$

- Thus, using Newton method, we reach to the global minima of a quadratic and strictly convex function in one step.

Convergence of Newton Method for General Functions

- Newton method requires that $\nabla^2 f(\mathbf{x})$ is positive definite for every $\mathbf{x}$ (strict convexity).
- Which implies that there exists a unique optimal solution $\mathbf{x}^*$.
- However, this is not enough to guarantee convergence.
- Consider the following example.

Example

- Consider the function $f(x) = \sqrt{1+x^2}$. The minimizer of f is $x = 0$.
- $f'(x) = \frac{x}{\sqrt{1+x^2}}$, $f''(x) = \frac{1}{(1+x^2)^{3/2}}$.
- Therefore, the Pure Newton method update equations are

$$x_{k+1} = x_k - (1+x_k^2)^{3/2} \frac{x_k}{\sqrt{1+x_k^2}} = x_k - x_k(1+x_k^2) = -x_k^3$$

- Newton method converges to $x^* = 0$ when $|x_0| < 1$. For $|x_0| > 1$, it diverges.

Quadratic Local Convergence of Newton's Method

Theorem

Let f be a twice continuously differentiable function defined over $\mathbb{R}^n$. Assume that

- There exists $m > 0$ for which $\nabla^2 f(\mathbf{x}) \succeq m\mathbf{I}$ for any $\mathbf{x} \in \mathbb{R}^n$,
- There exists $L > 0$ for which $\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\| \leq \|\mathbf{x} - \mathbf{y}\|$ for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

Let $\{\mathbf{x}_k\}_{k \geq 0}$ be the sequence generated by Newton's Method, and let $\mathbf{x}^*$ be the unique minimizer of f over $\mathbb{R}^n$. Then for any $k = 0, 1, 2, \ldots$ the inequality

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\| \leq \frac{L}{2m} \|\mathbf{x}_k - \mathbf{x}^*\|^2$$

holds. In addition, if $\|\mathbf{x}^* - \mathbf{x}_0\| \leq \frac{2m}{L}$, then

$$\|\mathbf{x}^* - \mathbf{x}_k\| \leq \frac{2m}{L} \left(\frac{1}{2}\right)^{2^k}, \quad k = 0, 1, 2, \ldots$$

- Thus, near the optimal solution, the error $e_k = \|\mathbf{x}^* - \mathbf{x}_k\|$ satisfies the inequality $e_{k+1} \leq M e_k^2$ for some positive $M > 0$.

Example 2: $\nabla f(\mathbf{x}) \succeq m\mathbf{I}$ not satisfied

- Consider the problem $\min_{x_1,x_2} \sqrt{1+x_1^2} + \sqrt{1+x_2^2}$. Optimal solution is $(0,0)$.
- Hessian of the function is $\nabla^2 f(\mathbf{x}) = \begin{pmatrix} \frac{1}{(1+x_1^2)^{3/2}} & 0 \\ 0 & \frac{1}{(1+x_2^2)^{3/2}} \end{pmatrix} \succeq \mathbf{0}$.
- Even though the hessian is positive definite, there does not exist an $m > 0$ for which $\nabla^2 f(\mathbf{x}) \succeq m\mathbf{I}$. As $x_1, x_2 \to \infty$, $\nabla^2 f(\mathbf{x})$ becomes a zero matrix.
- Basic assumption for convergence is not satisfied.
- This is reflected in implementation also.
- Newton's method with initial vector $\mathbf{x}_0 = (1,1)$ and tolerance parameter $\epsilon = 10^{-8}$ we obtain convergence after 37 iterations.
- Newton's method with initial vector $\mathbf{x}_0 = (10,10)$ diverges.

`iter= 1 f(x)=2.8284271247`
`iter= 2 f(x)=2.8284271247`
`: :`
`iter= 30 f(x)=2.8105247315`
`iter= 31 f(x)=2.7757389625`
`iter= 32 f(x)=2.6791717153`
`iter= 33 f(x)=2.4507092918`
`iter= 34 f(x)=2.1223796622`
`iter= 35 f(x)=2.0020052756`
`iter= 36 f(x)=2.0000000081`
`iter= 37 f(x)=2.0000000000`
	`iter= 1 f(x)=2000.0009999997`
	`iter= 2 f(x)=1999999999.9999990000`
	`iter= 3 f(x)=1999999999999999730000000000.0000000`
	`iter= 4 f(x)=19999999999999992300000000000000000000....`
	`iter= 5 f(x)= Inf`

(a) Starting point (1,1). Not much progress in 30 iterations. Converges in 37 iterations.	(b) Starting point (10,10). Newton method diverges.

Issues with the Newton Method

- Requires computing inverse of hessian in each iteration. Can be computationally intensive if the number of variables are large.
- No guarantee that $\mathbf{d}_N = -\nabla^2 f(\mathbf{x})^{-1} \nabla f(\mathbf{x})$ is descent direction as the algorithm does not check if the hessian is positive definite.
- Problem happens when hessian in singular in some iteration.
- No guarantee that the function value decreases in each iteration (as no line search is used).
- Convergence is sensitive to initial point.

Newton method with Backtracking Line Search

Damped Newton Method

- **Input:** $\alpha, \beta \in (0,1)$ - parameters for the backtracking procedure. $\epsilon > 0$ -tolerance parameter
- **Initialization:** Pick $\mathbf{x}_0 \in \mathbb{R}^n$ arbitrarily
- **General Step :** For any $k = 0, 1, 2, \ldots$ execute the following steps:
  1. Compute the Newton's direction, which is the solution to the linear system: $\nabla^2 f(\mathbf{x}_k)\mathbf{d}_k = -\nabla f(\mathbf{x}_k)$.
  2. Set $t_k = 1$. While,
  $$f(\mathbf{x}_k) - f(\mathbf{x}_k + t_k\mathbf{d}_k) < -\alpha t_k \nabla f(\mathbf{x}_k)^T \mathbf{d}_k$$
  Set $t_k = \beta t_k$.
  3. $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{d}_k$
  4. If $\|\nabla f(\mathbf{x}_{k+1})\| \leq \epsilon$, then STOP and output $\mathbf{x}_{k+1}$.

One can also use other step size selection methods.

Newton Method with Backtracking on Example 2

- Consider the problem $\min_{x_1,x_2} \sqrt{1+x_1^2} + \sqrt{1+x_2^2}$. Optimal solution is $(0,0)$.
- Take initial point (10,10).
- Using backtracking line search with $\alpha = \beta = 0.5$ and $\epsilon = 10^{-8}$ Newton method converges in 17 iterations.

Levenberg Marquardt Algorithm

- If the hessian matrix $\nabla^2 f(\mathbf{x}_k)^{-1}\nabla f(\mathbf{x}_k)$ is not positive definite, the Newton direction $\mathbf{d}_N = -\nabla^2 f(\mathbf{x})^{-1}\nabla f(\mathbf{x})$ may not remain a descent direction.
- This issue can be resolved by updating the Newton update in the following way.

$$\mathbf{x}_{k+1} = \mathbf{x}_k - (\nabla^2 f(\mathbf{x}_k) + \mu_k \mathbf{I})^{-1}\nabla f(\mathbf{x}_k)$$

  where $\mu_k \geq 0$.
- The idea is as follows.
  - Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of $\nabla^2 f(\mathbf{x}_k)$ and $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be the corresponding eigenvectors.
  - If $\nabla^2 f(\mathbf{x}_k)$ is not positive definite, then some of the eigenvalues of it are negative.
  - Matrix $\nabla^2 f(\mathbf{x}_k) + \mu_k \mathbf{I}$ has eigenvalues $\lambda_1 + \mu_k, \ldots, \lambda_n + \mu_k$ with $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be the corresponding eigenvectors.
  - if μ_k is chosen sufficiently large, all eigenvalues of $\nabla^2 f(\mathbf{x}_k) + \mu_k \mathbf{I}$ can become positive.
  - In that case $-(\nabla^2 f(\mathbf{x}_k) + \mu_k \mathbf{I})^{-1}\nabla f(\mathbf{x}_k)$ becomes a descent direction.

Choosing μ_k

Choosing μ_k

1. Start with some μ_k (a small value)
2. Do the Cholesky factorization of $\nabla^2 f(\mathbf{x}_k) + \mu_k \mathbf{I}$.
3. If Unsuccessful, increase the value of μ_k and go to step 2,

- If μ_k is very large, then this method becomes same as steepest descent.
- If μ_k is very small, then this method becomes same as Newton method.

Levenberg Marquardt Algorithm

- **Input:** Tolerance parameter $\epsilon > 0$, lower bound on minimum eigenvalue $\delta > 0$
- **Initialization:** Pick $\mathbf{x}_0 \in \mathbb{R}^n$ arbitrarily. Set $k = 0$.
- While $(\|\nabla f(\mathbf{x}_k)\| > \epsilon)$
  1. Find the smallest $\mu_k \geq 0$ such that the smallest eigenvalue of $\nabla^2 f(\mathbf{x}_k) + \mu_k \mathbf{I}$ is greater than δ.
  2. Set $\mathbf{d}_k = -(\nabla^2 f(\mathbf{x}_k) + \mu_k \mathbf{I})^{-1} \nabla f(\mathbf{x}_k)$
  3. Find $\alpha_k > 0$ using backtracking
  4. Update $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k$
  5. $k = k + 1$
- **Output:** $\mathbf{x}^* = \mathbf{x}_k$ as stationary point of $f(\mathbf{x})$.

Need for Cholesky Factorization

- In Levenberg Marquardt algorithm, it is required to validate the positive definiteness of the matrix $\nabla^2 f(\mathbf{x}_k) + \mu_k \mathbf{I}$.
- Another issue is to solve the equation $\nabla^2 f(\mathbf{x}_k)\mathbf{d} = -\nabla f(\mathbf{x}_k)$ in general for Newton method.
- These two issues are resolved using Cholesky factorization.

Solving $\mathbf{Ax} = \mathbf{b}$ using Cholesky Factorization

- Let $\mathbf{A}$ be $n \times n$ positive definite matrix. Cholesky factorization of $\mathbf{A}$ has the form $\mathbf{A} = \mathbf{L}\mathbf{L}^T$, where $\mathbf{L}$ is a lower triangular $n \times n$ matrix whose diagonal is positive
- Given the Cholesky factorization, equation $\mathbf{Ax} = \mathbf{b}$ can be solved in following two steps.
    - Find the solution $\mathbf{u}$ of $\mathbf{Lu} = \mathbf{b}$
    - Find the solution $\mathbf{x}$ of $\mathbf{L}^T\mathbf{x} = \mathbf{u}$.

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD