Gradient Descent and Newton's Method

Samuel Goodwin and Srinidhi Pawar

Center for Quantum Information and Control
University of New Mexico

12 June 2024

Motivation

- An optimization problem typically involves finding the minimum (or maximum) of a function $f(x)$ where x is a vector in $\mathbb{R}^n$.
- Gradient vanishes at optimal points. Search through all stationary points for the one with minimal function value.

Descent Direction Methods

- Iterative algorithm is of the form:

$$\boldsymbol{x_{k+1} = x_k + t_k d_k},\ k = 0, 1, 2, \cdots \tag{1}$$

where d_k is the direction and t_k is the stepsize.

Definition

Descent Direction: Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous differentiable function over $\mathbb{R}^n$. A vector $0 \neq d \in \mathbb{R}^n$ is called a descent direction of f at x if the directional derivative $f'(x; d)$ is negative

$$\boldsymbol{f'(x; d) = \nabla f(x)^T d < 0} \tag{2}$$

Lemma

***Descent property of descent directions**: Let f be a continuously differentiable function over $\mathbb{R}^n$, and let $x \in \mathbb{R}^n$. Suppose that d is a descent direction of f at x then there exists $\epsilon > 0$ such that*

$$\boldsymbol{f(x + td) < f(x)} \tag{3}$$

for any $t \in (0, \epsilon]$

Proof: Since $f'(x; d) < 0$, it follows that

$$\lim_{t \to 0^+} \frac{f(x + td) - f(x)}{t} = f'(x; d) < 0$$

$$\therefore \exists \text{ an } \epsilon > 0 \text{ such that } f(x + td) < f(x) \text{ for any } t \in (0, \epsilon].$$

Descent Directions Method

Initialization: Pick $x_0 \in \mathbb{R}^n$ arbitrarily.
General step: For any $k = 0, 1, 2, \cdots$ set

1. Pick a descent direction d_k.
2. Find a stepsize t_k satisfying $f(x_k + t_k d_k) < f(x_k)$.
3. Set $x_{k+1} = x_k + t_k d_k$.
4. If a stopping criterion is satisfied, then STOP and x_{k+1} is the output.

Descent Directions Method: Questions

- What is the starting point?
  - Chosen arbitrarily in the absence of an educated guess.
- What stepsize should be taken?
  - $f(x_{k+1}) < f(x_k)$
  - Process of finding step size t_k is called **line search**.
- What is the stopping criterion?

$$\|\nabla f(x_{k+1})\| \leq \epsilon \tag{4}$$

- How to choose the descent direction?
  - Main difference between different methods.

Stepsize Selection Rules

- **Constant stepsize**: $t_k = t'$ for any k.
- **Exact line search**: t_k is a minimizer of f along the ray $x_k + t_k d_k$:

$$t_k \in \text{argmin}_{t \geq 0} f(x_k + t_k d_k). \tag{5}$$

- **Backtracking**: The method requires three parameters: $s > 0, \alpha \in (0,1), \beta \in (0,1)$.
  - Set t_k to be equal to initial guess 's'.

$$f(x_k) - f(x_k + t_k d_k) < -\alpha t_k \nabla f(x_k)^T d_k \tag{6}$$

  - Set $t_k \leftarrow \beta t_k$ or $t_k = s\beta^{i_k}$ where i_k is the smallest nonnegative integer s.t.

$$f(x_k) - f(x_k + s\beta^{i_k} d_k) \geq -\alpha s \beta^{i_k} \nabla f(x_k)^T d_k \tag{7}$$

Sufficient Decrease Condition

The sufficient decrease condition is always satisfied for small enough t_k.

Lemma

Validity of the sufficient decrease condition: *Let f be a continuously differentiable function over $\mathbb{R}^n$, and let $x \in \mathbb{R}^n$. Suppose that $0 \neq d \in \mathbb{R}^n$ is a descent direction of f at x and let $\alpha \in (0, 1)$. Then there exists $\epsilon > 0$ such that*

$$\boldsymbol{f(x) - f(x + td) \geq -\alpha t \nabla f(x)^T d} \tag{8}$$

for any $t \in (0, \epsilon]$

Sufficient Decrease Condition

Proof: Since f is continuously differentiable,

$$f(x+td) = f(x) + t\nabla f(x)^T d + o(t\|d\|)$$

$$f(x) - f(x+td) = -\alpha t\nabla f(x)^T d - (1-\alpha)t\nabla f(x)^T d - o(t\|d\|) \tag{9}$$

Since d is a descent direction of f at x we have

$$\lim_{t\to 0^+} \frac{(1-\alpha)t\nabla f(x)^T d + o(t\|d\|)}{t} = (1-\alpha)\nabla f(x)^T d < 0.$$

Hence, there exists $\varepsilon > 0$ such that for all $t \in (0, \varepsilon]$ the inequality

$$(1-\alpha)t\nabla f(x)^T d + o(t\|d\|) < 0 \tag{10}$$

holds, which combined with (9) implies the desired result.

Example: Exact line search for quadratic functions

Let $f(x) = x^T Ax + 2b^T x + c$, where A is an $n \times n$ positive definite matrix, $b \in \mathbb{R}^n$, and $c \in R$. Let $x \in \mathbb{R}^n$ and $d \in R^n$ be a descent direction of f at x. Find an explicit formula for stepsize using line search.

Soln: Find solution of

$$\min_{t \geq 0} f(x + td)$$

$$\begin{aligned} g(t) = f(x+td) &= (x+td)^T A(x+td) + 2b^T(x+td) + c \\ &= (d^T Ad)t^2 + 2(d^T Ax + d^T b)t + f(x) \\ \text{Since, } g'(t) &= 2(d^T Ad)t + 2d^T(Ax+b) \\ \text{and, } \nabla f(x) &= 2(Ax+b) \end{aligned}$$

$g'(t) = 0$ only iff

$$\bar{t} = -\frac{d^T \nabla f(x)}{2d^T Ad}$$

$\because d^T \nabla f(x) < 0$, we have $\bar{t} > 0$.

Choice of descent direction: $\boldsymbol{d_k = -\nabla f(x_k)}$ because for $\|\nabla f(x_k)\| \neq 0$,

$$f'(x_k; -\nabla f(x_k)) = -\nabla f(x_k)^T \nabla f(x_k) = -\|\nabla f(x_k)\|^2 < 0$$

Lemma

Let f be a continuously differentianle function, and let $x \in \mathbb{R}^n$ be a non-stationary point ($\nabla f(x) \neq 0$). Then an optimal solution of

$$\min_{d \in \mathbb{R}^n} \{f'(x; d) : \|d\| = 1\} \tag{11}$$

is $d = -\frac{\nabla f(x)}{\|\nabla f(x)\|}$

Proof: Using Cauchy-Schwarz inequality,

$$\nabla f(x)^T d \geq -\|\nabla f(x)\| \cdot \|d\| = -\|\nabla f(x)\|. \tag{12}$$

Thus, $-\|\nabla f(x)\|$ is a lower bound on (11).
Plugging

$$d = -\frac{\nabla f(x)}{\|\nabla f(x)\|}$$

we obtain

$$f'\left(x, -\frac{\nabla f(x)}{\|\nabla f(x)\|}\right) = -\nabla f(x)^T \left(\frac{\nabla f(x)}{\|\nabla f(x)\|}\right) = -\|\nabla f(x)\|, \tag{13}$$

$\therefore$ the lower bound $-\|\nabla f(x)\|$ is attained at $d = -\frac{\nabla f(x)}{\|\nabla f(x)\|}$, which implies that this is an optimal solution of (11).

Gradient Method

Input: $\epsilon > 0$ tolerance parameter.
Initialization: Pick $x_0 \in \mathbf{R}^n$ arbitrarily.
General step: For any $k = 0, 1, 2, \cdots$ set

1. Pick a stepsize t_k using line search on $g(t) = f(x_k - t\nabla f(x_k))$.
2. Set $x_{k+1} = x_k - t_k \nabla f(x_k)$.
3. If $\|\nabla f(x_{k+1}\| \leq \epsilon$, then STOP and x_{k+1} is the output.

Find optimal solution of quadratic function

$$\min_{\boldsymbol{x} \in \mathbb{R}^{\boldsymbol{n}}} \{\boldsymbol{x^T A x + 2 b^T x}\} \tag{14}$$

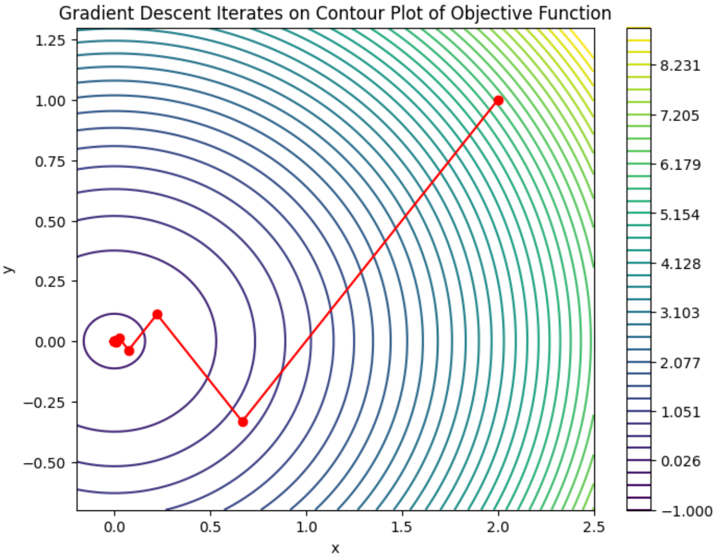
where $A \in \mathbb{R}^{n \times s}$ positive definite and $b \in \mathbb{R}^n$.

Consider the 2D minimization problem

$$\min_{x,y} x^2 + 2y^2 \tag{15}$$

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Zig-Zag Effect of Gradient Method



Condition Number

Definition

Let A be an $n \times n$ positive definite matrix. Then the condition number of A is defined by

$$\chi(A) = \frac{\lambda_{max}(A)}{\lambda_{min}(A)} \tag{16}$$

Gradient method applied to problems with large condition number might require large number of iterations and vice versa.

- Matrices with large condition number are called **ill-conditioned**.
- Matrices with small condition number are called **well-conditioned**.

Example with Code: Role of Condition Number

The *Rosenbrock function* is the following function:

$$f(x_1, x_2) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2. \tag{17}$$

The optimal solution is $(x_1, x_2) = (1, 1)$ with optimal value 0. The Rosenbrock function is extremely ill-conditioned at the optimal solution.

$$\nabla f(x) = \begin{pmatrix} -400x_1(x_2 - x_1^2) - 2(1 - x_1) \\ 200(x_2 - x_1^2) \end{pmatrix}, \tag{18}$$

$$\nabla^2 f(x) = \begin{pmatrix} -400x_2 + 1200x_1^2 + 2 & -400x_1 \\ -400x_1 & 200 \end{pmatrix}. \tag{19}$$

At $(x_1, x_2) = (1, 1)$,

$$\nabla^2 f(1, 1) = \begin{pmatrix} 802 & -400 \\ -400 & 200 \end{pmatrix} \tag{20}$$

Sensitivity of Solutions to Linear Systems

Sensitivity of the solution of the linear system to right-hand-side perturbations depends on the condition number of the coefficients matrix.

Consider a linear system $Ax = b$, and assume that A is positive definite. The solution is $x = A^{-1}b$.

Consider a perturbation $b + \Delta b$. Solution of the new system is

$$x + \Delta x = A^{-1}(b + \Delta b) = x + A^{-1}\Delta b,$$

so that $\Delta x = A^{-1}\Delta b$. Find a bound on the relative error $\frac{\|\Delta x\|}{\|x\|}$ in terms of $\frac{\|\Delta b\|}{\|b\|}$:

$$\frac{\|\Delta x\|}{\|x\|} = \frac{\|A^{-1}\Delta b\|}{\|x\|} \leq \frac{\|A^{-1}\|\|\Delta b\|}{\|x\|} = \frac{\lambda_{\max}(A^{-1})\|\Delta b\|}{\|x\|}, \tag{21}$$

the last equality follows from the fact that the spectral norm of a positive definite matrix D is $\|D\| = \lambda_{\max}(D)$. By the positive definiteness of A, it follows that $\lambda_{\max}(A^{-1}) = \frac{1}{\lambda_{\min}(A)}$:

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{1}{\lambda_{\min}(A)} \frac{\|\Delta b\|}{\|x\|} = \frac{1}{\lambda_{\min}(A)} \frac{\|\Delta b\|}{\|A^{-1}b\|} \leq \frac{1}{\lambda_{\min}(A)} \frac{\|\Delta b\|}{\lambda_{\min}(A^{-1})\|b\|} \tag{22}$$

$$= \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} \frac{\|\Delta b\|}{\|b\|} = \kappa(A) \frac{\|\Delta b\|}{\|b\|}, \tag{23}$$

Example with Code - Gradient Method

Consider the problem

$$\min\{1000x_1^2 + 40x_1x_2 + x_2^2\}$$

$$A = \begin{bmatrix} 1000 & 20 \\ 20 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Diagonal Scaling

Condition the problem by making an appropriate linear transformation of decision variables. Consider the unconstrained minimization problem

$$\min\{f(x) : x \in \mathbb{R}^n\}. \tag{24}$$

For a given nonsingular matrix $S \in \mathbb{R}^{n\times n}$, make the linear transformation $x = Sy$ and the equivalent problem is

$$\min\{g(y) \equiv f(Sy) : y \in \mathbb{R}^n\}. \tag{25}$$

Since $\nabla g(y) = S^T \nabla f(Sy)$, the gradient method applied to the transformed problem is

$$y_{k+1} = y_k - t_k S^T \nabla f(Sy_k). \tag{26}$$

Multiplying by S from the left, and using $x_k = Sy_k$, we obtain

$$x_{k+1} = x_k - t_k SS^T \nabla f(x_k). \tag{27}$$

Diagonal Scaling

Define $D = SS^T$, we obtain the scaled gradient method with scaling matrix D:

$$\boldsymbol{x_{k+1} = x_k - t_k D \nabla f(x_k)}. \tag{28}$$

By its definition, D is positive definite. The direction $-D\nabla f(x_k)$ is a descent direction of f at x_k when $\nabla f(x_k) \neq 0$ since

$$f'(x_k; -D\nabla f(x_k)) = -\nabla f(x_k)^T D \nabla f(x_k) < 0, \tag{29}$$

because of positive definiteness of D.

The scaled gradient method with scaling matrix D is equivalent to the gradient method employed on the function $g(y) = f(D^{1/2}y)$.
The gradient and Hessian of g are given by

$$\nabla g(y) = D^{1/2} \nabla f(D^{1/2}y) = D^{1/2} \nabla f(x), \tag{30}$$

$$\nabla^2 g(y) = D^{1/2} \nabla^2 f(D^{1/2}y) D^{1/2} = D^{1/2} \nabla^2 f(x) D^{1/2}, \tag{31}$$

where $x = D^{1/2}y$.

Scaled Gradient Method

Input: ϵ - tolerance parameter.
Initialization: Pick $x_0 \in \mathbb{R}^n$ arbitrarily.
General step: For any $k = 0, 1, 2, \ldots$ execute the following steps:

1. Pick a scaling matrix $D_k \succ 0$.
2. Pick a stepsize t_k by a line search procedure on the function

$$g(t) = f(x_k - tD_k \nabla f(x_k)). \tag{32}$$

3. Set $x_{k+1} = x_k - t_k D_k \nabla f(x_k)$.
4. If $\|\nabla f(x_{k+1})\| \leq \epsilon$, then STOP, and x_{k+1} is the output.

Diagonal Scaling

The main question is how to choose the scaling matrix D_k.

To accelerate the rate of convergence: Make scaled Hessian $D_k^{1/2}\nabla^2 f(x_k)D_k^{1/2}$ to be as close as possible to the identity matrix.

When $\nabla^2 f(x_k) > 0$, we can choose $D_k = (\nabla^2 f(x_k))^{-1}$ and the scaled Hessian becomes the identity matrix. The resulting method

$$x_{k+1} = x_k - t_k(\nabla^2 f(x_k))^{-1}\nabla f(x_k) \tag{33}$$

is the **Newton's method**.

Example with Code - Scaled Gradient Method

Consider the problem

$$\min\{1000x_1^2 + 40x_1x_2 + x_2^2\}$$

$$A = \begin{bmatrix} 1000 & 20 \\ 20 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Scaled gradient method with diagonal scaling matrix

$$A = \begin{bmatrix} \frac{1}{1000} & 0 \\ 0 & 1 \end{bmatrix}$$

Convergence Analysis of the Gradient Method

Lipschitz Property of the Gradient:
Given the unconstrained minimization problem

$$\min\{f(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^n\}$$

In order for gradient descent to work, we have to assume the object function f is continuously differentiable and its gradient ∇f is **Lipschitz continuous** over $\mathbb{R}^n$

Definition

A gradient ∇f is **Lipschitz continuous** over $\mathbb{R}^n$ when, for some $L \geq 0$:

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\| \text{ for any } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$$

Convergence Analysis of the Gradient Method

Definition

A gradient ∇f is **Lipschitz continuous** over $\mathbb{R}^n$ when, for some $L \geq 0$:

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\| \text{ for any } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$$

This L is called the **Lipschitz constant**

- If ∇f is Lipschitz with constant L, then it must also be Lipschitz with constant $\tilde{L}$ for all $\tilde{L} \geq L$
- There are an infinite number of Lipschitz constants, but we are usually only concerned with the smallest one.
- The class of functions with Lipschitz gradient with constant L is denoted by $C_L^{1,1}(\mathbb{R}^n)$ or $C_L^{1,1}$
    - $C^{k,\alpha}$ denotes a Hölder space
    - k - the left-hand side contains kth-order partial derivatives
    - α - the norm on the right-hand side is raised to the power α

Examples

- **Linear functions** Given $\mathbf{a} \in \mathbb{R}^n$, the function $f(\mathbf{x}) = \mathbf{a}^T\mathbf{x}$ is in $C_0^{1,1}$

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| = \mathbf{a} - \mathbf{a} = 0 \leq 0\|\mathbf{x} - \mathbf{y}\|$$

- **Quadratic functions** Let $\mathbf{A}$ be an n x n symmetric matrix, $\mathbf{b} \in \mathbb{R}^n$, and $c \in \mathbb{R}$. Then the function $f(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x} + 2\mathbf{b}^T\mathbf{x} + c$ is a $C_L^{1,1}$ function, where $L = 2\|\mathbf{A}\|$

Convergence Analysis of the Gradient Method

Theorem

Let f be a twice continuously differentiable function over $\mathbb{R}^n$. Then the following two claims are equivalent:

1. $f \in C_L^{1,1}(\mathbb{R}^n)$
2. $\|\nabla^2 f(\mathbf{x})\| \leq L$ *for any* $\mathbf{x} \in \mathbb{R}^n$

In other words, the gradient of f is Lipschitz continuous with Lipschitz constant L iff the norm of the Hessian of f is less than or equal to L

Example 4.21

Let $f : \mathbb{R} \to \mathbb{R}$ be given by $f(x) = \sqrt{1 + x^2}$. Then

$$0 \leq f''(x) = \frac{1}{(1 + x^2)^{3/2}} \leq 1$$

for any $x \in \mathbb{R}$, so $f \in C_1^{1,1}$

The Descent Lemma

$C^{1,1}$ functions can be bounded above by a quadratic function over the entire space, which is fundamental in convergence proofs of gradient-based methods

Lemma

Let $f \in C_L^{1,1}(\mathbb{R}^n)$. *Then for any* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2$$

Proof.

By the fundamental theorem of calculus,

$$f(\mathbf{y}) - f(\mathbf{x}) = \int_0^1 \langle \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})), \mathbf{y} - \mathbf{x} \rangle dt$$

Therefore,

$$f(\mathbf{y}) - f(\mathbf{x}) = \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \int_0^1 \langle \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle dt$$

Thus,

$$\begin{aligned}
|f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle| &= \left| \int_0^1 \langle \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle dt \right| \\
&\leq \int_0^1 |\langle \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle| dt \\
&\leq \int_0^1 \|\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x})\| \cdot \|\mathbf{y} - \mathbf{x}\| dt \\
&\leq \int_0^1 tL\|\mathbf{y} - \mathbf{x}\|^2 dt = \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2
\end{aligned}$$

Sufficient Decrease Lemma

Lemma

Sufficient Decrease Lemma: *Suppose that* $f \in C_L^{1,1}(\mathbb{R}^n)$. *Then for any* $\mathbf{x} \in \mathbb{R}^n$ *and* $t > 0$

$$f(\mathbf{x}) - f(\mathbf{x} - t\nabla f(\mathbf{x})) \geq t\left(1 - \frac{Lt}{2}\right)\|\nabla f(\mathbf{x})\|^2$$

A sufficient decrease property occurs in each of the stepsize selection strategies:

- constant
- exact line search
- backtracking

Lemma

Sufficient Decrease of the Gradient Method: *Let* $f \in C_L^{1,1}(\mathbb{R}^n)$. *Let* $\{\mathbf{x}_k\}_{k \geq 0}$ *be the sequence generated by the gradient method for solving*

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x})$$

with one of the following stepsize strategies:

- *constant stepsize* $\bar{t} \in (0, \frac{2}{L})$
- *exact line search*
- *backtracking procedure with parameters* $s \in \mathbb{R}_{++}, \alpha \in (0,1), \beta \in (0,1)$

Then,

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1}) \geq M \|\nabla f(\mathbf{x}_k)\|^2 \geq 0$$

Where

$$M = \begin{cases} \bar{t}(1 - \frac{\bar{t}L}{2}) & \text{constant stepsize} \\ \frac{1}{2L} & \text{exact line search} \\ \alpha \min\{s, \frac{2(1-\alpha)\beta}{L}\} & \text{backtracking} \end{cases}$$

Convergence of the Gradient Method

Theorem

Let $f \in C_L^{1,1}(\mathbb{R}^n)$ *and let* $\{\mathbf{x}_k\}_{k\geq 0}$ *be the sequence generated by the gradient method for solving*

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$$

With one of the following stepsize strategies

- *constant stepsize* $\bar{t} \in (0, \frac{2}{L})$
- *exact line search*
- *backtracking procedure with parameters* $s \in \mathbb{R}_{++}, \alpha \in (0,1), \beta \in (0,1)$

Assume that f *is bounded below over* $\mathbb{R}^n$*, that is, there exists* $m \in \mathbb{R}$ *such that* $f(\mathbf{x}) > m$ *for all* $\mathbf{x} \in \mathbb{R}^n$*. Then we have the following:*

1. *The sequence* $\{f(\mathbf{x}_k)\}_{k\geq 0}$ *is nonincreasing. In addition, for any* $k \geq 0$, $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_k)$ *unless* $\nabla f(\mathbf{x}_k) = 0$
2. $\nabla f(\mathbf{x}_k) \to 0$ *as* $k \to \infty$

Rate of Convergence of Gradient Norms

Theorem

Under the setting of the previous theorem, let f^* be the limit of the convergent sequence $\{f(\mathbf{x}_k)\}_{k\geq 0}$. Then for any $n = 0, 1, 2, \ldots$

$$\min_{k=0,1,\ldots,n} \|\nabla f(\mathbf{x}_k)\| \leq \sqrt{\frac{f(\mathbf{x_0}) - f^*}{M(n+1)}}$$

Where

$$M = \begin{cases} \bar{t}(1 - \frac{\bar{t}L}{2}) & \textit{constant stepsize} \\ \frac{1}{2L} & \textit{exact line search} \\ \alpha \min\{s, \frac{2(1-\alpha)\beta}{L}\} & \textit{backtracking} \end{cases}$$

Newton's Method

$$\mathbf{x}_{k+1} = \text{argmin}_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^T(\mathbf{x}-\mathbf{x}_k) + \frac{1}{2}(\mathbf{x}-\mathbf{x}_k)^T\nabla^2 f(\mathbf{x}_k)(\mathbf{x}-\mathbf{x_k})$$

- While gradient descent has linear convergence (locally), Newton's method has quadratic convergence (locally)
- This formula is not well defined unless we assume $\nabla^2 f(\mathbf{x}_k)$ is positive definite
  - When this is the case, we get Pure Newton's Method
- Each iteration is expensive computationally because it requires solving a system of linear equations.

Pure Newton's Method

Definition

Pure Newton's Method: Newton's Method when $\nabla^2 f(\mathbf{x}_k)$ is positive definite. The unique stationary point that minimizes this minimization problem is:

$$\nabla f(\mathbf{x}_k) + \nabla^2 f(\mathbf{x}_k)(\mathbf{x}_{k+1} - \mathbf{x}_k) = 0$$

Which is more useful when written as:

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

Definition

Newton Direction: The direction $\mathbf{d}_k$ the update formula steps in for each iteration.

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

When $\nabla^2 f(\mathbf{x}_k)$ is positive definite for any k, pure Newton's method is just a scaled gradient method and Newton's directions are descent directions.

Pure Newton's Method - Algorithm

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 direction $\mathbf{d}_k$, 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 $\mathbf{x}_{k+1}$ is the output.

Example 5.1

This example shows how $\nabla^2 f(\mathbf{x})$ being positive definite is not enough to guarantee convergence. The choice of $\mathbf{x}_0$ can also matter.

Consider the function $f(x) = \sqrt{1+x^2}$ defined over the real line. The minimizer of f over $\mathbb{R}$ is at $x = 0$. The first and second derivatives of f are

$$f'(x) = \frac{x}{\sqrt{1+x^2}}, f''(x) = \frac{1}{(1+x^2)^{3/2}}$$

So Pure Newton's Method has the form

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

- When $|x_0| \geq 1$, the method diverges
- When $|x_0| < 1$, the method converges to $x^* = 0$

Quadratic Local Convergence of Newton's Method

Let f be a twice continuously differntiable 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 $\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$ 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 $k = 0, 1, \ldots$ the inequality

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\| \leq \frac{L}{2m}\|\mathbf{x}_k - \mathbf{x}^*\|^2 \tag{34}$$

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

$$\|\mathbf{x}_k - \mathbf{x}^*\| \leq \frac{2m}{L}\left(\frac{1}{2}\right)^{2^k}, \ k = 0, 1, 2, \ldots \tag{35}$$

Example 5.3

$$f(x, y) = 100 * x^4 + 0.01 * y^4$$

$$(x_0, y_0) = (1, 1)$$

Damped Newton's Method - Algorithm

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 direction $\mathbf{d}_k$, 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. $\mathbf{x}_{k+1} = \mathbf{x}_k + t_k \mathbf{d}_k$.
4. If $\|\nabla f(\mathbf{x}_{k+1})\| \leq \epsilon$, then STOP, and $\mathbf{x}_{k+1}$ is the output.

Example 5.5

$$f(x, y) = \sqrt{x^2 + 1} + \sqrt{y^2 + 1}$$
$$(x_0, y_0) = (10, 10)$$

Hybrid Gradient-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. If $\nabla^2 f(\mathbf{x}_k) \succ 0$, then take $\mathbf{d}_k$ as the Newton direction $\mathbf{d}_k$, which is the solution to the linear system $\nabla^2 f(\mathbf{x}_k)\mathbf{d}_k = -\nabla f(\mathbf{x}_k)$. Otherwise, set $\mathbf{d}_k = -\nabla f(\mathbf{x}_k)$
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. $\mathbf{x}_{k+1} = \mathbf{x}_k + t_k \mathbf{d}_k$.
4. If $\|\nabla f(\mathbf{x}_{k+1})\| \le \epsilon$, then STOP, and $\mathbf{x}_{k+1}$ is the output.

Example 5.8 - Rosenbrock Function

$$f(x_1, x_2) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$$

- When a minimum is found with backtracking, it takes about 6900 iterations.
- With the Hybrid-Gradient Newton Method, it only takes 17 iterations!

Exercises

Beck 4.2, 4.3, 4.7, 5.2