Introduction to Optimization

K. R. Sahasranand

Data Science

sahasranand@iitpkd.ac.in

A general *iterative* algorithm

for finding the minimum of $f : \mathbb{R}^n \to \mathbb{R}$

i. Pick a starting point $x^{(0)} \in \mathbb{R}^n$.

A general *iterative* algorithm

for finding the minimum of $f : \mathbb{R}^n \to \mathbb{R}$

i. Pick a starting point $x^{(0)} \in \mathbb{R}^n$.

ii. For $k \geq 0$, let

$$x^{(k+1)} = x^{(k)} + \alpha d^{(k)}$$

where $d^{(k)} \in \mathbb{R}^n$.

A general *iterative* algorithm

for finding the minimum of $f : \mathbb{R}^n \to \mathbb{R}$

i. Pick a starting point $x^{(0)} \in \mathbb{R}^n$.

ii. For $k \geq 0$, let

$$x^{(k+1)} = x^{(k)} + \alpha d^{(k)}$$

where $d^{(k)} \in \mathbb{R}^n$.

$d^{(k)}$: which direction to move in?

α : by how much? "step size"

A general *iterative* algorithm

for finding the minimum of $f : \mathbb{R}^n \to \mathbb{R}$

i. Pick a starting point $x^{(0)} \in \mathbb{R}^n$.

ii. For $k \geq 0$, let

$$x^{(k+1)} = x^{(k)} + \alpha d^{(k)}$$

where $d^{(k)} \in \mathbb{R}^n$.

$d^{(k)}$: which direction to move in?

α : by how much? "step size"

Which direction yields the "most reduction in $f(\cdot)$"?

Enter Taylor's theorem

$f(x) = f(x_0) + Df(x_0)(x - x_0) + o(\|x - x_0\|)$

- First order approximation:

$$f(x^{(k+1)}) = f(x^{(k)} + \alpha d^{(k)})$$

Enter Taylor's theorem

$f(x) = f(x_0) + Df(x_0)(x - x_0) + o(\|x - x_0\|)$

- First order approximation:

$$\begin{aligned} f(x^{(k+1)}) &= f(x^{(k)} + \alpha d^{(k)}) \\ &= f(x^{(k)}) + \alpha \cdot \nabla f(x^{(k)})^\top d^{(k)} + o(\alpha \|d^{(k)}\|). \end{aligned}$$

Enter Taylor's theorem

$f(x) = f(x_0) + Df(x_0)(x - x_0) + o(\|x - x_0\|)$

- First order approximation:

$$\begin{aligned} f(x^{(k+1)}) &= f(x^{(k)} + \alpha d^{(k)}) \\ &= f(x^{(k)}) + \alpha \cdot \nabla f(x^{(k)})^\top d^{(k)} + o(\alpha \|d^{(k)}\|). \end{aligned}$$

- For what $d^{(k)}$ is $\left\langle \nabla f(x^{(k)}), d^{(k)} \right\rangle$ the smallest?

Enter Taylor's theorem

$f(x) = f(x_0) + Df(x_0)(x - x_0) + o(\|x - x_0\|)$

- First order approximation:

$$\begin{aligned} f(x^{(k+1)}) &= f(x^{(k)} + \alpha d^{(k)}) \\ &= f(x^{(k)}) + \alpha \cdot \nabla f(x^{(k)})^\top d^{(k)} + o(\alpha \|d^{(k)}\|). \end{aligned}$$

- For what $d^{(k)}$ is $\left\langle \nabla f(x^{(k)}), d^{(k)} \right\rangle$ the smallest?

- By Cauchy-Schwarz,

$$-\|\nabla f(x^{(k)})\| \cdot \|d^{(k)}\| \leq \left\langle \nabla f(x^{(k)}), d^{(k)} \right\rangle$$

Enter Taylor's theorem

$f(x) = f(x_0) + Df(x_0)(x - x_0) + o(\|x - x_0\|)$

- First order approximation:

$$\begin{aligned} f(x^{(k+1)}) &= f(x^{(k)} + \alpha d^{(k)}) \\ &= f(x^{(k)}) + \alpha \cdot \nabla f(x^{(k)})^\top d^{(k)} + o(\alpha \|d^{(k)}\|). \end{aligned}$$

- For what $d^{(k)}$ is $\left\langle \nabla f(x^{(k)}), d^{(k)} \right\rangle$ the smallest?

- By Cauchy-Schwarz,

$$-\|\nabla f(x^{(k)})\| \cdot \|d^{(k)}\| \leq \left\langle \nabla f(x^{(k)}), d^{(k)} \right\rangle$$

- Equality when

$$d^{(k)} = -\nabla f(x^{(k)}).$$

Enter Taylor's theorem

$f(x) = f(x_0) + Df(x_0)(x - x_0) + o(\|x - x_0\|)$

- First order approximation:

$$\begin{aligned} f(x^{(k+1)}) &= f(x^{(k)} + \alpha d^{(k)}) \\ &= f(x^{(k)}) + \alpha \cdot \nabla f(x^{(k)})^\top d^{(k)} + o(\alpha \|d^{(k)}\|). \end{aligned}$$

- For what $d^{(k)}$ is $\left\langle \nabla f(x^{(k)}), d^{(k)} \right\rangle$ the smallest?
- By Cauchy-Schwarz,

$$-\|\nabla f(x^{(k)})\| \cdot \|d^{(k)}\| \leq \left\langle \nabla f(x^{(k)}), d^{(k)} \right\rangle$$

- Equality when

$$d^{(k)} = -\nabla f(x^{(k)}).$$

- For $\alpha > 0$ small enough, $f(x^{(k+1)}) < f(x^{(k)})$.

Enter Taylor's theorem

$f(x) = f(x_0) + Df(x_0)(x - x_0) + o(\|x - x_0\|)$

- First order approximation:

$$\begin{aligned} f(x^{(k+1)}) &= f(x^{(k)} + \alpha d^{(k)}) \\ &= f(x^{(k)}) + \alpha \cdot \nabla f(x^{(k)})^\top d^{(k)} + o(\alpha \|d^{(k)}\|). \end{aligned}$$

- For what $d^{(k)}$ is $\left\langle \nabla f(x^{(k)}), d^{(k)} \right\rangle$ the smallest?

- By Cauchy-Schwarz,

$$-\|\nabla f(x^{(k)})\| \cdot \|d^{(k)}\| \leq \left\langle \nabla f(x^{(k)}), d^{(k)} \right\rangle$$

- Equality when

$$d^{(k)} = -\nabla f(x^{(k)}).$$

- For $\alpha > 0$ small enough, $f(x^{(k+1)}) < f(x^{(k)})$. $\boxed{\text{Gradient descent}}$

How to choose α?

Step size

Possible to use a different α in each iteration

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)}).$$

How to choose α?

Step size

Possible to use a different α in each iteration

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)}).$$

We choose α_k to minimize

$$\phi_k(\alpha) := f(x^{(k)} - \alpha \nabla f(x^{(k)}))$$

How to choose α?

Step size

Possible to use a different α in each iteration

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)}).$$

We choose α_k to minimize

$$\phi_k(\alpha) := f(x^{(k)} - \alpha \nabla f(x^{(k)}))$$

That is,

$$\alpha_k = \arg\min_{\alpha \geq 0} \phi_k(\alpha).$$

How to choose α?

Step size

Possible to use a different α in each iteration

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)}).$$

We choose α_k to minimize

$$\phi_k(\alpha) := f(x^{(k)} - \alpha \nabla f(x^{(k)}))$$

That is,

$$\alpha_k = \arg\min_{\alpha \geq 0} \phi_k(\alpha).$$

How to calculate?

How to choose α?

Step size

Possible to use a different α in each iteration

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)}).$$

We choose α_k to minimize

$$\phi_k(\alpha) := f(x^{(k)} - \alpha \nabla f(x^{(k)}))$$

That is,

$$\alpha_k = \arg\min_{\alpha \geq 0} \phi_k(\alpha).$$

How to calculate? *Line search on $\phi_k(\alpha)$*

Steepest descent

Gradient descent with optimized step size

i. Pick a starting point $x^{(0)} \in \mathbb{R}^n$.

ii. For $k \geq 0$, let

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)})$$

where

$$\alpha_k = \arg\min_{\alpha \geq 0} f(x^{(k)} - \alpha \nabla f(x^{(k)})).$$

Steepest descent

Gradient descent with optimized step size

i. Pick a starting point $x^{(0)} \in \mathbb{R}^n$.

ii. For $k \geq 0$, let

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)})$$

where

$$\alpha_k = \arg\min_{\alpha \geq 0} f(x^{(k)} - \alpha \nabla f(x^{(k)})).$$

(Descent property) If $\nabla f(x^{(k)}) \neq 0$, then $f(x^{(k+1)}) < f(x^{(k)})$.

Steepest descent

Gradient descent with optimized step size

i. Pick a starting point $x^{(0)} \in \mathbb{R}^n$.

ii. For $k \geq 0$, let

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)})$$

where

$$\alpha_k = \arg\min_{\alpha \geq 0} f(x^{(k)} - \alpha \nabla f(x^{(k)})).$$

(Descent property) If $\nabla f(x^{(k)}) \neq 0$, then $f(x^{(k+1)}) < f(x^{(k)})$.

Proof sketch: We have $\phi_k(\alpha_k) \leq \phi_k(\alpha)$.

Steepest descent

Gradient descent with optimized step size

i. Pick a starting point $x^{(0)} \in \mathbb{R}^n$.

ii. For $k \geq 0$, let

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)})$$

where

$$\alpha_k = \arg\min_{\alpha \geq 0} f(x^{(k)} - \alpha \nabla f(x^{(k)})).$$

(Descent property) If $\nabla f(x^{(k)}) \neq 0$, then $f(x^{(k+1)}) < f(x^{(k)})$.

Proof sketch: We have $\phi_k(\alpha_k) \leq \phi_k(\alpha)$. By chain rule,

$$\phi_k'(0) = \frac{d\phi_k}{d\alpha}(0)$$

Steepest descent

Gradient descent with optimized step size

i. Pick a starting point $x^{(0)} \in \mathbb{R}^n$.

ii. For $k \geq 0$, let

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)})$$

where

$$\alpha_k = \arg\min_{\alpha \geq 0} f(x^{(k)} - \alpha \nabla f(x^{(k)})).$$

(Descent property) If $\nabla f(x^{(k)}) \neq 0$, then $f(x^{(k+1)}) < f(x^{(k)})$.

Proof sketch: We have $\phi_k(\alpha_k) \leq \phi_k(\alpha)$. By chain rule,

$$\phi_k'(0) = \frac{d\phi_k}{d\alpha}(0) = \nabla f(x^{(k)} - 0\nabla f(x^{(k)}))^\top(-\nabla f(x^{(k)}))$$

Steepest descent

Gradient descent with optimized step size

i. Pick a starting point $x^{(0)} \in \mathbb{R}^n$.

ii. For $k \geq 0$, let

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)})$$

where

$$\alpha_k = \arg\min_{\alpha \geq 0} f(x^{(k)} - \alpha \nabla f(x^{(k)})).$$

(Descent property) If $\nabla f(x^{(k)}) \neq 0$, then $f(x^{(k+1)}) < f(x^{(k)})$.

Proof sketch: We have $\phi_k(\alpha_k) \leq \phi_k(\alpha)$. By chain rule,

$$\phi_k'(0) = \frac{d\phi_k}{d\alpha}(0) = \nabla f(x^{(k)} - 0\nabla f(x^{(k)}))^\top(-\nabla f(x^{(k)})) = -\|\nabla f(x^{(k)})\|^2 < 0.$$

Steepest descent

Gradient descent with optimized step size

i. Pick a starting point $x^{(0)} \in \mathbb{R}^n$.

ii. For $k \geq 0$, let

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)})$$

where

$$\alpha_k = \arg\min_{\alpha \geq 0} f(x^{(k)} - \alpha \nabla f(x^{(k)})).$$

(Descent property) If $\nabla f(x^{(k)}) \neq 0$, then $f(x^{(k+1)}) < f(x^{(k)})$.

Proof sketch: We have $\phi_k(\alpha_k) \leq \phi_k(\alpha)$. By chain rule,

$$\phi_k'(0) = \frac{d\phi_k}{d\alpha}(0) = \nabla f(x^{(k)} - 0\nabla f(x^{(k)}))^\top(-\nabla f(x^{(k)})) = -\|\nabla f(x^{(k)})\|^2 < 0.$$

Then, $\exists\ \overline{\alpha}$ such that $\phi_k(0) > \phi_k(\alpha)$ for all $\alpha \in (0, \overline{\alpha}]$.

Steepest descent

Gradient descent with optimized step size

i. Pick a starting point $x^{(0)} \in \mathbb{R}^n$.

ii. For $k \geq 0$, let

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)})$$

where

$$\alpha_k = \arg\min_{\alpha \geq 0} f(x^{(k)} - \alpha \nabla f(x^{(k)})).$$

(Descent property) If $\nabla f(x^{(k)}) \neq 0$, then $f(x^{(k+1)}) < f(x^{(k)})$.

Proof sketch: We have $\phi_k(\alpha_k) \leq \phi_k(\alpha)$. By chain rule,

$$\phi_k'(0) = \frac{d\phi_k}{d\alpha}(0) = \nabla f(x^{(k)} - 0\nabla f(x^{(k)}))^\top(-\nabla f(x^{(k)})) = -\|\nabla f(x^{(k)})\|^2 < 0.$$

Then, $\exists\ \overline{\alpha}$ such that $\phi_k(0) > \phi_k(\alpha)$ for all $\alpha \in (0, \overline{\alpha}]$.

Therefore,

$$f(x^{(k+1)}) = \phi_k(\alpha_k)$$

Steepest descent

Gradient descent with optimized step size

i. Pick a starting point $x^{(0)} \in \mathbb{R}^n$.

ii. For $k \geq 0$, let

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)})$$

where

$$\alpha_k = \arg\min_{\alpha \geq 0} f(x^{(k)} - \alpha \nabla f(x^{(k)})).$$

(Descent property) If $\nabla f(x^{(k)}) \neq 0$, then $f(x^{(k+1)}) < f(x^{(k)})$.

Proof sketch: We have $\phi_k(\alpha_k) \leq \phi_k(\alpha)$. By chain rule,

$$\phi_k'(0) = \frac{d\phi_k}{d\alpha}(0) = \nabla f(x^{(k)} - 0\nabla f(x^{(k)}))^\top(-\nabla f(x^{(k)})) = -\|\nabla f(x^{(k)})\|^2 < 0.$$

Then, $\exists\ \overline{\alpha}$ such that $\phi_k(0) > \phi_k(\alpha)$ for all $\alpha \in (0, \overline{\alpha}]$.

Therefore,

$$f(x^{(k+1)}) = \phi_k(\alpha_k) \leq \phi_k(\overline{\alpha})$$

Steepest descent

Gradient descent with optimized step size

i. Pick a starting point $x^{(0)} \in \mathbb{R}^n$.

ii. For $k \geq 0$, let

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)})$$

where

$$\alpha_k = \arg\min_{\alpha \geq 0} f(x^{(k)} - \alpha \nabla f(x^{(k)})).$$

(Descent property) If $\nabla f(x^{(k)}) \neq 0$, then $f(x^{(k+1)}) < f(x^{(k)})$.

Proof sketch: We have $\phi_k(\alpha_k) \leq \phi_k(\alpha)$. By chain rule,

$$\phi_k'(0) = \frac{d\phi_k}{d\alpha}(0) = \nabla f(x^{(k)} - 0\nabla f(x^{(k)}))^\top(-\nabla f(x^{(k)})) = -\|\nabla f(x^{(k)})\|^2 < 0.$$

Then, $\exists\ \overline{\alpha}$ such that $\phi_k(0) > \phi_k(\alpha)$ for all $\alpha \in (0, \overline{\alpha}]$.

Therefore,

$$f(x^{(k+1)}) = \phi_k(\alpha_k) \leq \phi_k(\overline{\alpha}) < \phi_k(0)$$

Steepest descent

Gradient descent with optimized step size

i. Pick a starting point $x^{(0)} \in \mathbb{R}^n$.

ii. For $k \geq 0$, let

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)})$$

where

$$\alpha_k = \arg\min_{\alpha \geq 0} f(x^{(k)} - \alpha \nabla f(x^{(k)})).$$

(Descent property) If $\nabla f(x^{(k)}) \neq 0$, then $f(x^{(k+1)}) < f(x^{(k)})$.

Proof sketch: We have $\phi_k(\alpha_k) \leq \phi_k(\alpha)$. By chain rule,

$$\phi_k'(0) = \frac{d\phi_k}{d\alpha}(0) = \nabla f(x^{(k)} - 0\nabla f(x^{(k)}))^\top(-\nabla f(x^{(k)})) = -\|\nabla f(x^{(k)})\|^2 < 0.$$

Then, $\exists\ \overline{\alpha}$ such that $\phi_k(0) > \phi_k(\alpha)$ for all $\alpha \in (0, \overline{\alpha}]$.

Therefore,

$$f(x^{(k+1)}) = \phi_k(\alpha_k) \leq \phi_k(\overline{\alpha}) < \phi_k(0) = f(x^{(k)}).$$

Steepest descent

Gradient descent with optimized step size

i. Pick a starting point $x^{(0)} \in \mathbb{R}^n$.

ii. For $k \geq 0$, let

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)})$$

where

$$\alpha_k = \arg\min_{\alpha \geq 0} f(x^{(k)} - \alpha \nabla f(x^{(k)})).$$

(Descent property) If $\nabla f(x^{(k)}) \neq 0$, then $f(x^{(k+1)}) < f(x^{(k)})$.

Proof sketch: We have $\phi_k(\alpha_k) \leq \phi_k(\alpha)$. By chain rule,

$$\phi_k'(0) = \frac{d\phi_k}{d\alpha}(0) = \nabla f(x^{(k)} - 0\nabla f(x^{(k)}))^\top(-\nabla f(x^{(k)})) = -\|\nabla f(x^{(k)})\|^2 < 0.$$

Then, $\exists\ \overline{\alpha}$ such that $\phi_k(0) > \phi_k(\alpha)$ for all $\alpha \in (0, \overline{\alpha}]$.

Therefore,

$$f(x^{(k+1)}) = \phi_k(\alpha_k) \leq \phi_k(\overline{\alpha}) < \phi_k(0) = f(x^{(k)}).$$

(see Theorem 8.2 in textbook) $\square$

Steepest descent

Gradient descent with optimized step size

i. Pick a starting point $x^{(0)} \in \mathbb{R}^n$.

ii. For $k \geq 0$, let

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)})$$

where

$$\alpha_k = \arg\min_{\alpha \geq 0} f(x^{(k)} - \alpha \nabla f(x^{(k)})).$$

(Orthogonal steps) For each k, the vector $x^{(k+1)} - x^{(k)}$ is orthogonal to the vector $x^{(k+2)} - x^{(k+1)}$.

Steepest descent

Gradient descent with optimized step size

i. Pick a starting point $x^{(0)} \in \mathbb{R}^n$.

ii. For $k \geq 0$, let

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)})$$

where

$$\alpha_k = \arg\min_{\alpha \geq 0} f(x^{(k)} - \alpha \nabla f(x^{(k)})).$$

(Orthogonal steps) For each k, the vector $x^{(k+1)} - x^{(k)}$ is orthogonal to the vector $x^{(k+2)} - x^{(k+1)}$.

Proof sketch: We have

$$\left\langle x^{(k+1)} - x^{(k)}, x^{(k+2)} - x^{(k+1)} \right\rangle = \alpha_k \alpha_{k+1} \left\langle \nabla f(x^{(k)}), \nabla f(x^{(k+1)}) \right\rangle$$

Steepest descent

Gradient descent with optimized step size

i. Pick a starting point $x^{(0)} \in \mathbb{R}^n$.

ii. For $k \geq 0$, let

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)})$$

where

$$\alpha_k = \arg\min_{\alpha \geq 0} f(x^{(k)} - \alpha \nabla f(x^{(k)})).$$

(Orthogonal steps) For each k, the vector $x^{(k+1)} - x^{(k)}$ is orthogonal to the vector $x^{(k+2)} - x^{(k+1)}$.

Proof sketch: We have

$$\left\langle x^{(k+1)} - x^{(k)}, x^{(k+2)} - x^{(k+1)} \right\rangle = \alpha_k \alpha_{k+1} \left\langle \nabla f(x^{(k)}), \nabla f(x^{(k+1)}) \right\rangle$$

By chain rule,

$$0 = \phi_k'(\alpha_k) = \frac{d\phi_k}{d\alpha}(\alpha_k)$$

Steepest descent

Gradient descent with optimized step size

i. Pick a starting point $x^{(0)} \in \mathbb{R}^n$.

ii. For $k \geq 0$, let

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)})$$

where

$$\alpha_k = \arg\min_{\alpha \geq 0} f(x^{(k)} - \alpha \nabla f(x^{(k)})).$$

(Orthogonal steps) For each k, the vector $x^{(k+1)} - x^{(k)}$ is orthogonal to the vector $x^{(k+2)} - x^{(k+1)}$.

Proof sketch: We have

$$\left\langle x^{(k+1)} - x^{(k)}, x^{(k+2)} - x^{(k+1)} \right\rangle = \alpha_k \alpha_{k+1} \left\langle \nabla f(x^{(k)}), \nabla f(x^{(k+1)}) \right\rangle$$

By chain rule,

$$0 = \phi_k'(\alpha_k) = \frac{d\phi_k}{d\alpha}(\alpha_k) = \nabla f(x^{(k)} - \alpha_k \nabla f(x^{(k)}))^\top (-\nabla f(x^{(k)}))$$

Steepest descent

Gradient descent with optimized step size

i. Pick a starting point $x^{(0)} \in \mathbb{R}^n$.

ii. For $k \geq 0$, let
$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)})$$
where
$$\alpha_k = \arg\min_{\alpha \geq 0} f(x^{(k)} - \alpha \nabla f(x^{(k)})).$$

(Orthogonal steps) For each k, the vector $x^{(k+1)} - x^{(k)}$ is orthogonal to the vector $x^{(k+2)} - x^{(k+1)}$.

Proof sketch: We have
$$\left\langle x^{(k+1)} - x^{(k)}, x^{(k+2)} - x^{(k+1)} \right\rangle = \alpha_k \alpha_{k+1} \left\langle \nabla f(x^{(k)}), \nabla f(x^{(k+1)}) \right\rangle$$

By chain rule,
$$0 = \phi_k'(\alpha_k) = \frac{d\phi_k}{d\alpha}(\alpha_k) = \nabla f(x^{(k)} - \alpha_k \nabla f(x^{(k)}))^\top (-\nabla f(x^{(k)}))$$
$$= -\left\langle \nabla f(x^{(k)}), \nabla f(x^{(k+1)}) \right\rangle.$$

Steepest descent

Gradient descent with optimized step size

i. Pick a starting point $x^{(0)} \in \mathbb{R}^n$.

ii. For $k \geq 0$, let

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)})$$

where

$$\alpha_k = \arg\min_{\alpha \geq 0} f(x^{(k)} - \alpha \nabla f(x^{(k)})).$$

(Orthogonal steps) For each k, the vector $x^{(k+1)} - x^{(k)}$ is orthogonal to the vector $x^{(k+2)} - x^{(k+1)}$.

Proof sketch: We have

$$\left\langle x^{(k+1)} - x^{(k)}, x^{(k+2)} - x^{(k+1)} \right\rangle = \alpha_k \alpha_{k+1} \left\langle \nabla f(x^{(k)}), \nabla f(x^{(k+1)}) \right\rangle$$

By chain rule,

$$0 = \phi_k'(\alpha_k) = \frac{d\phi_k}{d\alpha}(\alpha_k) = \nabla f(x^{(k)} - \alpha_k \nabla f(x^{(k)}))^\top (-\nabla f(x^{(k)}))$$
$$= -\left\langle \nabla f(x^{(k)}), \nabla f(x^{(k+1)}) \right\rangle.$$

(see Theorem 8.1 in textbook) $\square$

Order of convergence

a.k.a. rate of convergence

Given a sequence

$$x^{(k)} \to x^*,$$

we say that the order of convergence is at least $p \in \mathbb{R}$ if

$$\|x^{(k+1)} - x^*\| = O\left(\|x^{(k)} - x^*\|^p\right).$$

Order of convergence

a.k.a. rate of convergence

Given a sequence

$$x^{(k)} \to x^*,$$

we say that the order of convergence is at least $p \in \mathbb{R}$ if

$$\|x^{(k+1)} - x^*\| = O\left(\|x^{(k)} - x^*\|^p\right).$$

Steepest descent: order of convergence is 1 (in the worst case).

Order of convergence

a.k.a. rate of convergence

Given a sequence

$$x^{(k)} \to x^*,$$

we say that the order of convergence is at least $p \in \mathbb{R}$ if

$$\|x^{(k+1)} - x^*\| = O\left(\|x^{(k)} - x^*\|^p\right).$$

Steepest descent: order of convergence is 1 (in the worst case).

Newton's method: order of convergence is 2
if the initial guess is near the solution.

Newton's method

Quadratic approximation using Taylor's theorem

Second-order approximation (ignoring the $o(\cdot)$ term):

$$\begin{aligned} f(x^{(k+1)}) = f(x^{(k)}) &+ (x^{(k+1)} - x^{(k)})^\top \nabla f(x^{(k)}) \\ &+ \frac{1}{2}(x^{(k+1)} - x^{(k)})^\top F(x^{(k)})(x^{(k+1)} - x^{(k)}). \end{aligned}$$

Newton's method

Quadratic approximation using Taylor's theorem

Second-order approximation (ignoring the $o(\cdot)$ term):

$$\begin{aligned} f(x^{(k+1)}) = f(x^{(k)}) &+ (x^{(k+1)} - x^{(k)})^\top \nabla f(x^{(k)}) \\ &+ \frac{1}{2}(x^{(k+1)} - x^{(k)})^\top F(x^{(k)})(x^{(k+1)} - x^{(k)}). \end{aligned}$$

Applying the FONC yields

$$0 = \nabla f(x^{(k)}) + F(x^{(k)})(x^{(k+1)} - x^{(k)}).$$

Newton's method

Quadratic approximation using Taylor's theorem

Second-order approximation (ignoring the $o(\cdot)$ term):

$$\begin{aligned} f(x^{(k+1)}) &= f(x^{(k)}) + (x^{(k+1)} - x^{(k)})^\top \nabla f(x^{(k)}) \\ &\qquad + \frac{1}{2}(x^{(k+1)} - x^{(k)})^\top F(x^{(k)})(x^{(k+1)} - x^{(k)}). \end{aligned}$$

Applying the FONC yields

$$0 = \nabla f(x^{(k)}) + F(x^{(k)})(x^{(k+1)} - x^{(k)}).$$

whereby

$$x^{(k+1)} = x^{(k)} - F(x^{(k)})^{-1} \nabla f(x^{(k)})$$

Newton's method

Quadratic approximation using Taylor's theorem

Second-order approximation (ignoring the $o(\cdot)$ term):

$$\begin{aligned} f(x^{(k+1)}) = f(x^{(k)}) &+ (x^{(k+1)} - x^{(k)})^\top \nabla f(x^{(k)}) \\ &+ \frac{1}{2}(x^{(k+1)} - x^{(k)})^\top F(x^{(k)})(x^{(k+1)} - x^{(k)}). \end{aligned}$$

Applying the FONC yields

$$0 = \nabla f(x^{(k)}) + F(x^{(k)})(x^{(k+1)} - x^{(k)}).$$

whereby

$$x^{(k+1)} = x^{(k)} - F(x^{(k)})^{-1} \nabla f(x^{(k)})$$

For $f : \mathbb{R} \to \mathbb{R}$,

Newton's method: $x^{(k+1)} = x^{(k)} - \frac{f'(x^{(k)})}{f''(x^{(k)})}$.

Newton's method

Quadratic approximation using Taylor's theorem

Second-order approximation (ignoring the $o(\cdot)$ term):

$$\begin{aligned} f(x^{(k+1)}) = f(x^{(k)}) &+ (x^{(k+1)} - x^{(k)})^\top \nabla f(x^{(k)}) \\ &+ \frac{1}{2}(x^{(k+1)} - x^{(k)})^\top F(x^{(k)})(x^{(k+1)} - x^{(k)}). \end{aligned}$$

Applying the FONC yields

$$0 = \nabla f(x^{(k)}) + F(x^{(k)})(x^{(k+1)} - x^{(k)}).$$

whereby

$$x^{(k+1)} = x^{(k)} - F(x^{(k)})^{-1} \nabla f(x^{(k)})$$

For $f : \mathbb{R} \to \mathbb{R}$,

Newton's method: $x^{(k+1)} = x^{(k)} - \frac{f'(x^{(k)})}{f''(x^{(k)})}$.

Secant method: $f''(x^{(k)}) \approx \frac{f'(x^{(k)}) - f'(x^{(k-1)})}{x^{(k)} - x^{(k-1)}}$

Newton's method

Quadratic approximation using Taylor's theorem

Second-order approximation (ignoring the $o(\cdot)$ term):

$$f(x^{(k+1)}) = f(x^{(k)}) + (x^{(k+1)} - x^{(k)})^\top \nabla f(x^{(k)}) + \frac{1}{2}(x^{(k+1)} - x^{(k)})^\top F(x^{(k)})(x^{(k+1)} - x^{(k)}).$$

Applying the FONC yields

$$0 = \nabla f(x^{(k)}) + F(x^{(k)})(x^{(k+1)} - x^{(k)}).$$

whereby

$$x^{(k+1)} = x^{(k)} - F(x^{(k)})^{-1} \nabla f(x^{(k)})$$

For $f : \mathbb{R} \to \mathbb{R}$,

Newton's method: $x^{(k+1)} = x^{(k)} - \frac{f'(x^{(k)})}{f''(x^{(k)})}$.

Secant method: $f''(x^{(k)}) \approx \frac{f'(x^{(k)}) - f'(x^{(k-1)})}{x^{(k)} - x^{(k-1)}}$ (see sections $7.5, 7.6$)

An example

Newton's method for finding the minimum of a quadratic function

Let $Q \in \mathbb{R}^{n \times n}$ be symmetric, positive definite. Let

$$f(x) = \frac{1}{2} x^\top Q x - b^\top x.$$

An example

Newton's method for finding the minimum of a quadratic function

Let $Q \in \mathbb{R}^{n \times n}$ be symmetric, positive definite. Let

$$f(x) = \frac{1}{2} x^\top Q x - b^\top x.$$

Then,

$$\nabla f(x) = Qx - b$$

An example

Newton's method for finding the minimum of a quadratic function

Let $Q \in \mathbb{R}^{n \times n}$ be symmetric, positive definite. Let

$$f(x) = \frac{1}{2} x^\top Q x - b^\top x.$$

Then,

$$\nabla f(x) = Qx - b \qquad \text{and} \qquad F(x) = Q.$$

An example

Newton's method for finding the minimum of a quadratic function

Let $Q \in \mathbb{R}^{n\times n}$ be symmetric, positive definite. Let

$$f(x) = \frac{1}{2}x^\top Q x - b^\top x.$$

Then,

$$\nabla f(x) = Qx - b \qquad \text{and} \qquad F(x) = Q.$$

Start at an initial point $x^{(0)}$.

An example

Newton's method for finding the minimum of a quadratic function

Let $Q \in \mathbb{R}^{n \times n}$ be symmetric, positive definite. Let

$$f(x) = \frac{1}{2} x^\top Q x - b^\top x.$$

Then,

$$\nabla f(x) = Qx - b \qquad \text{and} \qquad F(x) = Q.$$

Start at an initial point $x^{(0)}$.

$$x^{(1)} = x^{(0)} - Q^{-1}(Qx^{(0)} - b) = Q^{-1} b = x^*.$$

An example

Newton's method for finding the minimum of a quadratic function

Let $Q \in \mathbb{R}^{n \times n}$ be symmetric, positive definite. Let

$$f(x) = \frac{1}{2} x^\top Q x - b^\top x.$$

Then,

$$\nabla f(x) = Qx - b \qquad \text{and} \qquad F(x) = Q.$$

Start at an initial point $x^{(0)}$.

$$x^{(1)} = x^{(0)} - Q^{-1}(Qx^{(0)} - b) = Q^{-1}b = x^*.$$

Verify that $\nabla f(x^*) = 0$.

Newton's method

Hiccups

- Quick convergence if the starting point is near the solution.

Newton's method

Hiccups

- Quick convergence if the starting point is near the solution.

- Not guaranteed to converge if we start far away from it.

Newton's method

Hiccups

- Quick convergence if the starting point is near the solution.

- Not guaranteed to converge if we start far away from it.

- May not even be well-defined – the Hessian may be singular.

Newton's method

Hiccups

- Quick convergence if the starting point is near the solution.

- Not guaranteed to converge if we start far away from it.

- May not even be well-defined – the Hessian may be singular.

- May not be a descent method; it is possible that

$$f(x^{(k+1)}) \geq f(x^{(k)}).$$

Descent property for Newton's method

Theorem 9.2 in textbook

Theorem – If $F(x^{(k)}) > 0$ and $\nabla f(x^{(k)}) \neq 0$,

Descent property for Newton's method

Theorem 9.2 in textbook

Theorem – If $F(x^{(k)}) > 0$ and $\nabla f(x^{(k)}) \neq 0$, then the search direction

$$d^{(k)} = -F(x^{(k)})^{-1} \nabla f(x^{(k)}) = x^{(k+1)} - x^{(k)}$$

from $x^{(k)}$ to $x^{(k+1)}$ is a descent direction

Descent property for Newton's method

Theorem 9.2 in textbook

Theorem – If $F(x^{(k)}) > 0$ and $\nabla f(x^{(k)}) \neq 0$, then the search direction

$$d^{(k)} = -F(x^{(k)})^{-1}\nabla f(x^{(k)}) = x^{(k+1)} - x^{(k)}$$

from $x^{(k)}$ to $x^{(k+1)}$ is a descent direction in the sense that there exists an $\overline{\alpha}$ such that for $\alpha \in (0, \overline{\alpha}]$,

$$f(x^{(k)} + \alpha d^{(k)}) < f(x^{(k)}).$$

Descent property for Newton's method

Theorem 9.2 in textbook

Theorem – If $F(x^{(k)}) > 0$ and $\nabla f(x^{(k)}) \neq 0$, then the search direction

$$d^{(k)} = -F(x^{(k)})^{-1} \nabla f(x^{(k)}) = x^{(k+1)} - x^{(k)}$$

from $x^{(k)}$ to $x^{(k+1)}$ is a descent direction in the sense that there exists an $\overline{\alpha}$ such that for $\alpha \in (0, \overline{\alpha}]$,

$$f(x^{(k)} + \alpha d^{(k)}) < f(x^{(k)}).$$

Proof sketch – Let

$$\phi(\alpha) = f(x^{(k)} + \alpha d^{(k)}).$$

Descent property for Newton's method

Theorem 9.2 in textbook

Theorem – If $F(x^{(k)}) > 0$ and $\nabla f(x^{(k)}) \neq 0$, then the search direction

$$d^{(k)} = -F(x^{(k)})^{-1}\nabla f(x^{(k)}) = x^{(k+1)} - x^{(k)}$$

from $x^{(k)}$ to $x^{(k+1)}$ is a descent direction in the sense that there exists an $\overline{\alpha}$ such that for $\alpha \in (0, \overline{\alpha}]$,

$$f(x^{(k)} + \alpha d^{(k)}) < f(x^{(k)}).$$

Proof sketch – Let

$$\phi(\alpha) = f(x^{(k)} + \alpha d^{(k)}).$$

Then, by chain rule,

$$\phi'(0) = \nabla f(x^{(k)})^\top d^{(k)}$$

Descent property for Newton's method

Theorem 9.2 in textbook

Theorem – If $F(x^{(k)}) > 0$ and $\nabla f(x^{(k)}) \neq 0$, then the search direction

$$d^{(k)} = -F(x^{(k)})^{-1}\nabla f(x^{(k)}) = x^{(k+1)} - x^{(k)}$$

from $x^{(k)}$ to $x^{(k+1)}$ is a descent direction in the sense that there exists an $\overline{\alpha}$ such that for $\alpha \in (0, \overline{\alpha}]$,

$$f(x^{(k)} + \alpha d^{(k)}) < f(x^{(k)}).$$

Proof sketch – Let

$$\phi(\alpha) = f(x^{(k)} + \alpha d^{(k)}).$$

Then, by chain rule,

$$\phi'(0) = \nabla f(x^{(k)})^\top d^{(k)} = -\nabla f(x^{(k)})^\top F(x^{(k)})^{-1}\nabla f(x^{(k)})$$

Descent property for Newton's method

Theorem 9.2 in textbook

Theorem – If $\boxed{F(x^{(k)}) > 0}$ and $\nabla f(x^{(k)}) \neq 0$, then the search direction

$$d^{(k)} = -F(x^{(k)})^{-1}\nabla f(x^{(k)}) = x^{(k+1)} - x^{(k)}$$

from $x^{(k)}$ to $x^{(k+1)}$ is a descent direction in the sense that there exists an $\overline{\alpha}$ such that for $\alpha \in (0, \overline{\alpha}]$,

$$f(x^{(k)} + \alpha d^{(k)}) < f(x^{(k)}).$$

Proof sketch – Let

$$\phi(\alpha) = f(x^{(k)} + \alpha d^{(k)}).$$

Then, by chain rule,

$$\phi'(0) = \nabla f(x^{(k)})^\top d^{(k)} = -\nabla f(x^{(k)})^\top F(x^{(k)})^{-1}\nabla f(x^{(k)}) < 0.$$

Descent property for Newton's method

Theorem 9.2 in textbook

Theorem – If $F(x^{(k)}) > 0$ and $\nabla f(x^{(k)}) \neq 0$, then the search direction

$$d^{(k)} = -F(x^{(k)})^{-1} \nabla f(x^{(k)}) = x^{(k+1)} - x^{(k)}$$

from $x^{(k)}$ to $x^{(k+1)}$ is a descent direction in the sense that there exists an $\overline{\alpha}$ such that for $\alpha \in (0, \overline{\alpha}]$,

$$f(x^{(k)} + \alpha d^{(k)}) < f(x^{(k)}).$$

Proof sketch – Let

$$\phi(\alpha) = f(x^{(k)} + \alpha d^{(k)}).$$

Then, by chain rule,

$$\phi'(0) = \nabla f(x^{(k)})^\top d^{(k)} = -\nabla f(x^{(k)})^\top F(x^{(k)})^{-1} \nabla f(x^{(k)}) \; < 0.$$

(complete the proof)

$\square$

Descent property for Newton's method

Theorem 9.2 in textbook

Theorem – If $F(x^{(k)}) > 0$ and $\nabla f(x^{(k)}) \neq 0$, then the search direction

$$d^{(k)} = -F(x^{(k)})^{-1}\nabla f(x^{(k)}) = x^{(k+1)} - x^{(k)}$$

from $x^{(k)}$ to $x^{(k+1)}$ is a descent direction in the sense that there exists an $\overline{\alpha}$ such that for $\alpha \in (0, \overline{\alpha}]$,

$$f(x^{(k)} + \alpha d^{(k)}) < f(x^{(k)}).$$

The theorem motivates the following Newton's descent algorithm:

$$\boxed{x^{(k+1)} = x^{(k)} - \alpha_k F(x^{(k)})^{-1}\nabla f(x^{(k)})}$$

where

$$\alpha_k = \arg\min_{\alpha \geq 0} f(x^{(k)} - \alpha F(x^{(k)})^{-1}\nabla f(x^{(k)})).$$

What if the Hessian is not positive definite?

Levenberg-Marquadt modification

- Suppose F is a symmetric matrix but not positive definite.

What if the Hessian is not positive definite?

Levenberg-Marquadt modification

- Suppose F is a symmetric matrix but not positive definite.

- Its eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ are real but not all positive.

What if the Hessian is not positive definite?

Levenberg-Marquadt modification

- Suppose F is a symmetric matrix but not positive definite.

- Its eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ are real but not all positive.

- Let $v_1, v_2, \ldots, v_n$ be the eigenvectors of F.

What if the Hessian is not positive definite?

Levenberg-Marquadt modification

- Suppose F is a symmetric matrix but not positive definite.

- Its eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ are real but not all positive.

- Let $v_1, v_2, \ldots, v_n$ be the eigenvectors of F.

- What about the eigenvalues of the matrix $F + \mu I$?

What if the Hessian is not positive definite?

Levenberg-Marquadt modification

- Suppose F is a symmetric matrix but not positive definite.
- Its eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ are real but not all positive.
- Let $v_1, v_2, \ldots, v_n$ be the eigenvectors of F.
- What about the eigenvalues of the matrix $F + \mu I$?

(explain on board)

What if the Hessian is not positive definite?

Levenberg-Marquadt modification

- Suppose F is a symmetric matrix but not positive definite.
- Its eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ are real but not all positive.
- Let $v_1, v_2, \ldots, v_n$ be the eigenvectors of F.
- What about the eigenvalues of the matrix $F + \mu I$?

(explain on board)

To ensure that the search direction is a descent direction, let:

$$x^{(k+1)} = x^{(k)} - \left\{ F(x^{(k)}) + \mu_k I \right\}^{-1} \nabla f(x^{(k)})$$

What if the Hessian is not positive definite?

Levenberg-Marquadt modification

- Suppose F is a symmetric matrix but not positive definite.
- Its eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$ are real but not all positive.
- Let $v_1, v_2, \dots, v_n$ be the eigenvectors of F.
- What about the eigenvalues of the matrix $F + \mu I$?

(explain on board)

To ensure that the search direction is a descent direction, let:

$$x^{(k+1)} = x^{(k)} - \left\{ F(x^{(k)}) + \mu_k I \right\}^{-1} \nabla f(x^{(k)})$$

To ensure descent property, use step size as in steepest/Newton's descent.

How to avoid computing the Hessian inverse?

Quasi-Newton methods for quadratic problems

Suppose we use the update

$$x^{(k+1)} = x^{(k)} - \alpha H_k \nabla f(x^{(k)}).$$

How to avoid computing the Hessian inverse?

Quasi-Newton methods for quadratic problems

Suppose we use the update

$$x^{(k+1)} = x^{(k)} - \alpha H_k \nabla f(x^{(k)}).$$

Then,

$$f(x^{(k+1)}) = f(x^{(k)}) + \nabla f(x^{(k)})^\top (x^{(k+1)} - x^{(k)}) + o(\|x^{(k+1)} - x^{(k)}\|)$$

How to avoid computing the Hessian inverse?

Quasi-Newton methods for quadratic problems

Suppose we use the update

$$x^{(k+1)} = x^{(k)} - \alpha H_k \nabla f(x^{(k)}).$$

Then,

$$\begin{aligned} f(x^{(k+1)}) &= f(x^{(k)}) + \nabla f(x^{(k)})^\top (x^{(k+1)} - x^{(k)}) + o(\|x^{(k+1)} - x^{(k)}\|) \\ &= f(x^{(k)}) - \alpha \nabla f(x^{(k)})^\top H_k \nabla f(x^{(k)}) + o(\|H_k \nabla f(x^{(k)})\| \alpha). \end{aligned}$$

How to avoid computing the Hessian inverse?

Quasi-Newton methods for quadratic problems

Suppose we use the update

$$x^{(k+1)} = x^{(k)} - \alpha H_k \nabla f(x^{(k)}).$$

Then,

$$\begin{aligned} f(x^{(k+1)}) &= f(x^{(k)}) + \nabla f(x^{(k)})^\top (x^{(k+1)} - x^{(k)}) + o(\|x^{(k+1)} - x^{(k)}\|) \\ &= f(x^{(k)}) - \alpha \nabla f(x^{(k)})^\top H_k \nabla f(x^{(k)}) + o(\|H_k \nabla f(x^{(k)})\| \alpha). \end{aligned}$$

To ensure *descent* for small α, we need

$$\nabla f(x^{(k)})^\top H_k \nabla f(x^{(k)}) > 0.$$

How to avoid computing the Hessian inverse?

Quasi-Newton methods for quadratic problems

Suppose we use the update

$$x^{(k+1)} = x^{(k)} - \alpha H_k \nabla f(x^{(k)}).$$

Then,

$$\begin{aligned} f(x^{(k+1)}) &= f(x^{(k)}) + \nabla f(x^{(k)})^\top (x^{(k+1)} - x^{(k)}) + o(\|x^{(k+1)} - x^{(k)}\|) \\ &= f(x^{(k)}) - \alpha \nabla f(x^{(k)})^\top H_k \nabla f(x^{(k)}) + o(\|H_k \nabla f(x^{(k)})\| \alpha). \end{aligned}$$

To ensure *descent* for small α, we need

$$\nabla f(x^{(k)})^\top H_k \nabla f(x^{(k)}) > 0.$$

$\rightarrow$ Ask for H_k to be positive definite.

Quasi-Newton algorithms

General structure

1. Set $k = 0$; select $x^{(0)}$ and a real symmetric p.d. H_0

Quasi-Newton algorithms

General structure

1. Set $k = 0$; select $x^{(0)}$ and a real symmetric p.d. H_0
2. If $\nabla f(x^{(k)}) = 0$, STOP. Else, $d^{(k)} = -H_k \nabla f(x^{(k)})$.

Quasi-Newton algorithms

General structure

1. Set $k = 0$; select $x^{(0)}$ and a real symmetric p.d. H_0
2. If $\nabla f(x^{(k)}) = 0$, STOP. Else, $d^{(k)} = -H_k \nabla f(x^{(k)})$.
3. Compute

$$\alpha_k = \arg\min_{\alpha \geq 0} f(x^{(k)} + \alpha d^{(k)})$$
$$x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)}.$$

Quasi-Newton algorithms

General structure

1. Set $k = 0$; select $x^{(0)}$ and a real symmetric p.d. H_0
2. If $\nabla f(x^{(k)}) = 0$, STOP. Else, $d^{(k)} = -H_k \nabla f(x^{(k)})$.
3. Compute

$$\begin{aligned} \alpha_k &= \arg\min_{\alpha \geq 0} f(x^{(k)} + \alpha d^{(k)}) \\ x^{(k+1)} &= x^{(k)} + \alpha_k d^{(k)}. \end{aligned}$$

4. Compute

$$\begin{aligned} \Delta x^{(k)} &= x^{(k+1)} - x^{(k)} \\ \Delta g^{(k)} &= \nabla f(x^{(k+1)}) - \nabla f(x^{(k)}) \\ H_{k+1} &= H_k + \left\{ \text{function of } \Delta x^{(k)}, \Delta g^{(k)}, H_k \right\} \end{aligned}$$

Quasi-Newton algorithms

General structure

1. Set $k = 0$; select $x^{(0)}$ and a real symmetric p.d. H_0
2. If $\nabla f(x^{(k)}) = 0$, STOP. Else, $d^{(k)} = -H_k \nabla f(x^{(k)})$.
3. Compute

$$\alpha_k = \arg\min_{\alpha \geq 0} f(x^{(k)} + \alpha d^{(k)})$$
$$x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)}.$$

4. Compute

$$\Delta x^{(k)} = x^{(k+1)} - x^{(k)}$$
$$\Delta g^{(k)} = \nabla f(x^{(k+1)}) - \nabla f(x^{(k)})$$
$$H_{k+1} = H_k + \left\{ \text{function of } \Delta x^{(k)}, \Delta g^{(k)}, H_k \right\}$$

- Rank one algorithm
- DFP algorithm
- BFGS algorithm

Conjugate Direction methods

Intermediate between Steepest descent and Newton's method

– Solve quadratics of n variables in n steps

$$f(x) = \frac{1}{2}x^\top Qx - x^\top b \; ; \; x \in \mathbb{R}^n, \; Q = Q^\top > 0.$$

Conjugate Direction methods

Intermediate between Steepest descent and Newton's method

– Solve quadratics of n variables in n steps

$$f(x) = \frac{1}{2}x^\top Qx - x^\top b \ ; \ x \in \mathbb{R}^n, \ Q = Q^\top > 0.$$

– No matrix inversion to arrive at

$$x^* = Q^{-1}b.$$

Q-conjugacy

Definition – Let $Q \in \mathbb{R}^{n \times n}$ be real, symmetric. The directions

$$d^{(0)}, d^{(1)}, \ldots, d^{(k)} \in \mathbb{R}^n$$

are Q-conjugate if for all $i \neq j$, we have

$$d^{(i)\top} Q d^{(j)} = 0.$$

Q-conjugacy

Definition – Let $Q \in \mathbb{R}^{n \times n}$ be real, symmetric. The directions

$$d^{(0)}, d^{(1)}, \ldots, d^{(k)} \in \mathbb{R}^n$$

are Q-conjugate if for all $i \neq j$, we have

$$d^{(i)\top} Q d^{(j)} = 0.$$

- $d^{(0)}, d^{(1)}, \ldots, d^{(k)}$ nonzero, Q-conjugate $\Longrightarrow$ linearly independent

(proof left as exercise)

Minimizing a quadratic function

$\min_{x \in \mathbb{R}^n} \frac{1}{2} x^\top Q x - x^\top b \; ; \; Q = Q^\top > 0$

Let

$$\boxed{g^{(k)} = \nabla f(x^{(k)}) = Qx^{(k)} - b}$$

Minimizing a quadratic function

$\min_{x \in \mathbb{R}^n} \frac{1}{2} x^\top Q x - x^\top b \; ; \; Q = Q^\top > 0$

Let

$$\boxed{g^{(k)} = \nabla f(x^{(k)}) = Q x^{(k)} - b}$$

Steepest descent:

$$x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)}$$

Minimizing a quadratic function

$\min_{x\in\mathbb{R}^n} \frac{1}{2}x^\top Qx - x^\top b \; ; \; Q = Q^\top > 0$

Let

$$\boxed{g^{(k)} = \nabla f(x^{(k)}) = Qx^{(k)} - b}$$

Steepest descent:

$$x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)}$$

where

$$\alpha_k = \arg\min_{\alpha\geq 0} f(x^{(k)} - \alpha g^{(k)})$$

Minimizing a quadratic function

$\min_{x \in \mathbb{R}^n} \frac{1}{2} x^\top Q x - x^\top b \; ; \; Q = Q^\top > 0$

Let

$$\boxed{g^{(k)} = \nabla f(x^{(k)}) = Qx^{(k)} - b}$$

Steepest descent:

$$x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)}$$

where

$$\alpha_k = \arg\min_{\alpha \geq 0} f(x^{(k)} - \alpha g^{(k)})$$

Applying FONC to $\phi(\alpha) = f(x^{(k)} - \alpha g^{(k)})$ yields

Minimizing a quadratic function

$\min_{x\in\mathbb{R}^n} \frac{1}{2}x^\top Qx - x^\top b \; ; \; Q = Q^\top > 0$

Let

$$\boxed{g^{(k)} = \nabla f(x^{(k)}) = Qx^{(k)} - b}$$

Steepest descent:

$$x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)}$$

where

$$\alpha_k = \arg\min_{\alpha\geq 0} f(x^{(k)} - \alpha g^{(k)})$$

Applying FONC to $\phi(\alpha) = f(x^{(k)} - \alpha g^{(k)})$ yields

$$0 = \phi'(\alpha) = (x^{(k)} - \alpha g^{(k)})^\top Q(-g^{(k)}) - b^\top(-g^{(k)})$$

Minimizing a quadratic function

$\min_{x\in\mathbb{R}^n} \frac{1}{2}x^\top Qx - x^\top b\ ;\ Q = Q^\top > 0$

Let

$$\boxed{g^{(k)} = \nabla f(x^{(k)}) = Qx^{(k)} - b}$$

Steepest descent:

$$x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)}$$

where

$$\alpha_k = \arg\min_{\alpha\geq 0} f(x^{(k)} - \alpha g^{(k)}) = \frac{g^{(k)\top} g^{(k)}}{g^{(k)\top} Q g^{(k)}}$$

Minimizing a quadratic function

$\min_{x\in\mathbb{R}^n} \frac{1}{2}x^\top Qx - x^\top b\ ;\ Q = Q^\top > 0$

Let

$$\boxed{g^{(k)} = \nabla f(x^{(k)}) = Qx^{(k)} - b}$$

Steepest descent:

$$x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)}$$

where

$$\alpha_k = \arg\min_{\alpha\geq 0} f(x^{(k)} - \alpha g^{(k)}) = \frac{g^{(k)\top} g^{(k)}}{g^{(k)\top} Q g^{(k)}}$$

Conjugate direction algorithm: Given $x^{(0)}$ and Q-conjugate directions $d^{(0)}, \ldots, d^{(n-1)}$; for $k \geq 0$,

$$x^{(k+1)} = x^{(k)} - \alpha_k d^{(k)},$$

$$\alpha_k = \frac{g^{(k)\top} d^{(k)}}{d^{(k)\top} Q d^{(k)}}.$$

Basic conjugate direction algorithm

$\min_{x\in\mathbb{R}^n} \frac{1}{2}x^\top Qx - x^\top b \; ; \; Q = Q^\top > 0$

Given $x^{(0)}$ and Q-conjugate directions $d^{(0)}, \ldots, d^{(n-1)}$; for $k \geq 0$,

$$\begin{aligned} g^{(k)} &= \nabla f(x^{(k)}) = Qx^{(k)} - b \\ x^{(k+1)} &= x^{(k)} - \alpha_k d^{(k)}, \\ \alpha_k &= \frac{g^{(k)\top} d^{(k)}}{d^{(k)\top} Q d^{(k)}}. \end{aligned}$$

Basic conjugate direction algorithm

$\min_{x\in\mathbb{R}^n} \frac{1}{2}x^\top Qx - x^\top b \; ; \; Q = Q^\top > 0$

Given $x^{(0)}$ and Q-conjugate directions $d^{(0)}, \ldots, d^{(n-1)}$; for $k \geq 0$,

$$
\begin{aligned}
g^{(k)} &= \nabla f(x^{(k)}) = Qx^{(k)} - b \\
x^{(k+1)} &= x^{(k)} - \alpha_k d^{(k)}, \\
\alpha_k &= \frac{g^{(k)\top} d^{(k)}}{d^{(k)\top} Q d^{(k)}}.
\end{aligned}
$$

Theorem – For any starting point $x^{(0)}$, the basic conjugate direction algorithm converges to the unique x^* in n steps.

Basic conjugate direction algorithm

$\min_{x\in\mathbb{R}^n} \frac{1}{2}x^\top Qx - x^\top b \ ; \ Q = Q^\top > 0$

Given $x^{(0)}$ and Q-conjugate directions $d^{(0)}, \ldots, d^{(n-1)}$; for $k \geq 0$,

$$\begin{aligned} g^{(k)} &= \nabla f(x^{(k)}) = Qx^{(k)} - b \\ x^{(k+1)} &= x^{(k)} - \alpha_k d^{(k)}, \\ \alpha_k &= \frac{g^{(k)\top} d^{(k)}}{d^{(k)\top} Q d^{(k)}}. \end{aligned}$$

Theorem – For any starting point $x^{(0)}$, the basic conjugate direction algorithm converges to the unique x^* in n steps.

Note: x^* satisfies $Qx^* = b$.

(proof on board)

Basic conjugate direction algorithm

$\min_{x \in \mathbb{R}^n} \frac{1}{2} x^\top Q x - x^\top b$; $Q = Q^\top > 0$

Given $x^{(0)}$ and Q-conjugate directions $d^{(0)}, \ldots, d^{(n-1)}$; for $k \geq 0$,

$$\begin{aligned} g^{(k)} &= \nabla f(x^{(k)}) = Qx^{(k)} - b \\ x^{(k+1)} &= x^{(k)} - \alpha_k d^{(k)}, \\ \alpha_k &= \frac{g^{(k)\top} d^{(k)}}{d^{(k)\top} Q d^{(k)}}. \end{aligned}$$

Theorem – For any starting point $x^{(0)}$, the basic conjugate direction algorithm converges to the unique x^* in n steps.

Note: x^* satisfies $Qx^* = b$.

(proof on board)

$\sim$ see Theorem 10.1 in textbook

How to choose the directions?

"compute as you go"

- Pick $x^{(0)}$ and let $d^{(0)} = -g^{(0)} = -\nabla f(x^{(0)})$.

How to choose the directions?

"compute as you go"

· Pick $x^{(0)}$ and let $d^{(0)} = -g^{(0)} = -\nabla f(x^{(0)})$.

· At each stage, the direction is calculated as a linear combination of the current gradient and the previous directions

How to choose the directions?

"compute as you go"

· Pick $x^{(0)}$ and let $d^{(0)} = -g^{(0)} = -\nabla f(x^{(0)})$.

· At each stage, the direction is calculated as a linear combination of the current gradient and the previous directions

$$d^{(k+1)} = -g^{(k+1)} + \beta_k d^{(k)}$$

How to choose the directions?

"compute as you go"

· Pick $x^{(0)}$ and let $d^{(0)} = -g^{(0)} = -\nabla f(x^{(0)})$.

· At each stage, the direction is calculated as a linear combination of the current gradient and the previous directions

$$d^{(k+1)} = -g^{(k+1)} + \beta_k d^{(k)}$$

· Need to ensure that $d^{(i)}$ are Q-conjugate.

How to choose the directions?

"compute as you go"

· Pick $x^{(0)}$ and let $d^{(0)} = -g^{(0)} = -\nabla f(x^{(0)})$.

· At each stage, the direction is calculated as a linear combination of the current gradient and the previous directions

$$d^{(k+1)} = -g^{(k+1)} + \beta_k d^{(k)}$$

· Need to ensure that $d^{(i)}$ are Q-conjugate. Choose β_k such that

$$\beta_k = \frac{g^{(k+1)\top} Q d^{(k)}}{d^{(k)\top} Q d^{(k)}}.$$

(see Proposition 10.1 and its proof)