Lenny Wu

# 10.4 Steepest Descent Methods

Recall: Newton → Quadratic convergence
Quasi-Newton → Superlinear convergence

However, those depend on a good initial guess
(If initial guess is bad, method might not converge!)

Today: Consider a method which converges independent of the initial guess.

Given $F: \mathbb{R}^n \to \mathbb{R}^n$, the solution to the system can be treated as the minimizer of the optimization problem:

$$F(\vec{x}^*) = 0 \iff \vec{x}^* = \underset{\vec{x} \in \mathbb{R}^n}{\operatorname{argmin}}\, g(\vec{x})$$

where $g(x) = \|F(\vec{x})^2\| = \sum_{i=1}^{n} f_i^2(\vec{x})$

**Def.** For $g : \mathbb{R}^n \to \mathbb{R}$, the gradient of $g$ at $\vec{x} \in \mathbb{R}^n$, denoted by $\nabla g(\vec{x})$, is defined by

$$\nabla g(\vec{x}) = \begin{bmatrix} \frac{\partial g}{\partial x_1}(\vec{x}) \\ \vdots \\ \frac{\partial g}{\partial x_n}(\vec{x}) \end{bmatrix} \in \mathbb{R}^n$$

The directional derivative of $g$ at $\vec{x}$ in the direction $\vec{v} \in \mathbb{R}^n$ $(\|\vec{v}\|_2 = 1)$ is defined by

$$D_v g(\vec{x}) = \lim_{h \to 0} \frac{g(\vec{x} + h\vec{v}) - g(\vec{x})}{h} = \vec{v}^T \cdot \nabla g(\vec{x})$$

**Rmk.** $D_v g(\vec{x})$ describes the rate of change of $g$ in the direction of $\vec{v}$.

**Ex.** If $g : \mathbb{R}^n \to \mathbb{R}$ is differentiable, show that $-\nabla g(\vec{x})$ is the direction of steepest descent in the value $g$ at $\vec{x}$.

**Pf.** Let $\vec{v}$ be a unit vector. Want to find direction $\vec{v}$ that minimizes $\nabla g(\vec{x})^T \vec{v}$.

$$\begin{aligned} \nabla g(\vec{x})^T \vec{v} &= \|\nabla g(\vec{x})\| \, \|\vec{v}\| \cos(\theta) \\ &= \|\nabla g(\vec{x})\| \cos(\theta) \end{aligned}$$

$\Rightarrow$ max occurs at $\theta = 0$

min occurs at $\theta = \pi$

$\Rightarrow$ $\nabla g(\vec{x})$ points in direction of steepest descent

$-\nabla g(\vec{x})$ points in direction of steepest descent

Ex. Let $F(\vec{x}) = A\vec{x} + \vec{b}$. Then finding the root of $F$ can be transformed into finding the minimum of $g(\vec{x}) = \|F(\vec{x})\|^2 = F(\vec{x})^T F(\vec{x})$, and $\nabla g(\vec{x}) = 2J_F(\vec{x}) \cdot F(\vec{x})$, where $J_F(\vec{x})$ is Jacobian of $F$.

More specifically,

$$g(\vec{x}) = \|A\vec{x} - \vec{b}\|_2^2 = (\vec{x}^T A^T - \vec{b}^T)(A\vec{x} - \vec{b})$$

$$= \vec{x}^T A^T A \vec{x} - 2\vec{x}^T A^T \vec{b} + \vec{b}^T \vec{b}$$

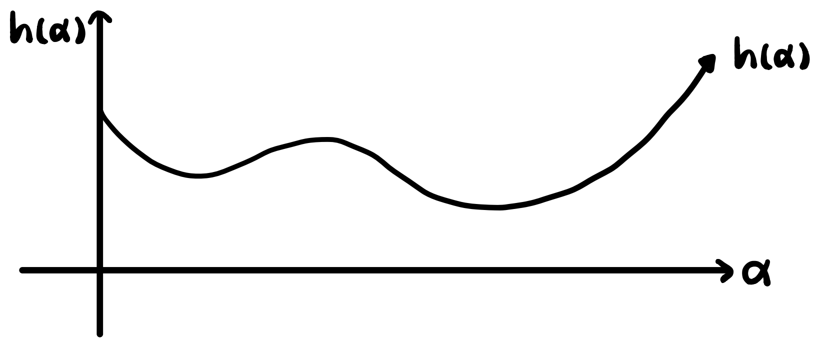
$$\nabla g(\vec{x}) = 2A^T A \vec{x} - 2A^T \vec{b} = 2A^T (A\vec{x} - \vec{b})$$

In the SD algorithm, update $\vec{x}$ as follows:

$$\vec{x}^{(k+1)} = \vec{x}^{(k)} - \alpha \nabla g(\vec{x}^{(k)})$$

where $\alpha > 0$ is a step size determined by minimizing the single-variable function

$$h(\alpha) = g(x^{(k)} - \alpha \nabla g(\vec{x}^{(k)}))$$




## Algorithm for SD

Give $\vec{x}^{(0)}$

For $k = 0, 1, \dots$ :

1. Find steepest descent direction of $g$ at $\vec{x}^{(k)}$ $(-\nabla g(\vec{x}^{(k)})$
2. Determine $\hat{\alpha} > 0$ which is / approximates the solution to $\hat{\alpha} = \underset{\alpha}{\operatorname{argmin}}\, g(\vec{x}^{(k)} - \alpha \nabla g(\vec{x}^{(k)}))$
3. Update the solution $\vec{x}^{(k+1)} = \vec{x}^{(k)} - \hat{\alpha}\nabla g(\vec{x}^{(k)})$

Rmk. 1. No need to invert Jacobian
(at most $O(n^2)$ per iteration)

2. Minimizing $h(\alpha)$ at each iteration can be expensive!
Instead, choose 3 numbers $\alpha_1 < \alpha_2 < \alpha_3$ and minimize a quadratic polynomial $p(\alpha)$ that interpolates $h$ at $\alpha_1, \alpha_2, \alpha_3$.
To choose $\alpha_1, \alpha_2, \alpha_3$, use backtracing:
   - choose $\alpha_1 = 0$, $\alpha_3 = 1$
   - if $h(\alpha_3) > h(\alpha_1)$, set $\alpha_3 \leftarrow \frac{\alpha_2}{2}$ until $h(\alpha_3) < h(\alpha_1)$
   - finally, choose $\alpha_2 = \frac{\alpha_3}{2}$

# 9.1-9.2 Eigenvalues, Orthonormal Matrices, and Similarity Transformations

**Def.** Let $A \in \mathbb{R}^{n \times n}$, $\lambda$ is called an **eigenvalue** of A if there exists $\vec{x} \in \mathbb{R}^n$ $(\vec{x} \neq 0)$ such that $A\vec{x} = \lambda\vec{x}$. $\vec{x}$ is the **eigenvector** corresponding to $\lambda$.

Define the **characteristic polynomial** of A as:

$$P_A(\lambda) = \det(A - \lambda I)$$

The eigenvalues are the roots of $P_A(\lambda)$.

The kernel of $A - \lambda I$ is called the **eigenspace** associated with $\lambda$, denoted by $E\lambda$:

$$E\lambda = \ker(A - \lambda I) = \{\vec{v} \in \mathbb{R}^n : A\vec{v} = \lambda\vec{v}\}$$

**Eigenbasis**: basis of A containing all linearly independent eigenvectors of A.

The dimension of $E\lambda$ is the **geometric multiplicity** of $\lambda$. The multiplicity of the root of $P_A(\lambda)$ is the **algebraic multiplicity**.

**Rmk.** 1. For $A \in \mathbb{R}^{n \times n}$, $\text{geom}(\lambda) \leq \text{alg}(\lambda)$

2. $\text{geom}(\lambda)$ = number of linearly independent eigenvectors of $A$ associated with $\lambda$.

**Ex.** Let $D$ be upper-triangular matrix. Find all eigenvalues of $D$.

**Sol.** Since $P_D(\lambda) = \prod_{i=1}^{n} (d_{ii} - \lambda) \Rightarrow d_{ii} = \lambda$.

**Def.** A matrix $Q \in \mathbb{R}^{n \times n}$ is **orthogonal** if its column vectors form an ON basis in $\mathbb{R}^n$.

**Thm. 9.10** Suppose $Q$ is an orthogonal $n \times n$ matrix.

Then (i) $Q$ is invertible with $Q^{-1} = Q^T$.

(ii) For any $\vec{x}, \vec{y} \in \mathbb{R}^n$, $(Q\vec{x})^T (Q\vec{y}) = \vec{x}^T \vec{y}$
(angle preserving)

(iii) For any $\vec{x} \in \mathbb{R}^n$, $\|Q\vec{x}\|_2 = \|\vec{x}\|_2$
(norm preserving)

**Rmk.** The converse of (i) holds:

$$Q^{-1} = Q^T \Rightarrow Q \text{ orthogonal}$$

**Ex.** The rotation matrix (CCW)

$$R = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \text{ is ON.}$$

**Def.** Two matrices are similar if nonsingular matrix exists with $A = S^{-1}BS$.

The map $B \mapsto S^{-1}BS$ is called a similarity transformation of B.

**Thm.** If A and B are similar with $A = S^{-1}BS$ and $\lambda$ is an eigenvalue of A with associated eigenvector $\vec{x}$, then $\lambda$ is an eigenvalue of B with associated eigenvector $S\vec{x}$.

**Pf.** By assumption, we have:

$$A\vec{x} = \lambda\vec{x} \quad \Rightarrow \quad S^{-1}BS\vec{x} = \lambda\vec{x}$$

$$\Rightarrow \quad B(S\vec{x}) = \lambda(S\vec{x}) \quad \square$$

**Def.** If $A \in \mathbb{R}^{n\times n}$ is similar to a diagonal matrix $D \in \mathbb{R}^{n\times n}$, there exists a nonsingular matrix $U \in \mathbb{C}^{n\times n}$ s.t.

$$T = U^{-1}AU$$

is an upper-triangular matrix whose diagonal matrix $D$ and an ON matrix $Q$ s.t.

$$A = QDQ^T.$$

**Rmk.** $A \in \mathbb{C}^{n\times n}$ is **Hermitian** if $A = A^*$, where $A^*$ is the conjugate transpose of $A$.
(MATLAB: "A' ")

**Fact** 1. $A \in \mathbb{R}^{n\times n}$ symmetric $\Rightarrow$ all $\lambda \in \mathbb{R}$

2. $A$ symmetric is positive definite $\Leftrightarrow$ all $\lambda > 0$.

## 9.3 The Power Method

**Def.** Let $\lambda_1, \lambda_2, \ldots \lambda_n$ be the eigenvalues of $A \in \mathbb{R}^{n\times n}$.
$\lambda_1$ is the **dominant eigenvalue** of $A$ if

$$|\lambda_1| > |\lambda_i|, \quad i = 2, 3, \ldots n$$

**Rmk.** Not every matrix has dominant eigenvalue.

**Goal:** Approximate the dominant eigenvalue.

Assume that $A$ has eigenvalues $\lambda_i$ satisfying $|\lambda_1| > |\lambda_2| \geq \ldots |\lambda_n|$ with associated eigenbasis $\{\vec{v}_i\}_{i=1}^{n}$.

Since $\vec{v}_i$ are linearly independent, $\vec{x} \neq 0$ can be represented

$$\vec{x} = a_1\vec{v}_1 + a_2\vec{v}_2 + \ldots + a_n\vec{v}_n$$

$$\begin{aligned}
\Rightarrow A^k\vec{x} &= A^k(a_1\vec{v}_1 + a_2\vec{v}_2 + \ldots + a_n\vec{v}_n) \\
&= a_1\lambda_1^k\vec{v}_1 + a_2\lambda_2^k\vec{v}_2 + \ldots + a_n\lambda_n^k\vec{v}_n \\
&= \lambda_1^k\left[a_1\vec{v}_1 + a_2\left(\frac{\lambda_2}{\lambda_1}\right)^k\vec{v}_2 + \ldots + a_n\left(\frac{\lambda_n}{\lambda_1}\right)^k\vec{v}_n\right]
\end{aligned}$$

Therefore, $\lim_{k\to\infty} A^k\vec{x} = \lim_{k\to\infty} \lambda_1^k a_1\vec{v}_1$.

**Thm.** If $\vec{v}$ if an eigenvector of $A$, then its corresponding eigenvalue is given by

$$\lambda = \frac{\vec{v}^T A\vec{v}}{\vec{v}^T\vec{v}}$$

called the Rayleigh Quotient

**Pf.** $A\vec{v} = \lambda\vec{v} \Rightarrow \vec{v}^T A\vec{v} = \lambda\vec{v}^T\vec{v}$

$$\Rightarrow \lambda = \frac{\vec{v}^T A\vec{v}}{\vec{v}^T\vec{v}}$$