TODO Notes

1 Linear Systems

Goal: solve $Ax = b$ for $A \in \mathbb{R}^{n \times n}$ (possibly positive semi-definite), $b \in \mathbb{R}^n$.

A fixed-point formulation for solving this system is:

$$x = x - M^{-1}(Ax - b), \tag{1}$$

which leads to the iteration:

$$x_{k+1} = x_k - M^{-1}(Ax_k - b), \tag{2}$$

for any invertible matrix $M \in \mathbb{R}^{n \times n}$. This is a generalization of the gradient method.

Connection to Gradient Descent

The standard gradient descent for minimizing the quadratic function

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

takes the form:

$$x_{k+1} = x_k - \alpha(Ax_k - b),$$

where $\alpha > 0$ is the stepsize. This corresponds to the iteration (2) with $M = \frac{1}{\alpha}I$, i.e.,

$$x_{k+1} = x_k - M^{-1}(Ax_k - b).$$

Thus, iteration (2) can be seen as a preconditioned gradient method, where M acts as a preconditioner. This connects fixed-point iterations for linear systems with optimization algorithms.

Why does this iteration converge when $\rho(I - M^{-1}A) < 1$?

We can write the iteration as:

$$x_{k+1} = (I - M^{-1}A)x_k + M^{-1}b.$$

This is a linear fixed-point iteration of the form:

$$x_{k+1} = Tx_k + c,$$

where $T = I - M^{-1}A$ and $c = M^{-1}b$. Standard theory says that such an iteration converges to the unique fixed point if and only if the spectral radius $\rho(T) < 1$. Hence, the method converges when $\rho(I - M^{-1}A) < 1$.

Theorem

The iteration $x_{k+1} = Mx_k + c$ converges to a unique fixed point for any initial guess x_0 if and only if the spectral radius of M satisfies

$$\rho(M) := \max\{|\lambda| : \lambda \text{ is an eigenvalue of } M\} < 1.$$

Proof

If $\rho(M) < 1$, then the iteration converges

It is a known result from matrix analysis that

$$\rho(M) = \inf_{\|\cdot\|} \|M\|,$$

where the infimum is taken over all induced matrix norms.

Hence, if $\rho(M) < 1$, there exists a norm $\|\cdot\|$ such that

$$\|M\| < 1.$$

Then for any $x, y \in \mathbb{R}^n$, we have:

$$\|\varphi(x) - \varphi(y)\| = \|Mx + c - (My + c)\| = \|M(x - y)\| \leq \|M\| \cdot \|x - y\|.$$

Since $\|M\| < 1$, the mapping $\varphi(x) = Mx + c$ is a contraction:

$$\|\varphi(x) - \varphi(y)\| \leq q\|x - y\| \quad \text{with } q = \|M\| < 1.$$

By the **Banach Fixed Point Theorem**, any contraction mapping on a complete normed space has a unique fixed point and the iteration $x_{k+1} = \varphi(x_k)$ converges to it for any initial point x_0.

If $\rho(M) \geq 1$, the iteration may not converge

Suppose λ is an eigenvalue of M with $|\lambda| \geq 1$, and let v be a corresponding eigenvector (possibly complex): $Mv = \lambda v$.

Let $x_0 = v$ and $c = 0$. Then the iteration becomes:

$$x_1 = Mx_0 = \lambda v, \quad x_2 = Mx_1 = \lambda^2 v, \quad \dots, \quad x_k = \lambda^k v.$$

If $|\lambda| \geq 1$, then $\|x_k\| = |\lambda|^k \cdot \|v\| \to \infty$ as $k \to \infty$.

Thus, the iteration diverges, and convergence is not guaranteed.

Step size in the quadratic case

We consider the objective function

$$f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle,$$

where $A \in \mathbb{R}^{d\times d}$ is a symmetric positive definite matrix. The gradient of f is given by

$$\nabla f(x) = Ax - b.$$

In the adaptive accelerated gradient descent algorithm, the step size λ_k is updated using the formula:

$$\lambda_k = \min\left\{\sqrt{1 + \frac{\theta_{k-1}}{2}} \cdot \lambda_{k-1},\ \frac{\|x^k - x^{k-1}\|}{2\|\nabla f(x^k) - \nabla f(x^{k-1})\|}\right\}.$$

For the quadratic case, we substitute $\nabla f(x) = Ax - b$ to obtain:

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

This implies:

$$\|\nabla f(x^k) - \nabla f(x^{k-1})\| = \|A(x^k - x^{k-1})\|.$$

Hence, the second term in the minimum becomes:

$$\frac{\|x^k - x^{k-1}\|}{2\|A(x^k - x^{k-1})\|}.$$

To simplify, we use the sub-multiplicative property of norms:

$$\|A(x^k - x^{k-1})\| \leq \|A\| \cdot \|x^k - x^{k-1}\|.$$

Therefore:

$$\frac{\|x^k - x^{k-1}\|}{2\|A(x^k - x^{k-1})\|} \geq \frac{1}{2\|A\|}.$$

This gives the lower bound:

$$\lambda_k \geq \min\left\{\sqrt{1 + \frac{\theta_{k-1}}{2}} \cdot \lambda_{k-1},\ \frac{1}{2\|A\|}\right\}.$$

About the matrix norm $\|A\|$: In this context, $\|A\|$ denotes the operator norm (also known as spectral norm), which is the norm induced by the Euclidean norm on $\mathbb{R}^d$:

$$\|A\| = \sup_{\|x\|=1} \|Ax\|.$$

Since A is symmetric and positive definite, it is diagonalizable and its spectral norm equals its largest eigenvalue:

$$\|A\| = \sqrt{\lambda_{\max}(A^\top A)} = \lambda_{\max}(A),$$

because $A = A^\top$. This norm governs how much the matrix A can stretch a vector, which makes it a natural choice when bounding expressions like $\|A(x^k - x^{k-1})\|$.

Thus, the step size is adaptively bounded from below by a quantity that depends on the curvature of the function, and in the quadratic case, that curvature is entirely captured by the spectral norm $\|A\|$.