How to evaluate an optimization method

- Does it converge to minimum?
- How fast?
- Practical issues: Is it easy to implement or tune?

We will see that all the methods we discussed converge to a minimum, but some of them require the function f to have additional good properties.

Remark: We say $x \in \mathbb{R}^n$ is a limit point of a sequence $\{x_k\}$, if there exists a subsequence of $\{x_k\}$ that converges to x.

Definition: Let $\{z_k\}$ converges to $\bar{z}$. We say the convergence of *order* $p (\geqslant 0)$ and with *factor* $\gamma\ (> 0)$, if $\exists k_0$ such that $\forall k \geqslant k_0$ we have

$$\|z_{k+1} - \bar{z}\| \leqslant \gamma \|z_k - \bar{z}\|^p.$$

- The larger the power p the faster the convergence.
- For the same p, the smaller γ, the faster the convergence.
- If $\{z_k\}$ converges with order p and factor γ, it also converges with order $\bar{p}$ for any $\bar{p} \leqslant p$.

Terminologies

- If $p = 1$, and $\gamma < 1$, we say convergence is *linear*: $\lim_{k\to\infty} \frac{\|z_{k+1} - \bar{z}\|}{\|z_k - \bar{z}\|} = \gamma < 1$
- If $p = 1$, and $\gamma = 1$, we say convergence is *sublinear*.
- If $p > 1$, we say that the convergence is *superlinear*: $\lim_{k\to\infty} \frac{\|z_{k+1} - \bar{z}\|}{\|z_k - \bar{z}\|} = 0$
- If $p = 2$, we say that the convergence is *quadratic*: $\lim_{k\to\infty} \frac{\|z_{k+1} - \bar{z}\|}{\|z_k - \bar{z}\|^2} < \infty$

The local convergence analysis approach

Basic ingredients of our local rate of convergence analysis approach

- Focus on a sequence $\{x_k\}$ that converges to a unique limit points $x^\star$
- Rate of convergence is evaluated using *error function* $E(x)$:

$$E : \mathbb{R}^n \to \mathbb{R} \text{ such that } E(x) \geqslant 0 \quad \forall x \in \mathbb{R}^n, \quad E(x^\star) = 0.$$

- Typical choices are
  - Euclidean distance: $E(x) = \|x - x^\star\|$
  - Cost difference: $E(x) = |f(x) - f(x^\star)|$
- Our analysis is asymptotic, i.e., we look at the rate of convergence of the tail of the error sequence $\{E(x_k)\}$
- Convergence type
  - *linear* convergence : $\lim_{k\to\infty} \frac{E(x_{k+1})}{E(x_k)} = \gamma < 1$
  - *superlinear* convergence : $\lim_{k\to\infty} \frac{E_{x_{k+1}}}{E(x_k)} = 0$
  - *quadratic*: $\lim_{k\to\infty} \frac{E(x_{k+1})}{E(x_k)^2} < \infty$

Proposition: Consider $f(x) = \frac{1}{2}x^\top Qx - b^\top x$ with $Q > 0$. For the steepest descent algorithm with exact line search, $\alpha_k = \text{argmin}\, f(x_k - \alpha_k \nabla f(x_k))$, we have $x_k \to x^\star$, starting from any $x_0 \in \mathbb{R}^n$ (this is called global convergence).

Proof: let $\lambda_1 = \lambda_{\min}(Q)$ and $\lambda_n = \lambda_{\max}(Q)$.

- Note that from $\nabla f(x) = Q\,x - b$. Therefore $x^\star = Q^{-1}\,b$. Because $Q > 0$, $f(x)$ is a strictly convex function. Therefore $x^\star = Q^{-1}b$ is the unique minimizer of $f(x)$, i.e, $E(x) = f(x) - f(x^\star) > 0$.
- $\alpha_k = \text{argmin}\, f(x_k - \alpha_k \nabla f(x_k)) = \frac{\nabla f(x_k)^\top \nabla f(x_k)}{\nabla f(x_k)^\top Q \nabla f(x_k)}$.
- we can write $f(x) = \underbrace{\frac{1}{2}(x - x^\star)^\top Q(x - x^\star)}_{E(x)} \underbrace{-\frac{1}{2}x^\star Q x^\star}_{f(x^\star)}$
- $E(x) = \frac{1}{2}\|x - x^\star\|_Q^2 = f(x) - f(x^\star)$
- Using $x_{k+1} = x_k - \frac{\nabla f(x_k)^\top \nabla f(x_k)}{\nabla f(x_k)^\top Q \nabla f(x_k)} \nabla f(x_k)$, we obtain

$$E(x_{k+1}) = \left(1 - \frac{\nabla f(x_k)^\top \nabla f(x_k)}{(\nabla f(x_k)^\top Q \nabla f(x_k))(\nabla f(x_k)^\top Q^{-1} \nabla f(x_k))}\right) E(x_k)$$

Convergence of steepest descent algorithm for quadratic cost functions

- Using Kantoraovich inequality

$$E(x_{k+1}) \leqslant \Big(1 - \frac{4\lambda_1\lambda_n}{(\lambda_1+\lambda_n)^2}\Big)E(x_k) = \underbrace{\Big(\frac{\lambda_n-\lambda_1}{\lambda_n+\lambda_1}\Big)^2}_{\beta} E(x_k).$$

- note that $\beta < 1$.
- $E(x_{k+1}) \leqslant \beta E(x_k)$ or equivalently $(f(x_{k+1}) - f(x^\star)) \leqslant \beta(f(x_k) - f(x^\star))$: linear rate of convergence with factor β
- if β is small, the rate of convergence is good.
- Rate of convergence and condition number: $\kappa(Q) = \frac{\lambda_n}{\lambda_1}$
  - $\beta = \Big(\frac{\frac{\lambda_n}{\lambda_1}-1}{\frac{\lambda_n}{\lambda_1}+1}\Big)^2 = \Big(\frac{\kappa(Q)-1}{\kappa(Q)+1}\Big)^2$
  - the problems with large κ are referred to as ill-conditioned
  - Steepest descent algorithm converges slowly for ill-conditioned problems

Convergence of steepest descent algorithm for quadratic cost functions

$$\beta = \left(\frac{\frac{\lambda_n}{\lambda_1} - 1}{\frac{\lambda_n}{\lambda_1} + 1}\right)^2 = \left(\frac{\kappa(Q) - 1}{\kappa(Q) + 1}\right)^2$$

$$\frac{f(\boldsymbol{x}^{k+1}) - f(\boldsymbol{x}^*)}{f(\boldsymbol{x}^k) - f(\boldsymbol{x}^*)} \leq \left(\frac{\kappa(\boldsymbol{Q}) - 1}{\kappa(\boldsymbol{Q}) + 1}\right)^2$$

$\kappa(Q) = \frac{\lambda_{max}}{\lambda_{min}}$	Upper Bound on Convergence Constant δ	Number of Iterations to Reduce the Optimality Gap by 0.10
1.1	0.0023	1
3.0	0.25	2
10.0	0.67	6
100.0	0.96	58
200.0	0.98	116
400.0	0.99	231

Convergence rate of steepest descent algorithm for non-quadratic cost functions

Consider cost function $f \in \mathcal{C}^2$ with a local minimizer $x^\star$. Let

- $\nabla^2 f(x^\star) > 0$
- $\lambda_n = \lambda_{\max}(\nabla^2 f(x^\star))$
- $\lambda_1 = \lambda_{\min}(\nabla^2 f(x^\star))$.

If $\{x_k\}$ converges to $x^\star$ and its is generated by steepest descent algorithm with stepsizes obtained from exact line search, then $f(x) \to f(x^\star)$, linearly with convergence ratio no greater than $\beta = (\frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1})^2$.

Proposition: Stationarity of Limit Points for Gradient Methods

Let $\{x_k\}$ be a sequence generated by a gradient method $x_{k+1} = x_k + \alpha_k \, d_k$, and assume that $\{d_k\}$ us gradient related $\nabla f(x_k)^\top d_k < 0$ and α_k is chosen by minimization rule, or the limited minimization rule, the Armijo rule or Goldstein rule. Then every limit point of $\{x_k\}$ is a stationary point.

Theorem. (Newton's method). Let $f \in \mathcal{C}^3$ on $\mathbb{R}^n$, and assume that at the local minimum point $x^\star$, the Hessian $\nabla^2 f(x^\star)$ is positive definite. Then if started sufficiently close to $x^\star$, the points generated by Newton's method $(x_{k+1} = x_k - (\nabla^2 f(x^\star))^{-1} \nabla f(x_k))$ converge to $x^\star$. The order of convergence is at least two.

proof see page 247 Ref[2]