Unconstrained Optimization I

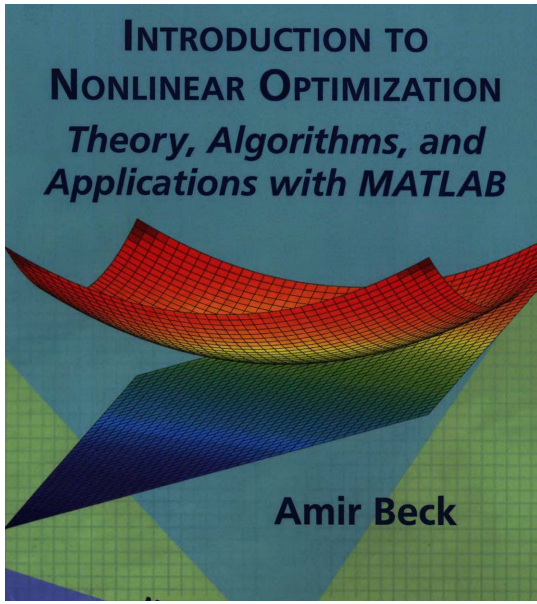
Note Title 3/25/2022

Let's recall some basic facts about unconstrained optimization:

Theorem 2.6 (first order optimality condition for local optima points). *Let $f : U \to \mathbb{R}$ be a function defined on a set $U \subseteq \mathbb{R}^n$. Suppose that $\mathbf{x}^* \in \text{int}(U)$ is a local optimum point and that all the partial derivatives of f exist at $\mathbf{x}^*$. Then $\nabla f(\mathbf{x}^*) = 0$.*

← called a stationary / critical pt of f.

Proof. Let $i \in \{1,2,\ldots,n\}$ and consider the one-dimensional function $g(t) = f(\mathbf{x}^* + t\mathbf{e}_i)$. Note that g is differentiable at $t = 0$ and that $g'(0) = \frac{\partial f}{\partial x_i}(\mathbf{x}^*)$. Since $\mathbf{x}^*$ is a local optimum point of f, it follows that $t = 0$ is a local optimum of g, which immediately implies that $g'(0) = 0$. The latter equality is exactly the same as $\frac{\partial f}{\partial x_i}(\mathbf{x}^*) = 0$. Since this is true for any $i \in \{1,2,\ldots,n\}$, the result $\nabla f(\mathbf{x}^*) = 0$ follows. □



Proposition 7.8 (sufficiency of stationarity under convexity). *Let f be a continuously differentiable function which is convex over a convex set $C \subseteq \mathbb{R}^n$. Suppose that $\nabla f(\mathbf{x}^*) = 0$ for some $\mathbf{x}^* \in C$. Then $\mathbf{x}^*$ is a global minimizer of f over C.*

Proof. Let $\mathbf{z} \in C$. Plugging $\mathbf{x} = \mathbf{x}^*$ and $\mathbf{y} = \mathbf{z}$ in the gradient inequality (7.6), we obtain that

$$f(\mathbf{z}) \geq f(\mathbf{x}^*) + \nabla f(\mathbf{x}^*)^T(\mathbf{z} - \mathbf{x}^*),$$

which by the fact that $\nabla f(\mathbf{x}^*) = 0$ implies that $f(\mathbf{z}) \geq f(\mathbf{x}^*)$, thus establishing that $\mathbf{x}^*$ is the global minimizer of f over C. □

Theorem 2.27 (sufficient second order optimality condition). *Let $f : U \to \mathbb{R}$ be a function defined on an open set $U \subseteq \mathbb{R}^n$. Suppose that f is twice continuously differentiable over U and that $\mathbf{x}^*$ is a stationary point. Then the following hold:*

(a) *If $\nabla^2 f(\mathbf{x}^*) \succ 0$, then $\mathbf{x}^*$ is a strict local minimum point of f over U.*

(b) *If $\nabla^2 f(\mathbf{x}^*) \prec 0$, then $\mathbf{x}^*$ is a strict local maximum point of f over U.*

Theorem 7.12 (second order characterization of convexity). *Let f be a twice continuously differentiable function over an open convex set $C \subseteq \mathbb{R}^n$. Then f is convex if and only if $\nabla^2 f(\mathbf{x}) \succeq 0$ for any $\mathbf{x} \in C$.*

Theorem 7.13 (sufficient second order condition for strict convexity). *Let f be a twice continuously differentiable function over a convex set $C \subseteq \mathbb{R}^n$, and suppose that $\nabla^2 f(\mathbf{x}) \succ 0$ for any $\mathbf{x} \in C$. Then f is strictly convex over C.*

From now on, assume that the objective function $f : \mathcal{U} \overset{\text{open}}{\subseteq} \mathbb{R}^n \to \mathbb{R}$ is at least C^1.

Our goal is to devolop algorithms, and the analysis of such algorithms, for solving

$$\min_{x \in \mathcal{U}} f(x).$$

Two strategies :

(i) line search : (I) choose a descent direction p_k at x_k (ie. a $p_k \in \mathbb{R}^n$ s.t. $\frac{d}{d\alpha} f(x_k + \alpha p_k) < 0$)

$$\frac{d}{d\alpha} f(x_k + \alpha p_k) = \nabla f(x_k)^T p_k$$

(II) then, choose a step size α_k that approximately solves

$$\min_{\alpha > 0} f(x_k + \alpha p_k)$$

(III) update $x_{k+1} = x_k + \alpha_k p_k$, and iterate.

(ii) trust region : (I) information gathered about f near x_k is used to construct a model function m_k, so that $f(x) \approx m_k(x)$ for $x \approx x_k$.

[Typically m_k is a quadratic approximation.]

(II) solve $\min_p m_k(x_k+p)$ st. $\|p\| \le \Delta$ approximately $\rightarrow p_k$ ("trust region radius")

Note: this "trust region subproblem" is a constraint optimization problem. (Something surprising here...)

(III) If $f(x_k + p_k)$ is significantly smaller than $f(x_k)$,
update $x_{k+1} = x_k + p_k$, and iterate.
else
reduces Δ, goto (II).

In effect, a trust region method chooses the descent direction and step size simultaneously.

E.g. $f(x) = \underbrace{10(x_2 - x_1^2)^2 + (1-x_1)^2}_{\text{a sum of squares, and non-convex}}$. At $x_k = (0,1)$, its 2nd order Taylor approximation is:

$$f(x) \approx \underbrace{11 + \underbrace{\begin{bmatrix}-2\\20\end{bmatrix}^T}_{=\nabla f(x_k)^T}\begin{bmatrix}x_1-0\\x_2-1\end{bmatrix} + \frac{1}{2}\begin{bmatrix}x_1-0\\x_2-1\end{bmatrix}^T\underbrace{\begin{bmatrix}-38 & 0\\0 & 20\end{bmatrix}}_{=\nabla^2 f(x_k)}\begin{bmatrix}x_1-0\\x_2-1\end{bmatrix}}_{m_k(x)}$$

Note: $\min f(x) = 0$, $\operatorname{argmin} f(x) = (1,1)$

$\min m_k(x) = -\infty$ (we definitely need to restrict x to a compact set in order to get a meaningful minimizer.)

Search direction for line search methods

How to choose $p \in \mathbb{R}^n$ s.t. $\frac{d}{d\alpha} f(x+\alpha p) = \nabla f(x)^T p < 0$?

The most obvious choice is $\boxed{p = -\nabla f(x)}$. If $\nabla f(x) = 0$, x is already a stationary point. Otherwise

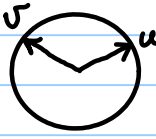
$$\nabla f(x)^T(-\nabla f(x)) = -\|\nabla f(x)\|^2 < 0.$$

moreover, it is obstensibly the best choice due to the Cauchy-Schwartz inequality:

$$|\langle \nabla f(x), p\rangle| \le \|\nabla f(x)\|_2 \|p\|_2 \quad \text{with equality holds} \Leftrightarrow \nabla f(x) \parallel p$$

so $\operatorname{argmin}_{\|p\|_2=1} \langle \nabla f(x), p\rangle = -\nabla f(x)/\|\nabla f(x)\|_2$

For this reason, $p = -\nabla f(x)$ is also called the "steepest descent direction".



If $\|u\|_2 = \|v\|_2 = 1$
$\langle u, v\rangle = -1$
$\Leftrightarrow u = -v$

But I want to convince you that this nomenclature can be misleading, and I prefer the name "gradient descent".

Here is one way to see it: For any non-singular matrix A,

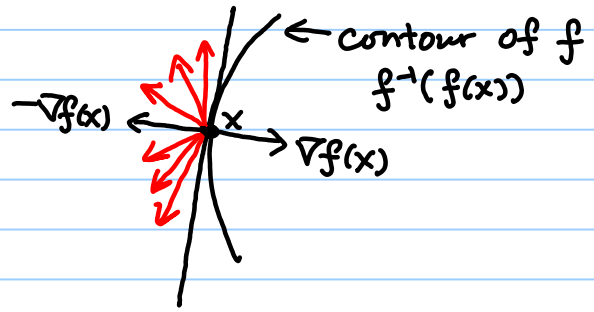
$$\frac{d}{d\alpha} f(x+\alpha p) = \nabla f(x)^T A A^{-1} p = \langle A^T \nabla f(x), A^{-1} p \rangle$$

so $\min_{\|A^{-1}p\|=1} \frac{d}{d\alpha} f(x+\alpha p) = -\| A^T \nabla f(x) \|_2$, the min is attained when and only when $A^{-1}p = -A^T \nabla f(x)$ and $\|A^{-1}p\| = 1$

$$\text{(✡)} \quad \operatorname*{argmin}_{p:\, \|A^{-1}p\|=1} \frac{d}{d\alpha} f(x+\alpha p) = \frac{-AA^T \nabla f(x)}{\| A^T \nabla f(x) \|_2}$$

Note: $\{ -AA^T \nabla f(x) : A \text{ nonsingular} \}$
$= \{ d : d^T \nabla f(x) < 0 \}$
when $\nabla f(x) \neq 0$.

Ex: Prove it.



By choosing different A, any descent direction is "steepest" in the sense of (✡).

Of course, you may find (✡) contrived when we change the round sphere $\{p : \|p\|_2 = 1\}$ into a ellipsoid $\{p : \|A^{-1}p\|_2 = 1\}$. But what it actually shows is that

the notion of "steepest descent direction" is not affine-invariant.

Imagine if we apply an affine change of coordinates to f and apply the "steepest descent direction" to $g(\bar{x}) := f(A^{-1}\bar{x})$ at $\bar{x} = Ax$. What it means is:

In the original coordinate system:

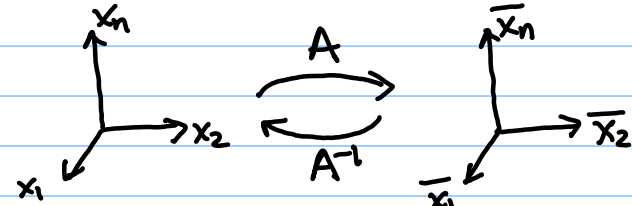
$$x_{k+1} = x_k - s\,\nabla f(x_k)$$

In the transformed coordinate system:

$$\bar{x}_{k+1} = \bar{x}_k - s\,\nabla g(\bar{x}_k)$$

$$\Downarrow$$

$$x_{k+1} = x_k - s\,A^{-1}A^{-T}\nabla f(x_k)$$



(assume constant size for now, other choices of step size do not really improve the situation by much as we shall see later.)

Chain rule:

$g = f \circ A^{-1}$

$dg(\bar{x}) = df(A^{-1}\bar{x}) \cdot A^{-1}$

$\nabla g(\bar{x}) = A^{-T}\nabla f(A^{-1}\bar{x})$

Steepest in one coordinate system is not the steepest system in another coordinate system!

Ex: Prove that the constant step size gradient descent method is "invariant under rigid transformation of coordinates." ($x \mapsto Ax + b$, $A \in \mathbb{R}^{n\times n}$, $AA^T = I$, $b \in \mathbb{R}^n$)

Here is yet another, arguably better, way to see why the "steepest descent" direction can be nowhere close to being steepest.

Consider $f(x) = \frac{1}{2}x^TAx + bx + c$, $A \succ 0$, which is pretty much (by Taylor's theorem) how any C^2 function looks like in the vicinity of a local minimizer x^* with $\nabla^2 f(x^*) \succ 0$.

In virtue of the rigid motion invariance of GD and the spectral theorem

$$A = U\begin{bmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_n\end{bmatrix}U^T$$

we may assume WLOG that $A = \operatorname{diag}(\lambda_1, \cdots, \lambda_n)$, $\lambda_i > 0$, $b = 0$, $c = 0$, i.e.

$$f(x) = \frac{1}{2}x^T\begin{bmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_n\end{bmatrix}x, \quad \lambda_i > 0.$$

In this case, the minimizer of f is at 0. So at x, the "true steepest descent" direction is $-x$, not $-\nabla f(x) = -Ax$. Moreover, we can prove:

Proposition 1.1 *The angle between $-\nabla f(x)$ and $-x$ can be arbitrarily close to $90°$. In other words, the negative gradient direction can be as far from the true steepest descent direction as possible.*

Proof: The angle θ between $-\nabla f(x)$ and $-x$ is maximized when $\cos(\theta)$ is minimized, so we seek the solution of:

$$\min_{x\in\mathbb{R}^n\setminus\{0\}} \frac{\langle -\nabla f(x), -x\rangle}{\|\nabla f(x)\|\|x\|} = \min_{x\in\mathbb{R}^n\setminus\{0\}} \frac{x^TAx}{\sqrt{x^TA^TAx}\sqrt{x^Tx}} = \sqrt{\min_{x\in\mathbb{R}^n\setminus\{0\}} \frac{(x^TAx)^2}{(x^TA^2x)(x^Tx)}}.$$

Apply a change of variable $y = \sqrt{A}x$. The last minimum inside the square-root can be written as

$$\min_{y\in\mathbb{R}^n\setminus\{0\}} \frac{(y^Ty)^2}{(y^TAy)(y^TA^{-1}y)}. \tag{1.8}$$

By the Kantorovich inequality, the minimum value above is

(see next page)

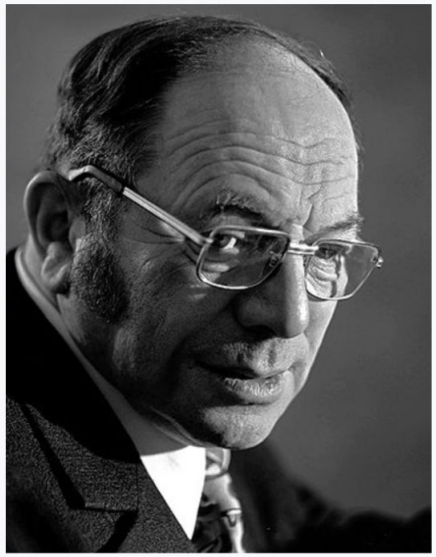
$$\frac{4\lambda_1\lambda_n}{(\lambda_1+\lambda_n)^2} = 1 - \frac{(\lambda_1-\lambda_n)^2}{(\lambda_1+\lambda_n)^2},$$

where λ_1 and λ_n are the largest and smallest eigenvalues of A. This shows that $\cos(\theta)$ gets arbitrarily close to 1, or equivalently θ gets arbitrarily close to $90°$, when A gets *increasingly ill-conditioned*, and x is chosen to be $\sqrt{A}y$ where y is a direction that realizes the minimum value in (1.8). ■

Kantorovich inequality: Let A be a positive definite $n \times n$ matrix. Then $\forall\, 0 \neq x \in \mathbb{R}^n$,

$$\frac{(x^T x)^2}{(x^T A x)(x^T A^{-1} x)} \geq \frac{4\,\lambda_{max}(A)\,\lambda_{min}(A)}{(\lambda_{max}(A) + \lambda_{min}(A))^2}. \qquad -(K)$$

Leonid Kantorovich
Леони́д Канторо́вич



Leonid Kantorovich in 1975

WLOG, can assume $A = \mathrm{diag}(\lambda_1, \cdots, \lambda_n)$, $\lambda_i > 0$.
($\lambda_1 = \lambda_{min}(A)$, $\lambda_n = \lambda_{max}(A)$)

Note that if (i) $x_i \neq 0 \Rightarrow \lambda_i = \lambda_1$ or λ_n and

(ii) $\sum_{i:\lambda_i=\lambda_1} x_i^2 = \sum_{i:\lambda_i=\lambda_n} x_n^2$ $(= A)$

then, $\dfrac{(x^T x)^2}{(x^T A x)(x^T A^{-1} x)} = \dfrac{(A+A)^2}{(\lambda_1 A + \lambda_n A)(\lambda_1^{-1} A + \lambda_n^{-1} A)}$

$= \dfrac{4}{(\lambda_1+\lambda_n)(\lambda_1^{-1}+\lambda_n^{-1})} = \dfrac{4\lambda_1\lambda_n}{(\lambda_1+\lambda_n)^2}$

The following proof of (K) shows that (i)+(ii) are not only sufficient for equality in (K), but also necessary. I learned this proof from:

P. Henrici. Two remarks on the Kantorovich inequality. *Amer. Math. Monthly*, 68:904–906, 1961.

Proof of the Kantorovich inequality:

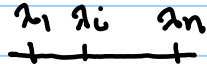
Besides the diagonalization simplication, we may also assume x on the LHS of (K) has unit length. And (K) is equivalent to:

$$(\xi_1\lambda_1+\cdots+\xi_n\lambda_n)(\xi_1\lambda_1^{-1}+\cdots+\xi_n\lambda_n^{-1}) \le (\lambda_1+\lambda_n)^2/4\lambda_1\lambda_n \quad \text{when } \sum\xi_i=1,\ \xi_i\ge 0$$

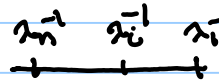
($\xi_i = x_i^2$)

If $\lambda_1=\lambda_n$, then equality holds trivially in (K). So, let's assume $\lambda_1<\lambda_n$.

A key trick is to observe that $\forall i$, we can always find $p_i, q_i \ge 0$ st.



$$\lambda_i = p_i\lambda_1 + q_i\lambda_n \quad \text{and} \quad \lambda_i^{-1} = p_i\lambda_1^{-1} + q_i\lambda_n^{-1}.$$



(Just solve the 2x2 linear system and see that the solution is non-negative.)

Then $1 = \lambda_i\lambda_i^{-1} = (p_i\lambda_1+q_i\lambda_n)(p_i\lambda_1^{-1}+q_i\lambda_n^{-1}) = (p_i+q_i)^2 + p_iq_i(\lambda_n-\lambda_1)^2/(\lambda_1\lambda_n)$, which also implies $p_i+q_i\le 1$.

Set $p := \sum\xi_i p_i$, $q = \sum\xi_i q_i$. So $p+q = \sum\xi_i(p_i+q_i) \le \sum\xi_i = 1$

$$p\lambda_1+q\lambda_n = \sum\xi_i\lambda_i$$
$$p\lambda_1^{-1}+q\lambda_n^{-1} = \sum\xi_i\lambda_i^{-1}.$$

$$\sum \xi_i \lambda_i \sum \xi_i \lambda_i^{-1} = (p\lambda_1 + q\lambda_n)(p\lambda_1^{-1} + q\lambda_n^{-1}) = (p+q)^2 + pq(\lambda_n - \lambda_1)^2/(\lambda_1\lambda_n),$$

$$\leq (p+q)^2\left[1 + (\lambda_n - \lambda_1)^2/4\lambda_1\lambda_n\right] = (p+q)^2\left[(\lambda_n+\lambda_1)^2/4\lambda_1\lambda_n\right] \leq (\lambda_n+\lambda_1)^2/4\lambda_1\lambda_n.$$

$(p+q)^2 \leq \left(\frac{p+q}{2}\right)^2$

Q.E.D.

Ex: see from the proof above that equality in (K) holds iff conditions (i) and (ii) (stated before the proof) hold.

Takeaway message: the "steepest descent" direction is far from steepest when $\nabla^2 f(x)$ is ill-conditioned.
(We shall establish rate-of-convergence results related to this.)

Newton direction is the direction that solves: $\min_{p \in \mathbb{R}^n} m_k(p)$, where

$m_k(p) = f(x_k) + \nabla f(x_k)^T p + \frac{1}{2} p^T [\nabla^2 f(x_k)] p$ is the 2nd order Taylor approximation to $f(x_k + p)$.

$m_k(p)$ has a unique minimizer when and only when $\nabla^2 f(x_k) > 0$ (so the Newton direction is well-defined if f is C^2 and strictly convex, or when x_k is close enough to a local minimizer x^* with $\nabla^2 f(x_*) > 0$.)

shorthand notation:
$f_k = f(x_k)$
$\nabla f_k = \nabla f(x_k)$
$\nabla^2 f_k = \nabla^2 f(x_k)$
etc.

In this case, $p_k^N := \operatorname{argmin}_p m_k(p) = -\nabla^2 f_k^{-1} \nabla f_k$

Facts about p_k^N:

(1) It is a descent direction: $\nabla f_k^T p_k^N = -\nabla f_k^T [\nabla^2 f_k]^{-1} \nabla f_k$

Again we need the assumption $\nabla^2 f_k \succ 0$, this time to guarantee that p_k^N is a descent direction.

(2) It is invariant under affine change of coordinates!

In the original coordinate system:

$$x_{k+1} = x_k - s \nabla^2 f_k^{-1} \nabla f_k$$

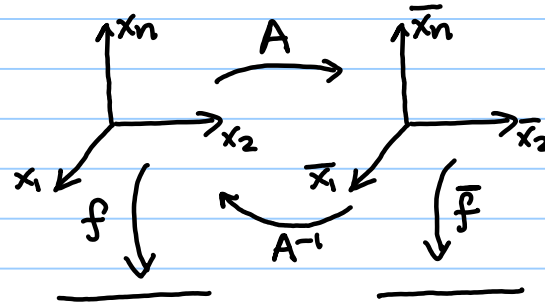
In the transformed coordinate system:

$$\bar{x}_{k+1} = \bar{x}_k - s \bar{\nabla}^2 \bar{f}_k^{-1} \bar{\nabla} \bar{f}_k$$

$$\downarrow \cdot A^{-1}$$

$$x_{k+1} = x_k - s A^{-1} A \nabla^2 f(x_k)^{-1} A^T A^{-T} \nabla f_k$$

$$= x_k - s \nabla^2 f_k^{-1} \nabla f_k$$



$\bar{f}(\bar{x}) = f(A^{-1}\bar{x})$
or $\bar{f} = f \circ A^{-1}$

Chain rule:
$\bar{f} = f \circ A^{-1}$
$d\bar{f}(\bar{x}) = df(A^{-1}\bar{x}) \cdot A^{-1}$
$\nabla \bar{f}(\bar{x}) = A^{-T} \nabla f(A^{-1}\bar{x})$
$d\nabla \bar{f}(\bar{x}) = A^{-T} d\nabla f(A^{-1}\bar{x}) \cdot A^{-1}$
(∇^2 on the left, ∇^2 on the right)

GD	Newton
• well-defined descent direction at all x (unless if x is already a critical point)	• not well-defined / not a descent dir when $\nabla^2 f(x) \not\succ 0$
• invariant under rigid transformation but not affine transformation	• affine invariant
• step size determined from approx. solving $\min_{\alpha} f(x + \alpha p)$	• "natural" step size 1 (more to say....)
• computation requires only $\underset{n \times 1}{\nabla f(x)}$	• computation requires $\underset{n \times 1}{\nabla f(x)}$, $\underset{n \times n}{\nabla^2 f(x)}$, and solving $-[\nabla^2 f(x)]p = \nabla f(x)$.
• rate of convergence : linear $\text{error}_k = O(\rho^k)$ (depends on choice of α_k and "conditioning" of f)	• rate of convergence : quadratic (a lot more to say...)

Quasi-Newton search directions provide an attractive alternative to Newton's method

- do not require computation of Hessian
- "superlinear" rate of convergence

Let me motivate this method.

The Newton method is based on

$$\min_x f(x) \xrightarrow{\text{local quadratic approx.}} \min_p \frac{1}{2} p^T \nabla^2 f_k p + \nabla f_k^T p + f_k \qquad \text{(Solve, and then iterate)}$$

It is also a method for solving nonlinear system of equations

$$\mathcal{F}(x) = 0 \xrightarrow{\text{local linear approx.}} \mathcal{F}(x_k) + D\mathcal{F}(x_k)(x - x_k) = 0 \qquad \text{(Solve, and then iterate)}$$

The latter applies to a general $\mathcal{F}: \mathbb{R}^n \to \mathbb{R}^n$. In this case, $d\mathcal{F}$ needs not be a symmetric $n \times n$ matrix.

When $f: \mathbb{R}^n \to \mathbb{R}^1$, $\mathcal{F} = \nabla f : \mathbb{R}^n \to \mathbb{R}^n$, the latter reduces to the former.

Ex: Check it. And recall that $d\nabla f(x) =$ (Hessian of f at x) is always symmetric.

Newton method for solving nonlinear equation in 1-D

x_{k+1} is the solution of $\underbrace{f_k + f_k'(x-x_k)}_{\approx f(x) \text{ near } x_k} = 0$. $\quad |x_{k+1}-x^*| \le C\,|x_k-x^*|^2$

Secant method:

x_{k+1} is the solution of $f_k + \frac{f_k - f_{k-1}}{x_k - x_{k-1}}(x-x_k) = 0$. $\quad |x_{k+1}-x^*| \le C\,|x_k-x^*|^{\alpha}$

$\alpha = (\sqrt{5}+1)/2 = 1.618.$

Quasi-Newton methods for minimization is based on generalizing the secant method to higher-dimensions for solving $\nabla f(x) = 0$.

In 1-D, $(f_k' - f_{k-1}')/(x_k - x_{k-1})$ is an approximation to f_k''

In higher-dimensions, a quasi-Newton method chooses a Hessian approximation B_k to $\nabla^2 f_k$ that satisfies the secant equation: $B_k(x_k - x_{k-1}) = \nabla f_k - \nabla f_{k-1}$.

In dimension $n>1$, such a B_k is under-determined. (In contrast, if we have available $x_{k-1}, \cdots, x_{k-n}$, $\nabla f_{k-1}, \cdots, \nabla f_{k-n}$, then (generically) $\exists!$ B_k st. $\overset{n\times n}{B_k}[\overset{n\times n}{x_k - x_{k-1}}, \cdots, x_k - x_{k-n}] = [\overset{n\times n}{\nabla f_k - \nabla f_{k-1}}, \cdots, \nabla f_k - \nabla f_{k-n}]$.

But this is too expensive to compute and quasi-Newton methods do not do that.)

We shall see that: with certain (economic) choices of B_k that satisfy the secant equation, the resulted iterative method enjoys superlinear convergence:

$$\|e_{k+1}\| / \|e_k\| \to 0. \qquad (SL)$$

Note that (SL) is weaker than saying

$$\|e_{k+1}\| \le C\|e_k\|^\alpha \text{ for some } \alpha > 1. \qquad (\text{strong-SL})$$

We know that when $n=1$, (strong-SL) holds with $\alpha = 1.618$.

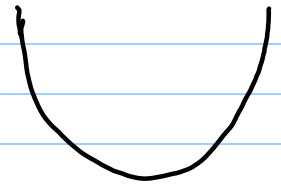
When $n>1$, even the existence of an exponent $\alpha>1$ is not guaranteed by the theory (and if it exists, one would expect it to deteriorate to 1 as $n \uparrow$.)

In contrast, Newton's method enjoys quadratic convergence $\|e_{k+1}\| \le C\|e_k\|^2$ in any dimension, at the cost of much more expensive computation per iteration.

Surprisingly, quasi-Newton methods are much more effective than GD.
Moreover, a version of Quasi-Newton method under the name of BFGS is widely used.

We shall present Quasi-Newton methods in details, several aspects (choice of B_k, choice of step size, etc) are quite tricky, and it is a tour de force to put all the pieces together.

Choice of step size



Let's begin with $f(x) = \frac{1}{2}x^2$ in 1-D.

$f'(x) = x$. $x_{k+1} = x_k - \alpha_k x_k$ ← GD with step size α_k

$= (1-\alpha_k)x_k$

$\alpha_k \in (-1, 1)$ guarantees a decrease in $|x_k|$, and hence also in $f(x_k)$, in each step.

But it is insufficient to guarantee that $x_n \to 0 = \operatorname{argmin} f(x)$.

$x_n = (1-\alpha_1)(1-\alpha_2)\cdots(1-\alpha_n)x_0$

When $\alpha_i \in [0, 1)$, $\prod_{i=1}^{\infty}(1-\alpha_i) > 0 \iff \sum_{i=1}^{\infty}\alpha_i$ converges

This means $\lim_{n\to\infty} x_n \neq 0$ if α_i is too small (e.g. $\alpha_i = 1/(i+1)^2$)

Three popular choices of step size

- constant stepsize $\alpha_k = \bar{\alpha} \quad \forall k$
- exact line search $\alpha_k \in \operatorname{argmin}_{\alpha \geq 0} f(x_k + \alpha p_k)$
- backtracking (see below)

A step size α_k satisfies the Sufficient decrease / Armijo condition if :

$$f(x_k + \alpha p_k) \leq f(x_k) + c\,\alpha\,\nabla f(x_k)^T p_k \qquad \text{for some } c \in (0,1)$$

If p_k is a descent direction, we can always find a $\alpha > 0$ s.t. $f(x_k + \alpha p_k) < f(x_k)$. The sufficient decrease condition is stronger than that. Still, such a step size is guaranteed to exist :

Lemma Let $f : \mathcal{U} \overset{\text{open}}{\subseteq} \mathbb{R}^n \to \mathbb{R}$ be C^1, p is a descent direction at x. For any $c \in (0,1)$, $\exists\, \varepsilon > 0$ st.

$$f(x) - f(x + \alpha p) \geq -c\,\alpha\,\nabla f(x)^T p \quad , \quad \forall\, \alpha \in [0, \varepsilon].$$

Proof: $f(x + \alpha p) = f(x) + \alpha \nabla f(x)^T p + o(\alpha \|p\|)$ (differentiability at x)

$$f(x) - f(x+\alpha p) = \underbrace{-c\,\alpha\nabla f(x)^T p}_{(0,1)} \underbrace{-(1-c)\alpha\nabla f(x)^T p}_{(0,1)} - o(\alpha\|p\|)$$

$$\underbrace{\qquad\qquad\qquad\qquad}_{>0 \text{ for } \alpha>0}$$

$$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}_{>0 \text{ for small } \alpha>0}$$

so $\quad f(x) - f(x+\alpha p) \geq -c\,\alpha\nabla f(x)^T p$

for small $\alpha > 0$. Q.E.D.

Algorithm 3.1 (Backtracking Line Search).

Choose $\bar{\alpha} > 0$, $\rho \in (0,1)$, $c \in (0,1)$; Set $\alpha \leftarrow \bar{\alpha}$;

repeat until $f(x_k + \alpha p_k) \leq f(x_k) + c\alpha\nabla f_k^T p_k$

$\alpha \leftarrow \rho\alpha$;

end (repeat)

Terminate with $\alpha_k = \alpha$.

Start with a big $\bar{\alpha}$, shrink it by a factor of ρ until the c-Armijo condition is satisfied.

A very subtle point: The "sufficient decrease condition" alone cannot guarantee sufficient decrease in the objective value, because, as the lemma above shows, any small enough step size satisfies it. Put differently, the condition alone does not prohibit the step size from getting arbitrarily close to 0.

But the backtracking algorithm does!

This means: In order to establish a convergence result for a line search method based on backtracking, the proof must exploit the specific feature of the backtracking algorithm instead of simply invoking the Armijo condition.

Turns out it can be done for GD, ie. when $p_k = -\nabla f(x_k)$. (Stay tuned!)

But for more general descent methods, it is probably easier/necessary to strengthen the Armijo condition so that the stronger condition prohibits the step size to get arbitrarily close to 0.

One such stronger condition requires:

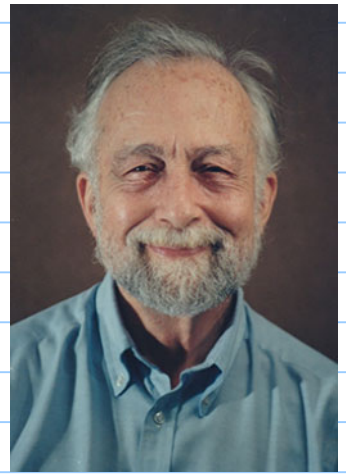
$$f(x_k + \alpha_k p_k) \leq f(x_k) + c_1 \alpha_k \nabla f_k^T p_k \quad \leftarrow \text{Armijo cond.}$$
$$\nabla f(x_k + \alpha_k p_k)^T p_k \leq c_2 \nabla f_k^T p_k \quad \leftarrow \text{"curvature cond."}$$

with

$$0 < c_1 < c_2 < 1$$

Known collectively as the Wolfe conditions.

P. Wolfe



Strong Wolfe conditions:

$$f(x_k + \alpha_k p_k) \leq f(x_k) + c_1 \alpha_k \nabla f_k^T p_k$$
$$|\nabla f(x_k + \alpha_k p_k)^T p_k| \leq c_2 |\nabla f_k^T p_k| \quad , \quad 0 < c_1 < c_2 < 1$$

Note: $\nabla f(x_k + \alpha p_k)^T p_k = \phi'(\alpha)$ where $\phi(\alpha) = f(x_k + \alpha p_k)$.

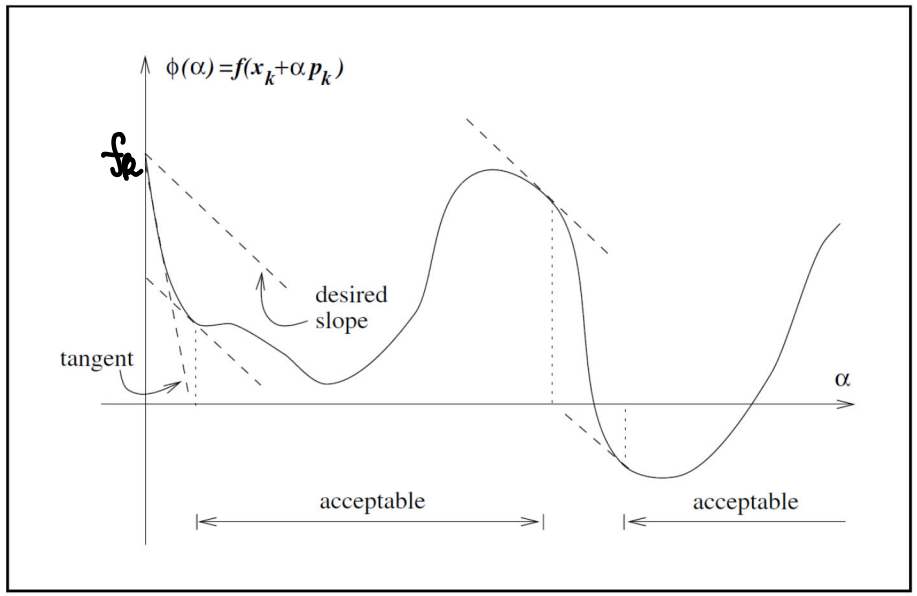


Figure 3.4 The curvature condition.

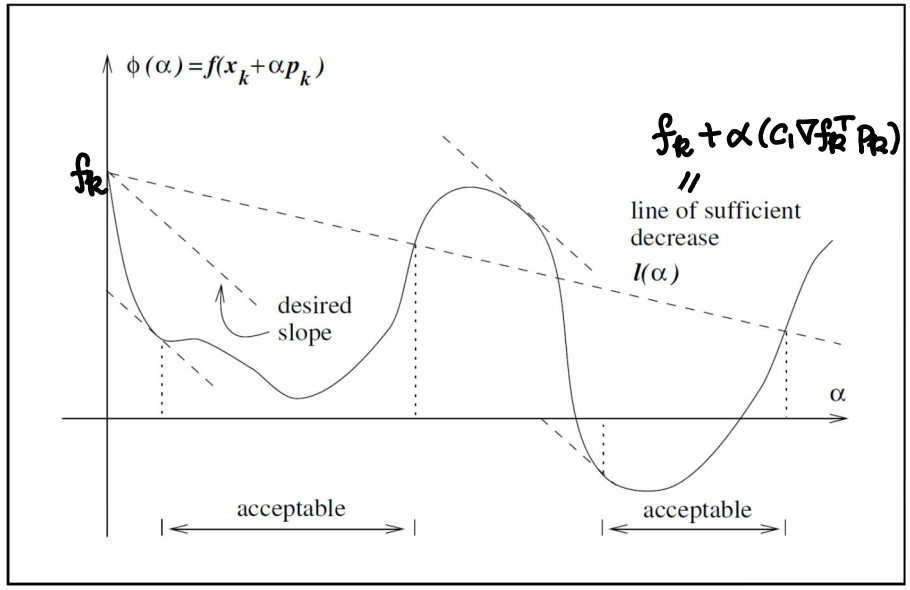


Figure 3.5 Step lengths satisfying the Wolfe conditions.

Lemma 3.1.

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable. Let p_k be a descent direction at x_k, and assume that f is bounded below along the ray $\{x_k + \alpha p_k | \alpha > 0\}$. Then if $0 < c_1 < c_2 < 1$, there exist intervals of step lengths satisfying the Wolfe conditions (3.6) and the strong Wolfe conditions (3.7).

Proof : see N&W, pg 35.

The proof is not hard, but constructing an algorithm for the Wolfe conditions is tricky.

We shall establish the following results :

- Convergence of GD, without assuming "strong convexity" near a critical point, for all 3 choices of step sizes (constant, exact, backtracking)
- rate of convergence results of GD under "strong convexity" assumption.
- a convergence result of general descent method with step size satisfying the Wolfe conditions
- various rate of convergence results for Newton and Quasi-Newton methods, after we

explained these methods. (We shall see, e.g., that backtracking does not work well with quasi-Newton methods and we need step size selection algorithms that satisfy Wolfe conditions.)