Notation :

s d alg. = search direction alg.
SD alg. = steepest-descent alg.

Properties of SD alg.
for standard quadratic V on $\mathbb{R}^n$

$$V(x) = a + b^T x + \tfrac{1}{2} x^T C x \quad , C^T = C > 0$$

Why consider a quadratic ?

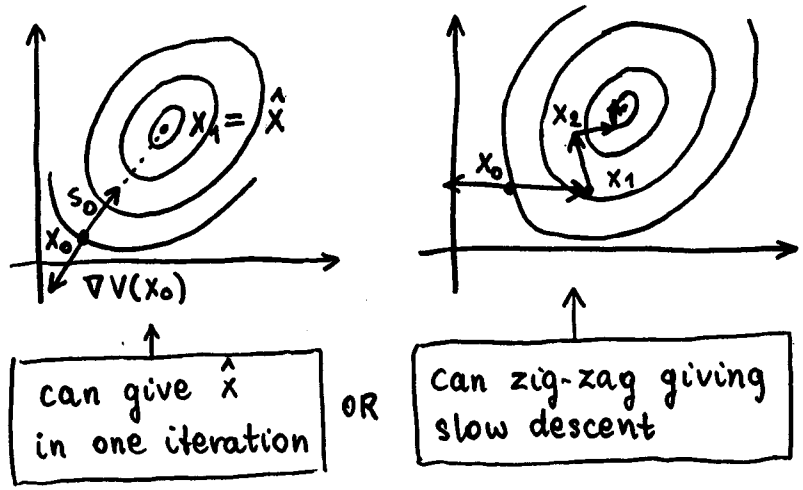
Often study algorithm behaviour on quadratics because :

- analysis easy
- [alg. good for quadratic V] $\Rightarrow$ [might be good for general V]
- [alg. bad for quadratic V] $\Rightarrow$ [alg. probably generally useless]

But, in practice, for a quadratic we would not usually find $\hat{x}$ using a s.d. alg. as $\hat{x} = -C^{-1}b$ for a quadratic

- so just solve $C\hat{x} = -b$

Characteristic of S.D. alg. with exact minimization along each s_j



can give $\hat{x}$ in one iteration OR Can zig-zag giving slow descent

depending on x_0

(3.1) Th For
- SD alg. with exact minimization along each s_j
- standard quadratic V on $\mathbb{R}^n$

(i) [$\nabla V(x_0)$ = eigenvector of C (#)] $\Rightarrow$ [$x_1 = \hat{x}$] - NICE

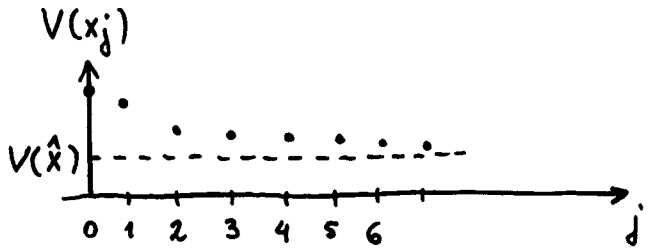
---

e is an eigenvector of C iff $\exists$ an eigenvalue λ_i of C and

$$(C - \lambda_i I)e = 0$$

(3.1) Th continued

(ii) $[\nabla V(x_0) \neq$ eigenvector of $C] \Rightarrow \begin{bmatrix} x_j \neq \hat{x}, \\ \forall \text{ finite } j \end{bmatrix}$ - BAD



(iii) $0 \leq [V(x_j) - V(\hat{x})] \leq \alpha^j [V(x_0) - V(\hat{x})]$

$$0 \leq \alpha = 1 - \frac{\lambda_{min}(C)}{\lambda_{max}(C)} < 1 \quad \$$$

$\Downarrow$

(iv) $V(x_j) \longrightarrow V(\hat{x})$

smaller α guarantees faster convergence

$\$$ Here $\lambda_{min}(C)$ is the smallest eigenvalue of C and $\lambda_{max}(C)$ is the largest

Justification of (3.1) Th. (i):

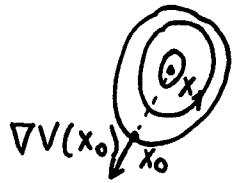
no proof for (ii) & (iii) here

for a quadratic $\hat{x} = -C^{-1}b$

so, for $x_1 = \hat{x} = -C^{-1}b$

needed that:

$x_1 - x_0 \parallel \nabla V(x_0)$ (parallel)



$\Downarrow$

$\exists \lambda_i$ s.t. $\neq 0$

$$\nabla V(x_0) + \lambda_i (x_1 - x_0) = 0$$

$$\nabla V(x_0) + \lambda_1 (-C^{-1}b - x_0) = 0$$

$$\nabla V(x_0) - \lambda_1 C^{-1}(\underbrace{Cx_0 + b}_{\nabla V(x_0)}) = 0$$

$\Downarrow$

$$C\nabla V(x_0) - \lambda_1 \nabla V(x_0) = 0$$

$\nabla V(x_0)$ is an eigenvector of C

QED

Remark: In view of the possibility of the zig-zag effect, SD is not a very good alg. BUT it is simple, robust

$\Rightarrow$ can be useful

⇒ an improved alg. for optimizing quadratics needed

First study:

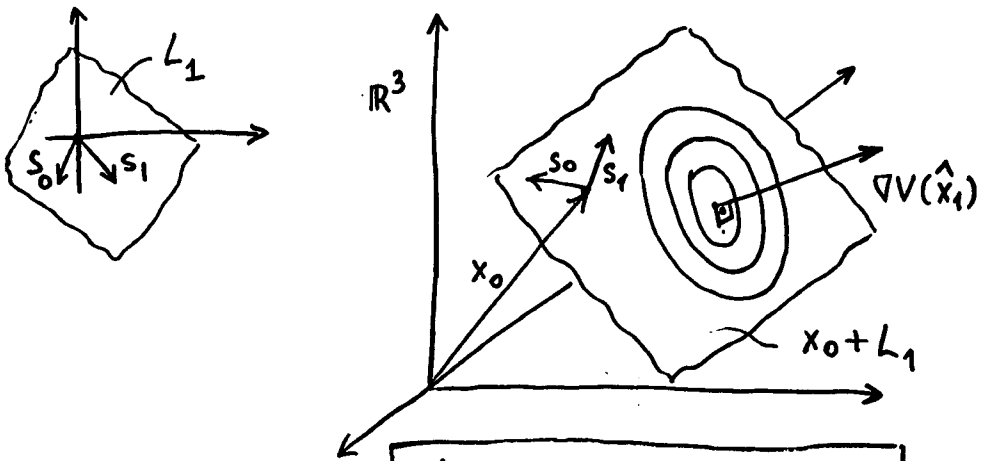
Optimization of a quadratic on a linear variety

$s_0, s_1, s_2, \ldots \in \mathbb{R}^n$ ← search directions

L.S.S. : $L_j := \mathcal{L}[s_0, s_1, \ldots, s_j]$

linear subspace

Linear variety : $x_0 + L_j := \{x_0 + y : y \in L_j\}$



unique for our standard quadratic

$\hat{x}_j$ denotes the global minimizer of V on variety $x_0 + L_j$

Suggests:

(3.2) Th: $\hat{x}_j$ minimizes V globally on $x_0 + L_j$ iff:

(i) $\hat{x}_j \in x_0 + L_j$

(ii) $\nabla V(\hat{x}_j)^T s_i = 0$, $i = 0, 1, \ldots, j$

(proof only "geometric")

Conjugate direction algorithm for standard quadratic on $\mathbb{R}^n$

CD alg = sd alg. when:

(i) exact minimization used along each s_j

(ii) s_i are C-conjugate, in that

$$s_i^T C s_j = 0 \quad , \quad i \neq j \qquad (\pounds)$$

(3.3) Th.: For a CD algorithm:

x_j minimizes V globally on any variety $x_0 + L_{j-1}$, $j = 1, 2, \ldots, n$

Proof: By Th (3.2) we just have to show that:

(a) $x_j \in x_0 + L_{j-1}$

(b) $\nabla V(x_j)^T s_i = 0 \;,\; i = 0,1,..,j-1$

Proof of (a): $(x_j \in x_0 + L_{j-1})$

$$x_1 = x_0 + \underbrace{w_0 s_0}_{\in \mathcal{L}[s_0] = L_0} \in x_0 + L_0$$

$$x_2 = x_1 + w_1 s_1 = x_0 + \underbrace{w_0 s_0 + w_1 s_1}_{\in \mathcal{L}[s_0 s_1] = L_1} \in x_0 + L_1$$

$$x_3 = x_2 + w_2 s_2 = x_0 + \underbrace{w_0 s_0 + w_1 s_1 + w_2 s_2}_{\in L[s_0, s_1, s_2] = L_2} \in x_0 + L_2$$

$$\Downarrow$$

$$x_j \in x_0 + L_{j-1} \qquad \text{Q.E.D.}$$

Proof of (b): $(\nabla V(x_j)^T s_i = 0,\; i = 0, 1, .., j-1)$

$$\Updownarrow$$

$(\nabla V(x_j)^T s_i = 0,\; i = 1, .., j-1)$

$$\nabla V(x) = b + Cx$$

$$\text{Th (2.5)}$$

$$\nabla V(x+\delta x) = b + C(x+\delta x)$$

$$= \underbrace{b + Cx}_{\nabla V(x)} + C\delta x$$

(3.4) $$\boxed{\nabla V(x+\delta x) = \nabla V(x) + C\delta x \;,\; \forall\, x, \delta x}$$

Also if V is minimized exactly along each s_j then:

$$\boxed{w_j = -\frac{\nabla V(x_j)^T s_j}{s_j^T C s_j} \;,\; \forall j}$$

this is because

$$w_j = \arg\min_{w \in \mathbb{R}} V(x_j + w s_j) \;,\; \forall j$$

$$\Downarrow$$

w_j must solve $\frac{d}{dw} V(x_j + w s_j)\big|_{w = w_j} = 0$

$$\Downarrow$$

$$\nabla V(x_j + w_j s_j)^T \cdot s_j = 0 \;,\; \forall j$$

$$\Updownarrow$$

$$s_j^T [b + C(x_j + s_j w_j)] = 0 \;,\; \forall j$$

$$\Updownarrow$$

$$s_j^T b + s_j^T C x_j + s_j^T C s_j \omega_j = 0 \quad , \forall i$$

$s_j^T C x_j = s_j^T \nabla V(x_j)$

$$s_j^T \nabla V(x_j) + s_j^T C s_j \omega_j = 0 \quad , \forall i \Rightarrow (3.4)$$

So:

(♯) $\nabla V(x_1)^T s_0 = \nabla V(x_0 + \omega_0 s_0)^T s_0$

$= \nabla V(x_0) + \omega_0 s_0^T C s_0 = 0$

(3.4)

as $\omega_0 = \dfrac{-\nabla V(x_0)^T s_0}{s_0^T C s_0}$

We will show next that, because the s_i are C-conjugate, the fact that

$\nabla V(x_1)^T s_0 = 0$

causes that

$\nabla V(x_i)^T s_0 = 0$ for all $i > 1$.

$$\nabla V(x_2)^T s_0 = \nabla V(x_1)^T s_0 + \omega_1 s_1^T C s_0 = 0 \qquad (*)$$

$\nabla V(x_2)^T s_0 = \nabla V(x_1 + \omega_1 s_1)^T s_0$ — $(\sharp)$ — 0

$\nabla V(x_1 + \omega_1 s_1) = \nabla V(x_1) + \omega_1 C s_1$

$\omega_1 s_1^T C s_0$ — $(\natural)$ — 0

$$\nabla V(x_3)^T s_0 = \nabla V(x_2)^T s_0 + \omega_2 s_2^T C s_0 = 0$$

$\nabla V(x_3)^T s_0 = \nabla V(x_2 + \omega_2 s_2)^T s_0$ — $(*)$ — 0

$\nabla V(x_2 + \omega_2 s_2) = \nabla V(x_2) + \omega_2 C s_2$

$\omega_2 s_2^T C s_0$ — $(\natural)$ — 0

. . . . . .

So (♯)

$\nabla V(x_1)^T s_0 = \nabla V(x_2)^T s_0 = \ldots = \nabla V(x_j)^T s_0 = 0$

Similarly

$\nabla V(x_2)^T s_1 = \ldots = \nabla V(x_j)^T s_1 = 0$

$\nabla V(x_3)^T s_2 = \ldots = \nabla V(x_j)^T s_2 = 0$

. . . . . .

$\nabla V(x_j)^T s_{j-1} = 0$

QED for Th (3.3)(b)

(3.5) Corollary For the CD alg.:

$$V(x_0) \geq V(x_1) \geq \cdots \geq V(x_n) = V(\hat{x})$$

as CD is a sd. alg.

special property of CD alg.

Actually, CD might give $V(x_j) = V(\hat{x})$ for $j < n$ but this shows that at most n iterations are needed to achieve $V(\hat{x})$.

very NICE

especially as SD usually never achieves $V(\hat{x})$ exactly.

Proof of Corollary: ($V(x_n) = V(\hat{x})$)

Th (3.3) $\Rightarrow$

x_n minimizes V on $x_0 + L_{n-1}$

$\|$

$\mathcal{L}[s_0, \ldots, s_{n-1}] \subset \mathbb{R}^n$

But $\mathcal{L}[s_0, \ldots, s_{n-1}] = \mathbb{R}^n$

because:

Fact: C-conjugate s_i $\Rightarrow$ linear independence of all s_i, $i = 0, \ldots, n-1$

for if not, then: $s_j^T C s_i = 0 \quad \forall\, i \neq j$

and $\sum_{i=1}^{n} \alpha_i s_i = 0$ for some α_i s.t. $\sum_{i=1}^{n} |\alpha_i| \neq 0$

$\Downarrow$

$\sum_{i=1}^{n} \alpha_i C s_i = 0$ -11- -11- -11-

$\Downarrow$

$s_j^T \cdot \left| \sum_{i=1}^{n} \alpha_i C s_i = 0 \right.$ -11- -11- -11-

$\Downarrow$

$\alpha_j \underbrace{s_j^T C s_j}_{>0 \text{ since } C>0} = 0$ -11- -11-, $\forall j$

$\Downarrow$

$\alpha_j = 0 \quad \forall j \Rightarrow \square$ — a contradiction.

$\Rightarrow$ $\Downarrow$

s_i, $i = 0, \ldots, n-1$ must be linearly independent

So: x_n minimizes V on $x_0 + \mathbb{R}^n = \mathbb{R}^n$

$\Rightarrow V(x_n) = V(\hat{x})$

One way to generate C-conjugate s_i:

(3.6) The Conjugate-Gradient Alg.

Choose $x_0 \in \mathbb{R}^n$

set $s_0 = -\nabla V(x_0)$, $j = 0$

1) Stop if $\|\nabla V(x_j)\| < \varepsilon$ ← as then $x_j \cong \hat{x}$

Choose $\omega_j \in \arg\min_{\omega \in \mathbb{R}} V(x_j + \omega s_j)$

Set $x_{j+1} = x_j + \omega_j s_j$

$$\beta_{j+1} := \frac{[\nabla V(x_{j+1}) - \nabla V(x_j)]^T \nabla V(x_{j+1})}{\|\nabla V(x_j)\|^2} \in \mathbb{R}$$

$$s_{j+1} := -\nabla V(x_{j+1}) + \beta_{j+1} s_j$$

$j := j+1$

Go to 1)

this term makes CG different from SD

(3.7) Th For CG alg. and quadratic V:

(i) the $s_0, s_1, \ldots$ generated before it stops are C-conjugate

not proved here

but easy to check eg. for s_0 and s_1:

$$\beta_1 = \frac{[\nabla V(x_1) - \nabla V(x_0)]^T \nabla V(x_1)}{\|\nabla V(x_0)\|^2} \quad (= \omega_0 C s_0 \text{ in the numerator}; \ \|\nabla V(x_0)\|^2 = \|s_0\|^2)$$

$$\Downarrow$$

$$s_0^T C s_1 = -s_0^T C \nabla V(x_1) + \omega_0 s_0^T C \frac{s_0^T C \nabla V(x_1)}{\|s_0\|^2} s_0$$

$$\frac{\|s_0\|^2}{s_0^T C s_0} = -\frac{\nabla V(x_0)^T s_0}{s_0^T C s_0}$$

$\omega_0 = 0$

$$\Downarrow$$

(ii) CG alg. = a CD alg.

$$\Downarrow$$ Th (3.3)

(iii) x_j minimizes V on $x_0 + L[s_0, \ldots, s_{j-1}]$ for each x_j found before it stops

$\Downarrow$

(iv) CG alg. stops after $k \leq n$ iterations with $x_k = \hat{x}$ ← NICE

Comparison of CG and SD:

(3.8) Th If, for a quadratic V, SD and CG start at the same x_0:

(a) $x_1 = x_1$ (CG, SD)

(b) either $x_1 = \hat{x}$ (CG, SD)

hence CG has a theoretical advantage over SD

or $V(x_j) < V(x_j)$, $\forall j > 1$ (CG, SD)

Remark: What usually happens is something like: