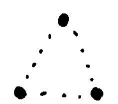
Algorithms for unconstrained optimization on $\mathbb{R}^n$

$F = \mathbb{R}^n$ — no constraints

Simplex algorithm

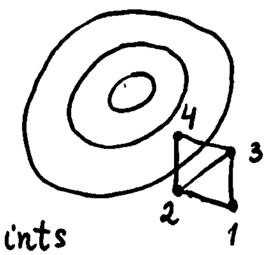
A simplex in $\mathbb{R}^n$ = set of $n+1$ equidistant points

e.g. in $\mathbb{R}^2$:



Basic simplex alg. — not the same as the Simplex Alg. of *Linear Programming*

- start with simplex (1,2,3)
- reflect the point in the simplex (1,2,3) which has greatest cost in the centroid of the remaining points giving a new simplex (2,3,4) "nearer" a l.m.

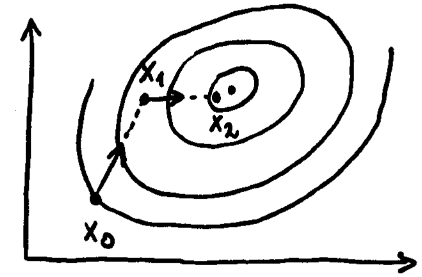


- repeat for simplex (2,3,4) etc.

Properties: simple in principle but needs modifications to help convergence to a l.m.

Search direction algorithm

$x_1 = x_0 + w_0 s_0$



- Choose $x_0 \in \mathbb{R}^n$ (an estimate of a g.m. $\hat{x}$)
- Set iteration index $j = 0$

1) [Iteration j] :

- Using info available at x_j, choose a search direction $s_j \in \mathbb{R}^n$ #
- choose a scalar $w_j > 0$ so
  $V(x_j + w_j s_j) < V(x_j)$
- Set
  $x_{j+1} := x_j + w_j s_j$
  $j := j+1$ and go to 1)

choose s_j so V should decrease as one moves along s_j, away from x_j

Remarks:

very important type of alg. as it can be developed & adapted a lot

Problems:

choice of s_j, ω_j

Concerning choice of s_j:

The Gradient

$\mathbb{R}^2$ case

Under weak conditions ← $\boxed{\text{see cond. (a) \& (b) on p. 2.4}}$

$$V\left[\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} \delta x_1 \\ \delta x_2 \end{pmatrix}\right] \overset{\Omega}{=} V\left[\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}\right] + \frac{\partial V}{\partial x_1}[x]\,\delta x_1 + \frac{\partial V}{\partial x_2}[x]\,\delta x_2$$

$$= V[x] + \underbrace{\begin{bmatrix} \partial V[x]/\partial x_1 \\ \partial V[x]/\partial x_2 \end{bmatrix}}_{\nabla V[x]}{}^{T} \underbrace{\begin{pmatrix} \delta x_1 \\ \delta x_2 \end{pmatrix}}_{\delta x}$$

$\boxed{\text{the gradient of } V \text{ at } x}$

i.e.

$$V[x+\delta x] \underset{\boxed{\text{for small } \delta x}}{\overset{\Omega}{=}} V[x] + \nabla V[x]^T \delta x$$

For $\mathbb{R}^n$ case:

(2.1) **Th** Consider $V: \mathbb{R}^n \to \mathbb{R}$ when the p.ds. $\frac{\partial V}{\partial x_i}[z]$:

(a) all exist $\forall z$ s.t. $\|z - x\| < \varepsilon$ for some $\varepsilon > 0$ #

(b) all are continuous at $z = x$

Then:

V is called differentiable at x and, for small δx:

$$V[x+\delta x] \overset{\Omega}{=} V[x] + \underbrace{\nabla V[x]}{}^{T} \delta x$$

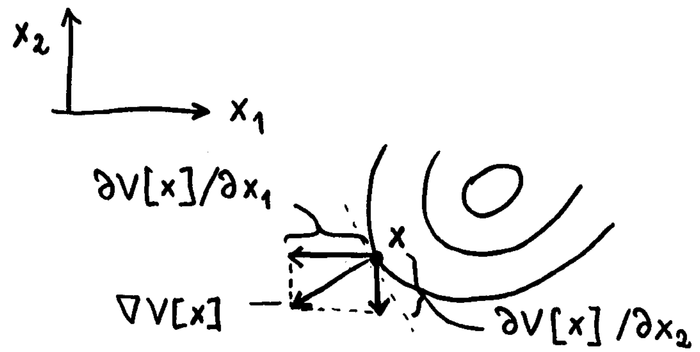
$$\nabla V[x] \triangleq \begin{bmatrix} \partial V[x]/\partial x_1 \\ \cdots\cdots \\ \partial V[x]/\partial x_n \end{bmatrix} \in \mathbb{R}^n$$

$\boxed{\text{the gradient of } V \text{ at } x}$

---

i.e. the p.ds. all exist everywhere near z

Graphical representation of ∇V

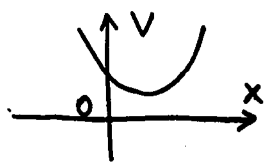


(2.2) Fact:

Even for the $\mathbb{R}^n$ case:
$\nabla V[x]$ is orthogonal to the constant V contour through x and points uphill

Quadratics on $\mathbb{R}^n$:

Quadratic of $x \in \mathbb{R}$



$$V(x) = \underset{\in \mathbb{R}}{a} + \underset{\in \mathbb{R}}{b}x + \tfrac{1}{2}x\,\underset{\in \mathbb{R}}{c}\,x, \quad V(x) \in \mathbb{R}$$

$\boxed{c \in \mathbb{R} \text{ \& symmetric}^{\#}}$

$\boxed{\text{for a min}: c > 0}$

in the sense that a scalar c can be thought of as a [x] matrix, which is symmetric

Standard quadratic of $x \in \mathbb{R}^n$:

$$V(x) = \underset{\in \mathbb{R}}{a} + \underset{\in \mathbb{R}^n}{b}^T x + \tfrac{1}{2}x^T \underset{\in \mathbb{R}^{n\times n}}{C}\, x, \quad V(x) \in \mathbb{R}$$

$C^T = C$ & $C > 0$

$\boxed{\begin{array}{l}\text{read as:}\\ C \text{ is positive}\\ \text{definite \& symmetric}\end{array}}$

$\boxed{\text{i.e } x^T C x > 0,\ \forall\, x \neq 0,\ x \in \mathbb{R}^n}$

Tests for $C^T = C > 0$

For symmetric C:

$$[C>0] \iff \left\{ \begin{array}{l} c_{11} > 0\ ; \\ \begin{vmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{vmatrix} > 0\ ; \\ \begin{vmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{vmatrix} > 0\ ; \\ \cdots\cdots \\ |C| > 0 \end{array} \right\}$$

Also:

$$[C>0] \Longleftrightarrow \lambda_i[C]>0\,,\ \forall i$$

the eigenvalues of C; since $C^T=C$, all the eigenvalues of C are real. This says that for C to be positive definite, all its eigenvalues must be strictly positive

By simple algebra:

$$V[x+\delta x] = V[x] + [b+Cx]^T\delta x + \underbrace{\tfrac{1}{2}\,\delta x^T C\,\delta x}_{\boxed{\text{second order}}} \quad (2.4)$$

$$\underset{\boxed{\text{for small } \delta x}}{\overset{\Omega}{=}} V[x] + [b+Cx]^T\delta x$$

Suggests #: $\nabla V[x] = b + Cx$ — check by partial differentiation

---

\# because we also have (from Th.(2.1)):

$$V(x+\delta x) \cong V(x) + \nabla V(x)^T\delta x$$

Hence:

(2.5) Th: For a standard quadratic on $\mathbb{R}^n$

(i) V is differentiable on $\mathbb{R}^n$
(i.e. V is differentiable at x, $\forall x \in \mathbb{R}^n$)

(ii) $\nabla V[x] = b + Cx$

(iii) $V[x+\delta x] = V[x] + \nabla V[x]^T\delta x + \frac{1}{2}\delta x^T C\,\delta x$

An optimality condition for general V

(2.6) Th If $V:\mathbb{R}^n \to \mathbb{R}$ is differentiable at $\hat{x}$:

$$[\underbrace{\hat{x} = \text{a l.m. of } V \text{ on } \mathbb{R}^n}_{\boxed{\text{unconstrained}}}] \Rightarrow [\underbrace{\nabla V(\hat{x}) = 0}_{\boxed{\text{necessary condition for } \hat{x} = \text{a l.m.}}}]$$

Remark: So, one way to find a l.m. is:
solve $\nabla V(\hat{x}) = 0$ for $\hat{x}$
check if $\hat{x}$ = a l.m.

<u>Proof of Th. (2.6):</u>

contrapositive i.e.:

$$\neg\,[\nabla V(\hat{x})=0] \Rightarrow \neg\,[\hat{x} = \text{a l.m.}]:$$

$$\neg\,[\nabla V(\hat{x})=0] \Rightarrow [\nabla V(\hat{x}) \neq 0]$$

$\Rightarrow$ for all sufficiently small real $\omega > 0$

(i) $\hat{x} + \omega[-\nabla V(\hat{x})] = \text{near } \hat{x}$

(ii) $V(\hat{x} + \underbrace{\omega[-\nabla V(\hat{x})]}_{\delta x}) =$

$$\overset{\#}{\cong} V(\hat{x}) + \nabla V(\hat{x})^T \delta x$$

$$= V(\hat{x}) - \omega \|\nabla V(\hat{x})\|^2 < V(\hat{x})$$

$$\Rightarrow [\hat{x} \neq \text{a l.m.}] \Rightarrow \neg\,[\hat{x} = \text{l.m.}]$$

QED

\# the main point here is that the fact that

$V(\hat{x}) + \nabla V(\hat{x})^T \delta x < V(\hat{x})$

guarantees that

$$V(\hat{x} + \delta x) < V(\hat{x})$$

<u>Application to minimization of a standard quadratic</u>

(I) Solve $\nabla V(\hat{x}) = 0$, $\nabla V(\hat{x}) = b + C\hat{x}$ — Th. (2.5 (ii))

$$\boxed{\text{Fact}: (C^T = C > 0) \Rightarrow C^{-1} \text{ exists}}$$

Hence: $\boxed{\hat{x} = -C^{-1}b}$

(II) Check: $\hat{x} = \text{a l.m.}$?

$$V(x) = V(\hat{x} + \underbrace{[x - \hat{x}]}_{\delta x}) \quad \text{— Th. 2.5 (iii)}$$

$$= V(\hat{x}) + \underbrace{\nabla V(\hat{x})}_{=0}\delta x + \tfrac{1}{2}\underbrace{\delta x^T C \delta x}_{\geq 0 \text{ as } C>0}$$

$$\geq V(\hat{x}), \quad \forall x \in \mathbb{R}^n$$

so $\underline{\hat{x} = \text{a g.m. of } V \text{ on } \mathbb{R}^n}$

$\Downarrow$

(2.7) <u>Th</u>: $\hat{x} = -C^{-1}b$ is a <u>unique g.m.</u> of a quadratic V on $\mathbb{R}^n$ and:

$$V(\hat{x}) = a - \tfrac{1}{2} b^T C^{-1} b$$

not just a l.m.

More about the Search Directions Alg.

A choice for s_j :

Assume that V is differentiable on $\mathbb{R}^n$

At x_j :

$\nabla V(x)$ points uphill

$-\nabla V(x_j)$ points downhill

choose : $s_j = \underbrace{-\nabla V(x_j)}$

steepest-descent search direction

as then s_j points downhill so that V should decrease as we move along s_j from x_j

The steepest-descent alg.

Choose $x_0 \in \mathbb{R}^n$. Set $j = 0$

1) [At iteration j]

- Stop if $\underbrace{\nabla V(x_j) = 0}$ — as then x_j satisfies a necessary condition for a l.m.

  in practice, use $\| \nabla V(x_j) \| < 10^{-8}$ say

- Set : $s_j := -\nabla V(x_j)$
- Choose a scalar $\omega_j > 0$ so

  $V(x_j + \omega_j s_j) < V(x_j)$

  often, ω_j chosen to minimize $V(x_j + \omega s_j)$ $\omega \in \mathbb{R}$ - at least approximately

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

  $j := j + 1$

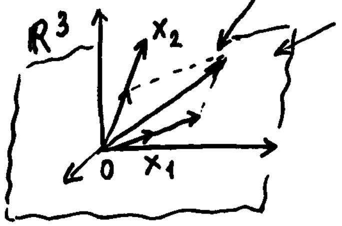
- Go to 1)

Some revision of linear algebra:

Consider $x^1, x^2, \dots, x^k \in \mathbb{R}^n$

A linear combination of these is $\sum_{i=1}^{k} \underbrace{\alpha_i}_{\boxed{\text{scalars}}} x^i$

e.g. $2x^1 + \frac{1}{2}x^2$



plane through $0, x^1, x^2$

= set of all linear combinations of x^1, x^2

$= \{\alpha_1 x^1 + \alpha_2 x^2 : \alpha_1, \alpha_2 \in \mathbb{R}\}$

the set of vectors $\alpha_1 x^1 + \alpha_2 x^2$ that one gets as α_1, α_2 vary all over the set of all real numbers

$\triangleq$

$\mathcal{L}[x^1, x^2]$

the linear subspace spanned by x^1, x^2

We can say: $x^1, \dots, x^k \in \mathbb{R}^n$ are linearly independent iff:

$$\{0\} \subsetneqq \mathcal{L}[x^1] \subsetneqq \mathcal{L}[x^1, x^2] \subsetneqq \cdots \subsetneqq \mathcal{L}[x^1, \dots, x^k]$$

Then we say:

$\mathcal{L}[x^1]$ has dimension 1

$\mathcal{L}[x^1, x^2]$ -"- -"- 2

$\mathcal{L}[x^1, x^2, x^3]$ -"- 3 etc.

.....

Further: $x^1, \dots, x^n \in \mathbb{R}^n$ are linearly indep. iff $\mathcal{L}[x^1, \dots, x^n] = \mathbb{R}^n$

In general:

$\mathcal{L}[x^1, \dots, x^k]$

$\equiv$

the linear subspace spanned by $x^1, \dots, x^k$

$\equiv$

$\{\sum_{i=1}^{k} \alpha_i x^i : \alpha_i \in \mathbb{R}, \forall i\} \subset \mathbb{R}^n$

$\equiv$

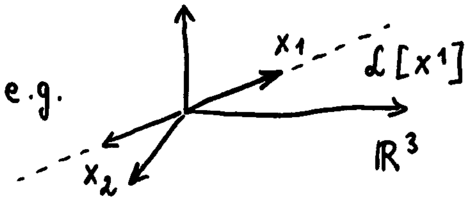
the plane through $0, x^1, \dots, x^k \subset \mathbb{R}^n$

Further: $x^1, \dots, x^k \in \mathbb{R}^n$ are linearly indep. iff:

$$\sum_{i=1}^{k} \alpha_i x^i = 0 \Rightarrow \alpha_i = 0 \ \forall i$$

Sometimes

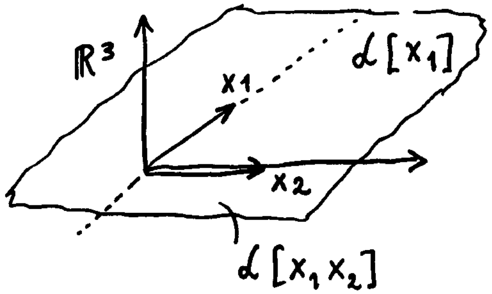
$\mathcal{L}[x^1] = \mathcal{L}[x^1, x^2]$ e.g.



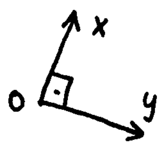
But usually

$\mathcal{L}[x^1] \subsetneq \mathcal{L}[x^1, x^2]$

e.g.



Vectors $x, y \in \mathbb{R}^n$ are orthogonal



iff $x^T y = 0$