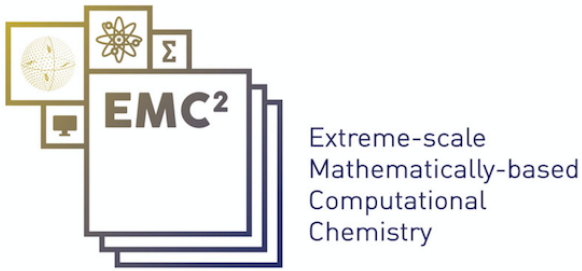
Mathematical aspects of electronic structure theory

Eric CANCES

Ecole des Ponts and Inria Paris, France

ISTPC 2024, Aussois, June 16-29



Question 1

A Fortran library for solving $A\mathbf{x} = \mathbf{b}$ gives the following results:

$$\mathbf{A} = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix} \qquad \textbf{Solution: } \mathbf{x} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

$$\mathbf{A} = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} 32.001 \\ 22.999 \\ 33.001 \\ 30.999 \end{pmatrix} \qquad \textbf{Solution: } \mathbf{x} = \begin{pmatrix} 1.082 \\ 0.862 \\ 1.035 \\ 0.979 \end{pmatrix}$$

$$\mathbf{A} = \begin{pmatrix} 10 & 7.021 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix} \qquad \textbf{Solution: } \mathbf{x} = \begin{pmatrix} -2.77\ldots \\ 7.19\ldots \\ -0.51\ldots \\ 1.90\ldots \end{pmatrix}$$

Should you trust this library?

Question 2

Constrained optimization is ubiquitous in quantum physics and chemistry (e.g. Hartree-Fock, DFT, etc.). In Physics and Chemistry textbooks, such problems are solved using the Lagrangian method.

Example: solve $\inf_{g(x)=0} E(x)$ **where** $E : \mathbb{R}^d \to \mathbb{R}$ **and** $g : \mathbb{R}^d \to \mathbb{R}^m$ **are regular.**

Introduce the Lagrangian $L : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$ **defined as**

$$L(x, \lambda) = E(x) + \lambda^T g(x).$$

Then, the minimizers are obtained by solving the system of equations

$$\begin{cases} \nabla_x L(x, \lambda) = 0 \\ \\ \nabla_\lambda L(x, \lambda) = 0, \end{cases}$$

Application: $d = 1$**,** $m = 1$**,** $E(x) = x$**,** $g(x) = x^2$

$$\begin{cases} 1 + 2\lambda x = 0 \\ x^2 = 0 \end{cases} \quad \Rightarrow$$

No solution, though $x = 0$ **is obviously a minimizer!**

What's the catch?

Question 3

Diagonalizing the translation operators $(T_{\mathbf{R}})_{\mathbf{R}\in\mathbb{Z}^3}$

$$(T_{\mathbf{R}}\psi)(\mathbf{r}) = \psi(\mathbf{r}-\mathbf{R})$$

Let $\psi \neq 0$ **be such that** $T_{\mathbf{R}}\psi = C(R)\psi$ **for all** $\mathbf{R} \in \mathbb{Z}^3$ **with** $C(R) \in \mathbb{C}$**. Since**

$$|C(\mathbf{R})|^2 \int |\psi(\mathbf{r})|^2\, d\mathbf{r} = \int |C(R)\psi(\mathbf{r})|^2\, d\mathbf{r} = \int |(T_{\mathbf{R}}\psi)(\mathbf{r})|^2\, d\mathbf{r} = \int |\psi(\mathbf{r}-\mathbf{R})|^2\, d\mathbf{r}$$
$$= \int |\psi(\mathbf{r})|^2\, d\mathbf{r}, \qquad (1)$$

then $|C(R)| = 1$ **and therefore** $C(R) = e^{i\alpha(R)}$**. Since** $T_{\mathbf{R}+\mathbf{R}'} = T_{\mathbf{R}}T_{\mathbf{R}'}$**, we get** $C(\mathbf{R}) = e^{i\mathbf{k}\cdot\mathbf{R}}$ **for some** $\mathbf{k} \in \mathbb{R}^3$**. And from there, we get**

$$\psi(\mathbf{r}) = e^{i\mathbf{k}\cdot\mathbf{r}}u(\mathbf{r}) \quad \textbf{where } u \textbf{ is an } \mathbb{Z}^3\textbf{-periodic function}$$

But then, $\int |\psi(\mathbf{r})|^2\, d\mathbf{r} = +\infty$**, and we can't infer from** (1) **that** $|C(R)| = 1$**.**

How to make this (physically correct) argument mathematically correct?

Question 4

The bound states of the hydrogen atom Hamiltonian are known

$$\hat{h} = -\frac{1}{2}\Delta - \frac{1}{r} \qquad \hat{h}\,\varphi_{n,\ell,m}(r,\theta,\phi) = E_n \varphi_{n,\ell,m}(r,\theta,\phi), \qquad \left| \begin{array}{l} n \in \mathbb{N}^* \\ 0 \leq \ell \leq n-1 \\ -\ell \leq m \leq \ell \end{array} \right.$$

When two hydrogen atoms are at distance $R \gg 1$ a.u., their interaction energy can be expanded as

$$\Delta E(R) = -\frac{C_6}{R^6} + \text{h.o.t.} \quad \textbf{(van der Waals interaction)}$$

The C_6 coefficient can be computed by perturbation theory

Using the "sum over state" technique in the basis $(\varphi_{n,\ell,m})$ we get

$C_6 \simeq 3.923$ u.a. **to be compared with the correct value** $C_6 \simeq 6.499$ u.a.

What's wrong in this approach?

Question 5

Which spin states can you actually represent with your two hands?

1. **Spectral theory of self-adjoint operators**

2. **From molecules to materials**

3. **A bit of numerical analysis**

4. **Constrained optimization and Lagrange multipliers**

1 - Spectral theory of self-adjoint operators

References:

- **E.B. Davies, *Linear operators and their spectra*, Cambridge University Press 2007.**
- **B. Helffer, *Spectral theory and its applications*, Cambridge University Press 2013.**
- **M. Reed and B. Simon, *Modern methods in mathematical physics*, in 4 volumes, 2nd edition, Academic Press 1972-1980.**
- **M. Lewin, *Théorie spectrale et mécanique quantique*, Springer 2022 (English version to appear soon).**

Notation: in this section, $\mathcal{H}$ denotes a separable complex Hilbert space, $\langle \cdot | \cdot \rangle$ its inner product, and $\| \cdot \|$ the associated norm.

Fundamental principles of quantum mechanics

1. **To each quantum system is associated a separable complex Hilbert space $\mathcal{H}$**

2. **If the state of the system at time t is completely known (pure state), it can be described by a normalized vector $\psi(t)$ of $\mathcal{H}$. The set of physically admissible pure states is the projective space $P(\mathcal{H})$.**

3. **Physical observables are represented by self-adjoint operators on $\mathcal{H}$.**

4. **Let a be a physical observable represented by the self-adjoint operator A. The outcome of a measurement of a is always in $\sigma(A)$, the spectrum of A.**

5. **If, just before the measurement, the system is in the pure state $\psi(t_0)$, then the probability that the outcome lays in the interval $I \subset \mathbb{R}$ is $\|\mathbb{1}_I(A)\psi(t_0)\|^2$, where $\mathbb{1}_I$ is the characteristic function of I and $\mathbb{1}_I(A)$ is defined by functional calculus.**

6. **If the system is isolated, its dynamics between two successive measures is given by $\psi(t) = U(t-t_0)\psi(t_0)$ where $U(\tau) = e^{-i\tau H/\hbar}$, H being the Hamiltonian, i.e. the self-adjoint operator associated with the energy.**

Definition (Hilbert space). **A Hilbert space is a real or complex vector space $\mathcal{H}$ endowed with a inner product $\langle\cdot|\cdot\rangle$ and complete for the associated norm $\|\cdot\|$.**

Definition (completeness). **A sequence $(\psi_n)_{n\in\mathbb{N}}$ of elements of a normed vector space $(\mathcal{H}, \|\cdot\|)$ is Cauchy if**

$$\forall \varepsilon > 0, \quad \exists N \in \mathbb{N} \quad \textbf{s.t.} \quad \forall q \geq p \geq N, \quad \|\psi_p - \psi_q\| \leq \varepsilon.$$

The normed vector space $(\mathcal{H}, \|\cdot\|)$ is called complete if any Cauchy sequence of elements of $\mathcal{H}$ converges in $\mathcal{H}$.

Example: all finite-dimensional normed $\mathbb{R}$- or $\mathbb{C}$-vector spaces are complete.

- **Endowed with the hermitian inner product, $\mathbb{C}^d$ is a Hilbert space:**

$$\langle \mathbf{x}|\mathbf{y}\rangle = \sum_{1\leq i\leq d} \overline{x_i}\, y_i, \qquad \|\mathbf{x}\| = \langle \mathbf{x}|\mathbf{s}\rangle^{1/2} = \left(\sum_{1\leq i\leq d} |x_i|^2\right)^{1/2}.$$

- **Let $S \in \mathbb{C}^{d\times d}$ be a positive definite hermitian matrix ($S_{ji} = \overline{S_{ij}}$ for all $1 \leq i, j \leq d$ and $\mathbf{x}^* S\mathbf{x} > 0$ for all $\mathbf{x} \in \mathbb{C}^d \setminus \{0\}$). Then $\langle \mathbf{x}|\mathbf{y}\rangle_S = \mathbf{x}^* S\mathbf{y}$ defines a inner product on $\mathbb{C}^d$ and**

$$\forall \mathbf{x} \in \mathbb{C}^d, \quad \lambda_1(S)\|\mathbf{x}\| \leq \|\mathbf{x}\|_S \leq \lambda_d(S)\|\mathbf{x}\|,$$

where $\lambda_1(S) \leq \lambda_2(S) \leq \cdots \leq \lambda_d(S)$ are the eigenvalues of S.

Fundamental examples: the Hilbert space $L^2(\mathbb{R}^d, \mathbb{C})$.

- **The sequilinear form**

$$\langle \varphi | \psi \rangle := \int_{\mathbb{R}^d} \overline{\varphi} \psi := \int_{\mathbb{R}^d} \overline{\varphi(\mathbf{r})}\, \psi(\mathbf{r})\, d\mathbf{r}$$

defines a inner product on

$$C_{\mathrm{c}}^{\infty}(\mathbb{R}^d, \mathbb{C}) := \left\{ \varphi \in C^{\infty}(\mathbb{R}^d, \mathbb{C}) \mid \ \varphi = 0 \text{ outside some bounded set} \right\},$$

but $C_{\mathrm{c}}^{\infty}(\mathbb{R}^d, \mathbb{C})$, endowed with the inner product $\langle \varphi | \psi \rangle$, is not a Hilbert space.

- **To obtain a Hilbert space, we have to "complete" it with "all the limits of the Cauchy sequences of elements of $C_{\mathrm{c}}^{\infty}(\mathbb{R}^d)$". We thus obtain the set**

$$L^2(\mathbb{R}^d, \mathbb{C}) := \left\{ \varphi : \mathbb{R}^d \to \mathbb{C} \mid \int_{\mathbb{R}^d} |\varphi|^2 < \infty \right\},$$

which, endowed with the inner product $\langle \varphi | \psi \rangle$, is a Hilbert space.

- **Technical details:**
  - **one must use the Lebesgue integral (doesn't work with Riemann integral);**
  - **the elements of $L^2(\mathbb{R}^d, \mathbb{C})$ are in fact equivalence classes of measurable functions (for the Lebesgue measure) for the equivalence relation $\varphi \sim \varphi'$ iff $\varphi = \varphi'$ everywhere except possibly on a set of zero Lebesgue measure.**

Fundamental examples: the Sobolev spaces $H^1(\mathbb{R}^d,\mathbb{C})$ and $H^2(\mathbb{R}^d,\mathbb{C})$.

- **The sets**

$$H^1(\mathbb{R}^d,\mathbb{C}) := \left\{\varphi \in L^2(\mathbb{R}^d,\mathbb{C}) \mid \nabla\varphi \in (L^2(\mathbb{R}^d,\mathbb{C}))^d\right\},$$
$$H^2(\mathbb{R}^d,\mathbb{C}) := \left\{\varphi \in L^2(\mathbb{R}^d,\mathbb{C}) \mid \nabla\varphi \in (L^2(\mathbb{R}^d,\mathbb{C}))^d \text{ and } D^2\varphi \in (L^2(\mathbb{R}^d,\mathbb{C}))^{d\times d}\right\}$$

are vector spaces. Respectively endowed with the inner products

$$\langle\varphi|\psi\rangle_{H^1} := \int_{\mathbb{R}^d} \overline{\varphi}\psi + \int_{\mathbb{R}^d} \overline{\nabla\varphi}\cdot\nabla\psi,$$
$$\langle\varphi|\psi\rangle_{H^2} := \int_{\mathbb{R}^d} \overline{\varphi}\psi + \int_{\mathbb{R}^d} \overline{\nabla\varphi}\cdot\nabla\psi + \int_{\mathbb{R}^d} \overline{D^2\varphi} : D^2\psi,$$

they are Hilbert spaces.

- **Technical detail: the gradient and the second derivatives are defined by means of distribution theory.**

Remark. Let $\varphi \in H^1(\mathbb{R}^d)$. A function $\widetilde{\varphi} \in H^1(\mathbb{R}^d)$ can be a very accurate approximation of φ in $L^2(\mathbb{R}^d)$ and a terrible approximation of φ in $H^1(\mathbb{R}^d)$.

For instance, let $\varphi(x) = \frac{1}{1+x^2}$ and $\varphi_n(x) = \left(1 + \frac{\sin(n^2x^2)}{n}\right)\varphi(x)$. The sequence $(\varphi_n)_{n\in\mathbb{N}^*}$ converges to φ in $L^2(\mathbb{R})$ and goes to infinity in $H^1(\mathbb{R})$.

Bounded linear operators on Hilbert spaces

Definition-Theorem (bounded linear operator). A bounded operator on $\mathcal{H}$ is a linear map $\hat{A} : \mathcal{H} \to \mathcal{H}$ such that

$$\|\hat{A}\| := \sup_{\Psi \in \mathcal{H} \setminus \{0\}} \frac{\|\hat{A}\Psi\|}{\|\Psi\|} < \infty.$$

The set $\mathcal{B}(\mathcal{H})$ of the bounded operators on $\mathcal{H}$ is a non-commutative algebra and $\|\cdot\|$ is a norm on $\mathcal{B}(\mathcal{H})$.

Remark. A bounded linear operator A is uniquely defined by the values of the sesquilinear form $\mathcal{H} \times \mathcal{H} \ni (\Psi_1, \Psi_2) \mapsto \langle \Psi_1 | \hat{A}\Psi_2 \rangle \in \mathbb{C}$.

Definition-Theorem (adjoint of a bounded linear operator). Let $A \in \mathcal{B}(\mathcal{H})$. The operator $\hat{A}^\dagger \in \mathcal{B}(\mathcal{H})$ defined by

$$\forall (u, v) \in \mathcal{H} \times \mathcal{H}, \quad \langle u | \hat{A}^\dagger v \rangle = \langle Au | v \rangle,$$

is called the adjoint of A. The operator A is called self-adjoint if $\hat{A}^\dagger = A$.

Endowed with its norm $\|\cdot\|$ and the $*$ operation, $\mathcal{B}(\mathcal{H})$ is a C*-algebra.

(Non necessarily bounded) linear operators on Hilbert spaces

Definition (linear operator). A linear operator on $\mathcal{H}$ is a linear map $\hat{A} : D(\hat{A}) \to \mathcal{H}$, where $D(\hat{A})$ is a subspace of $\mathcal{H}$ called the domain of $\hat{A}$. Note that bounded linear operators are particular linear operators.

Definition (extensions of operators). Let $\hat{A}_1$ and $\hat{A}_2$ be operators on $\mathcal{H}$. $\hat{A}_2$ is called an extension of $\hat{A}_1$ if $D(\hat{A}_1) \subset D(\hat{A}_2)$ and if $\forall u \in D(\hat{A}_1)$, $\hat{A}_2 u = \hat{A}_1 u$.

Definition (unbounded linear operator). An operator $\hat{A}$ on $\mathcal{H}$ which does not possess a bounded extension is called an unbounded operator on $\mathcal{H}$.

Definition (symmetric operator). A linear operator $\hat{A}$ on $\mathcal{H}$ with dense domain $D(\hat{A})$ is called symmetric if

$$\forall \Psi_1, \Psi_2 \in D(\hat{A}) \times D(\hat{A}), \quad \langle \Psi_1 | \hat{A} \Psi_2 \rangle = \langle \hat{A} \Psi_1 | \Psi_2 \rangle.$$

Symmetric operators are not very interesting. Only self-adjoint operators represent physical observables and have nice mathematical properties:

- **real spectrum;**
- **spectral decomposition and functional calculus.**

Definition (adjoint of a linear operator with dense domain). Let A be a linear operator on $\mathcal{H}$ with dense domain $D(\hat{A})$, and $D(\hat{A}^\dagger)$ the vector space defined as

$$D(\hat{A}^\dagger) = \left\{ v \in \mathcal{H} \mid \exists w_v \in \mathcal{H} \text{ s.t. } \forall u \in D(\hat{A}),\ \langle Au|v\rangle = \langle u|w_v\rangle \right\}.$$

The linear operator $\hat{A}^\dagger$ on $\mathcal{H}$, with domain $D(\hat{A}^\dagger)$, defined by

$$\forall v \in D(\hat{A}^\dagger), \quad \hat{A}^\dagger v = w_v,$$

(if w_v exists, it is unique since $D(\hat{A})$ is dense) is called the adjoint of A.

This definition agrees with the previous one for bounded operators.

Definition (self-adjoint operator). A linear operator $\hat{A}$ with dense domain is called self-adjoint if $\hat{A}^\dagger = \hat{A}$ (that is if $\hat{A}$ symmetric and $D(\hat{A}^\dagger) = D(\hat{A})$).

Case of bounded operators:

symmetric $\Leftrightarrow$ self-adjoint.

Case of unbounded operators:

symmetric (easy to check) $\begin{matrix} \Rightarrow \\ \Leftarrow \end{matrix}$ **self-adjoint (sometimes difficult to check)**

Some unbounded self-adjoint operators arising in quantum mechanics

- **position operator along the j axis:**
  - $\mathcal{H} = L^2(\mathbb{R}^d)$,
  - $D(\widehat{r}_j) = \left\{u \in L^2(\mathbb{R}^d) \mid r_j u \in L^2(\mathbb{R}^d)\right\}$, $(\widehat{r}_j\phi)(\mathbf{r}) = r_j\phi(\mathbf{r})$;
- **momentum operator along the j axis:**
  - $\mathcal{H} = L^2(\mathbb{R}^d)$,
  - $D(\widehat{p}_j) = \left\{u \in L^2(\mathbb{R}^d) \mid \partial_{r_j} u \in L^2(\mathbb{R}^d)\right\}$, $(\widehat{p}_j\phi)(\mathbf{r}) = -i\partial_{r_j}\phi(\mathbf{r})$;
- **kinetic energy operator:**
  - $\mathcal{H} = L^2(\mathbb{R}^d)$,
  - $D(T) = H^2(\mathbb{R}^d) := \left\{u \in L^2(\mathbb{R}^d) \mid \Delta u \in L^2(\mathbb{R}^d)\right\}$, $T = -\frac{1}{2}\nabla^2 = -\frac{1}{2}\Delta$;
- **Schrödinger operators in 3D: let $V \in L^2_{\text{unif}}(\mathbb{R}^3, \mathbb{R})$ ($V(\mathbf{r}) = -\frac{Z}{|\mathbf{r}|}$ OK)**
  - $\mathcal{H} = L^2(\mathbb{R}^3)$,
  - $D(H) = H^2(\mathbb{R}^3)$, $H = -\frac{1}{2}\Delta + V$.

Definition-Theorem (spectrum of a linear operator). Let A be a closed[1] linear operator on $\mathcal{H}$.

- **The open set $\rho(A) = \left\{ z \in \mathbb{C} \mid (z - A) \,:\, D(\hat{A}) \to \mathcal{H} \text{ invertible} \right\}$ is called the resolvent set of A. The analytic function**

$$\rho(A) \ni z \mapsto R_z(A) := (z - A)^{-1} \in \mathcal{B}(\mathcal{H})$$

is called the resolvent of A. It holds $R_z(A) - R_{z'}(A) = (z' - z) R_z(A) R_{z'}(A)$.

- **The closed set $\sigma(A) = \mathbb{C} \setminus \rho(A)$ is called the spectrum of A.**

- **If A is self-adjoint, then $\sigma(A) \subset \mathbb{R}$ and it holds $\sigma(A) = \sigma_{\rm p}(A) \cup \sigma_{\rm c}(A)$, where $\sigma_{\rm p}(A)$ and $\sigma_{\rm c}(A)$ are respectively the point spectrum and the continuous spectrum of A defined as**

$$\sigma_{\rm p}(A) \;=\; \left\{ z \in \mathbb{C} \mid (z - A) \,:\, D(\hat{A}) \to \mathcal{H} \text{ non-injective} \right\} = \{\text{eigenvalues of } A\}$$

$$\sigma_{\rm c}(A) \;=\; \overline{\left\{ z \in \mathbb{C} \mid (z - A) \,:\, D(\hat{A}) \to \mathcal{H} \text{ injective but non surjective} \right\}}.$$

[1] **The operator A is called closed if its graph $\Gamma(A) := \left\{ (u, Au),\, u \in D(\hat{A}) \right\}$ is a closed subspace of $\mathcal{H} \times \mathcal{H}$.**

On the physical meaning of point and continuous spectra

Theorem (RAGE, Ruelle '69, Amrein and Georgescu '73, Enss '78).

Let H be a locally compact self-adjoint operator on $L^2(\mathbb{R}^d)$ with no singular continuous spectrum. [Ex.: the Hamiltonian of the hydrogen atom.]

Let $\mathcal{H}_p = \overline{\text{Span}\,\{\text{eigenvectors of } H\}}$ and $\mathcal{H}_c = \mathcal{H}_p^{\perp}$.
[Ex.: for the Hamiltonian of the hydrogen atom, $\dim(\mathcal{H}_p) = \dim(\mathcal{H}_c) = \infty$.]

Let χ_{B_R} be the characteristic function of the ball $B_R = \{\mathbf{r} \in \mathbb{R}^d \mid |\mathbf{r}| < R\}$.

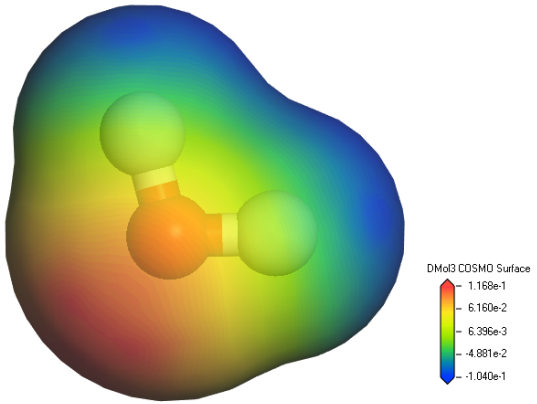
Then

$$(\phi_0 \in \mathcal{H}_p) \;\Leftrightarrow\; \forall \varepsilon > 0,\ \exists R > 0,\ \forall t \geq 0,\ \left\| (1 - \chi_{B_R}) e^{-itH} \phi_0 \right\|_{L^2}^2 \leq \varepsilon;$$

$$(\phi_0 \in \mathcal{H}_c) \;\Leftrightarrow\; \forall R > 0,\ \lim_{t \to +\infty} \left\| \chi_{B_R} e^{-itH} \phi_0 \right\|_{L^2}^2 = 0.$$

$\mathcal{H}_{\rm p}$: **subspace of localized states,** $\mathcal{H}_{\rm c}$: **subspace of scattering states**

Electronic problem for a given nuclear configuration $\{\mathbf{R}_A\}_{1\le A\le M}$



Ex: water molecule H_2O

$M = 3$, $N = 10$, $z_1 = 8$, $z_2 = 1$, $z_3 = 1$

$$v_{\mathrm{nuc}}(\mathbf{r}) = -\sum_{k=1}^{M} \frac{z_A}{|\mathbf{r} - \mathbf{R}_A|}$$

$$\left(-\frac{1}{2}\sum_{i=1}^{N} \Delta_{\mathbf{r}_i} + \sum_{i=1}^{N} v_{\mathrm{nuc}}(\mathbf{r}_i) + \sum_{1\le i<j\le N} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} \right) \Psi(\mathbf{x}_1, \cdots, \mathbf{x}_N) = E\, \Psi(\mathbf{x}_1, \cdots, \mathbf{x}_N)$$

$$\forall p \in \mathfrak{S}_N, \quad \Psi(\mathbf{x}_{p(1)}, \cdots, \mathbf{x}_{p(N)}) = \varepsilon(p)\Psi(\mathbf{x}_1, \cdots, \mathbf{x}_N), \qquad \textbf{(Pauli principle)}$$

$$\Psi \in \mathcal{H}_N = \bigwedge^N \mathcal{H}_1, \qquad \mathcal{H}_1 = L^2(\mathbb{R}^3 \times \{\uparrow, \downarrow\}; \mathbb{C})$$

Theorem (Kato '51). The operator $\hat{H}_N := -\frac{1}{2}\sum_{i=1}^{N} \Delta_{\mathbf{r}_i} + \sum_{i=1}^{N} v_{\mathrm{ext}}(\mathbf{r}_i) + \sum_{1\le i<j\le N} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|}$

with domain $D(\hat{H}_N) := \mathcal{H}_N \cap H^2((\mathbb{R}^3 \times \{\uparrow, \downarrow\})^N; \mathbb{C})$ **is self-adjoint on** $\mathcal{H}_N$**.**

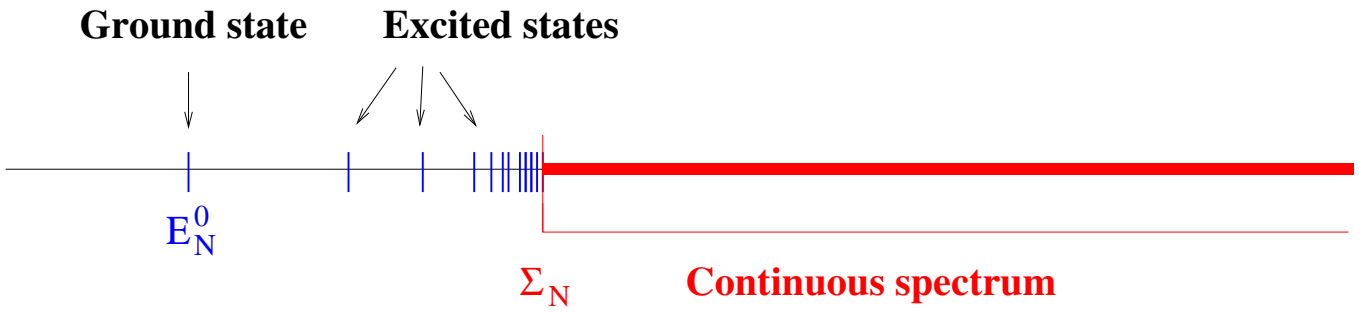
Theorem (spectrum of $\hat{H}_N$).

1. HVZ theorem (Hunziger '66, van Winten '60, Zhislin '60)

$\sigma_{\rm c}(\hat{H}_N) = [\Sigma_N, +\infty)$ **with** $\Sigma_N = \min \sigma(\hat{H}_{N-1}) \leq 0$ **and** $\Sigma_N < 0$ **iff** $N \geq 2$.

2. Bound states of neutral molecules and positive ions (Zhislin '61)

If $N \leq Z := \sum_{A=1}^{M} z_A$**, then** $\hat{H}_N$ **has an infinite number of bound states.**



3. Bound states of negative ions (Yafaev '72)

If $N \geq Z + 1$**, then** $\hat{H}_N$ **has at most a finite number of bound states.**

Spectra of Schrödinger operators with confining potentials

$$\mathcal{H} = L^2(\mathbb{R}^d), \qquad V \in C^0(\mathbb{R}^d), \qquad \lim_{|\mathbf{r}|\to+\infty} V(\mathbf{r}) = +\infty \text{ (confining potential)}$$

$$D(\hat{H}) = \left\{ \varphi \in L^2(\mathbb{R}^d) \mid -\frac{1}{2}\Delta\varphi + V\varphi \in L^2(\mathbb{R}^d) \right\}, \quad \forall \varphi \in D(\hat{H}),\ \hat{H}\varphi = -\frac{1}{2}\Delta\varphi + V\varphi.$$

$\hat{H}$ is bounded below and its spectrum is purely discrete ($\sigma_{\mathrm{d}}(\hat{H}) = \sigma(\hat{H})$, $\sigma_{\mathrm{ess}}(\hat{H}) = \emptyset$).

As a consequence, H is diagonalizable in a orthonormal basis: there exist

- **a non-decreasing sequence $(E_n)_{n\in\mathbb{N}}$ of real numbers going to $+\infty$;**
- **an orthonormal basis $(\psi_n)_{n\in\mathbb{N}}$ of $\mathcal{H}$ composed of vectors of $D(H)$,**

such that

$$\forall n \in \mathbb{N}, \quad \hat{H}\psi_n = E_n\psi_n.$$

In addition, the ground state eigenvalue E_0 is non-degenerate and the corresponding eigenvector can be chosen positive on $\mathbb{R}^d$.

Spectra of 3D Schrödinger operators with potentials decaying at infinity

V **such that** $\forall \varepsilon > 0,\ \exists (V_2, V_\infty) \in L^2(\mathbb{R}^3) \times L^\infty(\mathbb{R}^3)$ **s.t.** $V = V_2 + V_\infty$ **and** $\|V_\infty\|_{L^\infty} \leq \varepsilon$,

$$\mathcal{H} = L^2(\mathbb{R}^3), \qquad D(\hat{h}) = H^2(\mathbb{R}^3), \qquad \forall \varphi \in D(\hat{h}),\ \hat{h}\varphi = -\frac{1}{2}\Delta\varphi + V\varphi.$$

The operator H is self-adjoint, bounded below, and $\sigma_{\rm c}(\hat{h}) = [0, +\infty)$.

Depending on V, the discrete spectrum of $\hat{h}$ may be

- **the empty set;**
- **a finite number of negative eigenvalues;**
- **a countable infinite number of negative eigenvalues accumulating at 0 (ex: Ridberg states).**

If $\hat{h}$ has a ground state, then its energy is a non-degenerate eigenvalue and the corresponding eigenvector can be chosen positive on $\mathbb{R}^d$.

The special case of Kohn-Sham LDA Hamiltonians

$$\hat{h}_\rho^{\mathrm{KS}} = -\frac{1}{2}\Delta + V_\rho^{\mathrm{KS}} \quad \textbf{with} \quad V_\rho^{\mathrm{KS}}(\mathbf{r}) = -\sum_{A=1}^{M} \frac{z_A}{|\mathbf{r}-\mathbf{R}_A|} + \int_{\mathbb{R}^3} \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\, d\mathbf{r}' + \frac{de_{\mathrm{xc}}^{\mathrm{LDA}}}{d\rho}(\rho(\mathbf{r}))$$

For any $\rho \in L^1(\mathbb{R}^3) \cap L^3(\mathbb{R}^3)$, the KS potential V_ρ^{KS} satisfies the assumptions of the previous slide. In particular H_ρ is bounded below and $\sigma_{\mathrm{c}}(\hat{h}_\rho) = [0, +\infty)$.

Let $Z = \sum_{A=1}^{M} z_A$ be the total nuclear charge of the molecular system and $N = \int_{\mathbb{R}^3} \rho$.

- **If $N < Z$ (positive ion), $\hat{h}_\rho^{\mathrm{KS}}$ has a countable infinite number of negative eigenvalues accumulating at 0.**
- **If $N = Z$ (neutral molecular system) and if ρ_{GS} is a ground state density of the system, then $\hat{h}_{\rho_{\mathrm{GS}}}^{\mathrm{KS}}$ has at least N non-positive eigenvalues.**

Spectra of (restricted) Hartree-Fock Hamiltonians

Let $\Phi = (\phi_1, \cdots, \phi_N) \in (H^1(\mathbb{R}^3))^N$ **be such that** $\displaystyle\int_{\mathbb{R}^3} \phi_i \phi_j = \delta_{ij}$**,**

$$\gamma(\mathbf{r}, \mathbf{r}') = \sum_{i=1}^{N} \phi_i(\mathbf{r}) \phi_i(\mathbf{r}'), \qquad \rho_\gamma(\mathbf{r}) = 2\gamma(\mathbf{r}, \mathbf{r}) = 2 \sum_{i=1}^{N} |\phi_i(\mathbf{r})|^2.$$

$$\mathcal{H} = L^2(\mathbb{R}^3), \quad D(H) = H^2(\mathbb{R}^3),$$

$$(\hat{h}_\gamma^{\mathrm{HF}} \phi)(\mathbf{r}) = -\frac{1}{2}\Delta\phi(\mathbf{r}) - \sum_{A=1}^{M} \frac{z_A}{|\mathbf{r} - \mathbf{R}_A|}\phi(\mathbf{r}) + \left(\int_{\mathbb{R}^3} \frac{\rho_\gamma(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}\, d\mathbf{r}' \right) \phi(\mathbf{r}) - \int_{\mathbb{R}^3} \frac{\gamma(\mathbf{r}, \mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}\, \phi(\mathbf{r}')\, d\mathbf{r}'$$

Let $Z := \sum_{A=1}^{M} z_A$**. The operator** $\hat{h}_\gamma^{\mathrm{HF}}$ **is self-adjoint, bounded below, and we have:**

- $\sigma_{\mathrm{c}} = [0, +\infty)$**;**
- **if** $N < Z$ **(positive ion),** $\hat{h}_\gamma^{\mathrm{HF}}$ **has a countable infinite number of negative eigenvalues accumulating at** 0**;**
- **if** $N = Z$ **(neutral molecular system) and if** Φ_{GS} **is a HF ground state, then** $\hat{h}_{\gamma_{\mathrm{GS}}}^{\mathrm{HF}}$ **has at least** N **negative eigenvalues (counting multiplicities).**

Spectra of Dirac Hamiltonians

$$\mathcal{H} = L^2(\mathbb{R}^3;\mathbb{C}^4), \qquad D(\hat{D}_0) = H^1(\mathbb{R}^3;\mathbb{C}^4), \qquad \hat{D}_0 = c\hat{\vec{p}}\cdot\vec{\alpha} + mc^2\beta$$

$$\hat{p}_j = -\mathrm{i}\hbar\partial_j, \qquad \alpha_j = \begin{pmatrix} 0 & \sigma_k \\ \sigma_k & 0 \end{pmatrix} \in \mathbb{C}^{4\times 4}, \qquad \beta = \begin{pmatrix} I_2 & 0 \\ 0 & -I_2 \end{pmatrix} \in \mathbb{C}^{4\times 4}$$

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad \sigma_2 = \begin{pmatrix} 0 & -\mathrm{i} \\ \mathrm{i} & 0 \end{pmatrix}, \qquad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \qquad \textbf{(Pauli matrices)}$$

The free Dirac operator $\hat{D}_0$ is self-adjoint and

$$\sigma(\hat{D}_0) = \sigma_{\mathrm{c}}(\hat{D}_0) = (-\infty, -mc^2] \cup [mc^2, +\infty).$$

Theorem. **Let** $\alpha := \frac{e^2}{4\pi\varepsilon_0\hbar c} \simeq 1/137.036$ **be the fine structure constant. Let**

$$\hat{D}_Z = \hat{D}_0 - \frac{Z}{|\mathbf{r}|}, \qquad Z \in \mathbb{R} \quad \textbf{(physical cases: } Z = 1, 2, 3, \cdots \textbf{)}.$$

- **if** $|Z| < \frac{\sqrt{3}}{2\alpha} \simeq 118.677$**, the Dirac operator** $\hat{D}_Z$ **is essentially self-adjoint (meaning that there exists a unique domain** $D(\hat{D}_Z)$ **containing** $C_c^\infty(\mathbb{R}^3; \mathbb{C}^4)$ **for which** $\hat{D}_Z$ **is self-adjoint);**
- **if** $|Z| > \frac{\sqrt{3}}{2\alpha} \simeq 118.677$**,** $\hat{D}_Z$ **has many self-adjoint extensions;**
- **if** $|Z| < \frac{1}{\alpha} \simeq 137.036$**,** $\hat{D}_Z$ **has a special self-adjoint extension, considered as the physical one. The essential spectrum of this self-adjoint extension is** $(-\infty, -mc^2] \cup [mc^2, +\infty)$ **and its discrete spectrum consist of the eigenvalues**

$$E_{nj} = mc^2 \left[1 + \left(\frac{Z\alpha}{n - j - \frac{1}{2} + \sqrt{(j+\frac{1}{2})^2 - Z^2\alpha^2}} \right)^2 \right]^{-1/2}, \quad n \in \mathbb{N}^*, \ j = \frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \cdots \leq n - \frac{1}{2}.$$

Many-body Dirac-Coulomb Hamiltonian are not understood mathematically.

Functional calculus for diagonalizable self-adjoint operators

Let $\hat{A}$ be a self-adjoint operator that can be diagonalized in an orthonormal basis $(\varphi_n)_{n\in\mathbb{N}}$ (this is not the case for many useful self-adjoint operators!).

Dirac's bra-ket notation: $\hat{A} = \sum_{n\in\mathbb{N}} \lambda_n |\varphi_n\rangle\langle\varphi_n|, \quad \lambda_n \in \mathbb{R}, \quad \langle\varphi_m|\varphi_n\rangle = \delta_{mn}.$

Then,

- **the operator $\hat{A}$ is bounded if and only if $\|\hat{A}\| = \sup_n |\lambda_n| < \infty$;**
- $D(\hat{A}) = \left\{|\psi\rangle = \sum_{n\in\mathbb{N}} c_n|\varphi_n\rangle \mid \sum_{n\in\mathbb{N}}(1+|\lambda_n|^2)|c_n|^2 < \infty\right\}$**;**
- $\sigma_{\mathrm{p}}(\hat{A}) = \{\lambda_n\}_{n\in\mathbb{N}}$ **and** $\sigma_{\mathrm{c}}(\hat{A}) =$ **accumulation points of** $\{\lambda_n\}_{n\in\mathbb{N}} \backslash \sigma_{\mathrm{p}}(\hat{A})$**;**
- $\mathcal{H}_{\mathrm{p}} = \mathcal{H}$ **and** $\mathcal{H}_{\mathrm{c}} = \{0\}$ **(if $\hat{A}$ is a Hamiltonian: no scattering states!);**
- **functional calculus for diagonalizable self-adjoint operators: for all $f : \mathbb{R} \to \mathbb{C}$, the operator $f(\hat{A})$ defined by**

$$D(f(\hat{A})) = \left\{ |\psi\rangle = \sum_{n\in\mathbb{N}} c_n|\varphi_n\rangle \mid \sum_{n\in\mathbb{N}}(1+|f(\lambda_n)|^2)|c_n|^2 < \infty \right\}, \quad f(\hat{A}) = \sum_{n\in\mathbb{N}} f(\lambda_n)|\varphi_n\rangle\langle\varphi_n|$$

is independent of the choice of the spectral decomposition of $\hat{A}$.

Theorem (functional calculus for bounded functions). Let $\mathfrak{B}(\mathbb{R},\mathbb{C})$ be the $*$-algebra of bounded $\mathbb{C}$-valued Borel functions on $\mathbb{R}$ and let $\hat{A}$ be any self-adjoint operator on $\mathcal{H}$. Then there exists a unique map

$$\Phi_A \,:\, \mathfrak{B}(\mathbb{R},\mathbb{C}) \ni f \mapsto f(\hat{A}) \in \mathcal{B}(\mathcal{H})$$

satisfies the following properties:

1. Φ_A is a homomorphism of $*$-algebras:

$$(\alpha f+\beta g)(\hat{A}) = \alpha f(\hat{A})+\beta g(\hat{A}), \quad (fg)(\hat{A}) = f(\hat{A})g(\hat{A}), \quad f^*(\hat{A}) = f(\hat{A})^\dagger;$$

2. $\|f(\hat{A})\| \leq \sup_{x\in\mathbb{R}} |f(x)|$**;**

3. if $f_n(x) \to x$ pointwise and $|f_n(x)| \leq |x|$ for all n and all $x \in \mathbb{R}$, then

$$\forall \psi \in D(\hat{A}), \quad f_n(\hat{A})\psi \to \hat{A}\psi \text{ in } \mathcal{H};$$

4. if $f_n(x) \to f(x)$ pointwise and $\sup_n \sup_{x\in\mathbb{R}} |f_n(x)| < \infty$, then

$$\forall \psi \in \mathcal{H}, \quad f_n(\hat{A})\psi \to f(\hat{A})\psi \text{ in } \mathcal{H};$$

In addition, if $\psi \in \mathcal{H}$ is such that $\hat{A}\psi = \lambda\psi$, then $f(\hat{A})\psi = f(\lambda)\psi$.

Theorem (spectral projections and functional calculus - general case -).

Let $\hat{A}$ be a self-adjoint operator on $\mathcal{H}$.

- **For all $\lambda \in \mathbb{R}$, the bounded operator $\hat{P}^A_\lambda := \mathbb{1}_{(-\infty,\lambda]}(\hat{A})$, where $\mathbb{1}_{(-\infty,\lambda]}(\cdot)$ is the characteristic function of $(-\infty, \lambda]$, is an orthogonal projection.**

- **Spectral decomposition of $\hat{A}$: for all $\psi \in D(\hat{A})$ and $\psi' \in \mathcal{H}$, it holds**

$$\langle \psi' | \hat{A}\psi \rangle = \int_{\mathbb{R}} \lambda \underbrace{d\langle \psi' | \hat{P}^A_\lambda \psi \rangle}_{\text{Bounded complex measure on } \mathbb{R}}, \quad \textbf{which we denote by} \quad \hat{A} = \int_{\mathbb{R}} \lambda \, d\hat{P}^A_\lambda.$$

- **Functional calculus: let f be a (not necessarily bounded) $\mathbb{C}$-valued Borel function on $\mathbb{R}$. The operator $f(A)$ can be defined by**

$$D(f(\hat{A})) := \left\{ \psi \in \mathcal{H} \mid \int_{\mathbb{R}} |f(\lambda)|^2 \underbrace{d\langle \psi | P^A_\lambda \psi \rangle}_{\text{Bounded positive measure on } \mathbb{R}} < \infty \right\}$$

and

$$\forall (\psi, \psi') \in D(f(\hat{A})) \times \mathcal{H}, \ \langle \psi' | f(\hat{A})\psi \rangle := \int_{\mathbb{R}} f(\lambda) \, d\langle \psi' | \hat{P}^A_\lambda \psi \rangle.$$

Application of spectral theory and functional calculus: one-body density matrices

1-RDM associated with an N-body wavefunction Ψ_N

$$\gamma_{\Psi_N}(\mathbf{x},\mathbf{x}') := \langle \psi_N | \hat{\Psi}^\dagger(\mathbf{r}) \hat{\Psi}(\mathbf{r}') | \Psi_N \rangle$$
$$= N \int_{(\mathbb{R}^3\times\{\uparrow,\downarrow\})^{N-1}} \Psi_N(\mathbf{x},\mathbf{x}_2,\cdots,\mathbf{x}_N)\Psi_N(\mathbf{x}',\mathbf{x}_2,\cdots,\mathbf{x}_N)^* \, d\mathbf{x}_2\cdots d\mathbf{x}_N$$

It is extremely fruitful to consider $\gamma_{\Psi_N}(\mathbf{x},\mathbf{x}')$ as the integral kernel of an operator $\hat{\gamma}_{\Psi_N}$ on $\mathcal{H}_1$ (also called 1-RDM or DM for short)

$$\forall \varphi \in \mathcal{H}_1, \quad (\hat{\gamma}_{\Psi_N}\varphi)(\mathbf{x}) = \int_{\mathbb{R}^3\times\{\uparrow,\downarrow\}} \gamma_{\Psi_N}(\mathbf{x},\mathbf{x}')\,\varphi(\mathbf{x}')\,d\mathbf{x}'$$

The operator $\hat{\gamma}_{\Psi_N}$ is self-adjoint, diagonalizable, $\sigma(\hat{\gamma}_{\Psi_N}) \subset [0,1]$, and $\mathrm{Tr}(\hat{\gamma}_{\Psi_N}) = N$

$$\hat{\gamma}_{\Psi_N} = \sum_{j=1}^{+\infty} n_j |\varphi_j\rangle\langle\varphi_j|, \quad \langle\varphi_j|\varphi_{j'}\rangle = \delta_{jj'}, \quad 0 \le n_j \le 1, \quad \sum_{j=1}^{+\infty} n_j = N$$

The φ_j's are called the natural orbitals (associated with Ψ_N), and the n_j's the natural occupation numbers

Application of spectral theory and functional calculus: one-body density matrices

When Ψ_N is the Slater determinant of orthonormal orbitals $(\varphi_1, \cdots, \varphi_N)$, then $\hat{\gamma}_{\Psi_N}$ is the orthogonal projector on span$(\varphi_1, \cdots, \varphi_N)$:

$$\hat{\gamma}_{\Psi_N} = \sum_{j=1}^{N} |\varphi_j\rangle\langle\varphi_j|, \qquad \hat{\gamma}_{\Psi_N}^2 = \hat{\gamma}_{\Psi_N} = \hat{\gamma}_{\Psi_N}^\dagger$$

Application of spectral theory and functional calculus: one-body density matrices

Consider a system of non-interacting "electrons" with one-body Hamiltonian $\hat{h}$

Assume that $\hat{h}$ has at least N eigenvalues $\varepsilon_1 \leq \varepsilon_2 \leq \cdots \leq \varepsilon_N$ (counting multiplicities) and that $\varepsilon_N < \varepsilon_{N+1}$ (energy gap). Then

- **NVE ground-state density matrix is**

$$\hat{\gamma}_{\mathrm{NVE}} = \mathbb{1}_{(-\infty,\mu_{\mathrm{F}}]}(\hat{h})$$

  where μ_{F} is any number in the range $[\varepsilon_N, \varepsilon_{N+1})$ (Fermi level)

Assume that $\hat{h}$ is diagonalizable: $\hat{h} = \sum_{j=1}^{+\infty} \varepsilon_j |\varphi_j\rangle\langle\varphi_j|, \quad \langle\varphi_j|\varphi_{j'}\rangle = \delta_{jj'}$

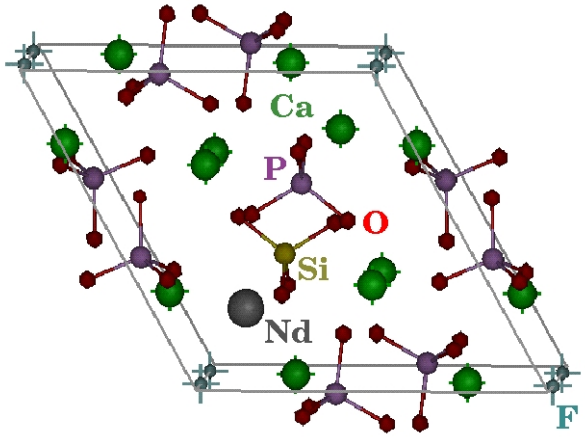
- **NVT (canonical) ground-state density matrix:**

$$\hat{\gamma}_{\mathrm{NVT}} = f_\beta(\hat{h} - \mu), \quad \mu \text{ such that } \mathrm{Tr}(\hat{\gamma}_{\mathrm{NVT}}) = N, \quad f_\beta(\varepsilon) = \frac{1}{1 + e^{\beta\varepsilon}}$$

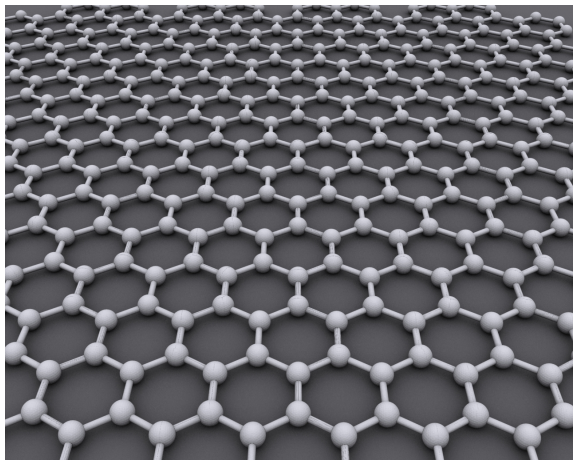
- **μVT (grand-canonical) ground-state density matrix:**

$$\hat{\gamma}_{\mu\mathrm{VT}} = f_\beta(\hat{h} - \mu)$$

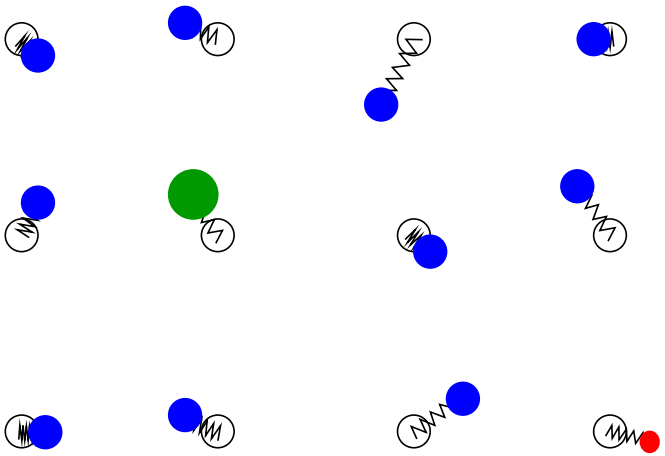
4 - From molecules to materials



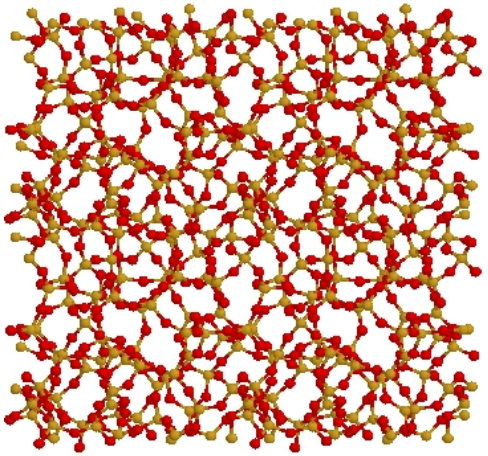
Periodic 3D system



Periodic 2D system



Alloy at finite temperature



Amorphous system

Preliminary remarks

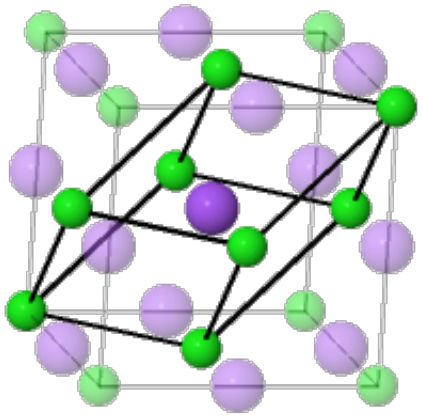
- **At the atomic scale, a material looks like an infinity system ($10^{23} \sim \infty$)**
- **There is no such thing as the wavefunction of a system with infinite number of electrons**
- **The way out is to only use n-particle density matrices and/or Green's functions, typically with $n = 1$ or $n = 1, 2$ and pass to the thermodynamic limit**

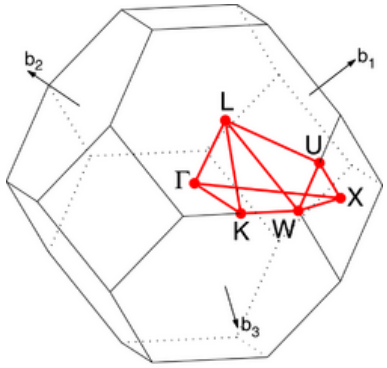
Bravais lattice $\mathbb{L}$, unit cell Ω, reciprocal lattice $\mathbb{L}^*$, and Brillouin zone $\mathcal{B}$

- **FCC 3D crystal (ex: aluminium, copper, gold...)**

$$\mathbb{L} = \mathbb{Z}a_1 + \mathbb{Z}a_2 + \mathbb{Z}a_3$$
$$\Omega = [0,1)a_1 + [0,1)a_2 + [0,1)a_3$$
$$\mathbb{L}^* = \mathbb{Z}b_1 + \mathbb{Z}b_2 + \mathbb{Z}b_3$$

$\mathcal{B}$: **truncated octahedron**





FCC path: Γ-X-W-K-Γ-L-U-W-L-K|U-X

[Setyawan & Curtarolo, DOI: 10.1016/j.commatsci.2010.05.010]

Bravais lattice $\mathbb{L}$, unit cell Ω, reciprocal lattice $\mathbb{L}^*$, and Brillouin zone $\mathcal{B}$

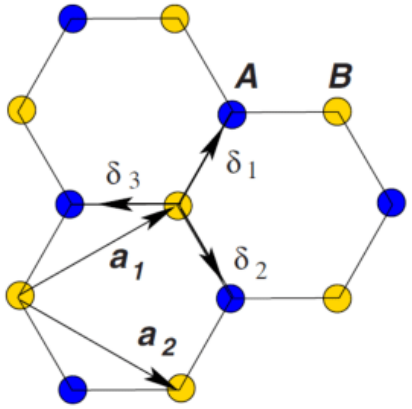
- **hexagonal 2D materials (e.g. graphene, hBN...)**

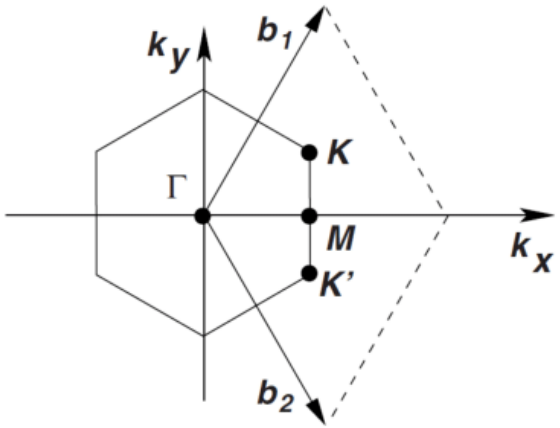
$\mathbb{L} = \mathbb{Z}\mathbf{a}_1 + \mathbb{Z}\mathbf{a}_2$

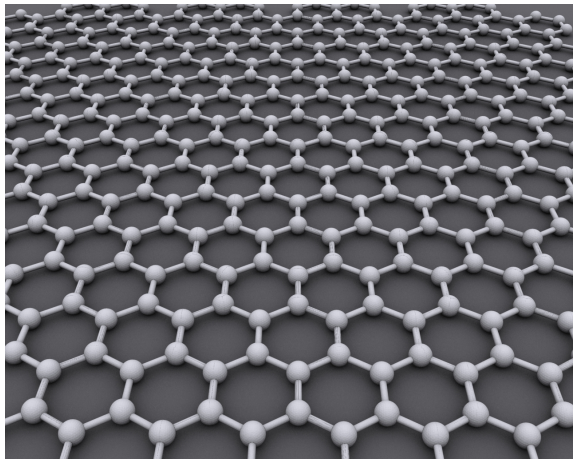
Ω: a cylinder

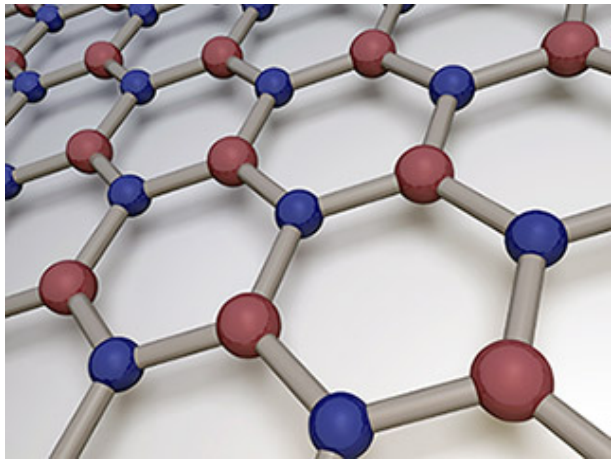
$\mathbb{L}^* = \mathbb{Z}\mathbf{b}_1 + \mathbb{Z}\mathbf{b}_2$

$\mathcal{B}$: an hexagon

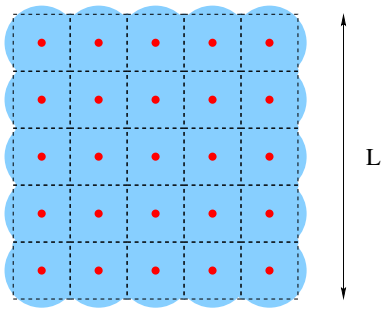








Thermodynamic limit (bulk limit) for perfect crystals ($\mathbb{L} = \mathbb{Z}^3$, simple cubic)



$$\begin{cases} \rho_L^{\mathrm{nuc}} = \displaystyle\sum_{\mathbf{R}\in\mathbb{Z}^3\cap(-L/2,L/2]^3} z\, m(\cdot - \mathbf{R}) \\ \\ zL^3 \text{ electrons} \end{cases} \longrightarrow \left| \begin{array}{ll} E_L^0 & \textbf{ground state total energy} \\ \\ \rho_L^0 & \textbf{(unique) ground state density} \\ \\ \gamma_L^0 & \textbf{a ground state density matrix} \end{array} \right.$$

Theorem (Catto-Le Bris-Lions, '01). **For the Hartree model (KS with no xc)**

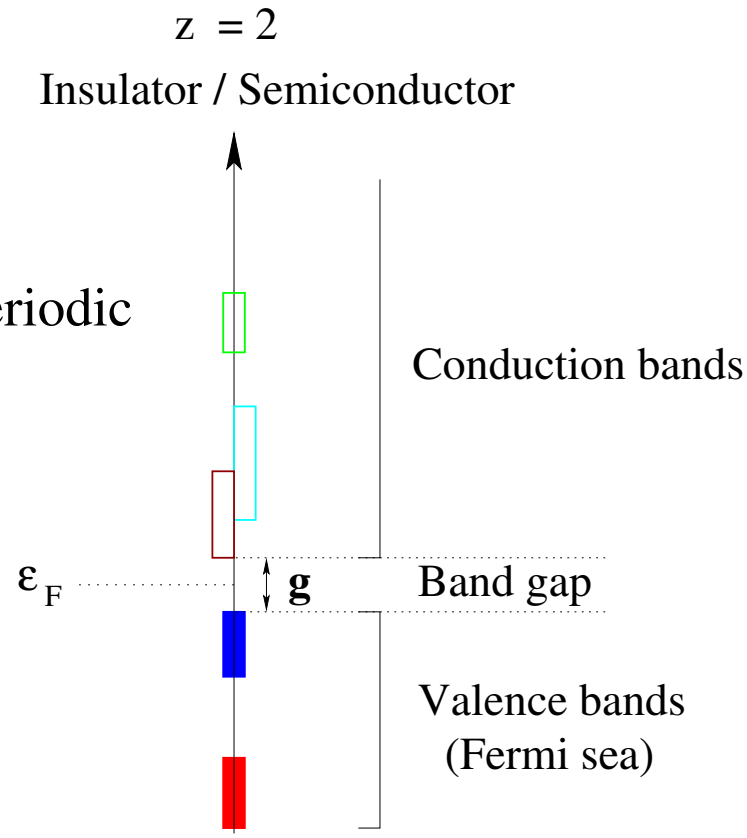
$$\lim_{L\to\infty} \frac{E_L^0}{L^3} = E_{\mathrm{per}}^0, \qquad \rho_L^0 \overset{\text{in some sense}}{\underset{L\to\infty}{\longrightarrow}} \rho_{\mathrm{per}}^0, \qquad \gamma_L^0 \overset{\text{in some sense}}{\underset{L\to\infty}{\longrightarrow}} \gamma_{\mathrm{per}}^0.$$

Periodic (spin-unpolarized) Kohn-Sham equations

$$
\begin{cases}
\hat{h}^0_{\mathrm{per}} = -\dfrac{1}{2}\Delta + \hat{V}^{\mathrm{Hartree}}_{\mathrm{per}} + \hat{V}^{\mathrm{xc}}_{\mathrm{per}} \quad \textbf{on } L^2(\mathbb{R}^3;\mathbb{C}) \\
-\Delta V^{\mathrm{Hartree}}_{\mathrm{per}}(\mathbf{r}) = 4\pi\left(\rho^{\mathrm{nuc}}_{\mathrm{per}}(\mathbf{r}) - \rho^0_{\mathrm{per}}(\mathbf{r})\right), \quad V^0_{\mathrm{per}}\ \mathbb{L}\text{-periodic} \\
\rho^0_{\mathrm{per}}(\mathbf{r}) = 2\gamma^0_{\mathrm{per}}(\mathbf{r},\mathbf{r}) \\
V^{\mathrm{xc}}_{\mathrm{per}}(\mathbf{r}) = \dfrac{de_{\mathrm{xc}}}{d\rho}(\rho^0_{\mathrm{per}}(\mathbf{r})) \quad \textbf{(LDA)} \\
\hat{\gamma}^0_{\mathrm{per}} = \mathbb{1}_{(-\infty,\varepsilon_{\mathrm{F}})}(H^0_{\mathrm{per}}), \qquad \displaystyle\int_\Omega \rho^0_{\mathrm{per}} = \int_\Omega \rho^{\mathrm{nuc}}_{\mathrm{per}}
\end{cases}
$$

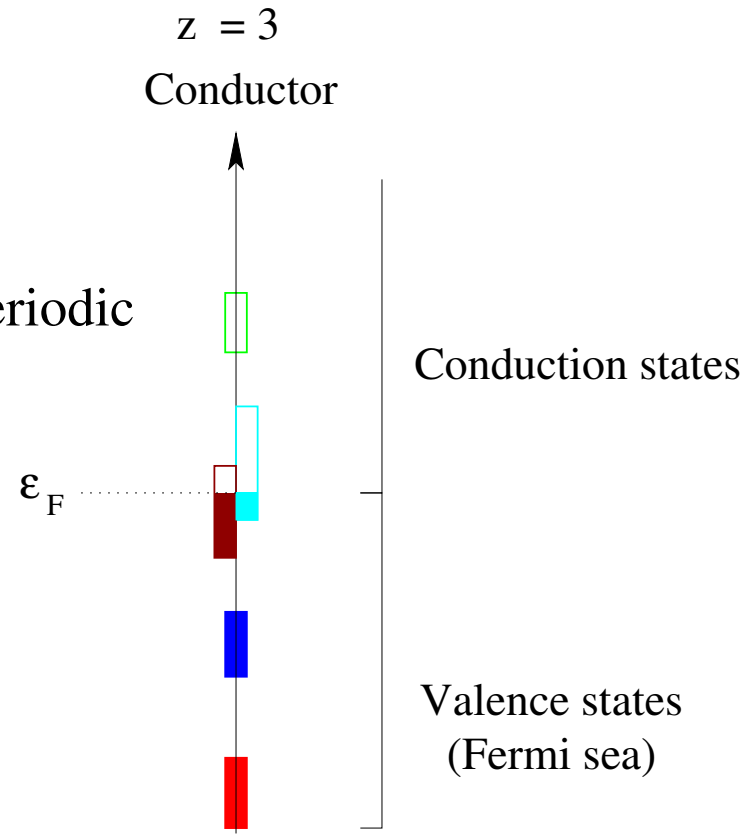
Periodic (spin-unpolarized) Kohn-Sham equations

$$
\begin{cases}
\hat{h}^0_{\mathrm{per}} = -\frac{1}{2}\Delta + \hat{V}^{\mathrm{Hartree}}_{\mathrm{per}} + \hat{V}^{\mathrm{xc}}_{\mathrm{per}} \quad \textbf{on } L^2(\mathbb{R}^3;\mathbb{C}) \\
-\Delta V^{\mathrm{Hartree}}_{\mathrm{per}}(\mathbf{r}) = 4\pi\left(\rho^{\mathrm{nuc}}_{\mathrm{per}}(\mathbf{r}) - \rho^0_{\mathrm{per}}(\mathbf{r})\right), \quad V^0_{\mathrm{per}}\ \mathbb{L}\text{-periodic} \\
\rho^0_{\mathrm{per}}(\mathbf{r}) = 2\gamma^0_{\mathrm{per}}(\mathbf{r},\mathbf{r}) \\
V^{\mathrm{xc}}_{\mathrm{per}}(\mathbf{r}) = \dfrac{de_{\mathrm{xc}}}{d\rho}(\rho^0_{\mathrm{per}}(\mathbf{r})) \quad \textbf{(LDA)} \\
\hat{\gamma}^0_{\mathrm{per}} = \mathbb{1}_{(-\infty,\varepsilon_{\mathrm{F}})}(H^0_{\mathrm{per}}), \qquad \displaystyle\int_\Omega \rho^0_{\mathrm{per}} = \int_\Omega \rho^{\mathrm{nuc}}_{\mathrm{per}}
\end{cases}
$$



Periodic (spin-unpolarized) Kohn-Sham equations

$$
\begin{cases}
\hat{h}^0_{\mathrm{per}} = -\frac{1}{2}\Delta + \hat{V}^{\mathrm{Hartree}}_{\mathrm{per}} + \hat{V}^{\mathrm{xc}}_{\mathrm{per}} \quad \textbf{on } L^2(\mathbb{R}^3;\mathbb{C}) \\
-\Delta V^{\mathrm{Hartree}}_{\mathrm{per}}(\mathbf{r}) = 4\pi\left(\rho^{\mathrm{nuc}}_{\mathrm{per}}(\mathbf{r}) - \rho^0_{\mathrm{per}}(\mathbf{r})\right), \quad V^0_{\mathrm{per}}\ \mathbb{L}\text{-periodic} \\
\rho^0_{\mathrm{per}}(\mathbf{r}) = 2\gamma^0_{\mathrm{per}}(\mathbf{r},\mathbf{r}) \\
V^{\mathrm{xc}}_{\mathrm{per}}(\mathbf{r}) = \frac{de_{\mathrm{xc}}}{d\rho}(\rho^0_{\mathrm{per}}(\mathbf{r})) \quad \textbf{(LDA)} \\
\hat{\gamma}^0_{\mathrm{per}} = \mathbb{1}_{(-\infty,\varepsilon_{\mathrm{F}})}(H^0_{\mathrm{per}}), \qquad \int_\Omega \rho^0_{\mathrm{per}} = \int_\Omega \rho^{\mathrm{nuc}}_{\mathrm{per}}
\end{cases}
$$

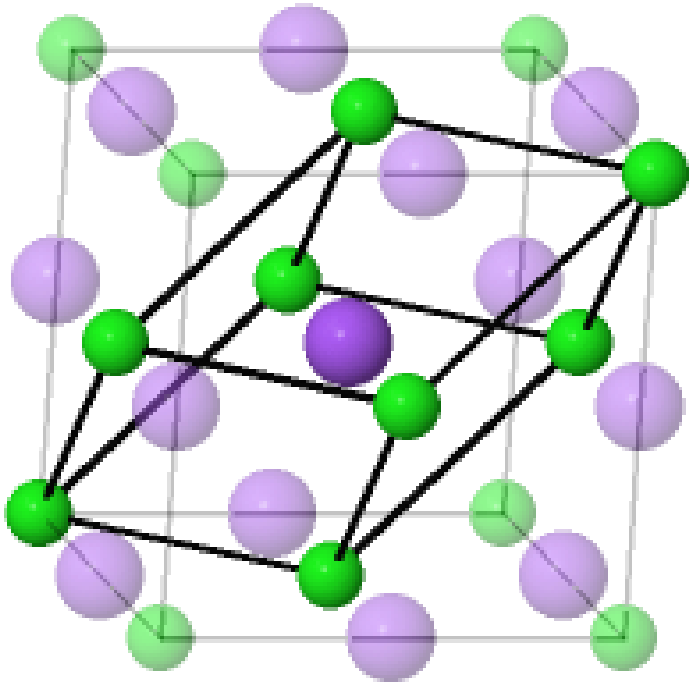


Bloch decomposition of periodic one-body Schrödinger operator $\hat{h} = -\frac{1}{2}\Delta + V$

$$\underbrace{\left(\frac{1}{2}(-i\nabla + \mathbf{k})^2 + V\right)}_{\hat{h}_{\mathbf{k}}} u_{n,\mathbf{k}} = \varepsilon_{n,\mathbf{k}} u_{n,\mathbf{k}}, \qquad (u_{n,\mathbf{k}})_{n\in\mathbb{N}^*} \text{ orthonormal basis of } L^2_{\mathrm{per}}(\Omega;\mathbb{C})$$

$\varepsilon_{1,\mathbf{k}} \leq \varepsilon_{2,\mathbf{k}} \leq \cdots$

$\mathbf{k} \mapsto \varepsilon_{n,\mathbf{k}}$ *-periodic from $\mathbb{R}^d$ to $\mathbb{R}$

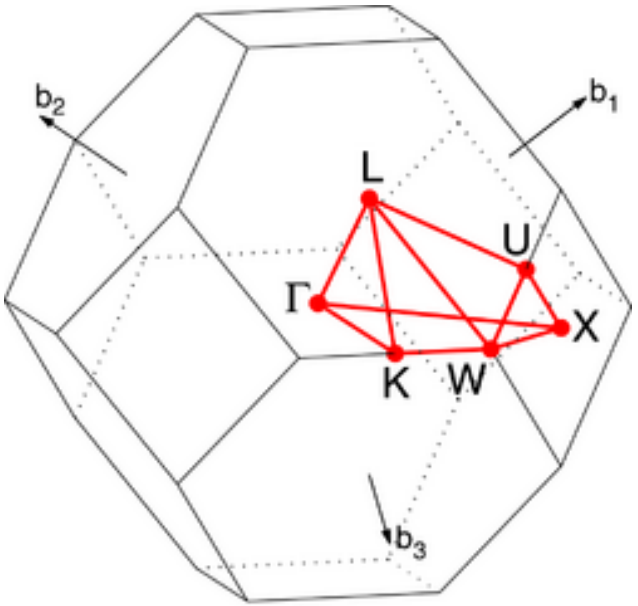


Bloch decomposition of periodic one-body Schrödinger operator $\hat{h} = -\frac{1}{2}\Delta + V$

$$\underbrace{\left(\frac{1}{2}(-i\nabla + \mathbf{k})^2 + V\right)}_{\hat{h}_{\mathbf{k}}} u_{n,\mathbf{k}} = \varepsilon_{n,\mathbf{k}} u_{n,\mathbf{k}}, \qquad (u_{n,\mathbf{k}})_{n\in\mathbb{N}^*} \text{ orthonormal basis of } L^2_{\mathrm{per}}(\Omega;\mathbb{C})$$

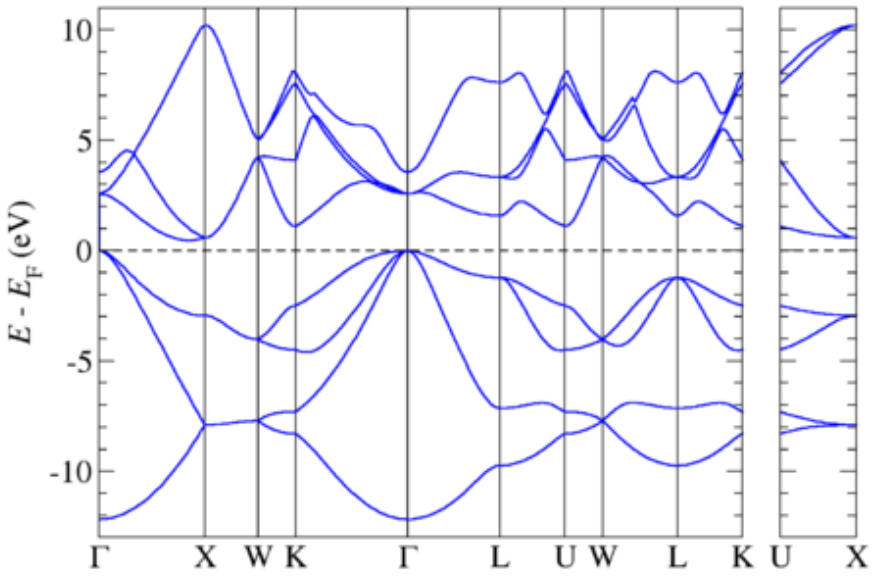
$\varepsilon_{1,\mathbf{k}} \leq \varepsilon_{2,\mathbf{k}} \leq \cdots$

$\mathbf{k} \mapsto \varepsilon_{n,\mathbf{k}}$ *-periodic from $\mathbb{R}^d$ to $\mathbb{R}$

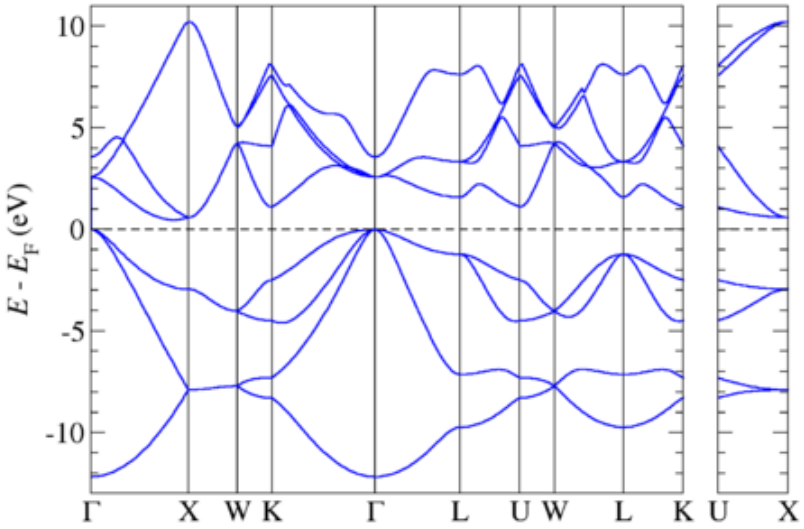


FCC path: Γ-X-W-K-Γ-L-U-W-L-K|U-X

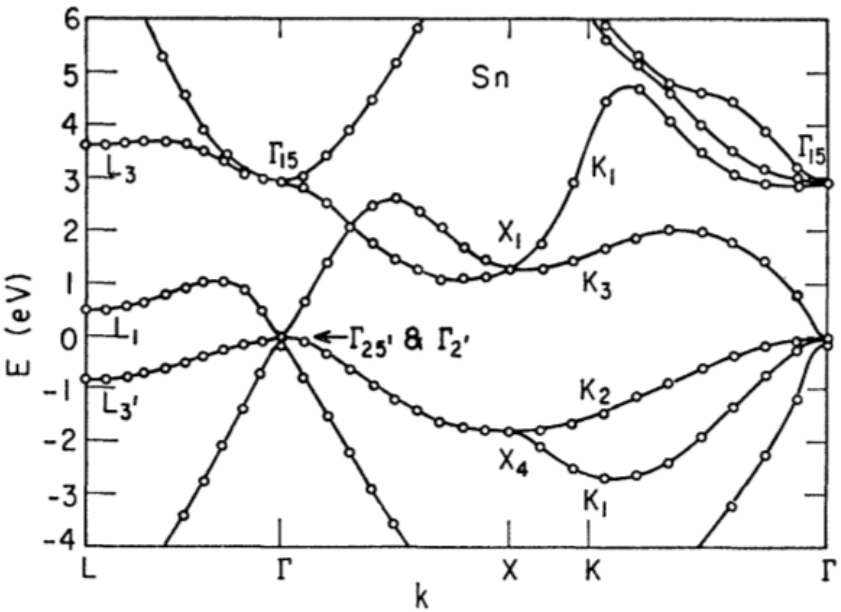
[Setyawan & Curtarolo, DOI: 10.1016/j.commatsci.2010.05.010]



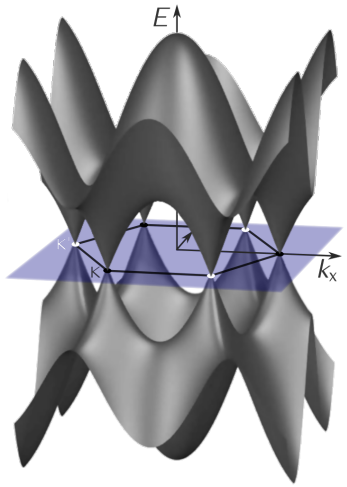
Silicon (semiconductor)

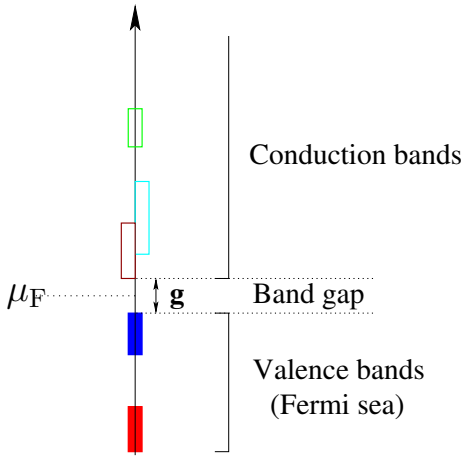


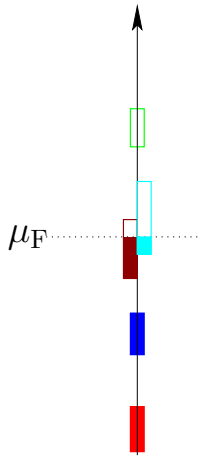
Tin (metal)



Graphene (semimetal)





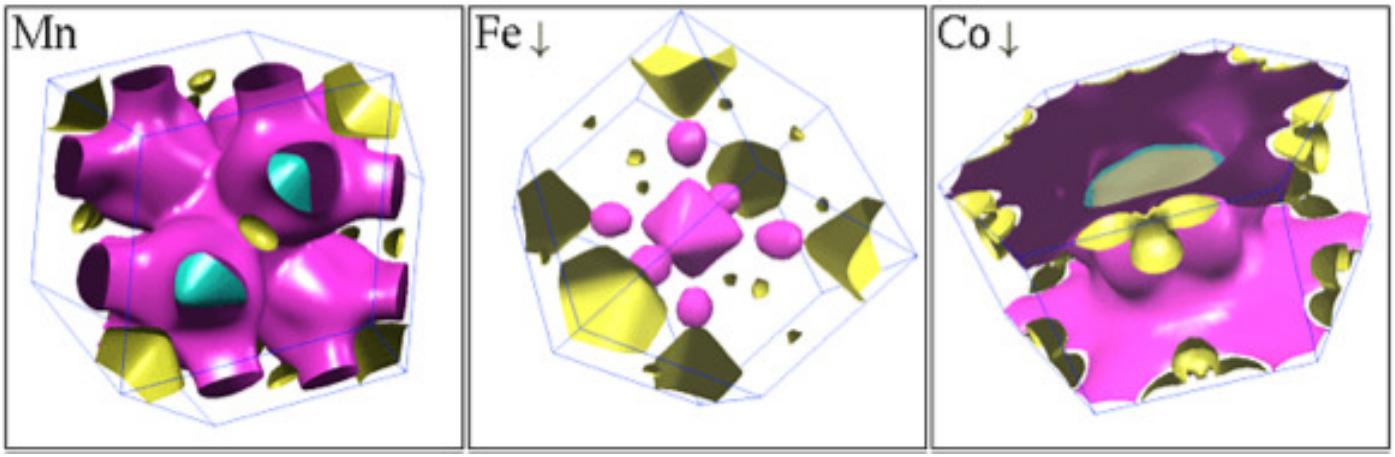


Materials classification in the independent-particle framework

- **Fermi surface and Fermi surface sheets**

$$\mathcal{S} := \{\mathbf{k} \in \mathcal{B} \mid \exists n \in \mathbb{N}^* \textbf{ s.t. } \varepsilon_{n,\mathbf{k}} = \mu_{\mathrm{F}}\} = \bigcup_{n \in \mathbb{N}^*} \mathcal{S}_n$$

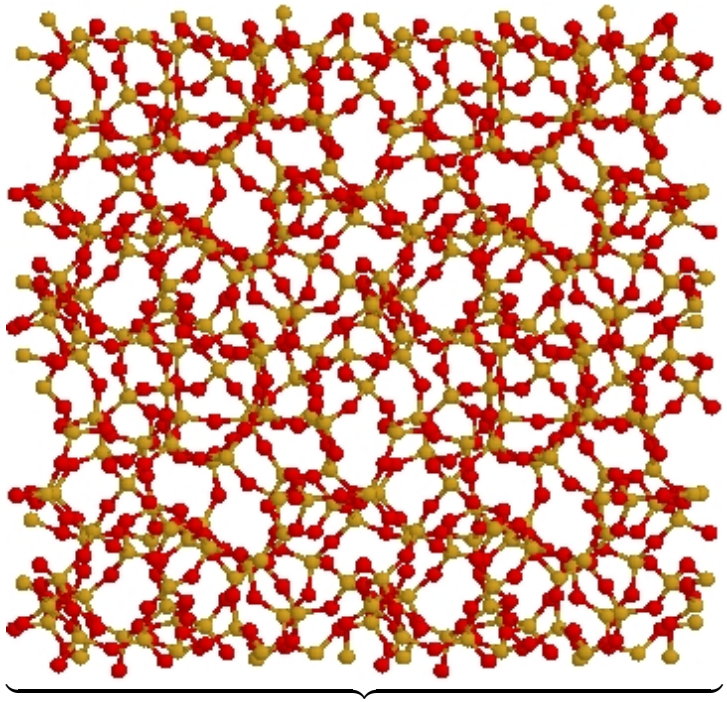
$$\mathcal{S}_n := \{\mathbf{k} \in \mathcal{B} \mid \varepsilon_{n,\mathbf{k}} = \mu_{\mathrm{F}}\}, \quad n \in \mathbb{N}^*$$



The Fermi surface database (`http://www.phys.ufl.edu/fermisurface/`)

- **Insulators/semiconductors:** $\mathcal{S} = \emptyset$
- **Non-degenerate metals:** $\mathcal{S} \neq \emptyset$, $\mathcal{S}_n \cap \mathcal{S}_{n+1} = \emptyset$, $\nabla_{\mathbf{k}} \varepsilon_{n,\mathbf{k}} \neq 0$ **on** $\mathcal{S}_n$
- **Semimetals:** $\mathcal{S} = \{$**a finite number of Dirac points**$\}$

Supercell method for Kohn-Sham simulations in the condensed phase



Size L

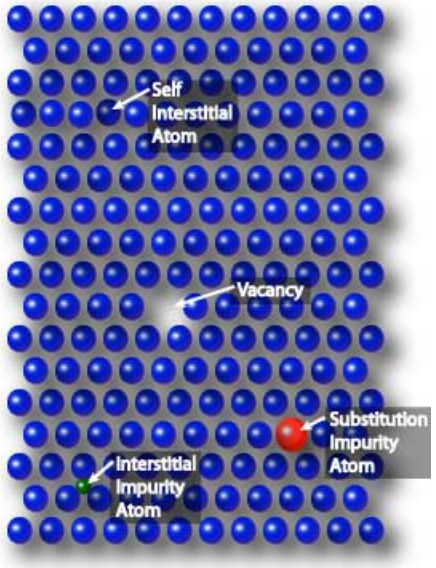
$$
\begin{cases}
\hat{h}^0_{L,\mathrm{per}} = -\frac{1}{2}\Delta + V^0_{L,\mathrm{per}} + V^{\mathrm{xc}}_{L,\mathrm{per}} \quad \textbf{on } L^2_{\mathrm{per}}\left([-\frac{L}{2},\frac{L}{2})^3\right) \\
-\Delta V^0_{L,\mathrm{per}} = 4\pi\left(\rho^{L,\mathrm{nuc}}_{\mathrm{per}} - \rho^0_{L,\mathrm{per}}\right), \quad V^0_{L,\mathrm{per}}\ L\mathbb{Z}^3\text{-periodic} \\
\rho^0_{L,\mathrm{per}}(\mathbf{r}) = 2\gamma^0_{L,\mathrm{per}}(\mathbf{r},\mathbf{r}) \\
\hat{\gamma}^0_{L,\mathrm{per}} = \mathbb{1}_{(-\infty,\varepsilon_{\mathrm{F}})}(\hat{h}^0_{L,\mathrm{per}}), \quad \int_{[-\frac{L}{2},\frac{L}{2})^3} \rho^0_{L,\mathrm{per}} = \int_{[-\frac{L}{2},\frac{L}{2})^3} \rho^{\mathrm{nuc}}_{\mathrm{per}}
\end{cases}
$$

For infinite, macroscopically homogeneous, systems:

supercell method $\sim$ representative volume method (RVP) of stochastic homogenization

Converges when $L \to \infty$ for the Hartree model for perfect crystals ($\Leftrightarrow$ uniform Brillouin zone discretization) and crystals with a single defect.

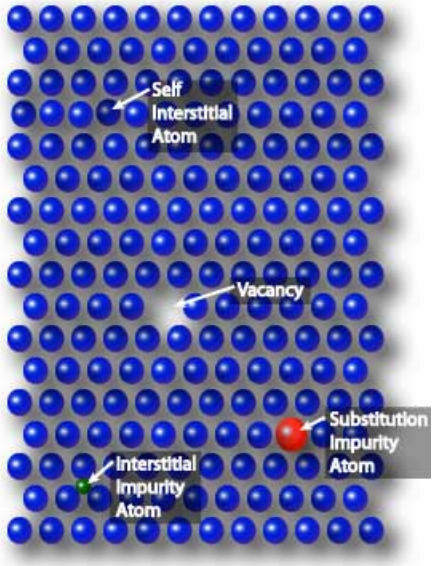
Thermodynamic limit for crystals with defects



Crystals are like people,
it is their defects
that make them interesting

(attributed to F. C. Franck)

Thermodynamic limit for crystals with defects



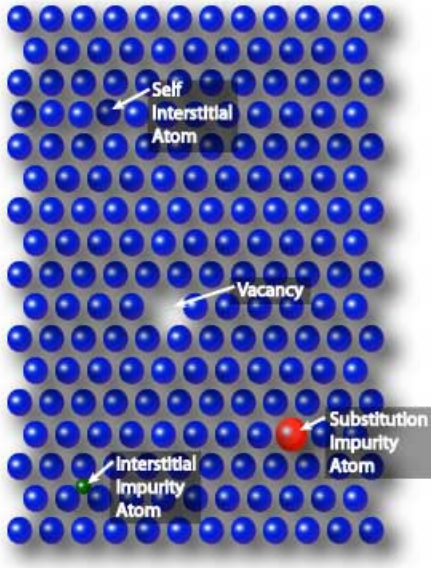
Crystals are like people,
it is their defects
that make them interesting

(attributed to F. C. Franck)

DFT models for a single defect (or a finite number of defects)

- **TF: Lieb-Simon ('77), TFW: Catto-Le Bris-Lions ('98)**
- **Hartree: EC, Deleurence, Lewin ('08), EC, Lewin ('10), Franck, Lewin, Lieb, Seiringer ('11), EC, Stoltz ('12), Gontier-Lahbabi ('16)**
- **LDA: EC, Deleurence, Lewin ('08)**

Thermodynamic limit for crystals with defects



Crystals are like people,
it is their defects
that make them interesting

(attributed to F. C. Franck)

DFT models for stationary random distributions of defects

- **TFW: Blanc, Le Bris, Lions ’07**
- **Hartree (short-range interaction only): EC, Lahbabi, Lewin, ’13**

3 - A bit of numerical analysis

The deterministic models used in quantum physics and chemistry give rise to

- **linear eigenvalue problems (N-body Schrödinger eq., LR-TDDFT, BSE, ...)**
- **constrained optimization problems (HF, DFT, MCSCF, ...)**
- **algebraic equations (CC, ...)**
- **time-dependent linear or nonlinear Schrödinger equations (RT-TDDFT, ...)**

Solving numerically all these problems eventually boils down to (cleverly!) performing numerical quadratures and matrix-vector products.

Example: let $F : \mathbb{R}^d \to \mathbb{R}^d$. A standard iterative algorithm to solve the equation $F(\mathrm{x}) = 0$ is the Newton algorithm:

x_k begin given, solve the linear system $F'(\mathrm{x}_k)\,\mathrm{y}_k = -F(\mathrm{x}_k)$, then set $\mathrm{x}_{k+1} = \mathrm{x}_k + \mathrm{y}_k$.

Linear systems can themselves be solved by iterative algorithms based on matrix-vector products.

A key concept: conditioning

Consider a problem consisting of computing an output s from an input y (the data). The problem is called

- **well-conditioned if a small variation of the input leads to a small variation of the output**
- **ill-conditioned otherwise.**

A key concept: conditioning

Consider a problem consisting of computing an output s from an input y (the data). The problem is called

- **well-conditioned if a small variation of the input leads to a small variation of the output**
- **ill-conditioned otherwise.**

Toy example of a very ill-conditioned problem:

$$\mathbf{y} = \begin{pmatrix} 2 & 10^{17} \\ 0 & 0.5 \end{pmatrix} \quad \longrightarrow \quad \mathbf{s} = \textbf{eigenvalues of } \mathbf{y} = (0.5; 2)$$

$$\mathbf{y}+\delta\mathbf{y} = \begin{pmatrix} 2 & 10^{17} \\ 10^{-17} & 0.5 \end{pmatrix} \quad \longrightarrow \quad \mathbf{s}+\delta\mathbf{s} = \textbf{eigenvalues of } \mathbf{y}+\delta\mathbf{y} = (0; 2.5)\,.$$

An apparently nicer problem: solve the linear system $\mathbf{A}\mathrm{x} = \mathrm{b}$ with

$$\mathbf{A} = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \quad \textbf{and} \quad \mathrm{b} = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix}$$

The matrix $\mathbf{A}$ is symmetric, $\det(\mathbf{A}) = 1$, and

$$\mathbf{A}^{-1} = \begin{pmatrix} 25 & -41 & 10 & -6 \\ -41 & 68 & -17 & 10 \\ 10 & -17 & 5 & -3 \\ -6 & 10 & -3 & 2 \end{pmatrix}$$

Reference linear system

$$\begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \begin{pmatrix} \\ \\ \\ \\ \end{pmatrix} = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix} \qquad \textbf{Solution} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

Slight perturbation of the right-hand side

$$\begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \begin{pmatrix} \\ \\ \\ \\ \end{pmatrix} = \begin{pmatrix} 32.001 \\ 22.999 \\ 33.001 \\ 30.999 \end{pmatrix} \qquad \textbf{Solution} = \begin{pmatrix} 1.082 \\ 0.862 \\ 1.035 \\ 0.979 \end{pmatrix}$$

Slight modification of the matrix A

$$\begin{pmatrix} 10 & 7.021 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \begin{pmatrix} \\ \\ \\ \\ \end{pmatrix} = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix} \qquad \textbf{Solution} = \begin{pmatrix} -2.77\ldots \\ 7.19\ldots \\ -0.51\ldots \\ 1.90\ldots \end{pmatrix}$$

This apparently nice problem is not so well-conditioned ...

l^p-norm of a vector $\mathbf{x} \in \mathbb{R}^n$

$$\|\mathbf{x}\|_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \quad \textbf{for } 1 \leq p < +\infty, \qquad \|\mathbf{x}\|_\infty = \max_{1 \leq i \leq n} |x_i|$$

l^p-norm of a vector $\mathbf{x} \in \mathbb{R}^n$

$$\|\mathbf{x}\|_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \quad \textbf{for } 1 \leq p < +\infty, \qquad \|\mathbf{x}\|_\infty = \max_{1 \leq i \leq n} |x_i|$$

l^p-norm of a matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$

$$\|\mathbf{A}\|_p := \sup_{\mathbf{x} \in \mathbb{R}^m \setminus \{0\}} \frac{\|\mathbf{A}\mathbf{x}\|_p}{\|\mathbf{x}\|_p}$$

l^p-norm of a vector $\mathbf{x} \in \mathbb{R}^n$

$$\|\mathbf{x}\|_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \quad \textbf{for } 1 \leq p < +\infty, \qquad \|\mathbf{x}\|_\infty = \max_{1 \leq i \leq n} |x_i|$$

l^p-norm of a matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$

$$\|\mathbf{A}\|_p := \sup_{\mathbf{x} \in \mathbb{R}^m \setminus \{0\}} \frac{\|\mathbf{A}\mathbf{x}\|_p}{\|\mathbf{x}\|_p}$$

Condition number: the condition number of the abstract problem $\mathbf{s} = f(\mathbf{y})$ at $\mathbf{y} = \mathbf{y}_0$ for the l^p-norm is ($\mathbf{s} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^m$) is

$$\boxed{\kappa_p(\mathbf{y}_0) = \frac{\|f'(\mathbf{y}_0)\|_p \, \|\mathbf{y}_0\|_p}{\|f(\mathbf{y}_0)\|_p}.}$$

l^p-norm of a vector $\mathbf{x} \in \mathbb{R}^n$

$$\|\mathbf{x}\|_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \quad \textbf{for } 1 \leq p < +\infty, \qquad \|\mathbf{x}\|_\infty = \max_{1 \leq i \leq n} |x_i|$$

l^p-norm of a matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$

$$\|\mathbf{A}\|_p := \sup_{\mathbf{x} \in \mathbb{R}^m \setminus \{0\}} \frac{\|\mathbf{A}\mathbf{x}\|_p}{\|\mathbf{x}\|_p}$$

Condition number: the condition number of the abstract problem $\mathbf{s} = f(\mathbf{y})$ at $\mathbf{y} = \mathbf{y}_0$ for the l^p-norm is ($\mathbf{s} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^m$) is

$$\boxed{\kappa_p(\mathbf{y}_0) = \frac{\|f'(\mathbf{y}_0)\|_p \, \|\mathbf{y}_0\|_p}{\|f(\mathbf{y}_0)\|_p}.}$$

Rule of thumb: if the condition number is $\sim 10^p$ and if you compute in double precision ($\varepsilon_{\text{machine}} = 10^{-16}$), you can only trust the first $16 - p$ digits of your result.

Condition number of an invertible square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ (for the l^p-norm)

$$\kappa_p(\mathbf{A}) := \|\mathbf{A}\|_p \, \|\mathbf{A}^{-1}\|_p$$

$\kappa_p(\mathbf{A})$ is the max. w.r.t. $\mathbf{x}$ of the condition numbers of the problems:

- **matrix-vector product:** $\mathbf{y} = (\mathbf{A}, \mathbf{x}) \mapsto \mathbf{s} = \mathbf{A}\mathbf{x}$
- **linear system solver:** $\mathbf{y} = (\mathbf{A}, \mathbf{x}) \mapsto \mathbf{s} = \mathbf{A}^{-1}\mathbf{x}$ **(solve $\mathbf{A}\mathbf{s} = \mathbf{x}$)**

Example:

$$\mathbf{A} = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \quad \longrightarrow \quad \kappa_2(\mathbf{A}) = 2984 \quad \textbf{and} \quad \kappa_\infty(\mathbf{A}) = 4488.$$

Theorem. **Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be an invertible matrix, and $\mathrm{b} \in \mathbb{R}^n$, $\mathrm{b} \neq 0$.**

- **Perturbation of the right-hand side**

$$\mathbf{A}\mathbf{x} = \mathrm{b}, \qquad \mathbf{A}\left(\mathbf{x} + \delta\mathbf{x}\right) = \mathrm{b} + \delta\mathrm{b} \qquad \Rightarrow \qquad \frac{\|\delta\mathbf{x}\|_p}{\|\mathbf{x}\|_p} \leq \kappa_p(\mathbf{A}) \frac{\|\delta\mathrm{b}\|_p}{\|\mathrm{b}\|_p}$$

and the inequality is optimal: $\mathbf{A}$ being given, there exists b and $\delta\mathrm{b}$ such that the inequality is an equality.

- **Perturbation of the matrix**

$$\mathbf{A}\mathbf{x} = \mathrm{b}, \qquad \left(\mathbf{A} + \delta\mathbf{A}\right)\left(\mathbf{x} + \delta\mathbf{x}'\right) = \mathrm{b} \qquad \Rightarrow \qquad \frac{\|\delta\mathbf{x}\|_p}{\|\mathbf{x} + \delta\mathbf{x}'\|_p} \leq \kappa_p(\mathbf{A}) \frac{\|\delta\mathbf{A}\|_p}{\|\mathbf{A}\|_p}$$

and the inequality is optimal: $\mathbf{A}$ being given, there exists b and $\delta\mathbf{A}$ such that the inequality is an equality.

Properties of the condition number $\kappa_p(A)$

- $\kappa_p(\mathbf{A}) \geq 1, \forall \mathbf{A} \in \mathrm{GL}_n(\mathbb{R})$ **(the set of invertible matrices)**
- $\kappa_2(\mathbf{U}) = 1$ **iff** $\mathbf{U}$ **is orthogonal** ($\mathbf{U}\mathbf{U}^T = \mathbf{U}^T\mathbf{U} = I_n$)
- $1/\kappa_p(\mathbf{A})$ **is a measure of the relative distance of the matrix** $\mathbf{A}$ **to the set of singular matrices:**

$$\frac{1}{\kappa_p(\mathbf{A})} = \min_{\mathbf{E} \mid (\mathbf{A}+\mathbf{E}) \notin \mathrm{GL}_n(\mathbb{R})} \frac{\|\mathbf{E}\|_p}{\|\mathbf{A}\|_p}.$$

- **If** $\mathbf{A}$ **is symmetric**

$$\kappa_2(\mathbf{A}) = \frac{\max_i |\lambda_i(\mathbf{A})|}{\min_i |\lambda_i(\mathbf{A})|}$$

$\lambda_1(A) \leq \lambda_2(A) \leq \cdots \leq \lambda_n(A)$ **denoting the eigenvalues of** $\mathbf{A}$**.**

An iterative algorithm for solving a problem P is a method for constructing, from an initial guess x_0, a sequence x_1, x_2, x_3, ... such that (hopefully)

$$\mathbf{x}_k \underset{k\to+\infty}{\longrightarrow} \mathbf{x}, \tag{2}$$

where x is a solution to the problem P (the solution if P is well-posed) .

An iterative algorithm for solving a problem P is a method for constructing, from an initial guess x_0, a sequence x_1, x_2, x_3, ... such that (hopefully)

$$\mathbf{x}_k \underset{k\to+\infty}{\longrightarrow} \mathbf{x}, \tag{1}$$

where x is a solution to the problem P (the solution if P is well-posed) .

The algorithm is called convergent if (1) holds. In practice, the algorithm is stopped when some stopping criteria are met. The efficiency of the algorithm heavily relies on the choice of the stopping criteria.

An iterative algorithm for solving a problem P is a method for constructing, from an initial guess x_0, a sequence x_1, x_2, x_3, ... such that (hopefully)

$$\mathbf{x}_k \underset{k\to+\infty}{\longrightarrow} \mathbf{x}, \tag{1}$$

where x is a solution to the problem P (the solution if P is well-posed) .

The algorithm is called convergent if (1) holds. In practice, the algorithm is stopped when some stopping criteria are met. The efficiency of the algorithm heavily relies on the choice of the stopping criteria.

Examples of stopping test for linear systems $\mathrm{Ax} = \mathrm{b}$:

- **a terrible one: maximum number of iterations $(k \geq k_{\max}) \Rightarrow$ STOP**
- **a good one: residual based error vector $(\|\mathrm{r}_k\|_2 \leq \varepsilon_k) \Rightarrow$ STOP, where**

  $\mathrm{r}_k = \mathrm{b}-\mathrm{A}\mathrm{x}_k = \mathrm{A}(\mathrm{x}-\mathrm{x}_k), \quad \varepsilon_k = \varepsilon_{\mathrm{tol}}(\|\mathrm{A}\|_1\|\mathrm{x}_k\|_\infty+\|\mathrm{b}\|_2)$ **(Oetli-Prager, 1963)**

  If A is symmetric, positive definite, then $\|\mathrm{r}_k\|_2 = \|\mathrm{x} - \mathrm{x}_k\|$ where $\|\cdot\|$ is the norm defined by $\|\mathrm{y}\| = \|\mathrm{A}\mathrm{y}\|_2$.

Reminder: **gradient of a differentiable function** $J : \mathbb{R}^d \to \mathbb{R}$

We have for all $\mathbf{x} \in \mathbb{R}^d$

$$\forall \mathbf{h} \in \mathbb{R}^d, \quad J(\mathbf{x}+\mathbf{h}) = J(\mathbf{x}) + \sum_{i=1}^{d} \frac{\partial J}{\partial x_i}(\mathbf{x})\, h_i + o(\mathbf{h}) = J(\mathbf{x}) + \nabla J(\mathbf{x}) \cdot \mathbf{h} + o(\mathbf{h})$$

Euclidean inner product

Euclidean gradient: $$\nabla J(\mathbf{x}) = \begin{pmatrix} \frac{\partial J}{\partial x_1}(\mathbf{x}) \\ \cdot \\ \cdot \\ \cdot \\ \frac{\partial J}{\partial x_d}(\mathbf{x}) \end{pmatrix}.$$

Reminder: gradient of a differentiable function $J : \mathbb{R}^d \to \mathbb{R}$

We have for all $\mathbf{x} \in \mathbb{R}^d$

$$\forall \mathbf{h} \in \mathbb{R}^d, \quad J(\mathbf{x}+\mathbf{h}) = J(\mathbf{x}) + \sum_{i=1}^{d} \frac{\partial J}{\partial x_i}(\mathbf{x})\, h_i + o(\mathbf{h}) = J(\mathbf{x}) + \nabla J(\mathbf{x}) \cdot \mathbf{h} + o(\mathbf{h})$$

Euclidean inner product

Euclidean gradient:
$$\nabla J(\mathbf{x}) = \begin{pmatrix} \frac{\partial J}{\partial x_1}(\mathbf{x}) \\ \cdot \\ \cdot \\ \cdot \\ \frac{\partial J}{\partial x_d}(\mathbf{x}) \end{pmatrix}.$$

If $\mathbb{R}^d$ is endowed with the inner product $(\mathbf{x}, \mathbf{y})_S := \mathbf{x}^T S \mathbf{y}$, where $S \in \mathbb{R}^{d \times d}$ is a positive definite symmetric matrix, then the gradient of J, which we will denote by $\nabla_S J(\mathbf{x})$, is related to the Euclidean gradient $\nabla J(\mathbf{x})$ by

$$\nabla_S J(\mathbf{x}) = S^{-1} \nabla J(\mathbf{x}).$$

Geometrical interpretation of the gradient

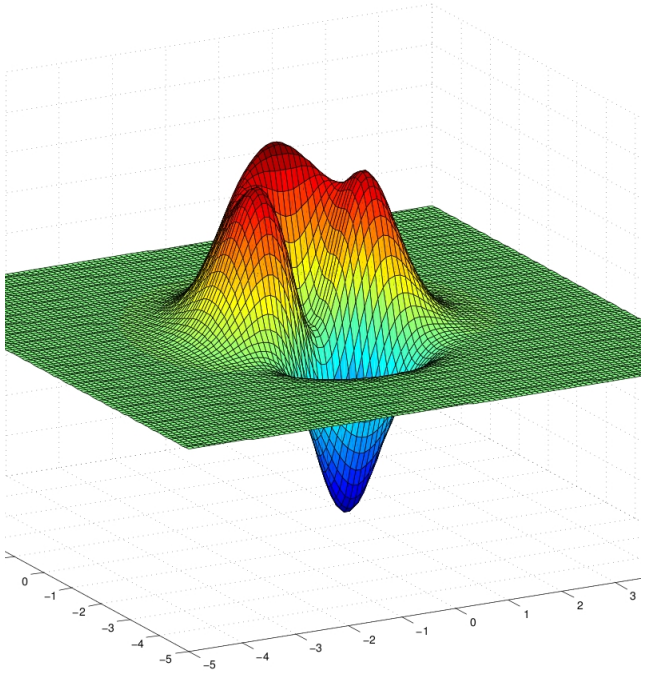
Let $J : \mathbb{R}^d \to \mathbb{R}$ of class C^1, $\mathbf{x}_0 \in \mathbb{R}^d$ and $\alpha = J(\mathbf{x}_0)$. If $\nabla J(\mathbf{x}_0) \neq 0$, then

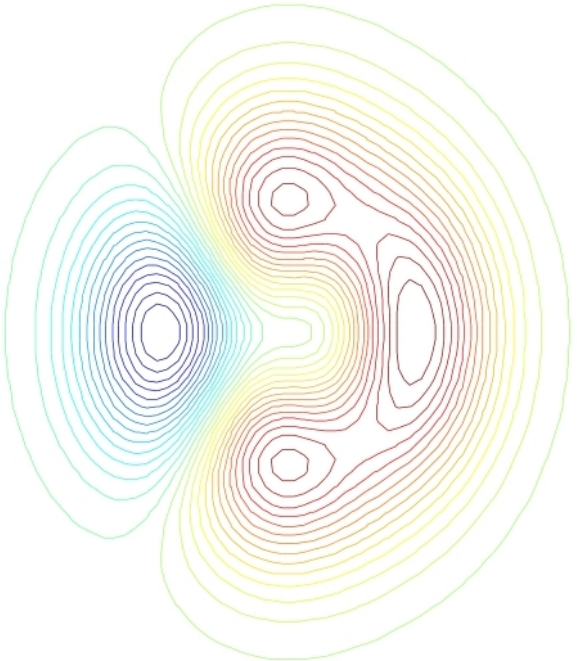
- **in the vicinity of $\mathbf{x}_0$, the level set**

$$\mathcal{C}_\alpha := \left\{ \mathbf{x} \in \mathbb{R}^d \mid \ J(\mathbf{x}) = \alpha \right\}$$

  is a C^1 hypersurface (a codimension 1 C^1 manifold);

- **the vector $\nabla J(\mathbf{x}_0)$ is orthogonal to the affine hyperplane tangent to $\mathcal{C}_\alpha$ at $\mathbf{x}_0$ and points toward the steepest ascent direction.**





Key remark: **if the matrix $\mathbf{A}$ is symmetric, positive definite, then**

solve $\mathbf{A}\mathbf{x} = \mathbf{b}$ $\quad\Leftrightarrow\quad$ **solve** $\min_{\mathbf{y}\in\mathbb{R}^d} J(\mathbf{y})$ **where** $J(\mathbf{y}) := \frac{1}{2}\mathbf{y}^T\mathbf{A}\mathbf{y} - \mathbf{b}^T\mathbf{y}.$

Key remark: **if the matrix $\mathbf{A}$ is symmetric, positive definite, then**

$$\textbf{solve } \mathbf{A}\mathbf{x} = \mathbf{b} \quad \Leftrightarrow \quad \textbf{solve } \min_{\mathbf{y}\in\mathbb{R}^d} J(\mathbf{y}) \quad \textbf{where} \quad J(\mathbf{y}) := \frac{1}{2}\mathbf{y}^T\mathbf{A}\mathbf{y} - \mathbf{b}^T\mathbf{y}.$$

Gradient methods **consist in choosing an initial guess $\mathbf{x}_0 \in \mathbb{R}^n$ and in building a sequence of iterates $(\mathbf{x}_k)_{k\in\mathbb{N}}$ of $\mathbb{R}^n$ such that**

$$J(\mathbf{x}_k) \underset{k\to+\infty}{\downarrow} \min_{\mathbb{R}^n} J \qquad \textbf{Note that} \quad \nabla J(\mathbf{y}) = \mathbf{A}\mathbf{y} - \mathbf{b}$$

Key remark: if the matrix A is symmetric, positive definite, then

$$\textbf{solve } \mathrm{A}\mathrm{x} = \mathrm{b} \quad \Leftrightarrow \quad \textbf{solve } \min_{\mathrm{y}\in\mathbb{R}^d} J(\mathrm{y}) \quad \textbf{where} \quad J(\mathrm{y}) := \frac{1}{2}\mathrm{y}^T\mathrm{A}\mathrm{y} - \mathrm{b}^T\mathrm{y}.$$

Gradient methods consist in choosing an initial guess $\mathrm{x}_0 \in \mathbb{R}^n$ and in building a sequence of iterates $(\mathrm{x}_k)_{k\in\mathbb{N}}$ of $\mathbb{R}^n$ such that

$$J(\mathrm{x}_k) \underset{k\to+\infty}{\downarrow} \min_{\mathbb{R}^n} J \qquad \textbf{Note that} \quad \nabla J(\mathrm{y}) = \mathrm{A}\mathrm{y} - \mathrm{b}$$

Gradient methods only involve matrix-vector and inner products. There are particularly efficient when

- **the matrix A cannot be stored (e.g. grid methods for Kohn-Sham)**
- **and/or matrix-vector products can be efficiently computed (sparse matrices, fast transforms such as FFT, ...)**

Key remark: if the matrix A is symmetric, positive definite, then

$$\textbf{solve } \mathrm{Ax} = \mathrm{b} \quad \Leftrightarrow \quad \textbf{solve } \min_{\mathrm{y}\in\mathbb{R}^d} J(\mathrm{y}) \quad \textbf{where} \quad J(\mathrm{y}) := \frac{1}{2}\mathrm{y}^T\mathrm{Ay} - \mathrm{b}^T\mathrm{y}.$$

Gradient methods consist in choosing an initial guess $\mathrm{x}_0 \in \mathbb{R}^n$ and in building a sequence of iterates $(\mathrm{x}_k)_{k\in\mathbb{N}}$ of $\mathbb{R}^n$ such that

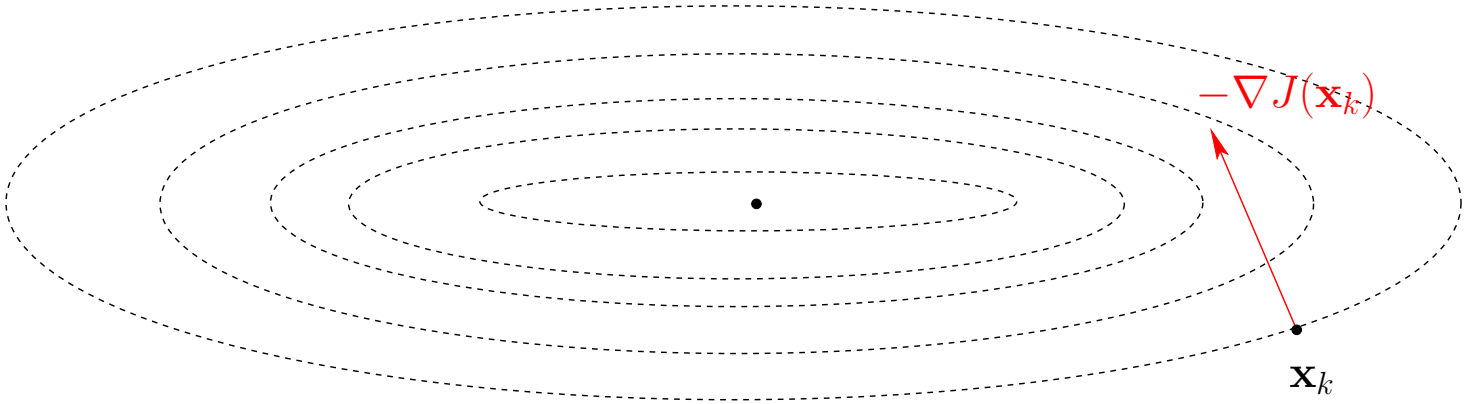
$$J(\mathrm{x}_k) \underset{k\to+\infty}{\downarrow} \min_{\mathbb{R}^n} J \qquad \textbf{Note that} \quad \nabla J(\mathrm{y}) = \mathrm{Ay} - \mathrm{b}$$

Gradient methods only involve matrix-vector and inner products. There are particularly efficient when

- **the matrix A cannot be stored (e.g. grid methods for Kohn-Sham)**
- **and/or matrix-vector products can be efficiently computed (sparse matrices, fast transforms such as FFT, ...)**

Remark: Extensions of gradient algorithms to general linear systems are available (MINRES - GMRES, 1986 - BiCGstab, 1992 - ...).

Fixed-step and optimal step gradient algorithms



The function J is decreasing in the direction

$$\mathbf{d}_k = -\nabla J(\mathbf{x}_k) = \mathbf{b} - \mathbf{A}\mathbf{x}_k \qquad \textbf{(residual)}$$

One then may choose

$$\mathbf{x}_{k+1} = \mathbf{x}_k + t_k \mathbf{d}_k$$

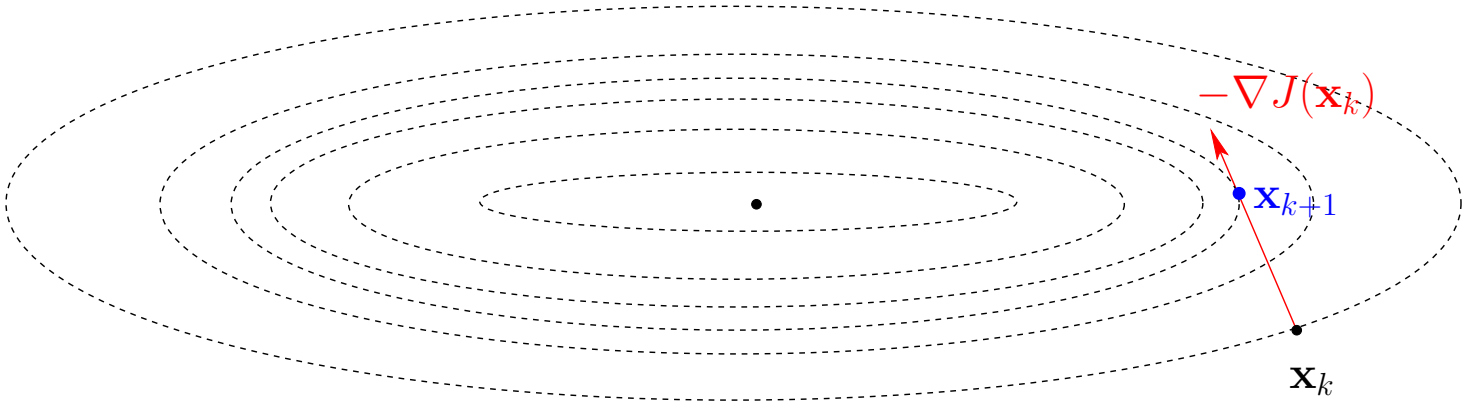
for some $t_k > 0$.

Fixed step: the step t is chosen once and for all

$$\begin{cases} \mathbf{r}_k = \mathbf{b} - \mathbf{A}\mathbf{x}_k \\ \mathbf{x}_{k+1} = \mathbf{x}_k + t\mathbf{r}_k \end{cases}$$

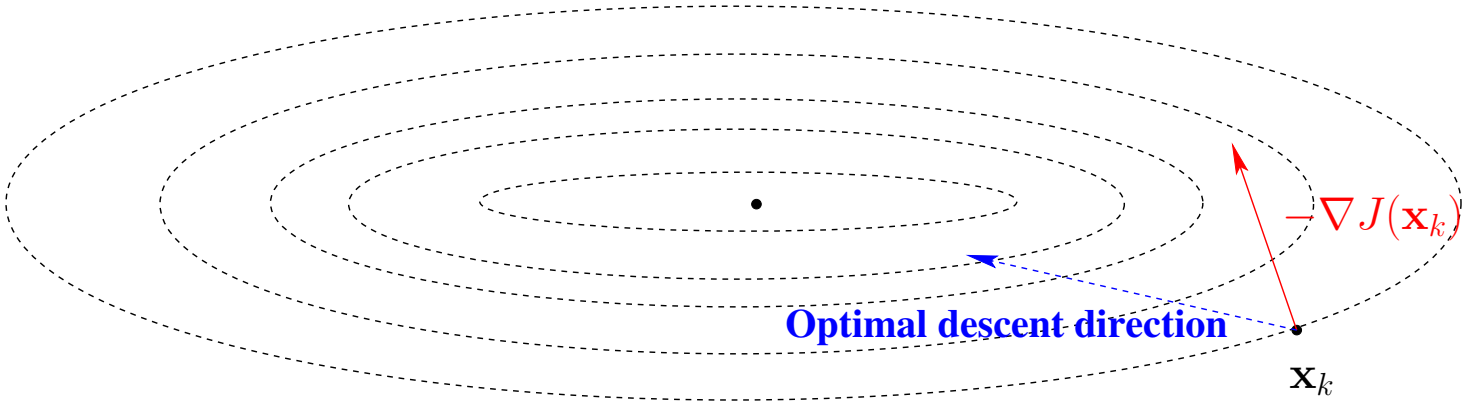
Optimal step: one chooses the "best" $\mathbf{x}_{k+1}$ on the half-line $\mathbf{x}_k - t\nabla J(x_k)$

$$\begin{cases} \mathbf{r}_k = \mathbf{b} - \mathbf{A}\mathbf{x}_k \\ t_k = \dfrac{\mathbf{r}_k^T \mathbf{r}_k}{\mathbf{r}_k^T \mathbf{A}\mathbf{r}_k} \\ \mathbf{x}_{k+1} = \mathbf{x}_k + t_k\mathbf{r}_k \end{cases}$$



Conjugate gradient algorithm (1952)

The descent direction $\mathrm{d}_k = -\nabla J(\mathrm{x}_k)$ is optimal for infinitesimal steps, but not in general for finite step.



The conjugate gradient algorithm provides better descent directions d_k.

Conjugate gradient algorithm:

- **Initialization. Choose** $\mathbf{x}_0 \in \mathbb{R}^n$ **and** ε_{tol}**, compute** $\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$ **and set** $\mathbf{d}_0 = \mathbf{r}_0$**. Set** $k = 0$**.**
- **Iterations.**

  1. Stopping test: if $\|\mathbf{r}_k\|_2 \leq \varepsilon_{\text{tol}}(\|\mathbf{A}\|_1\|\mathbf{x}_k\|_\infty + \|\mathbf{b}\|_2)$**, stop.**

  2. Update $\mathbf{x}_k$ **and the residual** $\mathbf{r}_k$ **:**

$$\mathbf{z}_k = \mathbf{A}\mathbf{d}_k, \qquad t_k = \frac{\mathbf{r}_k^T\mathbf{r}_k}{\mathbf{d}_k^T\mathbf{z}_k},$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k + t_k\mathbf{d}_k, \qquad \mathbf{r}_{k+1} = \mathbf{r}_k - t_k\mathbf{z}_k,$$

  3. Update the descent direction $\mathbf{d}_k$ **:**

$$\beta_k = \frac{\mathbf{r}_{k+1}^T\mathbf{r}_{k+1}}{\mathbf{r}_k^T\mathbf{r}_k}, \qquad \mathbf{d}_{k+1} = \mathbf{r}_{k+1} + \beta_k\mathbf{d}_k.$$

  4. Set $k = k + 1$ **and go to step 1.**

Krylov subspaces

The Krylov subspaces $(\mathcal{K}_k(\mathbf{y}))$ **associated with a matrix** $\mathbf{A} \in \mathbb{R}^{n\times n}$ **and a vector** $\mathbf{y}$ **are defined by**

$$\mathcal{K}_k(\mathbf{y}) = \mathbf{Span}(\mathbf{y}, \mathbf{A}\mathbf{y}, \cdots, \mathbf{A}^k\mathbf{y})$$

Application to linear systems

$$\begin{aligned}
\mathbf{x} &= \mathbf{A}^{-1}\mathbf{b} \\
&= \mathbf{A}^{-1}(\mathbf{A}\mathbf{x}_0 + \mathbf{b} - \mathbf{A}\mathbf{x}_0) \\
&= \mathbf{x}_0 + A^{-1}\mathbf{r}_0 \\
&= \mathbf{x}_0 + Q(\mathbf{A})\mathbf{r}_0 \qquad \textbf{with } Q \textbf{ polynomial of degree } m \leq n-1 \textbf{ (Hamilton-Cayley)} \\
&\in \mathbf{x}_0 + \mathcal{K}_m(\mathbf{r}_0).
\end{aligned}$$

Theorem. **Let $(\mathbf{x}_k)$ the sequence generated by the conjugate gradient algorithm (with $\varepsilon_{\text{tol}} = 0$).**

1. For all $k \geq 0$,

$$\mathbf{x}_k = \underset{\mathbf{y} \in \mathbf{x}_0 + \mathcal{K}_k(\mathbf{r}_0)}{\textbf{arginf}} J(\mathbf{y}), \qquad J(\mathbf{y}) = \frac{1}{2}\mathbf{y}^T \mathbf{A}\mathbf{y} - \mathbf{b}^T \mathbf{y}$$

Theorem. **Let $(\mathbf{x}_k)$ the sequence generated by the conjugate gradient algorithm (with $\varepsilon_{\text{tol}} = 0$).**

1. For all $k \geq 0$,

$$\mathbf{x}_k = \underset{\mathbf{y} \in \mathbf{x}_0 + \mathcal{K}_k(\mathbf{r}_0)}{\textbf{arginf}} \; J(\mathbf{y}), \qquad J(\mathbf{y}) = \frac{1}{2}\mathbf{y}^T \mathbf{A}\mathbf{y} - \mathbf{b}^T\mathbf{y}$$

2. The sequence of Krylov subspace $\mathcal{K}_k(\mathbf{r}_0)$ is strictly increasing until the algorithm has converged: if $\mathbf{x}_k \neq \mathbf{x}$, dim $\mathcal{K}_k(\mathbf{r}_0) = k + 1$. Consequently, the conjugate gradient algorithm converges in at most n iterations

Theorem. **Let $(\mathbf{x}_k)$ the sequence generated by the conjugate gradient algorithm (with $\varepsilon_{\text{tol}} = 0$).**

1. For all $k \geq 0$,

$$\mathbf{x}_k = \underset{\mathbf{y} \in \mathbf{x}_0 + \mathcal{K}_k(\mathbf{r}_0)}{\textbf{arginf}} \; J(\mathbf{y}), \qquad J(\mathbf{y}) = \frac{1}{2}\mathbf{y}^T \mathbf{A}\mathbf{y} - \mathbf{b}^T\mathbf{y}$$

2. The sequence of Krylov subspace $\mathcal{K}_k(\mathbf{r}_0)$ is strictly increasing until the algorithm has converged: if $\mathbf{x}_k \neq \mathbf{x}$, dim $\mathcal{K}_k(\mathbf{r}_0) = k+1$. Consequently, the conjugate gradient algorithm converges in at most n iterations

3. If the conjugate gradient algorithm converges in m iterations, then $\forall 0 \leq k \leq m-1$,

- **$(\mathbf{r}_0, \mathbf{r}_1, \cdots, \mathbf{r}_k)$ is an orthogonal basis of $\mathcal{K}_k(\mathbf{r}_0)$: $\mathbf{r}_i^T \mathbf{r}_j = \delta_{ij}$**
- **$(\mathbf{d}_0, \mathbf{d}_1, \cdots, \mathbf{d}_k)$ is an A-orthogonal basis of $\mathcal{K}_k(\mathbf{r}_0)$: $\mathbf{d}_i^T \mathbf{A}\mathbf{d}_j = \delta_{ij}$**

$\longrightarrow$ **The descent directions $\mathbf{d}_k$ are A-conjugate**

Theorem. **Let A a symmetric positive definite matrix, $\mathrm{b} \in \mathbb{R}^n$ and $\mathrm{x} \in \mathbb{R}^n$ the solution of $\mathrm{Ax} = \mathrm{b}$. Let (x_k) the sequence generated by the conjugate gradient algorithm with (avec $\varepsilon = 0$) from the initial guess x_0.**

The conjugate gradient algorithm converges at least linearly

$$\|\mathbf{x}_k - \mathbf{x}\|_{\mathbf{A}} \leq \rho^k \|\mathbf{x}_0 - \mathbf{x}\|_{\mathbf{A}} \qquad \textbf{with} \qquad 0 \leq \rho = \left(\frac{\sqrt{\kappa_2(\mathbf{A})} - 1}{\sqrt{\kappa_2(\mathbf{A})} + 1} \right) < 1,$$

where $\kappa_2(\mathrm{A}) = \dfrac{\lambda_n(\mathbf{A})}{\lambda_1(\mathbf{A})} \geq 1$ is the condition number of A for the l^2-norm, and where $\| \cdot \|_{\mathrm{A}}$ is the energy norm on $\mathbb{R}^n$ defined by $\|\mathrm{y}\|_{\mathrm{A}} = (\mathrm{Ay}, \mathrm{y})^{1/2}$.

Theorem. **Let A a symmetric positive definite matrix, $\mathrm{b} \in \mathbb{R}^n$ and $\mathrm{x} \in \mathbb{R}^n$ the solution of $\mathrm{Ax} = \mathrm{b}$. Let (x_k) the sequence generated by the conjugate gradient algorithm with (avec $\varepsilon = 0$) from the initial guess x_0.**

The conjugate gradient algorithm converges at least linearly

$$\|\mathbf{x}_k - \mathbf{x}\|_{\mathbf{A}} \leq \rho^k \|\mathbf{x}_0 - \mathbf{x}\|_{\mathbf{A}} \qquad \textbf{with} \qquad 0 \leq \rho = \left(\frac{\sqrt{\kappa_2(\mathbf{A})} - 1}{\sqrt{\kappa_2(\mathbf{A})} + 1} \right) < 1,$$

where $\kappa_2(\mathrm{A}) = \dfrac{\lambda_n(\mathbf{A})}{\lambda_1(\mathbf{A})} \geq 1$ is the condition number of A for the l^2-norm, and where $\| \cdot \|_{\mathrm{A}}$ is the energy norm on $\mathbb{R}^n$ defined by $\|\mathrm{y}\|_{\mathrm{A}} = (\mathrm{Ay}, \mathrm{y})^{1/2}$.

Remarks

- **This estimate is not optimal (convergence in at most n iterations)**

Theorem. **Let A a symmetric positive definite matrix, $\mathrm{b} \in \mathbb{R}^n$ and $\mathrm{x} \in \mathbb{R}^n$ the solution of $\mathrm{Ax} = \mathrm{b}$. Let (x_k) the sequence generated by the conjugate gradient algorithm with (avec $\varepsilon = 0$) from the initial guess x_0.**

The conjugate gradient algorithm converges at least linearly

$$\|\mathbf{x}_k - \mathbf{x}\|_{\mathbf{A}} \leq \rho^k \|\mathbf{x}_0 - \mathbf{x}\|_{\mathbf{A}} \qquad \textbf{with} \qquad 0 \leq \rho = \left(\frac{\sqrt{\kappa_2(\mathbf{A})} - 1}{\sqrt{\kappa_2(\mathbf{A})} + 1} \right) < 1,$$

where $\kappa_2(\mathrm{A}) = \dfrac{\lambda_n(\mathrm{A})}{\lambda_1(\mathrm{A})} \geq 1$ is the condition number of A for the l^2-norm, and where $\| \cdot \|_{\mathrm{A}}$ is the energy norm on $\mathbb{R}^n$ defined by $\|\mathrm{y}\|_{\mathrm{A}} = (\mathrm{Ay}, \mathrm{y})^{1/2}$.

Remarks

- **This estimate is not optimal (convergence in at most n iterations)**
- **The actual performance of the CG algorithm depends on the distribution of the eigenvalues of A**

Theorem. **Let A a symmetric positive definite matrix, $\mathrm{b} \in \mathbb{R}^n$ and $\mathrm{x} \in \mathbb{R}^n$ the solution of $\mathrm{Ax} = \mathrm{b}$. Let (x_k) the sequence generated by the conjugate gradient algorithm with (avec $\varepsilon = 0$) from the initial guess x_0.**

The conjugate gradient algorithm converges at least linearly

$$\|\mathbf{x}_k - \mathbf{x}\|_{\mathbf{A}} \leq \rho^k \|\mathbf{x}_0 - \mathbf{x}\|_{\mathbf{A}} \qquad \textbf{with} \qquad 0 \leq \rho = \left(\frac{\sqrt{\kappa_2(\mathbf{A})} - 1}{\sqrt{\kappa_2(\mathbf{A})} + 1} \right) < 1,$$

where $\kappa_2(\mathrm{A}) = \dfrac{\lambda_n(\mathrm{A})}{\lambda_1(\mathrm{A})} \geq 1$ is the condition number of A for the l^2-norm, and where $\| \cdot \|_{\mathrm{A}}$ is the energy norm on $\mathbb{R}^n$ defined by $\|\mathrm{y}\|_{\mathrm{A}} = (\mathrm{Ay}, \mathrm{y})^{1/2}$.

Remarks

- **This estimate is not optimal (convergence in at most n iterations)**
- **The actual performance of the CG algorithm depends on the distribution of the eigenvalues of A**
- **The smaller the condition number, the faster the algorithm**

Theorem. **Let A a symmetric positive definite matrix, $\mathrm{b} \in \mathbb{R}^n$ and $\mathrm{x} \in \mathbb{R}^n$ the solution of $\mathrm{Ax} = \mathrm{b}$. Let (x_k) the sequence generated by the conjugate gradient algorithm with (avec $\varepsilon = 0$) from the initial guess x_0.**

The conjugate gradient algorithm converges at least linearly

$$\|\mathbf{x}_k - \mathbf{x}\|_{\mathbf{A}} \leq \rho^k \|\mathbf{x}_0 - \mathbf{x}\|_{\mathbf{A}} \qquad \textbf{with} \qquad 0 \leq \rho = \left(\frac{\sqrt{\kappa_2(\mathbf{A})} - 1}{\sqrt{\kappa_2(\mathbf{A})} + 1} \right) < 1,$$

where $\kappa_2(\mathrm{A}) = \dfrac{\lambda_n(\mathrm{A})}{\lambda_1(\mathrm{A})} \geq 1$ is the condition number of A for the l^2-norm, and where $\| \cdot \|_{\mathrm{A}}$ is the energy norm on $\mathbb{R}^n$ defined by $\|\mathrm{y}\|_{\mathrm{A}} = (\mathrm{Ay}, \mathrm{y})^{1/2}$.

Remarks

- **This estimate is not optimal (convergence in at most n iterations)**
- **The actual performance of the CG algorithm depends on the distribution of the eigenvalues of A**
- **The smaller the condition number, the faster the algorithm**

$\longrightarrow$ **Preconditioning can (often must) be used to reduced the cond. numb.**

Iterative algorithms are usually totally inefficient without preconditioning.

Preconditioning of linear systems:

Basic idea: instead of solving

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

solve

$$\begin{cases} \mathbf{P}^{-1/2}\mathbf{A}\mathbf{P}^{-1/2}\mathbf{z} = \mathbf{P}^{-1/2}\mathbf{b}, \\ \mathbf{P}^{1/2}\mathbf{x} = \mathbf{z}. \end{cases}$$

for some symmetric matrix P such that

$$\kappa_2(\mathbf{P}^{-1/2}\mathbf{A}\mathbf{P}^{-1/2}) \ll \kappa_2(\mathbf{A})$$

This replacement can be done implicitely: no need to compute $\mathbf{P}^{-1/2}$.

Preconditioned conjugate gradient algorithm

- **Initialisation. Choose** $\mathbf{x}_0 \in \mathbb{R}^n$ **and a threshold** ε_{tol}**, compute** $\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0$**, and the solution** $\mathbf{y}_0$ **to** $\mathbf{P}\mathbf{y}_0 = \mathbf{r}_0$**. Set** $\mathbf{d}_0 = \mathbf{y}_0$ **and** $k = 0$**. ;**
- **Iterations.**

  1. Stopping test: if $\|\mathbf{r}_k\|_2 \leq \varepsilon_{\text{tol}}(\|A\|_1\|x_k\|_\infty + \|b\|_2)$**, stop.**

  2. Update $\mathbf{x}_k$ **and** $\mathbf{r}_k$

$$\mathbf{z}_k = \mathbf{A}\mathbf{d}_k, \qquad t_k = \frac{\mathbf{y}_k^T\mathbf{r}_k}{\mathbf{d}_k^T\mathbf{z}_k},$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k + t_k\mathbf{d}_k, \qquad \mathbf{r}_{k+1} = \mathbf{r}_k - t_k\mathbf{z}_k,$$

$$\textbf{Solve} \quad \mathbf{P}\mathbf{y}_{k+1} = \mathbf{r}_{k+1}$$

  3. Updated the descent direction $\mathbf{d}_k$

$$\beta_k = \frac{\mathbf{y}_{k+1}^T\mathbf{r}_{k+1}}{\mathbf{y}_k^T\mathbf{r}_k}, \qquad \mathbf{d}_{k+1} = \mathbf{y}_{k+1} + \beta_k\mathbf{d}_k.$$

  4. Set $k = k + 1$ **and go to step 1.**

For the preconditioning technique to be efficient, the preconditioner P must fulfill two conditions

1. $\kappa_2(\mathbf{P}^{-1/2}\mathbf{A}\mathbf{P}^{-1/2}) \ll \kappa_2(\mathbf{A})$

2. linear systems of the form $\mathbf{P}\mathbf{y} = \mathbf{r}$ **are easy to solve.**

$\longrightarrow$ **A trade-off has to be made.**

- **"Algebraic preconditioners"**
  - **diagonal preconditioner**
  - **SSOR preconditioner**
  - **incomplete** LU **or Cholesky decomposition**
- **"Physical preconditioners"**
  - **multigrid methods**
  - **simplified model**

Example: **planewave discretization of periodic Schrödinger operators**

$$H = -\frac{1}{2}\frac{d^2}{dx^2} + V, \quad V(x) = |\cos(\pi x)|, \quad e_k(x) = e^{2i\pi kx}, \quad X_N = \textbf{Span}(e_k,\ |k| \leq N)$$

$$H_{kl} = \langle e_k|H|e_l\rangle = 2\pi^2|k|^2\delta_{kl} + \hat{V}_{kl}, \quad \hat{V}_{kl} = \int_0^1 V(x)\, e^{2i\pi(l-k)x}\, dx, \quad -N \leq k, l \leq N$$

$$\textbf{Solve } \mathbf{Hx} = \mathbf{b}, \quad \textbf{with} \quad \mathbf{b} = (1, \cdots, 1)^T$$

$$\longrightarrow \qquad \textbf{Possible preconditioner: } \mathbf{P} \quad \textbf{s.t.} \quad P_{kl} = (1 + 2\pi^2|k|^2)\delta_{kl}$$

Stopping criterion: $\|\mathbf{r}_k\|_2 \leq 10^{-10}$ **where** $\mathbf{r}_k = \mathbf{b} - \mathbf{Hx}_k$

N	Size of the matrix H	# CG iter.	# PCG iter.
50	**101**	**71**	**5**
100	**201**	**98**	**5**
200	**401**	**304**	**5**
400	**801**	**613**	**5**

4 - Constrained optimization and Lagrange multipliers

Let $E : \mathbb{R}^d \to \mathbb{R}$ and $g : \mathbb{R}^d \to \mathbb{R}^m$ be two differentiable functions and consider the optimization problem

$$\boxed{\inf_{\mathbf{x}\in K} E(\mathbf{x}) \quad \textbf{where} \quad K = \left\{\mathbf{x} \in \mathbb{R}^d \mid g(\mathbf{x}) = 0\right\}.}$$

Let $E : \mathbb{R}^d \to \mathbb{R}$ and $g : \mathbb{R}^d \to \mathbb{R}^m$ be two differentiable functions and consider the optimization problem

$$\boxed{\inf_{\mathbf{x}\in K} E(\mathbf{x}) \quad \textbf{where} \quad K = \left\{\mathbf{x} \in \mathbb{R}^d \mid g(\mathbf{x}) = 0\right\}.}$$

Definition (qualification of the constraints). The equality constraints $g = 0$ are called qualified at $\mathbf{x}_0 \in K$ if $g'(\mathbf{x}_0) \in \mathbb{R}^{m\times d}$ is surjective (i.e. Ran$(g'(\mathbf{x}_0)) = \mathbb{R}^m$).

Let $E : \mathbb{R}^d \to \mathbb{R}$ and $g : \mathbb{R}^d \to \mathbb{R}^m$ be two differentiable functions and consider the optimization problem

$$\boxed{\inf_{\mathbf{x}\in K} E(\mathbf{x}) \quad \textbf{where} \quad K = \left\{\mathbf{x} \in \mathbb{R}^d \mid g(\mathbf{x}) = 0\right\}.}$$

Definition (qualification of the constraints). The equality constraints $g = 0$ are called qualified at $\mathbf{x}_0 \in K$ if $g'(\mathbf{x}_0) \in \mathbb{R}^{m\times d}$ is surjective (i.e. Ran$(g'(\mathbf{x}_0)) = \mathbb{R}^m$).

Theorem (Euler-Lagrange theorem). Let $\mathbf{x}_0 \in K$ be a local minimum of E on K. Assume that

1. **$\mathbf{x} \mapsto g'(\mathbf{x})$ is continuous in the vicinity of $\mathbf{x}_0$;**
2. **the equality constraints $g = 0$ is qualified at $\mathbf{x}_0$.**

Then, there exists a unique $\lambda \in \mathbb{R}^m$ such that

$$\nabla E(\mathbf{x}_0) + g'(\mathbf{x}_0)^T\lambda = 0,$$

where $g'(\mathbf{x}_0)^T$ is the transpose of $g'(\mathbf{x}_0)$. The vector λ is called the Lagrange multiplier of the constraint $g = 0$.

Euler-Lagrange equations

Assume that the constraints are qualified at any point of K. Then solving

$$\begin{cases} \textbf{seek } (\mathbf{x}, \lambda) \in \mathbb{R}^d \times \mathbb{R}^m \textbf{ such that} \\ \nabla E(\mathbf{x}) + g'(\mathbf{x})^T \lambda = 0 \\ g(\mathbf{x}) = 0 \end{cases} \tag{2}$$

allows one to find all the critical points (among which the local minimizers and the local maximizers) of E on K.

Remark : the above problem consists of $(d + m)$ scalar equations with $(d + m)$ scalar unknowns.

Euler-Lagrange equations

Assume that the constraints are qualified at any point of K. Then solving

$$\begin{cases} \textbf{seek } (\mathbf{x}, \lambda) \in \mathbb{R}^d \times \mathbb{R}^m \textbf{ such that} \\ \nabla E(\mathbf{x}) + g'(\mathbf{x})^T \lambda = 0 \\ g(\mathbf{x}) = 0 \end{cases} \tag{3}$$

allows one to find all the critical points (among which the local minimizers and the local maximizers) of E on K.

Remark : the above problem consists of $(d + m)$ scalar equations with $(d + m)$ scalar unknowns.

The solutions of the Euler-Lagrange equations (4) are called the critical points of E on K.

Euler-Lagrange equations

Assume that the constraints are qualified at any point of K. Then solving

$$\begin{cases} \textbf{seek } (\mathbf{x}, \lambda) \in \mathbb{R}^d \times \mathbb{R}^m \textbf{ such that} \\ \nabla E(\mathbf{x}) + g'(\mathbf{x})^T \lambda = 0 \\ g(\mathbf{x}) = 0 \end{cases} \tag{4}$$

allows one to find all the critical points (among which the local minimizers and the local maximizers) of E on K.

Remark : the above problem consists of $(d + m)$ scalar equations with $(d + m)$ scalar unknowns.

The solutions of the Euler-Lagrange equations (4) are called the critical points of E on K.

Remark. Equations (4) are equivalent to seeking $(\mathbf{x}, \lambda) \in \mathbb{R}^d \times \mathbb{R}^m$ s.t.

$\nabla_{\mathbf{x}} L(\mathbf{x}, \lambda) = 0,\ \nabla_{\lambda} L(\mathbf{x}, \lambda) = 0,$ **where** $L(\mathbf{x}, \lambda) := E(\mathbf{x}) + \lambda \cdot g(\mathbf{x})$ **(Lagrangian).**

Very important take-home messages

A mathematical theorem consists of

- **a list of assumptions;**
- **one of more results following from these assumptions.**

Do not forget to check the assumptions before using the results!

Very important take-home messages

A mathematical theorem consists of

- **a list of assumptions;**
- **one of more results following from these assumptions.**

Do not forget to check the assumptions before using the results!

Back to the example $d = 1$, $m = 1$, $E(x) = x$, $g(x) = x^2$. **Then**

$$K = \{x \in \mathbb{R} \mid g(x) = 0\} = \{0\} \qquad \textbf{and} \qquad g'(0) = 0.$$

The constraint $g = 0$ **is therefore not qualified, and this is the reason why the Lagragian method fails!**

Very important take-home messages

A mathematical theorem consists of

- **a list of assumptions;**
- **one of more results following from these assumptions.**

Do not forget to check the assumptions before using the results!

Back to the example $d = 1$, $m = 1$, $E(x) = x$, $g(x) = x^2$**. Then**

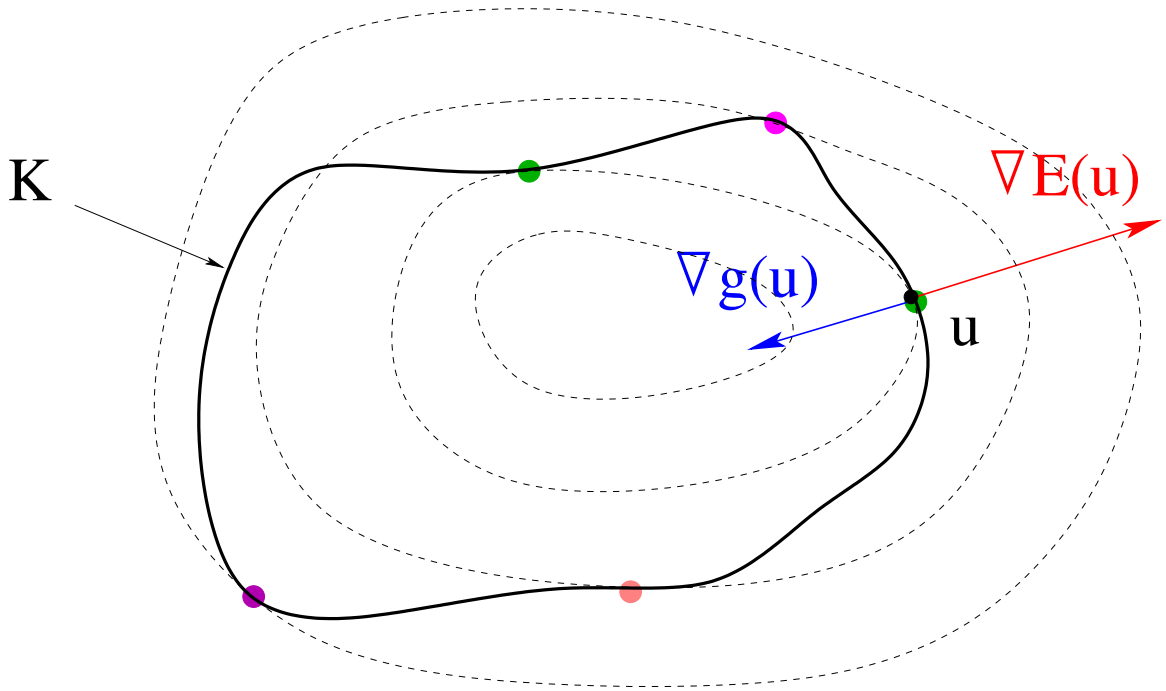
$$K = \{x \in \mathbb{R} \mid g(x) = 0\} = \{0\} \qquad \textbf{and} \qquad g'(0) = 0.$$

The constraint $g = 0$ **is therefore not qualified, and this is the reason why the Lagragian method fails!**

Be all the more careful that
not every "reasonable" mathematical statement is true!

Example: let $\mathcal{H}$ **be a Hilbert space. A continuous function** $E : \mathcal{H} \to \mathbb{R}$ **going to** $+\infty$ **at infinity does not necessarily have a minimizer.**

A simple 2D example ($d = 2$, $m = 1$)



On $K = g^{-1}(0) = \{\mathbf{x} \in \mathbb{R}^2 \,|\, g(\mathbf{x}) = 0\}$, the function E possesses

- **two local minimizers, all global**
- **two local maximizers, among which the global maximizer**
- **one critical point which is neither a local minimizer not a local maximizer.**

Sketch of the proof

- **Let $\mathbf{x}_0$ be a local minimizer of E on $K = g^{-1}(0) = \{\mathbf{x} \in \mathbb{R}^d \,|\, g(\mathbf{x}) = 0\}$ and $\alpha = E(\mathbf{x}_0)$.**

- **If the constraint $g = 0$ is qualified at $\mathbf{x}_0$ (i.e. if $g'(\mathbf{x}_0) : \mathcal{H} \to \mathcal{K}$ is surjective), then, in the vicinity of $\mathbf{x}_0$, K is a C^1 manifold with tangent space**
$$T_{\mathbf{x}_0}K = \{\mathbf{h} \in \mathbb{R}^d \mid g'(\mathbf{x}_0)\mathbf{h} = 0\} = \mathbf{Ker}(g'(\mathbf{x}_0)).$$

- **Since $\mathbf{x}_0$ is a minimizer of E on K, the vector $\nabla E(\mathbf{x}_0)$ must be orthogonal to $T_{\mathbf{x}_0}K$. Indeed, for any $\mathbf{h} \in T_{\mathbf{x}_0}K$, there exists a C^1 curve $\phi : [-1, 1] \to \mathbb{R}^d$ drawn on K such that $\phi(0) = \mathbf{x}_0$ et $\phi'(0) = \mathbf{h}$, and we have**
$$0 \leq E(\phi(t)) - E(\mathbf{x}_0) = E(\mathbf{x}_0 + t\mathbf{h} + o(t)) - E(\mathbf{x}_0) = t\nabla E(\mathbf{x}_0) \cdot \mathbf{h} + o(t).$$

- **We have**
$$\nabla E(\mathbf{x}_0) \in (T_{\mathbf{x}_0}K)^\perp = (\mathbf{Ker}(g'(\mathbf{x}_0)))^\perp = \mathbf{Ran}(g'(\mathbf{x}_0)^T).$$

- **Therefore, there exists $\lambda \in \mathbb{R}^m$ such that $\nabla E(\mathbf{x}_0) + g'(\mathbf{x}_0)^T\lambda = 0$.**

Remarks

- **The above results can be extended to the case when $E : \mathcal{H} \to \mathbb{R}$ and $g : \mathcal{H} \to \mathcal{K}$ where $\mathcal{H}$ and $\mathcal{K}$ are Hilbert spaces.**

- **Most often, Lagrange multipliers have a "physical" interpretation**

  – **statistical mechanics, the equilibrium state of a chemical system interacting with its environment is obtained by maximizing the entropy under the constraints that the energy, the volume and the concentration of chemical species are given on average:**

  $\to$ the Lagrange multipliers are respectively $1/T$, P/T and μ_i/T (T: temperature, P: pressure, μ_i chemical potential of species i)

  – **fluid mechanics, the admissible dynamics of an incompressible fluid are the critical points of the action under the constraint that the density of the fluid remains constant ($\operatorname{div}(u) = 0$)**

  $\to$ the Lagrange multiplier of the incompressibility constraint is the pressure field.

Analytical derivatives

$$\forall \mathbf{R} \in \mathbb{R}^k, \quad W(\mathbf{R}) = \inf \left\{ E(\mathbf{R}, \mathbf{x}),\ \mathbf{x} \in \mathbb{R}^d,\ g(\mathbf{R}, \mathbf{x}) = 0 \right\} \tag{5}$$

with $E : \mathbb{R}^k \times \mathbb{R}^d \to \mathbb{R}$, $g : \mathbb{R}^k \times \mathbb{R}^d \to \mathbb{R}^m$.

Assume (5) has a unique minimizer $\mathbf{x}(\mathbf{R})$ and $\mathbf{R} \mapsto \mathbf{x}(\mathbf{R})$ is regular. Then,

$$W(\mathbf{R}) = E(\mathbf{R}, \mathbf{x}(\mathbf{R})) \quad \Rightarrow \quad \frac{\partial W}{\partial R_k}(\mathbf{R}) = \frac{\partial E}{\partial R_k}(\mathbf{R}, \mathbf{x}(\mathbf{R})) + \nabla_{\mathbf{x}} E(\mathbf{R}, \mathbf{x}(\mathbf{R})) \cdot \frac{\partial \mathbf{x}}{\partial R_k}(\mathbf{R}),$$

$$g(\mathbf{R}, \mathbf{x}(\mathbf{R})) = 0 \quad \Rightarrow \quad \frac{\partial g}{\partial R_k}(\mathbf{R}, \mathbf{x}(\mathbf{R})) + g'_{\mathbf{x}}(\mathbf{R}, \mathbf{x}(\mathbf{R})) \frac{\partial \mathbf{x}}{\partial R_k}(\mathbf{R}) = 0.$$

Euler-Lagrange equation: $\nabla_{\mathbf{x}} E(\mathbf{R}, \mathbf{x}(\mathbf{R})) + g'_{\mathbf{x}}(\mathbf{R}, \mathbf{x}(\mathbf{R}))^T \lambda(\mathbf{R}) = 0.$

Therefore $$\frac{\partial W}{\partial R_k}(\mathbf{R}) = \frac{\partial E}{\partial R_k}(\mathbf{R}, \mathbf{x}(\mathbf{R})) + \left(\frac{\partial g}{\partial R_k}(\mathbf{R}, \mathbf{x}(\mathbf{R})), \lambda(\mathbf{R}) \right).$$