An in-depth look at SCF algorithms

Eric CANCES

Ecole des Ponts and INRIA Paris, France

Heidelberg, October 2-6, 2017

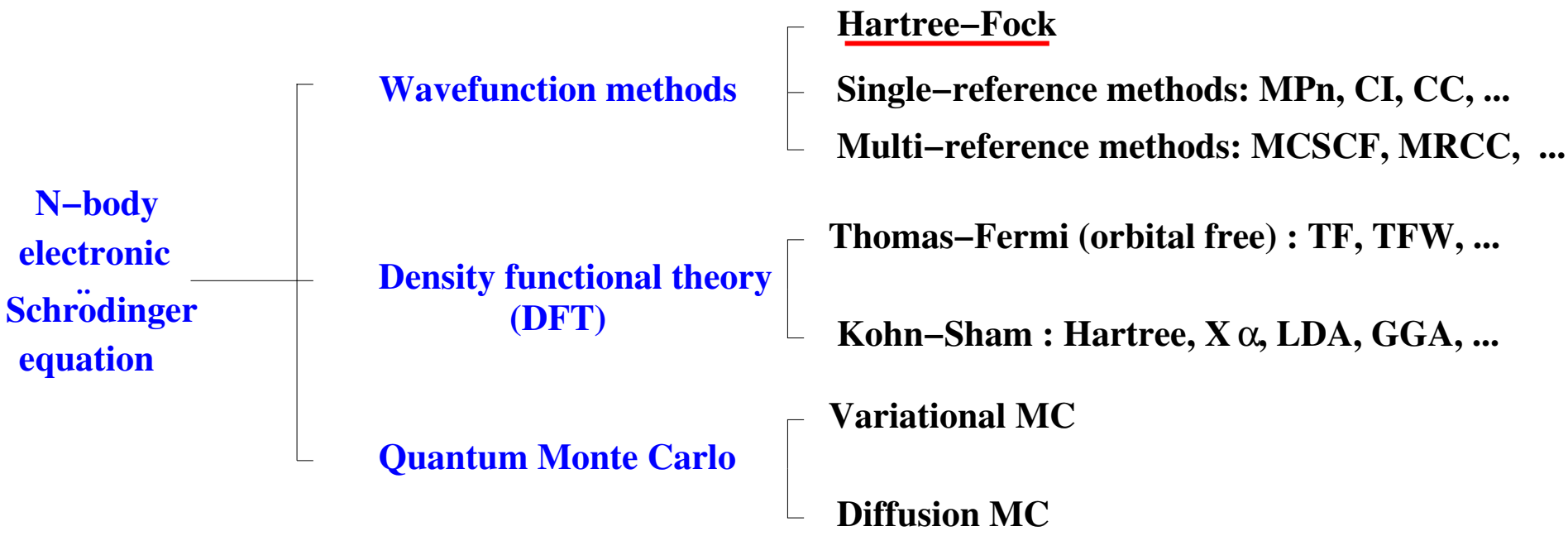
1. Mathematical derivation and properties of the Hartree-Fock model

2. Galerkin approximation of the Hartree-Fock model

3. SCF algorithms

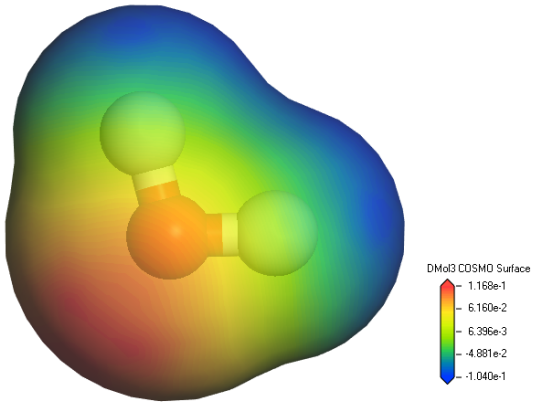
Appendix: Constrained optimization and Lagrange multipliers

1 - Mathematical derivation and properties of the HF model



Here the functions of $L^2(\mathbb{R}^3)$, $H^1(\mathbb{R}^3)$, $H^2(\mathbb{R}^3)$ are supposed real-valued. This is legitimate in the absence of external magnetic field or spin-orbit coupling and significantly simplifies the formalism and the numerical methods.

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



Ex: water molecule H_2O

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

$$V^{\mathrm{ne}}(\mathbf{r}) = -\sum_{k=1}^{M} \frac{z_k}{|\mathbf{r} - \mathbf{R}_k|}$$

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

$|\Psi(\mathbf{r}_1, \cdots, \mathbf{r}_N)|^2$ **probability density of observing electron 1 at $\mathbf{r}_1$, electron 2 at $\mathbf{r}_2$, ...**

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

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

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

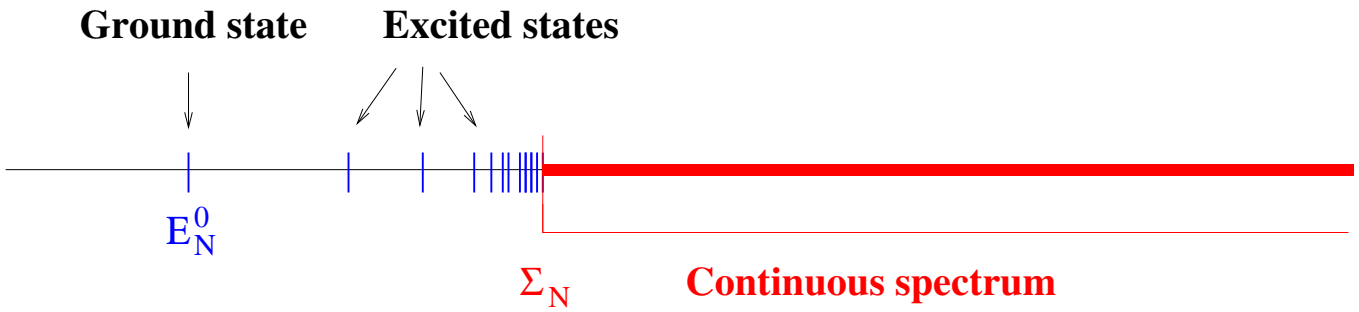
Theorem (spectrum of H_N).

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

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

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

If $N \leq Z := \sum_{k=1}^{M} z_k$**, then** H_N **has an infinite number of bound states.**



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

If $N \geq Z + 1$**, then** H_N **has at most a finite number of bound states.**

Variational expression of the ground state energy

$$E_0 = \inf \{\langle \psi | H_N | \psi \rangle,\ \psi \in \mathcal{W}_N\} \quad \mathcal{W}_N = \left\{ \psi \in \bigwedge^N L^2(\mathbb{R}^3) \cap H^1(\mathbb{R}^{3N}),\ \|\psi\|_{L^2} = 1 \right\}.$$

Variational expression of the ground state energy

$$E_0 = \inf\{\langle\psi|H_N|\psi\rangle,\ \psi \in \mathcal{W}_N\} \quad \mathcal{W}_N = \left\{\psi \in \bigwedge^N L^2(\mathbb{R}^3) \cap H^1(\mathbb{R}^{3N}),\ \|\psi\|_{L^2} = 1\right\}.$$

The Hartree-Fock approximation is variational. It consists in minimizing the exact energy functional $\langle\psi|H_N|\psi\rangle$ on the subset of $\mathcal{W}_N$ defined as

$$\left\{\psi_\Phi = \phi_1 \wedge \cdots \wedge \phi_N,\ \Phi = (\phi_1, \cdots, \phi_N) \in (H^1(\mathbb{R}^3))^N,\ \int_{\mathbb{R}^3} \phi_i\phi_j = \delta_{ij}\right\}$$

$$\psi_\Phi(\mathbf{r}_1, \cdots, \mathbf{r}_N) \stackrel{\text{def}}{=} \frac{1}{\sqrt{N!}} \begin{vmatrix} \phi_1(\mathbf{r}_1) & \phi_1(\mathbf{r}_2) & \cdots & \phi_1(\mathbf{r}_N) \\ \phi_2(\mathbf{r}_1) & \phi_2(\mathbf{r}_2) & \cdots & \phi_2(\mathbf{r}_N) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \phi_N(\mathbf{r}_1) & \phi_N(\mathbf{r}_2) & \cdots & \phi_N(\mathbf{r}_N) \end{vmatrix} \quad \textbf{(Slater determinant)}$$

Molecular orbital formulation of the Hartree-Fock model

$$E_0 \leq E_0^{\mathrm{HF}} = \inf \left\{ E^{\mathrm{HF}}(\Phi),\ \Phi = (\phi_1, \cdots, \phi_N) \in (H^1(\mathbb{R}^3))^N,\ \int_{\mathbb{R}^3} \phi_i \phi_j = \delta_{ij} \right\}$$

$$E^{\mathrm{HF}}(\Phi) = \frac{1}{2} \sum_{i=1}^{N} \int_{\mathbb{R}^3} |\nabla \phi_i|^2 + \int_{\mathbb{R}^3} \rho_\Phi V^{\mathrm{ne}} + \underbrace{\frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho_\Phi(\mathbf{r})\, \rho_\Phi(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}\, d\mathbf{r}\, d\mathbf{r}'}_{\textbf{Coulomb term}} \underbrace{- \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{|\gamma_\Phi(\mathbf{r}, \mathbf{r}')|^2}{|\mathbf{r} - \mathbf{r}'|}\, d\mathbf{r}\, d\mathbf{r}'}_{\textbf{exchange term}}$$

$$V^{\mathrm{ne}}(\mathbf{r}) = -\sum_{k=1}^{M} \frac{z_k}{|\mathbf{r} - \mathbf{R}_k|}, \qquad \gamma_\Phi(\mathbf{r}, \mathbf{r}') = \sum_{i=1}^{N} \phi_i(\mathbf{r})\, \phi_i(\mathbf{r}'), \qquad \rho_\Phi(\mathbf{r}) = \sum_{i=1}^{N} |\phi_i(\mathbf{r})|^2.$$

Molecular orbital formulation of the Hartree-Fock model

$$E_0 \leq E_0^{\mathrm{HF}} = \inf \left\{ E^{\mathrm{HF}}(\Phi),\ \Phi = (\phi_1, \cdots, \phi_N) \in (H^1(\mathbb{R}^3))^N,\ \int_{\mathbb{R}^3} \phi_i \phi_j = \delta_{ij} \right\}$$

$$E^{\mathrm{HF}}(\Phi) = \frac{1}{2} \sum_{i=1}^{N} \int_{\mathbb{R}^3} |\nabla \phi_i|^2 + \int_{\mathbb{R}^3} \rho_\Phi V^{\mathrm{ne}} + \underbrace{\frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho_\Phi(\mathbf{r})\, \rho_\Phi(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}\, d\mathbf{r}\, d\mathbf{r}'}_{\textbf{Coulomb term}} \underbrace{- \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{|\gamma_\Phi(\mathbf{r}, \mathbf{r}')|^2}{|\mathbf{r} - \mathbf{r}'|}\, d\mathbf{r}\, d\mathbf{r}'}_{\textbf{exchange term}}$$

$$V^{\mathrm{ne}}(\mathbf{r}) = -\sum_{k=1}^{M} \frac{z_k}{|\mathbf{r} - \mathbf{R}_k|}, \qquad \gamma_\Phi(\mathbf{r}, \mathbf{r}') = \sum_{i=1}^{N} \phi_i(\mathbf{r})\, \phi_i(\mathbf{r}'), \qquad \rho_\Phi(\mathbf{r}) = \sum_{i=1}^{N} |\phi_i(\mathbf{r})|^2.$$

Invariance property: if $\Phi \in (H^1(\mathbb{R}^3))^N$ satisfies the L^2-orthonormality constraints, then so does ΦU for all $U \in \mathrm{O}(N)$ (i.e. $U \in \mathbb{R}^{N \times N}$, $U^T U = I_N$), and

$$\psi_{\Phi U} = \det(U) \psi_\Phi, \quad \gamma_{\Phi U} = \gamma_\Phi, \quad \rho_{\Phi U} = \rho_\Phi, \quad E(\Phi U) = E(\Phi).$$

Theorem. **Assume that $N \leq Z := \sum_{k=1}^{M} z_k$ (neutral or positively charged molecular system). Then**

1. the HF model has a ground state $\Phi^0 = (\phi_1^0, \cdots, \phi_N^0)$ (Lieb & Simon '77);

Theorem. **Assume that** $N \leq Z := \sum_{k=1}^{M} z_k$ **(neutral or positively charged molecular system). Then**

1. the HF model has a ground state $\Phi^0 = (\phi_1^0, \cdots, \phi_N^0)$ **(Lieb & Simon '77);**

2. Euler-Lagrange equations: there exists $\lambda \in \mathbb{R}^{N \times N}$ **symmetric such that**

$$\begin{cases} \Phi^0 = (\phi_1^0, \cdots, \phi_N^0) \in (H^1(\mathbb{R}^3))^N \\ \\ -\frac{1}{2}\Delta\phi_i^0 + V^{\mathrm{ne}}\phi_i^0 + \left(\rho_{\Phi^0} \star |\cdot|^{-1}\right)\phi_i^0 - \displaystyle\int_{\mathbb{R}^3} \frac{\gamma_{\Phi^0}(\cdot, \mathbf{r}')}{|\cdot - \mathbf{r}'|}\,\phi_i^0(\mathbf{r}')\,d\mathbf{r}' = \sum_{j=1} \lambda_{ij}\phi_j^0 \\ \\ \displaystyle\int_{\mathbb{R}^3} \phi_i^0\phi_j^0 = \delta_{ij}; \end{cases}$$

Theorem. **Assume that** $N \leq Z := \sum_{k=1}^{M} z_k$ **(neutral or positively charged molecular system). Then**

1. the HF model has a ground state $\Phi^0 = (\phi_1^0, \cdots, \phi_N^0)$ **(Lieb & Simon '77);**

2. Euler-Lagrange equations: there exists $\lambda \in \mathbb{R}^{N\times N}$ **symmetric such that**

$$\begin{cases} \Phi^0 = (\phi_1^0, \cdots, \phi_N^0) \in (H^1(\mathbb{R}^3))^N \\ \\ -\frac{1}{2}\Delta\phi_i^0 + V^{\rm ne}\phi_i^0 + \left(\rho_{\Phi^0} \star |\cdot|^{-1}\right)\phi_i^0 - \displaystyle\int_{\mathbb{R}^3} \frac{\gamma_{\Phi^0}(\cdot, \mathbf{r}')}{|\cdot - \mathbf{r}'|}\,\phi_i^0(\mathbf{r}')\,d\mathbf{r}' = \sum_{j=1} \lambda_{ij}\phi_j^0 \\ \\ \displaystyle\int_{\mathbb{R}^3} \phi_i^0\phi_j^0 = \delta_{ij}; \end{cases}$$

3. Elliptic regularity: $\phi_i^0 \in H^2(\mathbb{R}^3) \cap C^\infty(\mathbb{R}^3 \setminus \{\mathbf{R}_k\})$**;**

Theorem (continued).

4. Fock operator:

$$\mathcal{F}_{\Phi^0} := -\frac{1}{2}\Delta + V^{\mathrm{ne}} + \rho_{\Phi^0} \star |\cdot|^{-1} + \mathcal{K}_{\Phi^0} \quad \textbf{where} \quad (\mathcal{K}_{\Phi^0}\phi)(\mathbf{r}) = -\int_{\mathbb{R}^3} \frac{\gamma_{\Phi^0}(\mathbf{r},\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,\phi(\mathbf{r}')\,d\mathbf{r}'$$

is a self-adjoint operator on $L^2(\mathbb{R}^3)$ with domain $H^2(\mathbb{R}^3)$. It is bounded below and $\sigma_{\mathrm{ess}}(H_0) = [0, +\infty)$;

Theorem (continued).

4. Fock operator:

$$\mathcal{F}_{\Phi^0} := -\frac{1}{2}\Delta + V^{\mathrm{ne}} + \rho_{\Phi^0} \star |\cdot|^{-1} + \mathcal{K}_{\Phi^0} \quad \textbf{where} \quad (\mathcal{K}_{\Phi^0}\phi)(\mathbf{r}) = -\int_{\mathbb{R}^3} \frac{\gamma_{\Phi^0}(\mathbf{r},\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,\phi(\mathbf{r}')\,d\mathbf{r}'$$

is a self-adjoint operator on $L^2(\mathbb{R}^3)$ with domain $H^2(\mathbb{R}^3)$. It is bounded below and $\sigma_{\mathrm{ess}}(H_0) = [0, +\infty)$;

5. Hartree-Fock equations: up to replacing Φ^0 by $\Phi^0 U$ for some $U \in O(N)$, it holds

$$\mathcal{F}_{\Phi^0}\phi_i^0 = \varepsilon_i \phi_i^0, \qquad \int_{\mathbb{R}^3} \phi_i^0 \phi_j^0 = \delta_{ij}, \qquad \varepsilon_1 \leq \cdots \leq \varepsilon_N < 0;$$

Theorem (continued).

4. Fock operator:

$$\mathcal{F}_{\Phi^0} := -\frac{1}{2}\Delta + V^{\mathrm{ne}} + \rho_{\Phi^0} \star |\cdot|^{-1} + \mathcal{K}_{\Phi^0} \quad \textbf{where} \quad (\mathcal{K}_{\Phi^0}\phi)(\mathbf{r}) = -\int_{\mathbb{R}^3} \frac{\gamma_{\Phi^0}(\mathbf{r},\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,\phi(\mathbf{r}')\,d\mathbf{r}'$$

is a self-adjoint operator on $L^2(\mathbb{R}^3)$ with domain $H^2(\mathbb{R}^3)$. It is bounded below and $\sigma_{\mathrm{ess}}(H_0) = [0, +\infty)$;

5. Hartree-Fock equations: up to replacing Φ^0 by $\Phi^0 U$ for some $U \in O(N)$, it holds

$$\mathcal{F}_{\Phi^0}\phi_i^0 = \varepsilon_i \phi_i^0, \qquad \int_{\mathbb{R}^3} \phi_i^0 \phi_j^0 = \delta_{ij}, \qquad \varepsilon_1 \leq \cdots \leq \varepsilon_N < 0;$$

6. *Aufbau* **principle: $\varepsilon_1 \leq \varepsilon_2 \leq \cdots \leq \varepsilon_N$ are the lowest N eigenvalues of $\mathcal{F}_{\Phi^0}$;**

Theorem (continued).

4. Fock operator:

$$\mathcal{F}_{\Phi^0} := -\frac{1}{2}\Delta + V^{\mathrm{ne}} + \rho_{\Phi^0} \star |\cdot|^{-1} + \mathcal{K}_{\Phi^0} \quad \textbf{where} \quad (\mathcal{K}_{\Phi^0}\phi)(\mathbf{r}) = -\int_{\mathbb{R}^3} \frac{\gamma_{\Phi^0}(\mathbf{r},\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,\phi(\mathbf{r}')\,d\mathbf{r}'$$

is a self-adjoint operator on $L^2(\mathbb{R}^3)$ with domain $H^2(\mathbb{R}^3)$. It is bounded below and $\sigma_{\mathrm{ess}}(H_0) = [0, +\infty)$;

5. Hartree-Fock equations: up to replacing Φ^0 by $\Phi^0 U$ for some $U \in O(N)$, it holds

$$\mathcal{F}_{\Phi^0}\phi_i^0 = \varepsilon_i \phi_i^0, \qquad \int_{\mathbb{R}^3} \phi_i^0 \phi_j^0 = \delta_{ij}, \qquad \varepsilon_1 \leq \cdots \leq \varepsilon_N < 0;$$

6. *Aufbau* **principle: $\varepsilon_1 \leq \varepsilon_2 \leq \cdots \leq \varepsilon_N$ are the lowest N eigenvalues of $\mathcal{F}_{\Phi^0}$;**

7. No unfilled-shell property (Bach, Lieb, Loss, Solovej '94): $\varepsilon_N < \varepsilon_{N+1}$ where ε_{N+1} is the $(N+1)^{\mathrm{st}}$ eigenvalue of $\mathcal{F}_{\Phi^0}$ (counting multiplicities) if $\mathcal{F}_{\Phi^0}$ has at least $(N+1)$ negative eigenvalues and 0 otherwise.

2 - Galerkin approximation of the HF model

Galerkin Approximation

$\mathcal{X} = \mathbf{Span}(\chi_1, \cdots, \chi_{N_b})$ **subspace of** $H^1(\mathbb{R}^3)$ **of finite dimension** N_b**.**

$$E_0 \leq E_0^{\mathrm{HF}} \leq E_{0,\mathcal{X}}^{\mathrm{HF}} = \inf \left\{ E^{\mathrm{HF}}(\Phi), \Phi = (\phi_1, \cdots, \phi_N) \in \mathcal{X}^N, \int_{\mathbb{R}^3} \phi_i \phi_j = \delta_{ij} \right\}$$

$$\Phi = (\phi_1, \cdots, \phi_N) \in \mathcal{X}^N \qquad \Rightarrow \qquad \phi_i(\mathbf{r}) = \sum_{\mu=1}^{N_b} C_{\mu i} \chi_\mu(\mathbf{r})$$

Discretized formulation of the Hartree-Fock model

$$E_{0,\mathcal{X}}^{\mathrm{HF}} = \inf\left\{E(CC^T),\ C \in \mathbb{R}^{N_b \times N},\ C^T S C = I_N\right\}$$

$$E(D) = \mathbf{Tr}(hD) + \frac{1}{2}\mathbf{Tr}(G(D)D), \qquad [G(D)]_{\mu\nu} = \sum_{\kappa\lambda} \left[(\mu\nu|\kappa\lambda) - (\mu\lambda|\kappa\nu)\right] D_{\kappa\lambda}$$

Electronic integrals

- **Overlap matrix:** $S_{\mu\nu} = \displaystyle\int_{\mathbb{R}^3} \chi_\mu \chi_\nu$
- **Core Hamiltonian matrix:** $h_{\mu\nu} = \dfrac{1}{2}\displaystyle\int_{\mathbb{R}^3} \nabla\chi_\mu \cdot \nabla\chi_\nu - \sum_{k=1}^{M} z_k \int_{\mathbb{R}^3} \frac{\chi_\mu(\mathbf{r})\chi_\nu(\mathbf{r})}{|\mathbf{r} - \mathbf{R}_k|}\, d\mathbf{r}$
- **Two-electron integrals:** $(\mu\nu|\kappa\lambda) = \displaystyle\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\chi_\mu(\mathbf{r})\chi_\nu(\mathbf{r})\chi_\kappa(\mathbf{r}')\chi_\lambda(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}\, d\mathbf{r}\, d\mathbf{r}'$

Important property: for all $D, D' \in \mathbb{R}^{N_b \times N_b}_{\mathrm{sym}}$**,** $\mathbf{Tr}(G(D)D') = \mathbf{Tr}(G(D')D)$**.**

For simplicity, we assume from now on that the basis $\{\chi_\mu\}_{1\leq\mu\leq N_b}$ is orthonormal. One can always get back to this case by the following changes of variables:

$$\widetilde{C} = S^{1/2}C, \qquad \widetilde{D} = \widetilde{C}\widetilde{C}^T = S^{1/2}CC^TS^{1/2} = S^{1/2}DS^{1/2}.$$

For simplicity, we assume from now on that the basis $\{\chi_\mu\}_{1\leq\mu\leq N_b}$ is orthonormal. One can always get back to this case by the following changes of variables:

$$\widetilde{C} = S^{1/2}C, \qquad \widetilde{D} = \widetilde{C}\widetilde{C}^T = S^{1/2}CC^TS^{1/2} = S^{1/2}DS^{1/2}.$$

Molecular orbital formulation of the HF problem

$$E^{\rm HF}_{0,\mathcal{X}} = \inf\left\{E(CC^T),\ C\in\mathcal{C}\right\}$$

$$\mathcal{C} = \left\{C\in\mathbb{R}^{N_b\times N},\ C^TC = I_N\right\} \qquad \textbf{(Stiefel manifold)}$$

$$E(D) = \mathbf{Tr}(hD) + \frac{1}{2}\mathbf{Tr}(G(D)D)$$

↑ linear ↑ quadratic

Density matrix formulation of the HF problem

When C varies in the set $\mathcal{C} = \left\{C \in \mathbb{R}^{N_b \times N},\ C^T C = I_N\right\}$, $D = CC^T$ spans

$$\mathcal{P} = \left\{D \in \mathbb{R}^{N_b \times N_b},\ D = D^T,\ \mathbf{Tr}(D) = N,\ D^2 = D\right\}$$

$$= \left\{ \text{ rank-}N \text{ orthogonal projectors of } \mathbb{R}^{N_b \times N_b} \right\} \qquad \textbf{(Grassmann manifold)}$$

Density matrix formulation of the HF problem

When C varies in the set $\mathcal{C} = \left\{C \in \mathbb{R}^{N_b \times N},\ C^T C = I_N\right\}$, $D = CC^T$ spans

$$\mathcal{P} = \left\{D \in \mathbb{R}^{N_b \times N_b},\ D = D^T,\ \mathbf{Tr}(D) = N,\ D^2 = D\right\}$$

$$= \left\{ \text{ rank-}N \text{ orthogonal projectors of } \mathbb{R}^{N_b \times N_b} \right\} \qquad \textbf{(Grassmann manifold)}$$

Therefore,

$$E_{0,\mathcal{X}}^{\mathrm{HF}} = \inf\left\{E(D),\ D \in P\right\},$$

$$E(D) = \mathbf{Tr}(hD) + \frac{1}{2}\mathbf{Tr}(G(D)D)$$

↑ **linear** ↑ **quadratic**

Discretized Hartree-Fock equations

$$\begin{cases} D = CC^T, \quad F = h + G(D) \\ F\Phi_i = \varepsilon_i \Phi_i, \quad \varepsilon_1 \leq \cdots \leq \varepsilon_N \textbf{ lowest } N \textbf{ eigenvalues of } F \\ C = (\Phi_1 \cdots \Phi_N), \quad \Phi_i^T \Phi_j = \delta_{ij} \end{cases}$$

$$D \in \mathbb{R}^{N_b \times N_b}_{\text{sym}}, \quad F \in \mathbb{R}^{N_b \times N_b}_{\text{sym}}, \quad \Phi_i \in \mathbb{R}^{N_b}, \quad C \in \mathbb{R}^{N_b \times N}$$

Discretized Hartree-Fock equations

$$\begin{cases} D = CC^T, \quad F = h + G(D) \\ \\ F\Phi_i = \varepsilon_i \Phi_i, \quad \varepsilon_1 \leq \cdots \leq \varepsilon_N \textbf{ lowest } N \textbf{ eigenvalues of } F \\ \\ C = (\Phi_1 \cdots \Phi_N), \quad \Phi_i^T \Phi_j = \delta_{ij} \end{cases}$$

$$D \in \mathbb{R}^{N_b \times N_b}_{\text{sym}}, \quad F \in \mathbb{R}^{N_b \times N_b}_{\text{sym}}, \quad \Phi_i \in \mathbb{R}^{N_b}, \quad C \in \mathbb{R}^{N_b \times N}$$

Solutions to the discretized Hartree-Fock problem can be obtained

- **either by solving a constrained optimization problem (on a Stiefel or a Grassmann manifold);**
- **or by solving the above equations by means of a self-consistent field (SCF) algorithm.**

The design of more efficient methods, in particular for very large molecular systems, is still an active field of research.

3 - SCF algorithms

Let us denote by $F(D) = h + G(D)$ the Fock matrix, i.e. the gradient of

$$E(D) = \mathbf{Tr}(hD) + \frac{1}{2}\mathbf{Tr}(G(D)D),$$

when $\mathbb{R}^{N_b \times N_b}_{\mathrm{sym}}$ is endowed with the Frobenius inner product:

$$(D, D')_{\mathrm{F}} = \mathbf{Tr}(D^T D') = \mathbf{Tr}(DD') = \sum_{\mu\nu=1}^{N_b} D_{\mu\nu} D'_{\mu\nu}.$$

Let us denote by $F(D) = h + G(D)$ the Fock matrix, i.e. the gradient of

$$E(D) = \mathbf{Tr}(hD) + \frac{1}{2}\mathbf{Tr}(G(D)D),$$

when $\mathbb{R}^{N_b \times N_b}_{\mathrm{sym}}$ is endowed with the Frobenius inner product:

$$(D, D')_{\mathrm{F}} = \mathbf{Tr}(D^T D') = \mathbf{Tr}(DD') = \sum_{\mu\nu=1}^{N_b} D_{\mu\nu} D'_{\mu\nu}.$$

Necessary conditions for D being a minimizer of the HF energy

$$\begin{cases} D = \displaystyle\sum_{i=1}^{N} \Phi_i \Phi_i^T \\ \\ F(D)\Phi_i = \varepsilon_i \, \Phi_i \\ \\ \Phi_i^T \Phi_j = \delta_{ij} \\ \\ \varepsilon_1 \leq \varepsilon_2 \leq \cdots \leq \varepsilon_N \text{ are the lowest } N \text{ eigenvalues of } F(D) \end{cases}$$

Lemma. **Let $H \in \mathbb{R}_{\mathrm{sym}}^{N_b \times N_b}$ be a symmetric matrix such that $\varepsilon_N < \varepsilon_{N+1}$, where ε_j is the j^{th} eigenvalue of H counting multiplicities.**

Lemma. **Let $H \in \mathbb{R}^{N_b \times N_b}_{\text{sym}}$ be a symmetric matrix such that $\varepsilon_N < \varepsilon_{N+1}$, where ε_j is the j^{th} eigenvalue of H counting multiplicities.**

Then, the solution D to the following problem

$$\begin{cases} D = \displaystyle\sum_{i=1}^{N} \Phi_i \Phi_i^T \\ H\Phi_i = \varepsilon_i \, \Phi_i \\ \Phi_i^T \Phi_j = \delta_{ij} \\ \varepsilon_1 \leq \varepsilon_2 \leq \cdots \leq \varepsilon_N \text{ are the lowest } N \text{ eigenvalues of } H \end{cases}$$

is unique and it holds

$$\boxed{D = \mathbb{1}_{(-\infty, \varepsilon_{\text{F}}]}(H)}$$

where ε_{F} is any number in the open interval $(\varepsilon_N, \varepsilon_{N+1})$.

Lemma. **Let $H \in \mathbb{R}^{N_b \times N_b}_{\text{sym}}$ be a symmetric matrix such that $\varepsilon_N < \varepsilon_{N+1}$, where ε_j is the j^{th} eigenvalue of H counting multiplicities.**

Then, the solution D to the following problem

$$\begin{cases} D = \displaystyle\sum_{i=1}^{N} \Phi_i \Phi_i^T \\ H\Phi_i = \varepsilon_i \, \Phi_i \\ \Phi_i^T \Phi_j = \delta_{ij} \\ \varepsilon_1 \leq \varepsilon_2 \leq \cdots \leq \varepsilon_N \text{ are the lowest } N \text{ eigenvalues of } H \end{cases}$$

is unique and it holds

$$\boxed{D = \mathbb{1}_{(-\infty, \varepsilon_{\text{F}}]}(H)}$$

where ε_{F} is any number in the open interval $(\varepsilon_N, \varepsilon_{N+1})$.

In addition,

$$\boxed{D = \textbf{argmin}\left\{\textbf{Tr}(H D'),\ D' \in \mathcal{P}\right\}}$$

Roothaan algorithm

$$\begin{cases} F(D_k)\Phi_i^{k+1} = \varepsilon_i^{k+1}\,\Phi_i^{k+1} \\ \Phi_i^{k+1\,T}\Phi_j^{k+1} = \delta_{ij} \\ \varepsilon_1^{k+1} \le \varepsilon_2^{k+1} \le \cdots \le \varepsilon_N^{k+1} \text{ are the lowest } N \text{ eigenvalues of } F(D^k) \\ D_{k+1} = \displaystyle\sum_{i=1}^{N} \Phi_i^{k+1}\Phi_i^{k+1\,T} \end{cases}$$

Roothaan algorithm

$$\begin{cases} F(D_k)\Phi_i^{k+1} = \varepsilon_i^{k+1}\,\Phi_i^{k+1} \\ \Phi_i^{k+1\,T}\Phi_j^{k+1} = \delta_{ij} \\ \varepsilon_1^{k+1} \leq \varepsilon_2^{k+1} \leq \cdots \leq \varepsilon_N^{k+1} \text{ are the lowest } N \text{ eigenvalues of } F(D^k) \\ D_{k+1} = \displaystyle\sum_{i=1}^{N} \Phi_i^{k+1}\Phi_i^{k+1\,T} \end{cases}$$

$$\boxed{D_{k+1} \in \textbf{argmin}\,\{\textbf{Tr}(F(D_k)D'),\ D' \in \mathcal{P}\}}$$

Roothaan algorithm

$$\begin{cases} F(D_k)\Phi_i^{k+1} = \varepsilon_i^{k+1}\,\Phi_i^{k+1} \\ \Phi_i^{k+1\,T}\Phi_j^{k+1} = \delta_{ij} \\ \varepsilon_1^{k+1} \leq \varepsilon_2^{k+1} \leq \cdots \leq \varepsilon_N^{k+1} \text{ are the lowest } N \text{ eigenvalues of } F(D^k) \\ D_{k+1} = \displaystyle\sum_{i=1}^{N} \Phi_i^{k+1}\Phi_i^{k+1\,T} \end{cases}$$

Theorem (E.C. & Le Bris '00, Levitt '12). **U.s.t.a., the sequence (D_k) generated by the Roothaan algorithm satisfies one of the following properties:**

- **either (D_k) converges towards an *Aufbau* solution to the HF equations;**
- **or (D_k) oscillates between two states, none of them being an *Aufbau* solution to the HF equations.**

Idea of the proof. **The sequence $(D_k)_{k\in\mathbb{N}}$ generated by the Roothaan algorithm coincides with the sequence $(D_k^{\mathrm{rel}})_{k\in\mathbb{N}}$ obtained by minimizing**

$$\mathcal{E}(D, D') = \mathbf{Tr}(hD) + \mathbf{Tr}(hD') + \mathbf{Tr}(G(D)D'),$$

on $\mathcal{P} \times \mathcal{P}$ starting from $D_0^{\mathrm{rel}} = D_0$, using the relaxation algorithm

Idea of the proof. **The sequence $(D_k)_{k\in\mathbb{N}}$ generated by the Roothaan algorithm coincides with the sequence $(D_k^{\rm rel})_{k\in\mathbb{N}}$ obtained by minimizing**

$$\mathcal{E}(D, D') = \mathbf{Tr}(hD) + \mathbf{Tr}(hD') + \mathbf{Tr}(G(D)D'),$$

on $\mathcal{P}\times\mathcal{P}$ starting from $D_0^{\rm rel} = D_0$, using the relaxation algorithm

$$\begin{aligned} D_1^{\rm rel} &= \textbf{arg inf}\,\{\mathcal{E}(D_0, D),\ D\in\mathcal{P}\} \\ &= \textbf{arg inf}\,\{\mathbf{Tr}(hD_0) + \mathbf{Tr}(hD) + \mathbf{Tr}(G(D_0)D),\ D\in\mathcal{P}\} \end{aligned}$$

Idea of the proof. **The sequence $(D_k)_{k\in\mathbb{N}}$ generated by the Roothaan algorithm coincides with the sequence $(D_k^{\rm rel})_{k\in\mathbb{N}}$ obtained by minimizing**

$$\mathcal{E}(D, D') = \mathbf{Tr}(hD) + \mathbf{Tr}(hD') + \mathbf{Tr}(G(D)D'),$$

on $\mathcal{P}\times\mathcal{P}$ starting from $D_0^{\rm rel} = D_0$, using the relaxation algorithm

$$\begin{aligned} D_1^{\rm rel} &= \textbf{arg inf}\,\{\mathcal{E}(D_0, D),\ D\in\mathcal{P}\} \\ &= \textbf{arg inf}\,\{\mathbf{Tr}(hD_0) + \mathbf{Tr}(hD) + \mathbf{Tr}(G(D_0)D),\ D\in\mathcal{P}\} \\ &= \textbf{arg inf}\,\{\mathbf{Tr}(F(D_0)D),\ D\in\mathcal{P}\} \end{aligned}$$

Idea of the proof. **The sequence $(D_k)_{k\in\mathbb{N}}$ generated by the Roothaan algorithm coincides with the sequence $(D_k^{\rm rel})_{k\in\mathbb{N}}$ obtained by minimizing**

$$\mathcal{E}(D, D') = \mathbf{Tr}(hD) + \mathbf{Tr}(hD') + \mathbf{Tr}(G(D)D'),$$

on $\mathcal{P}\times\mathcal{P}$ starting from $D_0^{\rm rel} = D_0$, using the relaxation algorithm

$$\begin{aligned} D_1^{\rm rel} &= \mathbf{arg\ inf}\,\{\mathcal{E}(D_0, D),\ D\in\mathcal{P}\} \\ &= \mathbf{arg\ inf}\,\{\mathbf{Tr}(hD_0) + \mathbf{Tr}(hD) + \mathbf{Tr}(G(D_0)D),\ D\in\mathcal{P}\} \\ &= \mathbf{arg\ inf}\,\{\mathbf{Tr}(F(D_0)D),\ D\in\mathcal{P}\} \\ &= D_1, \end{aligned}$$

Idea of the proof. **The sequence $(D_k)_{k\in\mathbb{N}}$ generated by the Roothaan algorithm coincides with the sequence $(D_k^{\rm rel})_{k\in\mathbb{N}}$ obtained by minimizing**

$$\mathcal{E}(D,D') = \mathbf{Tr}(hD) + \mathbf{Tr}(hD') + \mathbf{Tr}(G(D)D'),$$

on $\mathcal{P}\times\mathcal{P}$ starting from $D_0^{\rm rel} = D_0$, using the relaxation algorithm

$$\begin{aligned} D_1^{\rm rel} &= \mathbf{arg\ inf}\,\{\mathcal{E}(D_0,D),\ D\in\mathcal{P}\} \\ &= \mathbf{arg\ inf}\,\{\mathbf{Tr}(hD_0) + \mathbf{Tr}(hD) + \mathbf{Tr}(G(D_0)D),\ D\in\mathcal{P}\} \\ &= \mathbf{arg\ inf}\,\{\mathbf{Tr}(F(D_0)D),\ D\in\mathcal{P}\} \\ &= D_1, \end{aligned}$$

$$\begin{aligned} D_2^{\rm rel} &= \mathbf{arg\ inf}\,\{\mathcal{E}(D,D_1),\ D\in\mathcal{P}\} \\ &= \mathbf{arg\ inf}\,\{\mathcal{E}(D_1,D),\ D\in\mathcal{P}\} \\ &= \mathbf{arg\ inf}\,\{\mathbf{Tr}(hD) + \mathbf{Tr}(hD_1) + \mathbf{Tr}(G(D_1)D),\ D\in\mathcal{P}\} \\ &= \mathbf{arg\ inf}\,\{\mathbf{Tr}(F(D_1)D),\ D\in\mathcal{P}\} \\ &= D_2, \end{aligned}$$

Idea of the proof. **The sequence $(D_k)_{k\in\mathbb{N}}$ generated by the Roothaan algorithm coincides with the sequence $(D_k^{\rm rel})_{k\in\mathbb{N}}$ obtained by minimizing**

$$\mathcal{E}(D, D') = \mathbf{Tr}(hD) + \mathbf{Tr}(hD') + \mathbf{Tr}(G(D)D'),$$

on $\mathcal{P}\times\mathcal{P}$ starting from $D_0^{\rm rel} = D_0$, using the relaxation algorithm

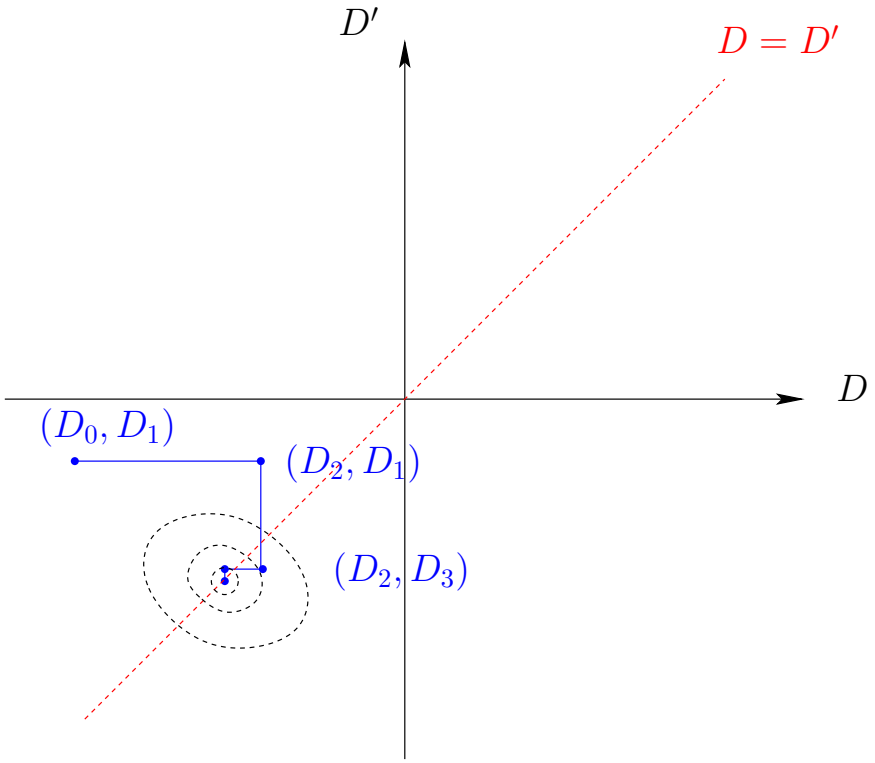
$$\begin{aligned} D_1^{\rm rel} &= \mathbf{arg\ inf}\,\{\mathcal{E}(D_0, D),\ D\in\mathcal{P}\} \\ &= \mathbf{arg\ inf}\,\{\mathbf{Tr}(hD_0) + \mathbf{Tr}(hD) + \mathbf{Tr}(G(D_0)D),\ D\in\mathcal{P}\} \\ &= \mathbf{arg\ inf}\,\{\mathbf{Tr}(F(D_0)D),\ D\in\mathcal{P}\} \\ &= D_1, \end{aligned}$$

$$\begin{aligned} D_2^{\rm rel} &= \mathbf{arg\ inf}\,\{\mathcal{E}(D, D_1),\ D\in\mathcal{P}\} \\ &= \mathbf{arg\ inf}\,\{\mathcal{E}(D_1, D),\ D\in\mathcal{P}\} \\ &= \mathbf{arg\ inf}\,\{\mathbf{Tr}(hD) + \mathbf{Tr}(hD_1) + \mathbf{Tr}(G(D_1)D),\ D\in\mathcal{P}\} \\ &= \mathbf{arg\ inf}\,\{\mathbf{Tr}(F(D_1)D),\ D\in\mathcal{P}\} \\ &= D_2, \end{aligned}$$

...

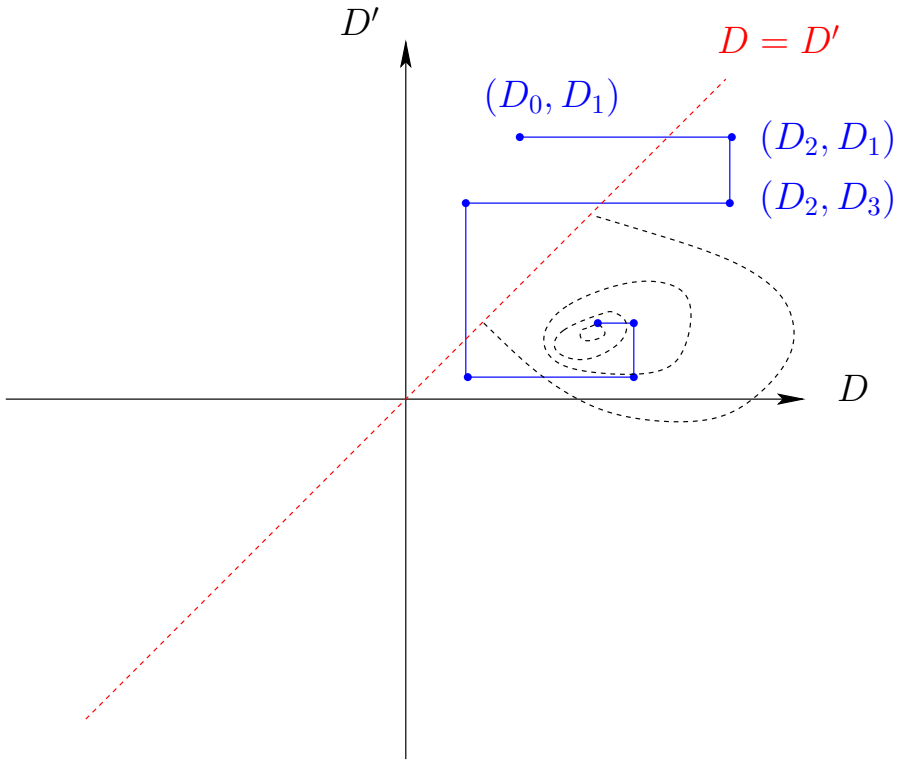
It follows that the sequence $(D_{2k}, D_{2k+1})_{k\in\mathbb{N}}$ converges to a local minimizer $(D_{\text{even}}, D_{\text{odd}})$ of $\mathcal{E}(D, D')$ on $\mathcal{P} \times \mathcal{P}$

Case $D_{\text{even}} = D_{\text{odd}}$



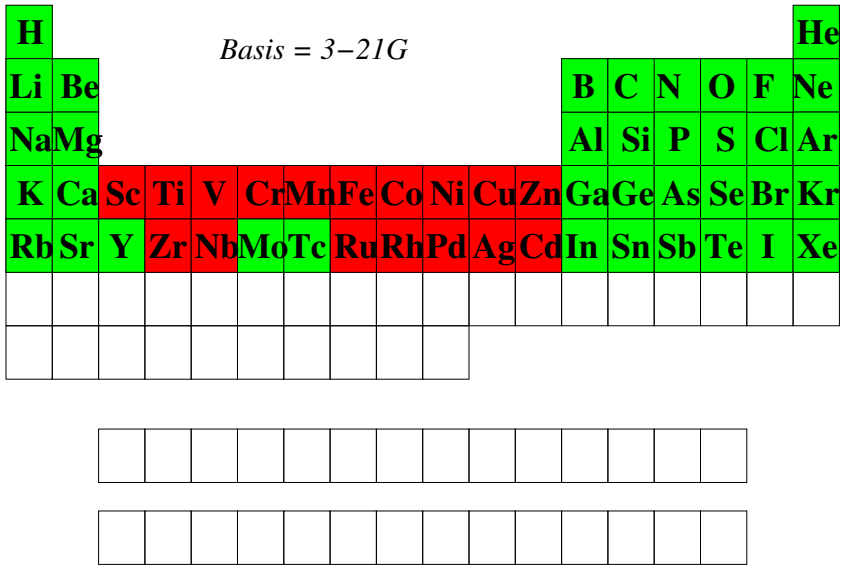
The Roothaan alg. converges

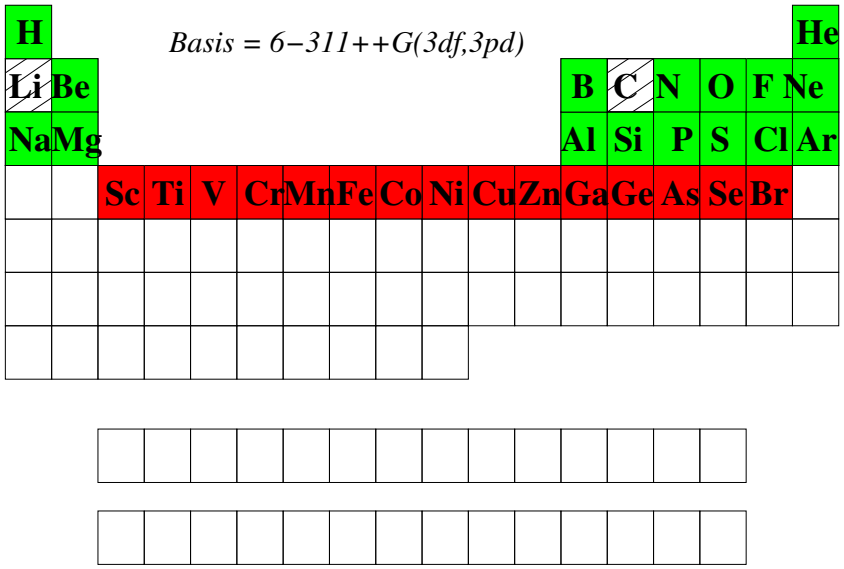
Case $D_{\text{even}} \neq D_{\text{odd}}$



The Roothaan alg. oscillates between two states: charge sloshing phenomenon

Ground state calculations of atoms with the Roothaan algorithm





Convergence to the ground state

Convergence to another solution to the HF equations

Oscillation

Not available

Calculations performed with the DIIS algorithm (Pulay 1982, default algorithm in most quantum chemistry codes until 2000)

System	Energy of D^0 (Ha)	Energy at convergence (Ha)
CH_3-NH-CH=CH-NO_2 6-31G	-374.0038 -322.2373	-375.3869 Does not CV
Cr_2 6-31G	-2069.5400 -2051.4339	-2085.5449 -2085.4042
$[Fe(H_2O)_6]^{2+}$ 178 AO	-1700.7596 -1538.7283	-1717.8928 -1717.7355

Relaxed Constrained Algorithm (EC, Le Bris, 2000)

Replace the Hartree-Fock problem

$$\inf \{E(D),\ D \in \mathcal{P}\} \tag{1}$$

$$\mathcal{P} = \left\{D \in \mathbb{R}^{N_b \times N_b},\ D^T = D,\ \mathbf{Tr}(D) = N,\ D^2 = D\right\}$$

by

$$\inf \left\{E(D),\ D \in \widetilde{\mathcal{P}}\right\} \tag{2}$$

$$\widetilde{\mathcal{P}} = \left\{D \in \mathbb{R}^{N_b \times N_b},\ D^T = D,\ \mathbf{Tr}(D) = N,\ D^2 \leq D\right\}$$

Fundamental property : (1) and (2) have the same local minima
(discrete counterpart of Lieb's variational principle, Lieb '81).

What is gained : the set $\widetilde{\mathcal{P}}$ is convex

Proof. **Assume that $\widetilde{D}$ is a minimizer of E on $\widetilde{\mathcal{P}}$ that does not verify the constraint $\widetilde{D}^2 = \widetilde{D}$.**

The optimality conditions lead to

$$\widetilde{D} = \sum_{\varepsilon_i < \varepsilon_F} \Phi_i \Phi_i^T + \sum_{\varepsilon_j = \varepsilon_F} n_j \Phi_j \Phi_j^T \qquad \textbf{with } 0 \leq n_j \leq 1.$$

Let Φ and Φ' two partially occupied orbitals ($0 < n, n' < 1$).
By transfering $0 < \delta n \ll 1$ electron from Φ to Φ', one obtains

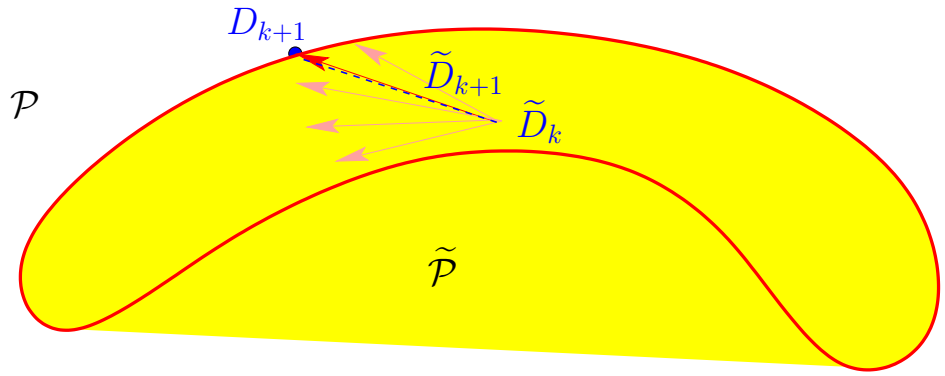
$$\widetilde{D}' = \widetilde{D} + \delta n \left(\Phi' \Phi'^T - \Phi \Phi^T \right) \quad \in \widetilde{\mathcal{P}}$$

and

$$\Delta E = E(\widetilde{D}') - E(\widetilde{D}) = -\frac{\delta n^2}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{|\phi(\mathbf{r})\,\phi'(\mathbf{r}') \;-\; \phi(\mathbf{r}')\,\phi'(\mathbf{r})|^2}{|\mathbf{r} - \mathbf{r}'|}\, d\mathbf{r}\, d\mathbf{r}' < 0,$$

where $\phi(\mathbf{r}) = \sum_{\mu=1}^{N_b} \Phi_\mu \chi_\mu(\mathbf{r})$ and $\phi'(\mathbf{r}) = \sum_{\mu=1}^{N_b} \Phi'_\mu \chi_\mu(\mathbf{r})$.

Optimal Damping Algorithm (ODA)



"Optimal step gradient" :

1. Calculation of a descent direction $d = D^{k+1} - \widetilde{D}^k$ **where**

$$D_{k+1} = \textbf{arginf} \left\{ \frac{d}{d\lambda} E\left(\widetilde{D}_k + t(\widetilde{D} - \widetilde{D}_k)\right)\Big|_{t=0}, \ \widetilde{D} \in \widetilde{\mathcal{P}} \right\} ;$$

2. Line search: set $\widetilde{D}_{k+1} = \textbf{arginf} \left\{ E(\widetilde{D}), \ \widetilde{D} \in \textbf{Seg}[\widetilde{D}_k, D_{k+1}] \right\}$ **where**

$$\textbf{Seg}[\widetilde{D}_k, D_{k+1}] = \left\{ (1-t)\widetilde{D}_k + \lambda D_{k+1}, \ t \in [0,1] \right\}.$$

Since $E(D)$ **is a second-degree polynomial in** D, $E((1-t)\widetilde{D}_k + \lambda D_{k+1})$ **is a second-degree polynomial in** t**: the line search pb can be easily solved.**

ODA steepest descent calculation

$$
\begin{aligned}
D_{k+1} &= \textbf{arginf} \left\{ \frac{d}{d\lambda} E\left(\widetilde{D}_k + \lambda(\widetilde{D} - \widetilde{D}_k) \right) \Big|_{\lambda=0}, \ \widetilde{D} \in \widetilde{\mathcal{P}} \right\} \\
&= \textbf{arginf} \Big\{ \textbf{Tr}(F(\widetilde{D}_k) D), \ D \in \mathcal{P} \Big\}
\end{aligned}
$$

$$
\begin{cases}
D_{k+1} = \displaystyle\sum_{i=1}^{N} \Phi_i^{k+1} \, \Phi_i^{k+1\,T} \\
F(\widetilde{D}_k) \Phi_i^{k+1} = \varepsilon_i^{k+1} \, \Phi_i^{k+1} \\
\Phi_i^{k+1\,T} \Phi_j^{k+1} = \delta_{ij} \\
\varepsilon_1^{k+1} \leq \varepsilon_2^{k+1} \leq \cdots \leq \varepsilon_N^{k+1} \ \textbf{are the lowest } N \textbf{ eigenvalues of } F(\widetilde{D}_k)
\end{cases}
$$

Comparison between DIIS (defaut algorithm in Gaussian 98) and ODA.

System	$E^{\mathrm{RHF}}(D_0)$	DIIS (Ha)	ODA (Ha)	ΔE (kcal/mol)
CH_3-NH-CH=CH-NO_2	-374.0038	-375.3869	-375.3869	0
6-31G	-322.2373	Does not CV	-375.3869	-
Cr_2	-2069.5400	-2085.5449	-2085.8060	163.71
6-31G	-2051.4339	-2085.4042	-2085.8060	251.93
$[Fe(H_2O)_6]^{2+}$	-1700.7596	-1717.8928	-1718.0151	76.68
178 AO	-1538.7283	-1717.7355	-1718.0151	175.31

DIIS (Pulay, 1981) vs EDIIS (EC, Kudin, Scuseria, 2002)

- **DIIS (Direct Inversion in the Iterative Space)**

$$\begin{cases} D_{k+1} = \textbf{arginf}\left\{\textbf{Tr}(F(\widetilde{D}_k)D),\ D \in \mathcal{P}\right\} \\ \widetilde{D}_{k+1} = \displaystyle\sum_{i=0}^{k+1} c_i^{\mathrm{DIIS}} D_i, \qquad (c_i^{\mathrm{DIIS}}) = \textbf{argmin}\left\{ \left\| \sum_{i=0}^{k+1} c_i [F(D_i), D_i] \right\|^2,\ \sum_{i=0}^{k+1} c_i = 1 \right\} \end{cases}$$

- **EDIIS (Energy DIIS)**

$$\begin{cases} D_{k+1} = \textbf{arginf}\left\{\textbf{Tr}(F(\widetilde{D}_k)D),\ D \in \mathcal{P}\right\} \\ \widetilde{D}_{k+1} = \textbf{argmin}\left\{ E(\widetilde{D}),\ \widetilde{D} = \displaystyle\sum_{i=0}^{k+1} c_i D_i,\ 0 \le c_i \le 1,\ \sum_{i=0}^{k+1} c_i = 1 \right\} \end{cases}$$

The default algorithm is Gaussian consists in first iterating with EDIIS, and switching to DIIS when some convergence criterion is met.

Appendix: Constrained optimization and Lagrange multipliers

For brevity, we will limit ourselves to the setting of real Hilbert spaces.

All the results presented here can be extended to complex Hilbert spaces.

Definition (derivative of a function at a point). **Let V, W be Hilbert spaces, $F : V \to W$, and $v \in V$. The function F is called differentiable at v, if there exists a continuous linear map $d_v F : V \to W$ such that in the vicinity of v,**

$$F(v+h) = F(v) + d_v F(h) + o(h),$$

which means

$$\forall \varepsilon > 0,\ \exists \eta > 0 \text{ s.t. } \forall h \in V \text{ s.t. } \|h\|_V \le \eta, \quad \|F(v+h) - F(v) - d_v F(h)\|_W \le \varepsilon \|h\|_V.$$

If such a linear map $d_v F$ exists, it is unique. It is called the derivative of F at v.

Definition (differentiable and C^1 functions). **F is called differentiable if F is differentiable at each point of V. In this case, the mapping**

$$\begin{array}{rcl} dF\ :\ V & \longrightarrow & \mathcal{B}(V;W) \\ v & \mapsto & d_v F \end{array}$$

is called the derivative of F. F is called of class C^1 on V if dF is continuous.

Remark. **One can similarly define the derivative of a function $F : U \to W$, where U is an open subset of V (that is $U = V \setminus F$, where F is a closed subset of V).**

Theorem (Riesz). **Let V be a Hilbert space and $l : V \to \mathbb{R}$ a continuous linear map. Then there exists a unique vector $w \in V$ such that**

$$\forall v \in V, \quad l(v) = (w, v)_V.$$

Definition (gradient). **Let V be a Hilbert space endowed with the inner product $(\cdot, \cdot)_V$, U an open subset of V and $E : U \to \mathbb{R}$ a function differentiable at $v \in U$.**

The unique vector of V denoted by $\nabla E(v)$ and defined by

$$\forall h \in V, \qquad d_v E(h) = (\nabla E(v), h)_V \qquad \text{(by means of Riesz theorem)}$$

is called the gradient of E at v.

Gradient of a function $E : \mathbb{R}^d \to \mathbb{R}$

The above abstract definition of the gradient agrees with the usual one when $V = \mathbb{R}^d$ endowed with the Euclidean inner product:

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

with

$$\nabla E(\mathbf{x}) = \begin{pmatrix} \dfrac{\partial E}{\partial x_1}(\mathbf{x}) \\ \cdot \\ \cdot \\ \cdot \\ \dfrac{\partial E}{\partial x_d}(\mathbf{x}) \end{pmatrix}.$$

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 E, which we will denote by $\nabla_S E(\mathbf{x})$, is related to the Euclidean gradient $\nabla E(\mathbf{x})$ by

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

Geometric interpretation of the gradient

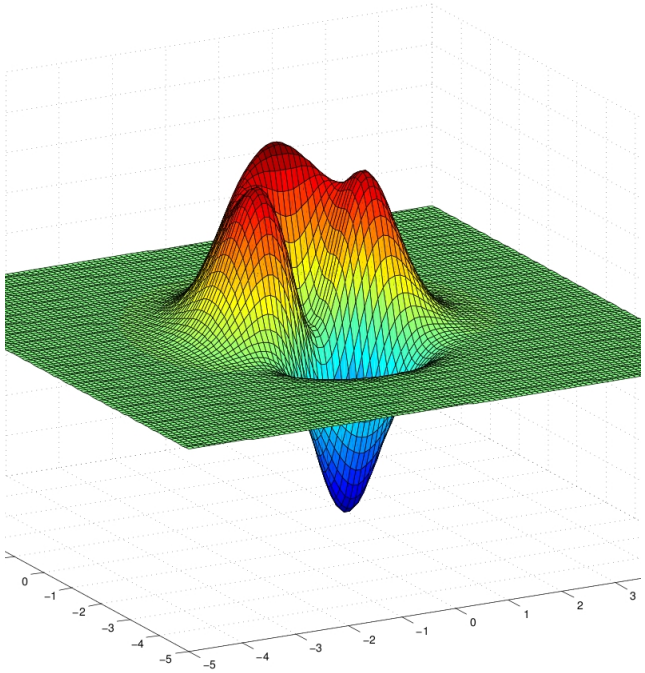
Let $E : V \to \mathbb{R}$ of class C^1, $v \in V$ and $\alpha = E(v)$. If $\nabla E(v) \neq 0$, then

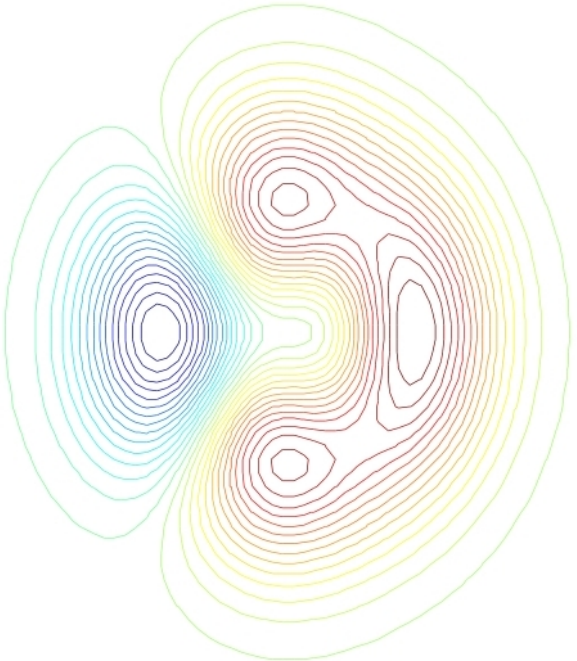
- **in the vicinity of v, the level set**

$$\mathcal{C}_\alpha := \{w \in V \mid \ E(w) = \alpha\}$$

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

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





Equality constrained optimization. **Let V and W be Hilbert spaces s.t. $\dim(W) < \infty$, $E : V \to \mathbb{R}$, $g : V \to W$. Consider the optimization problem**

$$\inf_{v \in K} E(v) \quad \textbf{where} \quad K = \{v \in V \mid g(v) = 0\}.$$

Definition (qualification of the constraints). **The equality constraints $g = 0$ are called qualified at $u \in K$ if $d_u g : V \to W$ is surjective (i.e. $\mathrm{Ran}(d_u g) = W$).**

Theorem (Euler-Lagrange theorem). **Let $u \in K$ be a local minimum of E on**

$$K = \{v \in V \mid g(v) = 0\}.$$

Assume that

1. **E is differentiable at u and g is C^1 in the vicinity of u;**
2. **the equality constraint $g(v) = 0$ is qualified at u.**

Then, there exists a unique $\lambda \in W$ such that

$$\forall h \in V,\ d_u E(h) = (\lambda, d_u g(h))_W \quad \textbf{or equivalently} \quad \nabla E(u) = d_u g^*(\lambda),$$

where $d_u g^*$ is the adjoint of $d_u g$. 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

$$\begin{cases} \textbf{seek } (u,\lambda) \in V \times W \textbf{ such that} \\ \nabla E(u) - d_u g^*(\lambda) = 0 \\ g(u) = 0 \end{cases} \tag{3}$$

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

The solutions of the Euler-Lagrange equations (3) are called the critical points of E on K.

Remark : if $\dim(V) = d < \infty$ and $\dim(W) = m < \infty$, then the above problem consists of $(d+m)$ scalar equations with $(d+m)$ scalar unknowns.

Remark. Equations (3) are equivalent to seeking $(u,\lambda) \in V \times W$ such that

$$\frac{\partial L}{\partial v}(u,\lambda) = 0, \ \frac{\partial L}{\partial \mu}(u,\lambda) = 0, \quad \textbf{where} \quad L(v,\mu) := E(v) - (\mu, g(v))_W \quad \textbf{(Lagrangian)}.$$

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!

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!

Example: $d = 1$, $m = 1$, $E(x) = x$, $g(x) = x^2$. **Then**

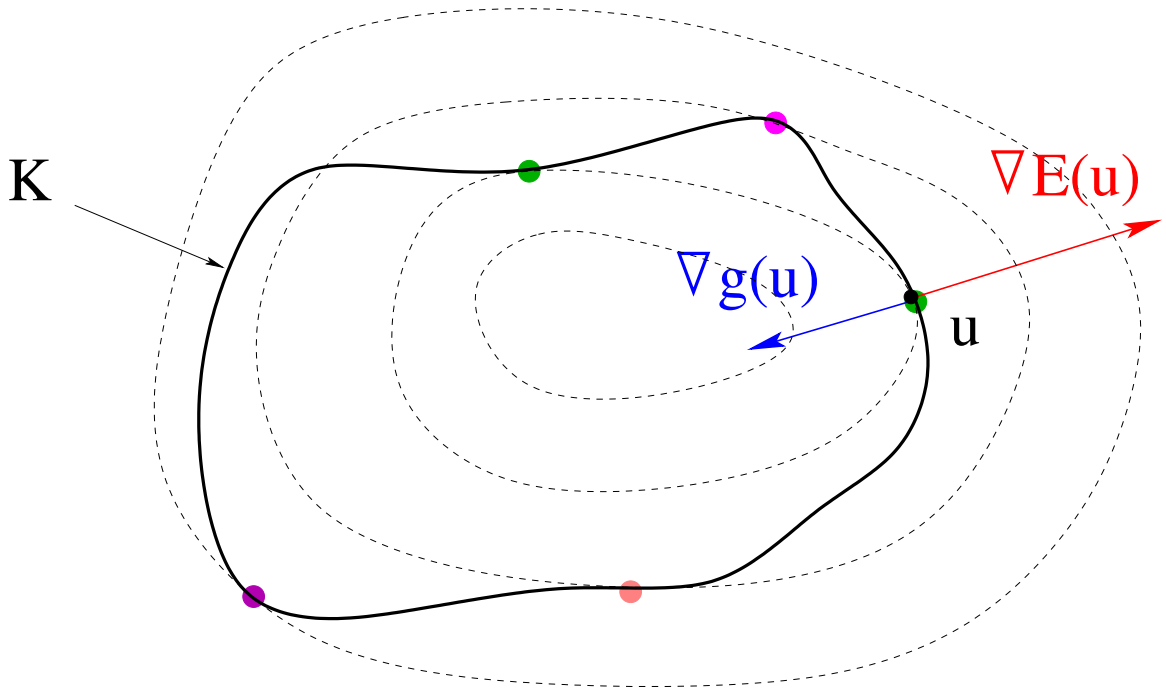
$$K = \{x \in \mathbb{R} \mid g(x) = 0\} = \{0\} \qquad \textbf{and} \qquad g'(0) = 0.$$

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 ($V = \mathbb{R}^2$, $W = \mathbb{R}$)



On $K = g^{-1}(0) = \{v \in V \mid g(v) = 0\}$, 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 u be a local minimizer of E on $K = g^{-1}(0) = \{v \in V \mid g(v) = 0\}$ and $\alpha = E(u)$.**

- **If the constraint $g = 0$ is qualified at u (i.e. if $d_u g : \mathcal{H} \to \mathcal{K}$ is surjective), then, in the vicinity of u, K is a C^1 manifold with tangent space**

$$T_u K = \{h \in \mathcal{H} \mid d_u g(h) = 0\} = \mathbf{Ker}(d_u g).$$

- **Since u is a minimizer of E on K, the vector $\nabla E(u)$ must be orthogonal to $T_u K$. Indeed, for any $h \in T_u K$, there exists a C^1 curve $\phi : [-1, 1] \to V$ drawn on K such that $\phi(0) = u$ et $\phi'(0) = h$, and we have**

$$0 \leq E(\phi(t)) - E(u) = E(u + th + o(t)) - E(u) = t\nabla E(u) \cdot h + o(t).$$

- **We have**

$$\nabla E(u) \in (T_u K)^\perp = (\mathbf{Ker}(d_u g))^\perp = \overline{\mathbf{Ran}(d_u g^*)} = \mathbf{Ran}(d_u g^*) \text{ since } \mathbf{dim}(W) < \infty.$$

- **Therefore, there exists $\lambda \in W$ such that $\nabla E(u) = d_u g^*(\lambda)$.**

Remark: a Lagrange multiplier often has 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**:

  $\rightarrow$ 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 ($\mathrm{div}\,(u) = 0$)

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

Analytical derivatives

$$\forall \mathbf{x} \in \mathbb{R}^d, \quad W(\mathbf{x}) = \inf \{E(\mathbf{x}, v),\ v \in V,\ g(\mathbf{x}, v) = 0\} \tag{4}$$

with $E : \mathbb{R} \times V \to \mathbb{R}$, $g : \mathbb{R} \times V \to W$, V, W **Hilbert spaces, dim**$(W) < \infty$**.**

Assume (4) has a unique minimizer $v(\mathbf{x})$ **and** $\mathbf{x} \mapsto v(\mathbf{x})$ **is regular. Then,**

$$W(\mathbf{x}) = E(\mathbf{x}, v(\mathbf{x})) \quad \Rightarrow \quad \frac{\partial W}{\partial x_i}(\mathbf{x}) = \frac{\partial E}{\partial x_i}(\mathbf{x}, v(\mathbf{x})) + \frac{\partial E}{\partial v}(\mathbf{x}, v(\mathbf{x})) \left(\frac{\partial v}{\partial x_i}(\mathbf{x}) \right),$$

$$g(\mathbf{x}, v(\mathbf{x})) = 0 \quad \Rightarrow \quad \frac{\partial g}{\partial x_i}(\mathbf{x}, v(\mathbf{x})) + \frac{\partial g}{\partial v}(\mathbf{x}, v(\mathbf{x})) \left(\frac{\partial v}{\partial x_i}(\mathbf{x}) \right) = 0.$$

Euler-Lagrange equation: $\forall h \in V, \quad \dfrac{\partial E}{\partial v}(\mathbf{x}, v(\mathbf{x}))\,(h) = \left(\dfrac{\partial g}{\partial v}(\mathbf{x}, v(\mathbf{x}))(h), \lambda(\mathbf{x}) \right)_W.$

Therefore $\dfrac{\partial W}{\partial x_i}(\mathbf{x}) = \dfrac{\partial E}{\partial x_i}(\mathbf{x}, v(\mathbf{x})) - \left(\dfrac{\partial g}{\partial x_i}(\mathbf{x}, v(\mathbf{x})), \lambda(\mathbf{x}) \right)_W.$