Module B2
Introduction à la méthode des moindres carrés

Dans ce module, E désigne l'espace euclidien $\mathbb{R}^n$ avec $n \in \mathbb{N}$, muni du produit scalaire usuel et de norme associée la norme euclidienne, notée $\|\cdot\|_2$. Pour plus de détails, le lecteur pourra se référer au module **A2 : Différentiabilité sur les espaces euclidiens**.

Dans ce module, on va s'intéresser à une première application concrète en optimisation non contrainte, à savoir le problème des moindres carrés. Il s'agit d'un problème que l'on rencontre couramment dans de nombreux domaines (analyse numérique, statistiques...). L'idée est d'appliquer les résultats établis dans les modules précédents (existence, unicité des minimiseurs, caractérisation au premier et au second ordre) sur ce problème.

1 Fonctions quadratiques généralisées

1.1 Définition

Définition 1 (Fonction quadratique généralisée)

Soient $A \in \mathcal{M}_{n,n}(\mathbb{R})$ une matrice réelle, $b \in E$ un vecteur et $c \in \mathbb{R}$ un scalaire. Alors on appelle *fonction quadratique généralisée* la fonction définie par

$$\left\{ \begin{array}{ccc} E & \to & \mathbb{R} \\ x & \mapsto & \frac{1}{2}\langle A\,x, x\rangle - \langle b, x\rangle + c \end{array} \right. \qquad (\star)$$

Autrement dit, toute fonction quadratique généralisée est la somme d'une forme quadratique, d'une forme linéaire et d'une constante.

Exemple

Quelques exemples de fonctions quadratiques généralisées

- Les fonctions constantes $x \mapsto c$ avec $c \in \mathbb{R}$;
- les formes linéaires ;
- les formes quadratiques.

Notons que, dans la définition $(\star)$, on a choisi dans ce module une forme particulière, à savoir adjoindre un facteur $1/2$ devant la forme quadratique et le signe $-$ devant la forme linéaire. Il est évident que toute fonction de la forme

$$x \mapsto \langle \tilde{A}\,x, x\rangle + \langle \tilde{b}, x\rangle + \tilde{c}$$

est une fonction quadratique généralisée ; il suffit de poser $A = 2\,\tilde{A}$, $\tilde{b} = -b$ et $\tilde{c} = c$. Ces choix spécifiques seront justifiés dans la suite de ce module.

Montrons qu'il est toujours possible de supposer que la forme quadratique est définie à l'aide d'une matrice symétrique :

Lemme 1 (Symétrisation d'une matrice)

Soit $A \in \mathcal{M}_{n,n}(\mathbb{R})$ une matrice réelle. On pose

$$\tilde{A} = \frac{1}{2}\left(A + A^\top\right)$$

Alors $\forall\, x \in E, \qquad \langle A\,x, x\rangle = \langle \tilde{A}\,x, x\rangle$

Par ailleurs, la matrice $\tilde{A}$ est symétrique.

Démonstration :

- Soit $x \in E$. On a

$$\langle A\,x, x\rangle = \frac{1}{2}\left(\langle A\,x, x\rangle + \langle A\,x, x\rangle\right) = \frac{1}{2}\left(\langle A\,x, x\rangle + \langle x, A^\top\,x\rangle\right)$$

  On en déduit l'identité annoncée.

- On a $\tilde{A}^\top = \left(\frac{1}{2}\left(A + A^\top\right)\right)^\top = \frac{1}{2}\left(A + A^\top\right)^\top = \frac{1}{2}\left(A^\top + (A^\top)^\top\right)$

  Autrement dit, $\tilde{A} = (\tilde{A})^\top$. ■

Cette proposition justifie que l'on suppose à présent que la matrice A qui apparaît dans la définition d'une fonction quadratique généralisée $(\star)$ est **symétrique**. En pratique, si A est une matrice quelconque, il suffit de la remplacer par $\tilde{A}$ dans tous les énoncés qui suivent.

1.2 Différentiabilité des fonctions quadratiques généralisées

Les fonctions quadratiques généralisées sont différentiables, comme le montre la proposition suivante :

Proposition 1

Soient $A \in \mathcal{M}_{n,n}(\mathbb{R})$ une matrice réelle **symétrique**, $b \in E$ un vecteur et $c \in \mathbb{R}$ un scalaire, et soit f la fonction quadratique généralisée définie dans $(\star)$. Alors f est différentiable et son gradient est donné pour tout $x \in E$ par

$$\nabla f(x) = A\,x - b$$

Démonstration : Soit $(x, h) \in E^2$. Calculons $f(x+h)$:

$$\begin{aligned} f(x+h) &= \frac{1}{2}\,\langle A\,(x+h), x+h\rangle - \langle b, x+h\rangle + c \\ &= \frac{1}{2}\,\langle A\,x, x\rangle + \frac{1}{2}\,\langle A\,h, x\rangle + \frac{1}{2}\,\langle A\,x, h\rangle + \frac{1}{2}\,\langle A\,h, h\rangle - \langle b, x\rangle - \langle b, h\rangle + c \end{aligned}$$

On reconnaît dans cette somme l'expression de $f(x)$. Par ailleurs, la symétrie de A permet de montrer que

$$\langle A\,h, x\rangle = \langle A^\top x, h\rangle = \langle A\,x, h\rangle$$

Aussi, on obtient après simplification et réarrangement des termes,

$$f(x+h) = f(x) + \langle A\,x, h\rangle - \langle b, h\rangle + \frac{1}{2}\langle A\,h, h\rangle$$

Puisqu'en dimension finie, toutes les applications linéaires sont bornées [a], l'inégalité de Cauchy-Schwarz assure que, si $h \neq 0$,

$$|\langle A\,h, h\rangle| \leq \|A\,h\|_2\,\|h\|_2 \leq \frac{\|A\,h\|_2}{\|h\|_2}\,\|h\|_2\,\|h\|_2 \leq \underbrace{|||A|||\,\|h\|_2}_{\xrightarrow[h\to 0]{} 0}\,\|h\|_2 = o(\|h\|_2)$$

On en déduit donc que f est différentiable en x, de gradient $A\,x - b$. ■

a. Cf. Compléments **Compléments C1 : Éléments d'algèbre linéaire**

Remarque : La symétrie de la matrice A est essentielle ici pour obtenir cette expression pour le gradient.

Proposition 2

Soient $A \in \mathcal{M}_{n,n}(\mathbb{R})$ une matrice réelle **symétrique**, $b \in E$ un vecteur et $c \in \mathbb{R}$ un scalaire, et soit f la fonction quadratique généralisée définie dans $(\star)$. Alors f est deux fois différentiable et

$$\forall\, x \in E, \qquad \mathrm{Hess}\, f(x) = A$$

Démonstration : Soit $(x, h) \in E^2$. Calculons $\nabla f(x+h)$:

$$\nabla f(x+h) = A\,(x+h) - b = A\,x - b + A\,h = \nabla f(x) + A\,h$$

Le terme linéaire est clairement continue ; ainsi ∇f est bien différentiable, de matrice jacobienne $J_{\nabla f}(x) = A$ pour tout $x \in E$. On conclut en utilisant le fait que $\mathrm{Hess}\, f(x) = J_{\nabla f}(x)$. ■

Remarque : On vérifie que la matrice hessienne est bien symétrique.

1.3 Fonctions quadratiques généralisées convexes

Les fonctions quadratiques généralisées étant deux fois différentiables, on peut utiliser la caractérisation des fonctions convexes deux fois différentiables pour établir les conditions sous lesquelles elles sont convexes (cf. Module **A3 : Fonctions convexes différentiables**).

Proposition 3

Soient $A \in \mathcal{M}_{n,n}(\mathbb{R})$ une matrice réelle **symétrique**, $b \in E$ un vecteur et $c \in \mathbb{R}$ un scalaire, et soit f la fonction quadratique généralisée définie dans $(\star)$. Alors

- f est convexe si et seulement si A est semi-définie positive ;
- f est strictement convexe si et seulement si A est définie positive.

Remarque : Le second point peut s'interpréter de la sorte : f est strictement convexe si et seulement si $\mathrm{Hess}\, f(x)$ est définie positive pour tout $x \in E$. Attention : en général, cette équivalence est fausse (il existe des fonctions strictement convexes dont la matrice hessienne n'est pas définie positive sur tout E).

Démonstration :

- Le premier point est l'application directe de la proposition 10 du module **A3 : Fonctions convexes différentiables**.
- La réciproque du second point est la conséquence de la proposition 11 du module **A3 : Fonctions convexes différentiables**. Démontrons le sens direct. On suppose que f est strictement convexe. Soit $(x, h) \in E^2$ avec $h \neq 0$. On applique la caractérisation d'une fonction strictement convexe différentiable (proposition 6 du module **A3 : Fonctions convexes différentiables**) aux points x et $x + h$, ce qui donne, puisque $x \neq x + h$,

$$0 < \langle \nabla f(x) - \nabla f(x+h), x - (x+h) \rangle = \langle -A\,h, -h \rangle = \langle A\,h, h \rangle \quad \blacksquare$$

Exemple

Cas réel Si $n = 1$, les fonctions quadratiques généralisées sont les polynômes de degré inférieur ou égal à 2 :

$$f(t) = \frac{1}{2}\,a\,t^2 - b\,t + c$$

- Si $a > 0$, alors f est une parabole strictement convexe ;
- si $a = 0$, alors f est une fonction affine (donc convexe, mais non strictement convexe) ;
- si $a < 0$, alors f est une parabole non convexe.

2 Optimisation quadratique non contrainte

On appelle *problème d'optimisation quadratique non contraint* la minimisation d'une fonction quadratique généralisée. L'objectif de cette section est d'étudier l'existence et l'unicité des minimiseurs de telles fonctions, en fonction des propriétés des éléments qui définissent les fonctions quadratiques généralisées (à savoir la matrice A, le vecteur b et le scalaire c) et, le cas échéant, caractériser ces minimiseurs.

2.1 Points critiques d'une fonction quadratique généralisée

Puisqu'on a montré que les fonctions quadratiques étaient différentiables, on peut énoncer le résultat suivant, qui n'est rien d'autre que la condition nécessaire d'optimalité de premier ordre :

Proposition 4

Soient $A \in \mathcal{M}_{n,n}(\mathbb{R})$ une matrice réelle **symétrique**, $b \in E$ un vecteur et $c \in \mathbb{R}$ un scalaire, et soit f la fonction quadratique généralisée définie dans $(\star)$. Alors les points critiques de f sont les solutions du système linéaire

$$A\,x = b$$

Démonstration : On rappelle que les points critiques de f sont les points qui annulent le gradient de f ; or, la proposition 1 assure que

$$\nabla f(x) = 0 \qquad \Longleftrightarrow \qquad A\,x - b = 0 \quad \blacksquare$$

2.2 Cas des fonctions quadratiques généralisées convexes

Proposition 5

Soient $A \in \mathcal{M}_{n,n}(\mathbb{R})$ une matrice réelle **symétrique** semi-définie positive, $b \in E$ un vecteur et $c \in \mathbb{R}$ un scalaire, et soit f la fonction quadratique généralisée définie dans $(\star)$. Alors les minimiseurs de f sont les solutions du système linéaire

$$A\,x = b$$

Démonstration : Il s'agit cette fois d'une conséquence directe de la proposition 14 du module **B1 : Minimisation d'une fonction. Conditions d'optimalité.**, qui établit que les minimiseurs globaux d'une fonction convexe sont exactement ses points critiques. ■

Pour établir des conditions suffisantes pour lesquelles un problème d'optimisation quadratique admet au moins une solution, on va établir une condition nécessaire et suffisante pour qu'une fonction quadratique généralisée soit infinie à l'infini. On verra alors que cette condition garantit automatiquement l'unicité de la solution optimale. On va commencer par rappeler qu'une matrice est semi-définie positive (resp. définie positive) si et seulement si ses valeurs propres sont positives (resp. strictement positives). On en déduit immédiatement que

Proposition 6

Soient $A \in \mathcal{M}_{n,n}(\mathbb{R})$ une matrice réelle **symétrique**, $b \in E$ un vecteur et $c \in \mathbb{R}$ un scalaire, et soit f la fonction quadratique généralisée définie dans $(\star)$.

- f est convexe si et seulement si les valeurs propres de A sont positives.
- f est strictement convexe si et seulement si les valeurs propres de A sont strictement positives.

Démonstration : Il suffit de traduire la proposition 3 en termes de valeurs propres de A. ■

On peut maintenant établir les conditions pour que la fonction soit infinie à l'infini :

Proposition 7

Soient $A \in \mathcal{M}_{n,n}(\mathbb{R})$ une matrice réelle **symétrique**, $b \in E$ un vecteur et $c \in \mathbb{R}$ un scalaire, et soit f la fonction quadratique généralisée définie dans $(\star)$. Alors f est infinie à l'infini si et seulement si elle est strictement convexe.

Démonstration : On commence par rappeler que A étant une matrice symétrique réelle, elle est diagonalisable. Notons que $(\lambda_i)_{1\leq i\leq n}$ ses valeurs propres, ordonnées de la manière suivante :

$$\lambda_1 \leq \cdots \leq \lambda_n$$

Par ailleurs, on sait qu'il existe une famille de vecteurs propres $(V_i)_{1\leq i\leq n}$ vérifiant

$$\forall\, i = 1, \ldots, n, \qquad A\,V_i = \lambda_i$$

et telle que $(V_i)_{1\leq i\leq n}$ forme une base orthonormée de E, c'est-à-dire que

$$\forall\, i,j=1,\dots,n, \qquad \langle V_i, V_j\rangle = \begin{cases} \|V_i\|_2^2 = 1 & \text{si } i=j \\ 0 & \text{sinon} \end{cases}$$

- **Sens direct.** On suppose que f est infinie à l'infini. Par l'absurde, on suppose également que f n'est pas strictement convexe, c'est-à-dire que $\lambda_1 \leq 0$. On définit alors $x = t\,V_1$ pour $t>0$. Dans ce cas,

  $$f(x) = f(t\,V_1) = \frac{1}{2}\,\langle A\,(t\,V_1), t\,V_1\rangle - \langle b, t\,V_1\rangle + c = \frac{t^2}{2}\,\langle A\,V_1, V_1\rangle - t\,\langle b, V_1\rangle + c$$

  Par définition de V_1, on a $\langle A\,V_1, V_1\rangle = \lambda_1$. On en déduit que

  $$f(x) = \frac{\lambda_1}{2}\,t^2 - t\,\langle b, V_1\rangle + c$$

  Par hypothèse, deux cas sont possibles :

  (i) soit $\lambda_1 = 0$, auquel cas $f(x)$ se simplifie en

  $$f(x) = -t\,\langle b, V_1\rangle + c$$

  Considérons les possibles valeurs de $\langle b, V_1\rangle$.

  1. Si $\langle b, V_1\rangle = 0$, alors f est une fonction constante, et ne peut être infinie à l'infini car pour toute suite $(x_n)_{n\in\mathbb{N}}$ telle que $\|x_n\|_2$ tend vers $+\infty$, la suite des $f(x_n)$ reste constante (et en particulier, ne diverge pas vers $+\infty$).
  2. Si $\langle b, V_1\rangle < 0$, alors $t \mapsto f(t\,V_1)$ est une fonction affine, dont le coefficient dominant est strictement positif. Il est alors aisé de vérifier qu'en choisissant $x_n = -n\,V_1$ pour tout $n\in\mathbb{N}$, la suite des $\|x_n\|_2$ diverge vers $+\infty$ mais que la suite de $f(x_n)$ diverge vers $-\infty$. Donc f n'est pas infinie à l'infini dans ce cas non plus.
  3. Enfin, si $\langle b, V_1\rangle > 0$, alors $t \mapsto f(t\,V_1)$ est une fonction affine de coefficient dominant strictement négatif et il suffit de considérer la suite des $x_n = n\,V_1$.

  (ii) soit $\lambda_1 < 0$, auquel cas $t \mapsto f(t\,V_1)$ est un polynôme du second degré, dont le coefficient dominant est strictement négatif. On sait alors que lorsque t tend vers $+\infty$, $f(t\,V_1)$ tend vers $-\infty$. On peut donc considérer par exemple la suite des $x_n = n\,V_1$ pour démontrer que f n'est pas infinie à l'infini.

  Dans les deux cas, f n'est pas infinie à l'infini, ce qui contredit l'hypothèse initiale. On en déduit alors que λ_1 ne peut être négative.

- **Réciproque.** On suppose maintenant que toutes les valeurs propres de A sont strictement positives. Posons $P = {}^t(V_1,\dots,V_n) \in \mathcal{M}_{n,n}(\mathbb{R})$. Alors P est une matrice orthonormée, et on a $P^{-1} = P^\top$. Par ailleurs,

  $$A = P^\top D\,P \qquad \text{avec} \qquad D = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \lambda_n \end{pmatrix}$$

  Calculons maintenant $\langle A\,x, x\rangle$ pour tout $x\in E$. En utilisant l'écriture introduite ci-dessus, on obtient que

  $$\langle A\,x, x\rangle = \langle P^\top D\,P\,x, x\rangle = \langle D\,P\,x, P\,x\rangle$$

  Écrivons ce produit scalaire sous forme étendue :

  $$\langle D\,P\,x, P\,x\rangle = \sum_{i=1}^{n} (D\,P\,x)_i\,(P\,x)_i$$

Or, la matrice D étant diagonale, on en déduit que, pour tout $i = 1, \dots, n$, on a $(D\,P\,x)_i = \lambda_i\,(P\,x)_i$. Ainsi, on obtient que

$$\langle D\,P\,x, P\,x\rangle = \sum_{i=1}^{n} \lambda_i\,(P\,x)_i\,(P\,x)_i = \sum_{i=1}^{n} \lambda_i\,(P\,x)_i^2$$

Puisque les $(P\,x)_i^2$ sont positifs, on peut minorer chacun des termes de cette somme par $\lambda_1\,(P\,x)_i^2$. En factorisant par λ_1, on obtient finalement que

$$\langle A\,x, x\rangle \geq \lambda_1 \sum_{i=1}^{n} (P\,x)_i^2 = \lambda_1\,\|P\,x\|_2^2 = \lambda_1\,\|x\|_2^2$$

la dernière égalité provenant du fait que P est orthonormée. En particulier,

$$\forall\, x \in E, \qquad f(x) = \frac{1}{2}\,\langle A\,x, x\rangle - \langle b, x\rangle + c \geq \frac{1}{2}\,\lambda_1\,\|x\|_2^2 - \langle b, x\rangle + c$$

En utilisant l'inégalité de Cauchy-Schwarz, on en déduit la minoration suivante

$$\forall\, x \in E, \qquad f(x) \geq \frac{1}{2}\,\lambda_1\,\|x\|_2^2 - \|b\|_2\,\|x\|_2 + c$$

Si on applique cette inégalité une suite $(x_k)_{k\in\mathbb{N}}$ divergeant vers $+\infty$, on voit que le membre de droite diverge vers $+\infty$. Par comparaison, c'est le cas $(f(x_k))_{k\in\mathbb{N}}$, ce qui prouve que f est infinie à l'infini. ■

Corollaire 1

Soient $A \in \mathcal{M}_{n,n}(\mathbb{R})$ une matrice réelle **symétrique**, $b \in E$ un vecteur et $c \in \mathbb{R}$ un scalaire, et soit f la fonction quadratique généralisée définie dans $(\star)$. Si A est définie positive, alors f admet un unique minimiseur.

Démonstration : D'après la proposition 3, si A est définie positive, alors f admet au plus un minimiseur (car f est strictement convexe). L'existence du minimiseur peut alors être démontrée en appliquant la proposition 9 du module **B1 : Minimisation d'une fonction. Conditions d'optimalité.**Or, f étant (strictement) convexe, elle est nécessairement continue. Par ailleurs, f est strictement convexe si et seulement si f est infinie à l'infini (proposition 7). ■

> En réalité, puisque les minimiseurs d'une fonction quadratique généralisée sont les solutions du système linéaire $A\,x = b$, on voit que, si A est définie positive, elle est en particulier inversible, donc le système linéaire $A\,x = b$ possède exactement une solution. On pouvait donc s'épargner les preuves ci-dessus.

Qu'en est-il du cas des fonctions quadratiques généralisées convexes mais non strictement convexes ? Le fait qu'elles ne soient pas infinies à l'infini ne permettent pas de conclure, puisque le caractère infini à l'infini est une condition suffisante mais non nécessaire à l'existence de minimiseurs. Puisque ces derniers sont les solutions du système linéaire $A\,x = b$, et que A admet nécessairement une valeur propre nulle, on en déduit que A est non inversible, donc le système linéaire concerné admet soit aucune solution, soit une infinité de solutions. Ainsi, la fonction quadratique généralisée associée admet soit aucun minimiseur (par exemple si elle est affine non constante), soit une infinité (par exemple si elle est constante).

2.3 Cas des fonctions quadratiques généralisées non convexes

Dans ce paragraphe, on va établir que seules les fonctions quadratiques généralisées convexes peuvent admettre un minimiseur. Ce n'est évidemment pas le cas en général : une fonction non convexe peut admettre des minimiseurs (il suffit de songer au cosinus).

Proposition 8

Soient $A \in \mathcal{M}_{n,n}(\mathbb{R})$ une matrice réelle **symétrique**, $b \in E$ un vecteur et $c \in \mathbb{R}$ un scalaire, et soit f la fonction quadratique généralisée définie dans $(\star)$. On suppose que A n'est pas semi-définie positive. Alors f n'admet aucun minimiseur.

DÉMONSTRATION : D'après la condition nécessaire d'optimalité du second ordre (proposition 16 du module **B1 : Minimisation d'une fonction. Conditions d'optimalité.**) les minimiseurs de f, s'ils existent, sont des points critiques x^* de f pour lesquels $\operatorname{Hess} f(x^*)$ sont semi-définies positives. Or, puisque $\operatorname{Hess} f(x^*) = A$ (proposition 2), et que celle-ci est supposée non semi-définie positive, on en déduit qu'aucun des points critiques de f n'est minimiseur de f. ■

3 Méthode des moindres carrés

3.1 Premier exemple : régression linéaire dans la loi d'OHM

Commençons par considérer un problème très simple. La loi d'OHM en électricité stipule qu'il est possible de déterminer la valeur $R_{\text{théorique}}$ d'une résistance en mesurant l'intensité I qui la traverse et la tension à ses bornes, la relation entre ces trois grandeurs étant donnée par la célèbre formule

$$U = R_{\text{théorique}}\, I$$

Ainsi, pour une valeur de résistance donnée, la relation entre l'intensité et la tension est linéaire. Supposons que l'on souhaite déterminer cette valeur. On demande à $p = 100$ étudiants, chacun équipé d'un voltmètre et d'une batterie délivrant une intensité réglable, de faire passer un courant d'une intensité I_i et de mesurer la tension U_i aux bornes de la résistance. On collecte toutes les données recueillies et on les représente dans la figure 1. On obtient donc un nuage de points, avec l'intensité en abscisse et la tension en ordonnée. On constate que les points ne sont pas tout à fait alignés (cf. Figure 1). Cela est dû au fait que les mesures sont entachées d'erreur, soit par exemple que l'étudiant ait mal recopié la valeur de la tension, soit que le voltmètre ne soit pas très précis. Par conséquent, au lieu d'avoir l'égalité prescrite par la loi d'OHM, on obtient plutôt pour tout $i = 1, \dots, p$

$$U_i \approx R_{\text{théorique}}\, I_i$$

Notons ε_i l'écart au modèle linéaire de la loi d'OHM pour tout $i = 1, \dots, p$, défini par

$$\varepsilon_i = R_{\text{théorique}}\, I_i - U_i$$

La question est alors la suivante : comment estimer $R_{\text{théorique}}$ à partir des données récoltées ? Il est évident que si on n'avait qu'une seule mesure, c'est-à-dire un unique couple (I_1, U_1), alors, sans autre information, la réponse la plus raisonnable, si I_1 est non nul, est de donner l'estimation suivante

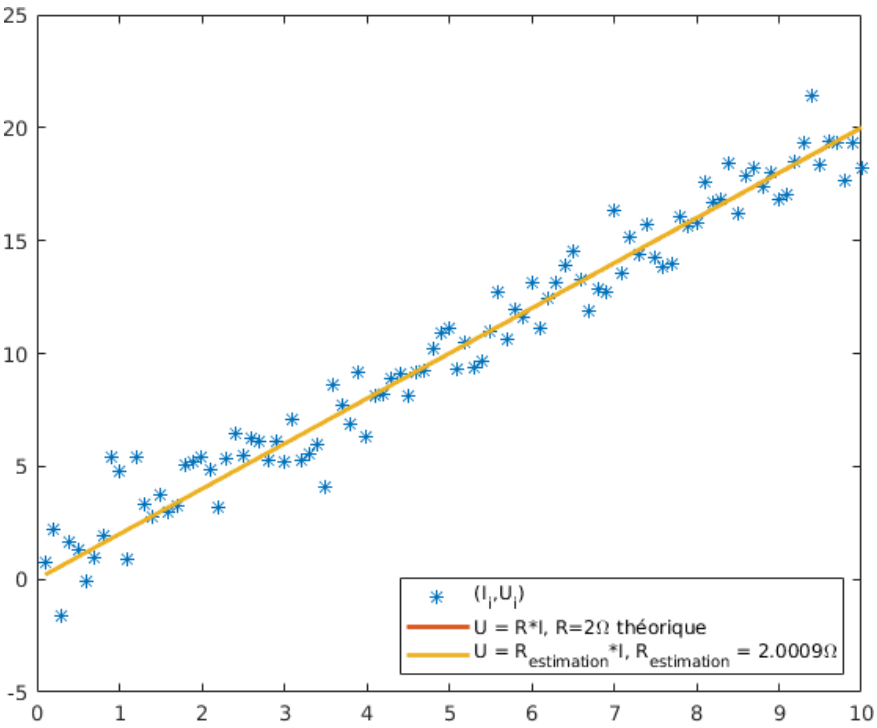

FIGURE 1 – Illustration de l'estimation au sens des moindres carrés de la valeur d'une résistance. En rouge, le modèle théorique, en jaune, l'estimation du modèle à partir des points bleus.

$$R_{\text{estimation}} = \frac{U_1}{I_1}$$

Dès lors qu'on possède plus de mesures, faire ce calcul pour chaque couple (I_i, U_i) peut *a priori* donner une multitude de valeurs possibles. On pourrait les moyenner ; mais quel sens physique donner à la valeur obtenue ?

Si les erreurs ε_i sont indépendantes les unes des autres, nulles en moyenne (en particulier, pas préférentiellement positives ou négatives), et qu'elles suivent la même loi gaussienne, alors (voir encart) une manière raisonnable de procéder est de trouver la valeur R qui va minimiser l'erreur quadratique moyenne, c'est-à-dire minimiser la fonction suivante

$$R \mapsto \frac{1}{p} \sum_{i=1}^{p} \varepsilon_i^2$$

(les ε_i dépendant de la valeur de R). On dit dans ce cas qu'on estime R **au sens des moindres carrés**. On notera que p étant strictement positif, les minimiseurs de la fonction précédente, s'ils existent, sont les mêmes que celles de la fonction

$$R \mapsto \frac{1}{2} \sum_{i=1}^{p} \varepsilon_i^2$$

(on a multiplié par $p/2 > 0$ la fonction précédente). On se ramène donc à minimiser la fonction

$$f : R \mapsto \frac{1}{2} \sum_{i=1}^{p} (R\, I_i - U_i)^2 = \frac{1}{2} \left\| R\, I - U \right\|_2^2$$

Pourquoi minimiser l'erreur quadratique ?

Lorsque les erreurs sont supposées indépendantes les unes les autres et suivre la même loi gaussienne de moyenne nulle, minimiser l'erreur quadratique revient à trouver la valeur de résistance R qui maximise la *vraisemblance* d'observer les erreurs ε_i associées. Il s'agit d'une notion hors-programme, aussi, contentons-nous d'en donner une idée générale. Supposons que les erreurs ε_i suivent la loi normale

$$\varepsilon_i \sim \mathcal{N}(0, \sigma)$$

de densité de probabilité donnée par la fonction

$$p(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{t^2}{2\,\sigma^2}\right)$$

La *vraisemblance* de R au vu des observations (I_i, U_i) est la densité de probabilité associée aux erreurs ε_i, c'est-à-dire

$$\prod_{i=1}^{p} p(R\,I_i - U_i) = \prod_{i=1}^{p} \left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(R\,I_i - U_i)^2}{2\,\sigma^2}\right)\right)$$

L'idée est alors de trouver une valeur R qui maximise cette quantité, c'est-à-dire de trouver un maximiseur de la fonction

$$R \mapsto \prod_{i=1}^{p} p(R\,I_i - U_i)$$

On parle d'estimer le maximum de vraisemblance. Simplifions l'expression de la vraisemblance :

$$\begin{aligned}
\prod_{i=1}^{p} p(R\,I_i - U_i) &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^p \prod_{i=1}^{p} \exp\left(-\frac{(R\,I_i - U_i)^2}{2\,\sigma^2}\right) \\
&= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^p \exp\left(-\sum_{i=1}^{p} \frac{(R\,I_i - U_i)^2}{2\,\sigma^2}\right) \\
&= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^p \exp\left(-\frac{1}{2\,\sigma^2} \sum_{i=1}^{p} (R\,I_i - U_i)^2\right)
\end{aligned}$$

Or, il est aisé de vérifier que les maximiseurs de $R \mapsto a\,g(R)$ sont exactement les maximiseurs de $R \mapsto g(R)$ lorsque $a > 0$; ainsi, le problème devient celui de la maximisation de la fonction

$$R \mapsto \exp\left(-\frac{1}{2\,\sigma^2} \sum_{i=1}^{p} (R\,I_i - U_i)^2\right)$$

Par ailleurs, les maximiseurs de $R \mapsto g(R)$ sont également exactement les maximiseurs de $R \mapsto \ln(g(R))$ car le logarithme est une fonction strictement croissante. Puisque $\ln \circ \exp = \mathrm{Id}$, cela revient à maximiser

$$R \mapsto -\frac{1}{2\,\sigma^2} \sum_{i=1}^{p} (R\,I_i - U_i)^2$$

Enfin, les maximiseurs de $R \mapsto g(R)$ étant exactement les minimiseurs de $R \mapsto -a\,g(R)$ si $a > 0$, on en déduit finalement que maximiser la vraisemblance revient à minimiser la fonction

$$R \mapsto \sum_{i=1}^{p} (R\,I_i - U_i)^2$$

c'est-à-dire à estimer R au sens des moindres carrés.

où $I = (I_i)_{1\leq i\leq p}$ et $U = (U_i)_{1\leq i\leq p}$. Développons la fonction f ; on obtient

$$f(R) = \frac{1}{2}\|R\,I\|_2^2 - \langle R\,I, U\rangle + \frac{1}{2}\|U\|_2^2 = \frac{R^2}{2}\|I\|_2^2 - R\,\langle I, U\rangle + \frac{1}{2}\|U\|_2^2$$

On est donc ramené à trouver le minimiseur d'un polynôme du second degré. Celui-ci est unique, et vaut

$$R_{\text{estimation}} = \frac{\langle I, U\rangle}{\|I\|_2^2}$$

Notons que, puisque $U = R_{\text{théorique}}\,I + \varepsilon$, nous avons

$$R_{\text{estimation}} = \frac{\langle I, R_{\text{théorique}}\,I + \varepsilon\rangle}{\|I\|_2^2} = \frac{R_{\text{théorique}}\,\|I\|_2^2 + \langle I, \varepsilon\rangle}{\|I\|_2^2} = R_{\text{théorique}} + \frac{\langle I, \varepsilon\rangle}{\|I\|_2^2}$$

Autrement dit,

$$|R_{\text{estimation}} - R_{\text{théorique}}| = \frac{|\langle I, \varepsilon\rangle|}{\|I\|_2^2} \leq \frac{\|I\|_2\|\varepsilon\|_2}{\|I\|_2^2} \leq \frac{\|\varepsilon\|_2}{\|I\|_2}$$

ce qui montre que, de manière attendue, l'erreur commise dans l'estimation de $R_{\text{théorique}}$ est d'autant plus petite que les erreurs de mesure sont petites devant les intensités injectées.

Avant de clore ce paragraphe, signalons que, dans la définition de $R_{\text{estimation}}$, les rôles de I et de U ne sont pas symétriques. Pourtant, il est possible de réécrire la loi d'Ohm en introduisant la conductance G définie comme l'inverse de la résistance :

$$I = \frac{U}{R} = G\,U$$

Avec les mêmes données, on peut estimer G au sens des moindres carrés, c'est-à-dire minimiser

$$G \mapsto \frac{1}{2}\|G\,U - I\|_2^2$$

On obtient alors $$G_{\text{estimation}} = \frac{\langle I, U\rangle}{\|U\|_2^2}$$

qui, en général, est différent de $1/R_{\text{estimation}}$. Autrement dit, alors qu'on utilise les mêmes données et que l'on cherche toujours à minimiser une erreur quadratique, on obtient deux estimations différentes de R. En réalité, selon la formulation de la loi d'Ohm, l'erreur quadratique minimisée n'est pas la même. Graphiquement (voir Figure 2), l'estimation (directe) de R revient à minimiser la distance *verticale* (au carré) des points à une droite, alors que l'estimation de G minimiser la distance *horizontale*. On peut interpréter ce choix de la manière suivante : dans le premier cas, on suppose implicitement que l'erreur ne porte que sur la mesure de la tension, mais pas de l'intensité (on suppose donc que le générateur a délivré exactement un courant d'intensité prescrite) ; c'est l'inverse pour le second cas. Aussi, suivant le cas pratique, l'une des deux formulations peut être plus raisonnable que l'autre.

3.2 Régression parabolique

On va maintenant considérer un exemple plus théorique où cette méthode est utilisée pour estimer les paramètres d'un modèle **non linéaire**. On suppose qu'on possède un nuage de points (X_i, Y_i) pour $i = 1, \ldots, p$, et on aimerait trouver la parabole

$$X \mapsto a\,X^2 + b\,X + c$$

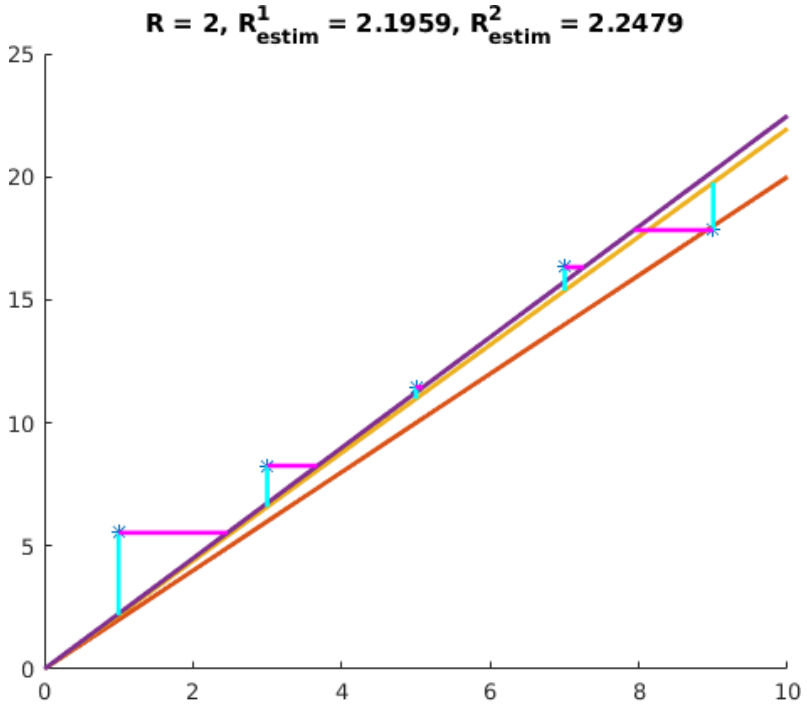

FIGURE 2 – Estimation directe de la résistance (en jaune) *versus* estimation de la résistance *via* l'estimation de la conductance (en violet). En rouge, le modèle théorique. L'estimation directe minimise la distance en cyan, tandis que l'estimation *via* celle de la conductance minimise la distance en magenta.

qui approche au mieux ce nuage au sens des moindres carrés, c'est-à-dire trouver trois réels a, b et c minimisant la fonction

$$f : (a, b, c) \mapsto \frac{1}{2} \sum_{i=1}^{p} (a\, X_i^2 + b\, X_i + c - Y_i)^2$$

En posant $P = \begin{pmatrix} X_1^2 & X_1 & 1 \\ \vdots & \vdots & \vdots \\ X_p^2 & X_p & 1 \end{pmatrix}$, $g = \begin{pmatrix} Y_1 \\ \vdots \\ Y_p \end{pmatrix}$ et $x = \begin{pmatrix} a \\ b \\ c \end{pmatrix}$

on voit qu'il s'agit de minimiser la fonction

$$f : x \mapsto \frac{1}{2} \|P\, x - g\|_2^2$$

> Dans l'exemple considéré, on suppose que les valeurs Y_i sont des **mesures** obtenues à partir des valeurs X_i. On peut voir l'analogie avec l'exemple précédent, où les tensions U_i sont mesurées en injectant l'intensité I_i dans le système. Les valeurs X_i (resp. I_i) sont choisies par l'expérimentateur, qui sont donc sous son contrôle (on les suppose de précision infinie en particulier) ; on peut donc les considérer comme des **paramètres** de l'expérience. Les valeurs Y_i (resp. U_i) sont obtenues par **mesure** ; ce sont elles qui sont affectées par une erreur.

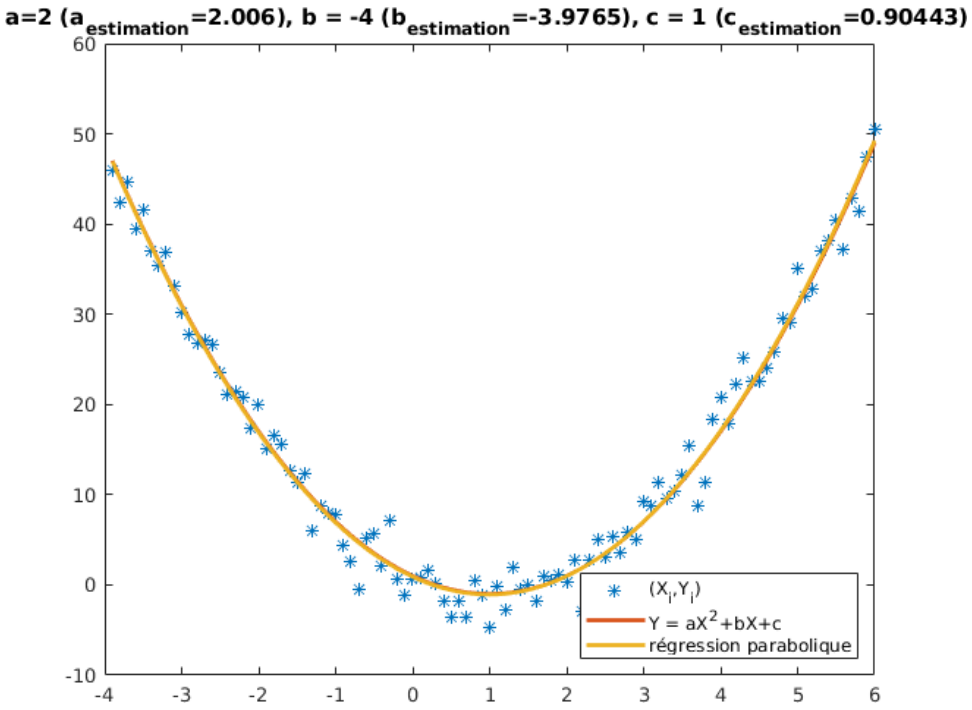

FIGURE 3 – Illustration de la régression parabolique. En bleu, le nuage de points, en rouge la parabole théorique, en jaune, l'estimation obtenue au sens des moindres carrés.

Puisque

$$f(x) = \frac{1}{2}\|Px\|_2^2 - \langle Px, g\rangle + \frac{1}{2}\|g\|_2^2 = \frac{1}{2}\langle P^\top Px, x\rangle - \langle x, P^\top g\rangle + \frac{1}{2}\|g\|_2^2$$

on voit que f est une fonction quadratique généralisée, convexe car $P^\top P$ est semi-définie positive. Ainsi, ses minimiseurs sont ses points critiques, données par les solutions du système linéaire

$$P^\top Px = P^\top g$$

soit

$$\begin{cases} a\sum_{i=1}^p X_i^4 + b\sum_{i=1}^p X_i^3 + c\sum_{i=1}^p X_i^2 = \sum_{i=1}^p X_i^2 Y_i \\ a\sum_{i=1}^p X_i^3 + b\sum_{i=1}^p X_i^2 + c\sum_{i=1}^p X_i = \sum_{i=1}^p X_i Y_i \\ a\sum_{i=1}^p X_i^2 + b\sum_{i=1}^p X_i + cp = \sum_{i=1}^p Y_i \end{cases}$$

Remarquons que la matrice $P^\top P$ est inversible si et seulement si P est injective. Par définition, c'est le cas si et seulement si $Px = 0$ pour $x = 0$ uniquement. Cette condition s'écrit ici

$$\begin{cases} aX_1^2 + bX_1 + c = 0 \\ \vdots \quad\quad \vdots \quad\quad \vdots \\ aX_p^2 + bX_p + c = 0 \end{cases} \iff a = b = c = 0$$

Puisque $(0,0,0)$ est une solution de ce système linéaire, il faut et il suffit qu'il existe trois indices $1 \leq i_1, i_2, i_3 \leq p$ tels que

$$\begin{cases} a\,X_{i_1}^2 + b\,X_{i_1} + c = 0 \\ a\,X_{i_2}^2 + b\,X_{i_2} + c = 0 \\ a\,X_{i_3}^2 + b\,X_{i_3} + c = 0 \end{cases} \iff a = b = c = 0$$

Cette condition est équivalente au fait que la matrice

$$\begin{pmatrix} X_{i_1}^2 & X_{i_1} & 1 \\ X_{i_2}^2 & X_{i_2} & 1 \\ X_{i_3}^2 & X_{i_3} & 1 \end{pmatrix}$$

soit inversible. On reconnaît là une matrice de VANDERMONDE, qui est inversible si et seulement si les nombres X_{i_1}, X_{i_2} et X_{i_3} sont deux à deux distincts. Une interprétation de ce résultat est la suivante : pour obtenir une estimation uniquement déterminée d'une parabole, il faut et il suffit d'avoir au moins trois points distincts dans le nuage de points.

3.3 Cas général

Dans les deux exemples précédents, on a un ensemble de **mesures** (les tensions U_i ou les images Y_i) effectuées en sortie d'un système (circuit électrique ou autre) en réponse à une entrée (les intensités I_i ou les arguments X_i), sous le contrôle de l'expérimentateur (et donc assimilées à des **paramètres** de l'expérience). Les mesures sont supposées entachées d'une erreur faible, indépendantes des mesures précédentes, et suivant la même loi normale de moyenne nulle. On a un à-priori sur la manière dont ont été produites ces mesures, c'est-à-dire que l'on a un modèle de génération : dans le premier cas, $U_i \approx R\,I_i$ et dans le second cas, $Y_i = P(X_i)$. Dans le cas de la régression linéaire, on pense (à tort ou à raison, mais cela n'est pas la question ici) que le modèle est linéaire, mais on ne connaît pas la pente de la droite $P(X) = R\,X$ recherchée ; dans le cas de la régression parabolique, on pense que le modèle est parabolique, mais on ne connaît pas les coefficients du polynôme $P(X) = a\,X^2 + b\,X + c$ recherché. Dans les deux cas, ce qui importe, c'est que les coefficients recherchés (R ou (a, b, c)) interviennent de manière **linéaire** dans le modèle.

Plaçons-nous à présent dans un cadre plus général. On a un système que l'expérimentateur peut contrôler dans le sens suivant : il injecte p entrées $I_i \in \mathbb{R}^m$ et il récolte en sortie des mesures $g_i \in \mathbb{R}$ associées, lesquelles sont supposées dégradées par une erreur suivant une loi normale centrée (chaque erreur étant indépendante des autres). En outre, l'expérimentateur dispose d'un modèle de génération de ses données : il sait que

$$\forall\, i = 1, \dots, p, \qquad g_i \approx M(I_i)$$

avec

$$M(I) = \sum_{j=1}^{n} a_j\, f_j(I)$$

où les $f_j : \mathbb{R}^m \to \mathbb{R}$ sont des fonctions connues (par exemple des monômes, ou n'importe quelle base d'espaces fonctionnels). Le modèle M est donc une combinaison **linéaire** de fonctions connues ; seules les valeurs des coefficients $x = (a_j)_{1 \le j \le n}$ sont inconnues. Puisque les f_j sont connues, les vecteurs $(f_j(I_i))_{1 \le j \le n}$ le sont également. On a alors la relation linéaire suivante

$$\forall\, i = 1, \dots, p, \qquad M(I_i) = \langle x, (f_j(I_i))_{1 \le j \le n} \rangle \approx g_i$$

L'ensemble des entrées permet donc de construire une matrice $P \in \mathcal{M}_{p,n}(\mathbb{R})$ de la forme

$$P = \begin{pmatrix} f_1(I_1) & \cdots & f_n(I_1) \\ \vdots & \vdots & \vdots \\ f_1(I_p) & \cdots & f_n(I_p) \end{pmatrix}$$

avec $$P\,x \approx g = (g_i)_{1 \leq i \leq p}$$

Estimer x au sens des moindres carrés revient donc à trouver

$$x^* \in \arg\min_{x \in E} \left\{ \frac{1}{2} \|P\,x - g\|_2^2 \right\}$$

Commençons par remarquer que ce problème est un problème d'optimisation quadratique ; en effet, la fonction que l'on cherche à minimiser s'écrit

$$f(x) = \frac{1}{2}\|P\,x - g\|_2^2 = \frac{1}{2}\|P\,x\|_2^2 - \langle P\,x, g\rangle + \frac{1}{2}\|g\|_2^2 = \frac{1}{2}\langle P^\top P\,x, x\rangle - \langle x, P^\top g\rangle + \frac{1}{2}\|g\|_2^2$$

Puisque $(P^\top P)^\top = P^\top (P^\top)^\top = P^\top P$, on en déduit que la fonction étudiée est une fonction quadratique généralisée. Elle est de plus convexe ; en effet,

$$\forall\, x \in E, \qquad \langle P^\top P\,x, x\rangle = \|P\,x\|_2^2 \geq 0$$

ce qui assure que la matrice hessienne de f est bien semi-définie positive. On voit d'ailleurs immédiatement que cette matrice est définie positive si et seulement si $P\,x \neq 0$ si $x \neq 0$, donc si et seulement si P est injective.

> Lorsque P est injective, on appelle *pseudo-inverse* de P la matrice $(P^\top P)^{-1}\,P$.

> D'une manière générale, le problème considéré peut s'interpréter comme un problème de projection orthogonale sur un convexe fermé (cf. module **B6 : Projection sur un convexe**) : en effet, puisque
>
> $$\mathrm{Im}(P) = \{z \in \mathbb{R}^m \mid \exists\, x \in E, z = P\,x\}$$
>
> on peut réécrire le problème sous la forme équivalente
>
> $$x^* \in E \text{ tel que } P\,x^* = z^* = \arg\min_{z \in \mathrm{Im}(P)} \left\{ \frac{1}{2}\|z - g\|_2^2 \right\}$$
>
> Le problème revient donc à trouver le projeté orthogonal de g sur $\mathrm{Im}(P)$. Or, celui-ci est un ensemble convexe car c'est un espace vectoriel ; par ailleurs, il est fermé car c'est un sous-espace vectoriel de $\mathbb{R}^p$, et est donc de dimension finie.

> Attention : l'unicité de la projection sur $\mathrm{Im}(P)$ n'assure pas l'unicité de x^* ; celle-ci n'est garantie que si P est injective, auquel cas on a déjà établi l'unicité du minimiseur de f par stricte convexité.

On peut alors démontrer les résultats suivants :

Théorème 1

Soient $P \in \mathcal{M}_{p,n}(\mathbb{R})$ et $g \in \mathbb{R}^p$. On s'intéresse à la fonction suivante :

$$f : \left\{ \begin{array}{ccc} E & \to & \mathbb{R} \\ x & \mapsto & \dfrac{1}{2} \|Px - g\|_2^2 \end{array} \right.$$

(i) La fonction f admet au moins un minimiseur.

(ii) Les minimiseurs de f sont les solutions du système linéaire

$$P^\top P\, x = P^\top g$$

(iii) De plus, si P est injective, alors l'unique minimiseur de f vaut

$$x^* = (P^\top P)^{-1} P^\top g$$

Démonstration :

(i) Admis ; il s'agit de la conséquence directe du théorème 1 du module **B6 : Projection sur un convexe**.

(ii) Il s'agit de la conséquence directe de la proposition 5.

(iii) Il s'agit de la conséquence directe du corollaire 1, car la fonction objectif est dans ce cas strictement convexe. ■