Optimisation-RO: Optimisation convexe

Clément Royer

Certificat Chef de Projet IA - Université Paris Dauphine-PSL

15 novembre 2021

Pour la séance de cet après-midi

Table des matières

Outline

Premières définitions

Ensemble convexe

Un ensemble $\mathcal{C} \in \mathbb{R}^d$ est dit **convexe** si

$$\forall (\boldsymbol{u}, \boldsymbol{v}) \in \mathcal{C}^2,\ \forall t \in [0,1], \qquad t\boldsymbol{u} + (1-t)\boldsymbol{v} \in \mathcal{C}.$$

Premières définitions

Ensemble convexe

Un ensemble $\mathcal{C} \in \mathbb{R}^d$ est dit **convexe** si

$$\forall (\boldsymbol{u}, \boldsymbol{v}) \in \mathcal{C}^2, \ \forall t \in [0,1], \qquad t\boldsymbol{u} + (1-t)\boldsymbol{v} \in \mathcal{C}.$$

Exemples :

- $\mathbb{R}^d$;
- Droite : $\{t\boldsymbol{w} | t \in \mathbb{R}\}$ pour tout $\boldsymbol{w} \in \mathbb{R}^d$;
- Boule : $\left\{\boldsymbol{w} \in \mathbb{R}^d | \|\boldsymbol{w}\|^2 = \sum_{i=1}^d [\boldsymbol{w}]_i^2 \leq 1\right\}$.

Définition

Une fonction $f : \mathbb{R}^d \to \mathbb{R}$ est dite **convexe** si

$$\forall (\boldsymbol{u}, \boldsymbol{v}) \in (\mathbb{R}^d)^2,\ \forall t \in [0,1], \qquad f(t\boldsymbol{u} + (1-t)\boldsymbol{v}) \leq t\, f(\boldsymbol{u}) + (1-t)\, f(\boldsymbol{v}).$$

Fonctions convexes

Définition

Une fonction $f : \mathbb{R}^d \to \mathbb{R}$ est dite **convexe** si

$$\forall (\boldsymbol{u}, \boldsymbol{v}) \in (\mathbb{R}^d)^2, \forall t \in [0, 1], \qquad f(t\boldsymbol{u} + (1-t)\boldsymbol{v}) \le t\, f(\boldsymbol{u}) + (1-t)\, f(\boldsymbol{v}).$$

Exemples :

- Fonction linéaire : $f(\boldsymbol{w}) = \boldsymbol{a}^{\mathrm{T}}\boldsymbol{w} + b$;
- Norme au carré : $f(\boldsymbol{w}) = \|\boldsymbol{w}\|^2 = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}$.

Fonctions convexes lisses/douces

Convexité et gradient

Une fonction $f : \mathbb{R}^d \to \mathbb{R}$ de classe $\mathcal{C}^1$ est convexe si et seulement si

$$\forall \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d, \quad f(v) \geq f(u) + \nabla f(u)^{\mathrm{T}}(v - u).$$

Fonctions convexes lisses/douces

Convexité et gradient

Une fonction $f : \mathbb{R}^d \to \mathbb{R}$ de classe $\mathcal{C}^1$ est convexe si et seulement si

$$\forall \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d, \quad f(v) \geq f(u) + \nabla f(u)^{\mathrm{T}}(v - u).$$

L'autre inégalité clé en optimisation.

Fonctions convexes lisses/douces

Convexité et gradient

Une fonction $f : \mathbb{R}^d \to \mathbb{R}$ de classe $\mathcal{C}^1$ est convexe si et seulement si

$$\forall \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d, \quad f(v) \geq f(u) + \nabla f(u)^{\mathrm{T}}(v - u).$$

L'autre inégalité clé en optimisation.

Convexité et matrice hessienne

Une fonction $f : \mathbb{R}^d \to \mathbb{R}$ de classe $\mathcal{C}^2$ est dite convexe si et seulement si $\nabla^2 f(\boldsymbol{w}) \succeq 0$ pour tout vecteur $\boldsymbol{w} \in \mathbb{R}^d$, .

Optimisation et fonction convexe

$$\operatorname*{minimiser}_{\boldsymbol{w} \in \mathbb{R}^d} f(\boldsymbol{w}), \ f \text{ convexe.}$$

$$\operatorname*{minimiser}_{\boldsymbol{w}\in\mathbb{R}^d} f(\boldsymbol{w}), f \text{ convexe.}$$

Théorème

Tout minimum local de f est un minimum global.

Optimisation et fonction convexe

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimiser}}\, f(\boldsymbol{w}),\ f \text{ convexe.}$$

Théorème

Tout minimum local de f est un minimum global.

Corollaire

Si f est de classe $\mathcal{C}^1$, tout point $\boldsymbol{w}^*$ tel que $\|\nabla f(\boldsymbol{w}^*)\| = 0$ est un minimum global de f.

Outline

Fonctions fortement convexes

Définition

Une fonction $f\ : \mathbb{R}^d \to \mathbb{R}$ est μ-fortement convexe si pour tous $(\boldsymbol{u}, \boldsymbol{v}) \in (\mathbb{R}^d)^2$ et $t \in [0,1]$, on a

$$f(t\boldsymbol{u} + (1-t)\boldsymbol{v}) \ \leq \ t\,f(\boldsymbol{u}) + (1-t)f(\boldsymbol{v}) - \frac{\mu}{2} t(1-t)\|\boldsymbol{v} - \boldsymbol{u}\|^2.$$

$\boldsymbol{w} \mapsto \frac{\mu}{2}\|\boldsymbol{w}\|^2$ est μ-fortement convexe.

Fonctions fortement convexes

Définition

Une fonction $f : \mathbb{R}^d \to \mathbb{R}$ est μ-fortement convexe si pour tous $(\boldsymbol{u}, \boldsymbol{v}) \in (\mathbb{R}^d)^2$ et $t \in [0,1]$, on a

$$f(t\boldsymbol{u} + (1-t)\boldsymbol{v}) \;\leq\; t\,f(\boldsymbol{u}) + (1-t)f(\boldsymbol{v}) - \frac{\mu}{2}t(1-t)\|\boldsymbol{v} - \boldsymbol{u}\|^2.$$

$\boldsymbol{w} \mapsto \frac{\mu}{2}\|\boldsymbol{w}\|^2$ est μ-fortement convexe.

Théorème

- Une fonction fortement convexe a au plus un minimum global.
- Une fonction continue fortement convexe a un unique minimum global.

Fonctions fortement convexes

Gradient et convexité forte

Soit $f : \mathbb{R}^d \to \mathbb{R}$ de classe $\mathcal{C}^1$. Alors,

$$\forall \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d, \quad f(\boldsymbol{v}) \geq f(\boldsymbol{u}) + \nabla f(\boldsymbol{u})^{\mathrm{T}}(\boldsymbol{v} - \boldsymbol{u}) + \frac{\mu}{2}\|\boldsymbol{v} - \boldsymbol{u}\|^2.$$

Hessienne et convexité forte

Soit $f : \mathbb{R}^d \to \mathbb{R}$ de classe $\mathcal{C}^2$. Alors

$$f \text{ est } \mu\text{-fortement convexe} \quad \Longleftrightarrow \quad \nabla^2 f(\boldsymbol{w}) \succeq \mu \mathrm{I} \;\; \forall \boldsymbol{w} \in \mathbb{R}^d.$$

Minimisation d'une quadratique convexe

$$\operatorname*{minimiser}_{\boldsymbol{w} \in \mathbb{R}^d} f(\boldsymbol{w}) := \frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{w} + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{w}, \quad \boldsymbol{A} \succeq 0.$$

- $\nabla^2 f(\boldsymbol{w}) = \boldsymbol{A}$;
- Fortement convexe si $\boldsymbol{A} \succ 0$ avec $\mu = \lambda_{\mathsf{min}}(\boldsymbol{A})$.

Exemples de problèmes fortement convexes

Minimisation d'une quadratique convexe

$$\text{minimiser}_{\boldsymbol{w} \in \mathbb{R}^d} f(\boldsymbol{w}) := \frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{w} + \boldsymbol{b}^{\mathrm{T}}\boldsymbol{w}, \quad \boldsymbol{A} \succeq 0.$$

- $\nabla^2 f(\boldsymbol{w}) = \boldsymbol{A}$;
- Fortement convexe si $\boldsymbol{A} \succ 0$ avec $\mu = \lambda_{\min}(\boldsymbol{A})$.

Projection sur un ensemble fermé convexe

$$\text{minimiser}_{\boldsymbol{w} \in \mathcal{X}} \frac{1}{2}\|\boldsymbol{w} - \boldsymbol{a}\|^2, \quad \mathcal{X} \text{ fermé convexe.}$$

- Generalise le cas $\mathcal{X} = \mathbb{R}^d$;
- L'objectif est 1-fortement convexe $\Rightarrow$ il existe une unique solution.

Outline

1. Convexité
   - Ensembles et fonctions convexes
   - Convexité forte
   - Analyse convexe
2. Programmation convexe
3. Algorithmes pour l'optimisation convexe

Sur les fonctions non lisses

Définition

Une fonction est dite **non lisse** si elle n'est pas dérivable partout.
NB: Non lisse $\neq$ Discontinue.

Exemples de fonctions non lisses

- $w \mapsto |w|$ de $\mathbb{R}$ dans $\mathbb{R}$;
- $w \mapsto \|w\|_1 = \sum_{i=1}^{d} |w_i|$ de $\mathbb{R}^d$ dans $\mathbb{R}$;
- ReLU: $w \mapsto \max\{w, 0\}$ de $\mathbb{R}^d$ dans $\mathbb{R}$.

Optimisation non lisse

- Un algorithme d'optimisation lisse est basé sur des dérivées;
- Non lisse $\Leftrightarrow$ Pas de dérivée en certains points;
- On utilise des notions plus générales de dérivée.

Alternatives

- Si possible, reformuler en un problème lisse :
  Ex) $\text{minimiser}_{w \in \mathbb{R}} |w|$ se ré-écrit

$$\underset{w, t^+, t^- \in \mathbb{R}}{\text{minimiser}} \; t^+ - t^- \quad \text{s. c.} \quad w = t^+ - t^-, t^+ \geq 0, t^- \geq 0.$$

- Si la *fonction* est lipschitzienne, elle possède un gradient en presque tous les points (mais souvent pas en les minima).
  Ex) Fonction d'activation des réseaux de neurones
  $ReLU(\boldsymbol{w}) = [\max\{w_i, 0\}]$.

Optimisation convexe non lisse

Définition

Soit $f : \mathbb{R}^d \rightarrow \mathbb{R}$ une fonction convexe. Un vecteur $\boldsymbol{g} \in \mathbb{R}^d$ est un **sous-gradient** de f en $\boldsymbol{w} \in \mathbb{R}^d$ si

$$\forall \boldsymbol{z} \in \mathbb{R}^d, \qquad f(\boldsymbol{z}) \ \geq \ f(\boldsymbol{w}) + \boldsymbol{g}^{\mathrm{T}}(\boldsymbol{z} - \boldsymbol{w}).$$

L'ensemble des sous-gradients de f en $\boldsymbol{w}$ s'appelle le *sous-différentiel* de f en $\boldsymbol{w}$: on le note $\partial f(\boldsymbol{w})$.

- Si f dérivable en $\boldsymbol{w}$, $\partial f(\boldsymbol{w}) = \{\nabla f(\boldsymbol{w})\}$;
- $0 \in \partial f(\boldsymbol{w}) \Leftrightarrow \boldsymbol{w}$ minimum de f!

Exemple : Soit $f : \mathbb{R} \rightarrow \mathbb{R}$, $f(w) = |w|$.

$$\partial f(w) \ = \ \begin{cases} -1 & \text{si } w < 0 \\ 1 & \text{si } w > 0 \\ [-1, 1] & \text{si } w = 0. \end{cases}$$

Méthode de sous-gradient

Itération pour minimiser$_{\mathbf{w}}$ $f(\mathbf{w})$, f convex nonsmooth

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \alpha_k \mathbf{g}_k, \qquad \mathbf{g}_k \in \partial f(\mathbf{w}_k).$$

- Dépend du choix du sous-gradient;
- Choix de α_k plus technique (f peut croître dans la direction d'un sous-gradient !).

À retenir

- Certaines méthodes dites "de gradient" se basent en fait sur des sous-gradients;
- Ceux-ci sont bien compris pour des problèmes simples, et dans le cas convexe.

1. La sphère $\{\boldsymbol{w} \mid \|\boldsymbol{w}\|^2 = 1\}$ est-elle un ensemble convexe ?
2. Quelle est la différence entre une fonction convexe et une fonction fortement convexe du point de vue des minima ?
3. Quel objet mathématique remplace le gradient en optimisation non lisse ?

Table des matières

Outline

Programmes convexes

Forme de base

$$
\left\{
\begin{array}{lll}
\text{minimiser}_{\boldsymbol{w} \in \mathbb{R}^d} & f(\boldsymbol{w}) & \\
\text{s. c.} & g_i(\boldsymbol{w}) \le 0, & i = 1, \dots, m \\
 & h_i(\boldsymbol{x}) = 0, & i = 1, \dots, \ell,
\end{array}
\right.
$$

avec $f : \mathbb{R}^d \to \mathbb{R}$ convexe, $g_i : \mathbb{R}^d \to \mathbb{R}$ convexe pour tout i et $h_i : \mathbb{R}^d \to \mathbb{R}$ affine pour tout i.

- Tout programme convexe peut être mis sous cette forme;
- On considère ici que les fonctions ne prennent que des valeurs finies (mais théorie plus générale).

Exemple : Programmes linéaires (LP)

$$
\begin{array}{ll}
\text{minimiser}_{\boldsymbol{w}\in\mathbb{R}^d} & \boldsymbol{c}^{\mathrm{T}}\boldsymbol{w} \\
\text{s. c.} & \boldsymbol{A}\boldsymbol{w} = \boldsymbol{b} \\
 & \boldsymbol{w} \geq 0,
\end{array}
$$

avec $\boldsymbol{c} \in \mathbb{R}^d$, $\boldsymbol{A} \in \mathbb{R}^{m\times d}$, $\boldsymbol{b} \in \mathbb{R}^m$.

- Problème canonique en optimisation;
- Possible de résoudre des problèmes avec des millions/milliards de variables.

Exemple : Programmes quadratiques (QP)

$$
\begin{array}{ll}
\text{minimiser}_{\boldsymbol{w}\in\mathbb{R}^d} & \frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{w} + \boldsymbol{c}^{\mathrm{T}}\boldsymbol{w} \\
\text{s. c.} & \boldsymbol{A}\boldsymbol{w} = \boldsymbol{b} \\
 & \boldsymbol{w} \geq 0,
\end{array}
$$

avec $\boldsymbol{H} \in \mathbb{R}^{d\times d}$, $\boldsymbol{c} \in \mathbb{R}^d$, $\boldsymbol{A} \in \mathbb{R}^{m\times d}$, $\boldsymbol{b} \in \mathbb{R}^m$.

- Extension du cas sans contraintes avec contraintes linéaires;
- Peut avoir une solution même si $\boldsymbol{H} \succeq 0$!

Exemples : Programmes coniques d'ordre deux (SOCP)

$$
\begin{array}{ll}
\text{minimiser}_{\boldsymbol{w} \in \mathbb{R}^d} & \boldsymbol{c}^{\mathrm{T}} \boldsymbol{w} \\
\text{s. c.} & \|\boldsymbol{A}_i \boldsymbol{w} + \boldsymbol{b}_i\| \le \boldsymbol{c}_i^{\mathrm{T}} \boldsymbol{w} + d_i, \quad i = 1, \dots, n.
\end{array}
$$

avec pour tout $i = 1, \dots, m$, $\boldsymbol{A}_i \in \mathbb{R}^{n_i \times d}$, $\boldsymbol{b}_i \in \mathbb{R}^{n_i}$, $\boldsymbol{c}_i \in \mathbb{R}^d$ et $d_i \in \mathbb{R}$.

- Généralisent LP et QP;
- Peuvent être résolus efficacement.

Exemples : Programmes semi-définis positifs (SDP)

$$
\begin{array}{ll}
\text{minimiser}_{\boldsymbol{W} \in \mathbb{R}^{d \times d}} & \text{trace}(\boldsymbol{C}^{\mathrm{T}} \boldsymbol{W}) \\
\text{s. c.} & \text{trace}(\boldsymbol{A}_i^{\mathrm{T}} \boldsymbol{X}) = b_i, \quad i = 1, \dots, n \\
& \boldsymbol{W} = \boldsymbol{W}^{\mathrm{T}} \succ 0,
\end{array}
$$

avec $\boldsymbol{C}, \boldsymbol{A}_1, \dots, \boldsymbol{A}_n \in \mathbb{R}^{d \times d}$ et $\boldsymbol{b} = [b_i] \in \mathbb{R}^n$.

- Généralisent LP et QP;
- Possible de les résoudre en temps polynomial, mais les calculs algébriques peuvent être coûteux;
- Problèmes souvent de grandes tailles (relaxations de problèmes continus ou combinatoires).

Outline

Dualité en général

Principe

- Définir une fonction qui combine l'objectif et les contraintes;
- Maximiser cette fonction conduit à un nouveau problème d'optimisation dit dual.

Définition

Soit le problème d'optimisation

$$\left\{ \begin{array}{ll} \text{minimiser}_{\boldsymbol{w} \in \mathbb{R}^d} & f(\boldsymbol{w}) \\ \text{s. c.} & g_i(\boldsymbol{w}) \leq 0, \quad i = 1, \dots, m \\ & h_i(\boldsymbol{w}) = 0, \quad i = 1, \dots, \ell, \end{array} \right.$$

Le **lagrangien** du problème est défini par

$$\begin{array}{rl} \multicolumn{2}{l}{\forall (\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{s}) \in \mathbb{R}^d \times \mathbb{R}^\ell \times (\mathbb{R}_+)^m,} \\ \mathcal{L}(\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{s}) & := f(\boldsymbol{w}) + \boldsymbol{v}^{\mathrm{T}} \boldsymbol{h}(\boldsymbol{w}). + \boldsymbol{s}^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{w}) \\ & = f(\boldsymbol{w}) + \sum_{i=1}^{\ell} v_i h_i(\boldsymbol{w}) + \sum_{i=1}^{m} s_i g_i(\boldsymbol{w}). \end{array}$$

Lagrangien

Définition

Soit le problème d'optimisation

$$\left\{\begin{array}{lll} \mathrm{minimiser}_{\boldsymbol{w}\in\mathbb{R}^d} & f(\boldsymbol{w}) & \\ \mathrm{s.\ c.} & g_i(\boldsymbol{w}) \le 0, & i = 1,\dots,m \\ & h_i(\boldsymbol{w}) = 0, & i = 1,\dots,\ell, \end{array}\right.$$

Le **lagrangien** du problème est défini par

$$\begin{aligned} &\forall (\boldsymbol{w},\boldsymbol{v},\boldsymbol{s}) \in \mathbb{R}^d \times \mathbb{R}^\ell \times (\mathbb{R}_+)^m, \\ &\qquad \mathcal{L}(\boldsymbol{w},\boldsymbol{v},\boldsymbol{s}) \quad := f(\boldsymbol{w}) + \boldsymbol{v}^{\mathrm{T}}\boldsymbol{h}(\boldsymbol{w}). + \boldsymbol{s}^{\mathrm{T}}\boldsymbol{g}(\boldsymbol{w}) \\ &\qquad\qquad\qquad\quad = f(\boldsymbol{w}) + \textstyle\sum_{i=1}^{\ell} v_i h_i(\boldsymbol{w}) + \sum_{i=1}^{m} s_i g_i(\boldsymbol{w}). \end{aligned}$$

- Agrégation de l'objectif et des contraintes;
- Cette fonction peut prendre des valeurs infinies !

Problème dual

Définition: $$\underset{\boldsymbol{v}\in\mathbb{R}^{\ell},\boldsymbol{s}\in(\mathbb{R}_+)^m}{\text{maximiser}}\ d(\boldsymbol{v},\boldsymbol{s}), \quad d(\boldsymbol{v},\boldsymbol{s}) := \inf_{\boldsymbol{w}\in\mathbb{R}^d} \mathcal{L}(\boldsymbol{w},\boldsymbol{v},\boldsymbol{s})$$

- Problème concave (opposé convexe) !
- Le problème originel est dit primal et s'écrit

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimiser}} \sup_{\boldsymbol{v}\in\mathbb{R}^{\ell},\boldsymbol{s}\in(\mathbb{R}_+)^m} \mathcal{L}(\boldsymbol{w},\boldsymbol{v},\boldsymbol{s})$$

Problème dual

Définition: $\underset{\boldsymbol{v}\in\mathbb{R}^{\ell},\boldsymbol{s}\in(\mathbb{R}_+)^m}{\text{maximiser}}\ d(\boldsymbol{v},\boldsymbol{s}), \quad d(\boldsymbol{v},\boldsymbol{s}) := \inf_{\boldsymbol{w}\in\mathbb{R}^d} \mathcal{L}(\boldsymbol{w},\boldsymbol{v},\boldsymbol{s})$

- Problème concave (opposé convexe) !
- Le problème originel est dit primal et s'écrit

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimiser}} \sup_{\boldsymbol{v}\in\mathbb{R}^{\ell},\boldsymbol{s}\in(\mathbb{R}_+)^m} \mathcal{L}(\boldsymbol{w},\boldsymbol{v},\boldsymbol{s})$$

À quoi sert le dual?

- Si on a dualité forte, les deux problèmes ont même valeur optimale;
- Un triplet $(\boldsymbol{w}^*,\boldsymbol{v}^*,\boldsymbol{s}^*)$ solution primale-duale vérifie alors :

$$\mathcal{L}(\boldsymbol{w}^*,\boldsymbol{v},\boldsymbol{s}) \le \mathcal{L}(\boldsymbol{w}^*,\boldsymbol{v}^*,\boldsymbol{s}^*) \le \mathcal{L}(\boldsymbol{w},\boldsymbol{v}^*,\boldsymbol{s}^*).$$

- Pour résoudre le problème, on cherche alors un point selle du lagrangien qui vérifie les contraintes.

Conditions de Karush-Kuhn-Tucker (KKT)

Problème: $\text{minimiser}_{\boldsymbol{w} \in \mathbb{R}^d} f(\boldsymbol{w}) \quad \text{s. c.} \;\; \boldsymbol{g}(\boldsymbol{w}) \le 0_{\mathbb{R}^m}, \boldsymbol{h}(\boldsymbol{w}) = 0_{\mathbb{R}^\ell}.$

Énoncé

On suppose qu'il y a dualité forte et que $f, \boldsymbol{g}, \boldsymbol{h}$ sont de classe $\mathcal{C}^1$. Alors, si $\boldsymbol{w}^* \in \mathbb{R}^d$ est une solution du problème, il existe $\boldsymbol{v}^* \in \mathbb{R}^\ell$ et $\boldsymbol{s}^* \in \mathbb{R}^m$ tels que

$$
\begin{array}{rcl}
\nabla f(\boldsymbol{w}^*) + \boldsymbol{J_h}(\boldsymbol{w}^*)\boldsymbol{v}^* + \boldsymbol{J_g}(\boldsymbol{w}^*)\boldsymbol{s}^* & = & 0 \\
\boldsymbol{h}(\boldsymbol{w}^*) & = & 0_{\mathbb{R}^\ell} \\
\boldsymbol{g}(\boldsymbol{w}^*) & \le & 0_{\mathbb{R}^m} \\
g_i(\boldsymbol{w}^*)s_i^* & = & 0 \ \forall i = 1, \ldots, m.
\end{array}
$$

Conditions de Karush-Kuhn-Tucker (KKT)

Problème: $\text{minimiser}_{\boldsymbol{w} \in \mathbb{R}^d} f(\boldsymbol{w}) \quad \text{s. c.} \ \boldsymbol{g}(\boldsymbol{w}) \le 0_{\mathbb{R}^m}, \boldsymbol{h}(\boldsymbol{w}) = 0_{\mathbb{R}^\ell}.$

Énoncé

On suppose qu'il y a dualité forte et que $f, \boldsymbol{g}, \boldsymbol{h}$ sont de classe $\mathcal{C}^1$. Alors, si $\boldsymbol{w}^* \in \mathbb{R}^d$ est une solution du problème, il existe $\boldsymbol{v}^* \in \mathbb{R}^\ell$ et $\boldsymbol{s}^* \in \mathbb{R}^m$ tels que

$$\begin{array}{rcl} \nabla f(\boldsymbol{w}^*) + \boldsymbol{J_h}(\boldsymbol{w}^*)\boldsymbol{v}^* + \boldsymbol{J_g}(\boldsymbol{w}^*)\boldsymbol{s}^* & = & 0 \\ \boldsymbol{h}(\boldsymbol{w}^*) & = & 0_{\mathbb{R}^\ell} \\ \boldsymbol{g}(\boldsymbol{w}^*) & \le & 0_{\mathbb{R}^m} \\ g_i(\boldsymbol{w}^*)s_i^* & = & 0 \ \forall i = 1, \dots, m. \end{array}$$

- Les trois premières conditions correspondent à annuler le gradient du lagrangien par rapport à $\boldsymbol{w}^*$, $\boldsymbol{v}^*$ et $\boldsymbol{s}^*$.
- $\boldsymbol{v}^*$ et $\boldsymbol{s}^*$ sont des variables duales aussi appelées multiplicateurs de Lagrange.

Outline

Schéma de résolution

Pour les classes majeures

- LP, QP, SDP, SOCP;
- D'autres classes non traitées ici.

Dans la suite

- On illustre les notions dans le cadre des programmes linéaires;
- On donne l'idée derrière les méthodes dites de points intérieurs.

Cas des programmes linéaires

Problème de base (dit primal)

$$\underset{\boldsymbol{w}\in\mathbb{R}^d}{\text{minimiser}}\, \boldsymbol{c}^{\mathrm{T}}\boldsymbol{w} \quad \text{s. c.} \quad \boldsymbol{A}\boldsymbol{w} = \boldsymbol{b},\ \boldsymbol{w} \geq 0,$$

$\boldsymbol{A} \in \mathbb{R}^{m\times d}$, $\boldsymbol{b} \in \mathbb{R}^m$.

Problème dual (aussi LP !)

$$\underset{\boldsymbol{v}\in\mathbb{R}^m, \boldsymbol{s}\in\mathbb{R}^d}{\text{maximiser}}\, \boldsymbol{b}^{\mathrm{T}}\boldsymbol{v} \quad \text{s. c.} \quad \boldsymbol{A}^{\mathrm{T}}\boldsymbol{v} + \boldsymbol{s} = \boldsymbol{c},\ \boldsymbol{s} \geq 0.$$

Conditions de KKT

Si $\boldsymbol{w}^* \in \mathbb{R}^d$ est une solution, il existe $\boldsymbol{v}^* \in \mathbb{R}^m$ et $\boldsymbol{s}^* \in \mathbb{R}^d$ tels que

$$\begin{array}{ll}
\boldsymbol{A}\boldsymbol{w}^* = \boldsymbol{b}, & \boldsymbol{w}^* \geq 0, \\
\boldsymbol{A}^{\mathrm{T}}\boldsymbol{v}^* + \boldsymbol{s}^* = \boldsymbol{c}, & \boldsymbol{s}^* \geq 0, \\
w_i s_i = 0\ \forall i = 1, \dots, d. &
\end{array}$$

Dualité forte en programmation linéaire

Théorème

Pour tout programme linéaire,

i) Soit ce problème et son dual ont chacun une solution, et l'ensemble vérifie les conditions de KKT;
ii) Soit un des deux problèmes n'a pas de solution, et l'ensemble réalisable de l'autre est vide (problème irréalisable);
iii) Soit les deux ensembles réalisables sont vides.

Dualité forte en programmation linéaire

Théorème

Pour tout programme linéaire,

i) Soit ce problème et son dual ont chacun une solution, et l'ensemble vérifie les conditions de KKT;
ii) Soit un des deux problèmes n'a pas de solution, et l'ensemble réalisable de l'autre est vide (problème irréalisable);
iii) Soit les deux ensembles réalisables sont vides.

Conséquences

- Trouver une solution revient à trouver un triplet vérifiant les contraintes de chaque problème !
- C'est ce que font les approches de points intérieurs, de manière primale-duale ou duale.

Points intérieurs

Principe

- Partir d'un triplet $(\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{s})$ réalisable pour le problème et son dual avec $\boldsymbol{w} > 0$ et $\boldsymbol{s} > 0$.
- Appliquer la méthode de Newton au système de KKT

$$\left\{ \begin{array}{lll} \boldsymbol{A}\boldsymbol{w}^* & = & \boldsymbol{b} \\ \boldsymbol{A}^{\mathrm{T}}\boldsymbol{v}^* + \boldsymbol{s}^* & = & \boldsymbol{c} \\ w_i s_i & = & 0 \ \forall i = 1, \dots, d. \end{array} \right.$$

- Obtenir un nouveau point "intérieur" avec une valeur $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{s}$ réduite.

Points intérieurs

Principe

- Partir d'un triplet $(\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{s})$ réalisable pour le problème et son dual avec $\boldsymbol{w} > 0$ et $\boldsymbol{s} > 0$.
- Appliquer la méthode de Newton au système de KKT

$$\left\{ \begin{array}{lcl} \boldsymbol{A}\boldsymbol{w}^* & = & \boldsymbol{b} \\ \boldsymbol{A}^{\mathrm{T}}\boldsymbol{v}^* + \boldsymbol{s}^* & = & \boldsymbol{c} \\ w_i s_i & = & 0 \ \forall i = 1, \dots, d. \end{array} \right.$$

- Obtenir un nouveau point "intérieur" avec une valeur $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{s}$ réduite.

Ça marche !

- Vitesses de convergence en théorie;
- Très bonnes performances en pratique;
- Implémentations très sophistiquées.

Révisons ! Programmation convexe

1. Citer deux classes de problèmes d'optimisation convexe.
2. Quel objet mathématique concatène l'objectif et les contraintes d'un problème d'optimisation ? Que représente une solution du problème par rapport à cet objet ?
3. Quelle technique s'applique aussi bien à la résolution de programmes linéaires qu'à la résolution de programmes quadratiques ?

Table des matières

Outline

$$\underset{\boldsymbol{x}\in\mathbb{R}^d}{\text{minimiser}}\, f(\boldsymbol{x}), \qquad f \in \mathcal{C}_L^{1,1}.$$

Descente de gradient

- Itération: $\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha_k \nabla f(\boldsymbol{w}_k)$, terminer si $\nabla f(\boldsymbol{w}_k) = 0$.
- Choix typique en théorie : $\alpha_k = \frac{1}{L}$.

Avec la convexité

Hypothèse : $f^* = \min_{\boldsymbol{w}\in\mathbb{R}^d} f(\boldsymbol{w})$ est atteint.

- Garanties relativement à un minimum global;
- On peut montrer que $f(\boldsymbol{w}_k) \to f^*$;
- On peut aussi montrer une convergence vers l'argmin

Complexités et vitesses de convergence

Cas non convexe

- Critère : $\|\nabla f(\boldsymbol{w}_k)\|$;
- Idée : être proche d'un point stationnaire.

Cas convexe/fortement convexe

- : $f(\boldsymbol{w}_k) - f^* \leq \epsilon$, avec $f^* = \min_{\boldsymbol{w} \in \mathbb{R}^d} f(\boldsymbol{w})$;
- Idée : être proche de la valeur à l'optimum.
- Valeur liée à $\|\boldsymbol{w}_k - \boldsymbol{w}^*\|$ dans le cas fortement convexe.

Résultats de complexité

Théorème

Si $f \in \mathcal{C}^{1,1}_L$ est convexe et $\alpha_k = \frac{1}{L}$, la descente de gradient calcule $\boldsymbol{w}_k$ tel que $f(\boldsymbol{w}_k) - f^* \leq \epsilon$ en au plus

- $\mathcal{O}(\epsilon^{-1})$ itérations;
- $\mathcal{O}\left(\frac{L}{\mu}\ln(\epsilon^{-1})\right)$ itérations si f est μ-fortement convexe.

- Cas non convexe : $\mathcal{O}(\epsilon^{-2})$ pour garantir $\|\nabla f(\boldsymbol{w}_k)\| \leq \epsilon$.
- On dit que la descente de gradient possède une meileure complexité dans le cas convexe/fortement convexe.

Vitesses de convergence

Théorème

Si $f \in \mathcal{C}^{1,1}_L$ est convexe et $\alpha_k = \frac{1}{L}$, pour tout $K \in \mathbb{N}$, on a:

$$f(\boldsymbol{w}_K) - f^* \leq \frac{L \max_{\boldsymbol{w} \in \operatorname{argmin}_{\boldsymbol{v}} f(\boldsymbol{v})} \|\boldsymbol{w}_0 - \boldsymbol{w}\|}{2} \frac{1}{K}$$

pour f convexe, et

$$f(\boldsymbol{w}_K) - f^* \leq \left(1 - \frac{\mu}{L}\right)^K (f(\boldsymbol{w}_0) - f^*).$$

pour f μ-fortement convexe.

- Cas non convexe : $\mathcal{O}(1/\sqrt{K})$ pour $\min_{0 \leq k \leq K-1} \|\nabla f(\boldsymbol{w}_k)\|$.
- On dit que la descente de gradient converge plus rapidement dans le cas fortement convexe que dans le cas convexe.

Outline

Accélération en optimisation convexe

Motivation

- En optimisation non convexe, $\mathcal{O}(1/\sqrt{K})$ est la meilleure vitesse de convergence pour une méthode type gradient;
- Dans le cas convexe, c'est $\mathcal{O}(1/K^2)$, mieux que la descente de gradient en $\mathcal{O}(1/K)$.

Comment obtenir cette meilleure vitesse ?

- Stratégies de **gradient accéléré**, basées sur l'idée de **momentum**;
- **Principe :** Réutiliser l'information de l'itération précédente.

Algorithme($\boldsymbol{w}_0, \boldsymbol{w}_{-1} = \boldsymbol{w}_0$)

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha_k \nabla f\left(\boldsymbol{w}_k + \beta_k(\boldsymbol{w}_k - \boldsymbol{w}_{k-1})\right) + \beta_k(\boldsymbol{w}_k - \boldsymbol{w}_{k-1}).$$

- Un appel de gradient par itération;
- Terme de momentum : $\boldsymbol{w}_k - \boldsymbol{w}_{k-1}$ (pas précédent).

Version à deux suites($\boldsymbol{w}_0, \boldsymbol{z}_0 = \boldsymbol{w}_0$)

$$\left\{ \begin{array}{lcl} \boldsymbol{w}_{k+1} & = & \boldsymbol{z}_k - \alpha_k \nabla f(\boldsymbol{z}_k) \\ \boldsymbol{z}_{k+1} & = & \boldsymbol{w}_{k+1} + \beta_{k+1}(\boldsymbol{w}_{k+1} - \boldsymbol{w}_k). \end{array} \right.$$

Choix des paramètres

Longueur de pas α_k

- $\alpha_k = \frac{1}{L}$;
- Autres : décroissante, recherche linéaire, etc.

Momentum β_k

- f μ-fortement convexe : $\beta_k = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$;
- f convexe : Utiliser deux suites

$$t_{k+1} = \frac{1}{2}(1 + \sqrt{1 + 4t_k^2}), t_0 = 0, \quad \beta_k = \frac{t_k - 1}{t_{k+1}}.$$

Vitesses de convergence

Fonctions convexes

- Descente de gradient : $f(\boldsymbol{w}_K) - f^* \leq \mathcal{O}\left(\frac{1}{K}\right)$;
- Gradient accéléré : $f(\boldsymbol{w}_K) - f^* \leq \mathcal{O}\left(\frac{1}{K^2}\right)$.

Fonctions μ-fortement convexes

- Descente de gradient :

$$f(\boldsymbol{w}_K) - f^* \leq \left(1 - \frac{\mu}{L}\right)^K (f(\boldsymbol{w}_0) - f^*).$$

- Gradient accéléré

$$f(\boldsymbol{w}_K) - f^* \leq \mathcal{C}\left(1 - \sqrt{\frac{\mu}{L}}\right)^K (f(\boldsymbol{w}_0) - f^*).$$

Autres algorithmes accélérés

Méthode de la boule lestée (*Heavy ball*, Polyak, 1964)

- $\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha \nabla f(\boldsymbol{w}_k) + \beta(\boldsymbol{w}_k - \boldsymbol{w}_{k+1})$;
- Optimale sur des quadratiques fortement convexes, mais ne converge pas toujours pour f fortement convexe!
- "Précurseure" du gradient accéléré.

Méthode du gradient conjugué (Hestenes and Stiefel, 1952)

- $\boldsymbol{w}_{k+1} = \boldsymbol{w}_k + \alpha_k p_k, \ p_k = -\nabla f(x_k) + \beta_k p_{k-1}$;
- Developée pour les quadratiques fortement convexes, optimale sans connaître L ou μ!
- D'autres versions pour les fonctions fortement convexes, convexes et non convexes efficaces en pratique.

Outline

$$\underset{\boldsymbol{w} \in R^d}{\text{minimiser}} \quad \underbrace{f(\boldsymbol{w})}_{attache\ \ donnes} \quad + \quad \underbrace{\lambda\Omega(\boldsymbol{w})}_{rgularisation} \ .$$

où $\lambda > 0$ est un paramètre de régularisation.

Exemple : Régularisation *ridge* ou écrêtée

$$\underset{\boldsymbol{w} \in \mathbb{R}^d}{\text{minimiser}}\, f(\boldsymbol{w}) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2.$$

Interprétations :

- Revient à ajouter une contrainte sur $\|\boldsymbol{w}\|^2 = \sum_{i=1}^{d} w_i^2$;
- Favorise les vecteurs $\boldsymbol{w}$ avec des composantes uniformément faibles en amplitude;
- Réduit la variance des solutions par rapport aux données;
- Problème fortement convexe pour λ suffisamment grand.

Problème régularisé et exemples (2)

Régularisation ℓ_0

- Objectif : Trouver $\boldsymbol{w} \in \mathbb{R}^d$ qui colle aux données avec le plus de coefficients nuls possible;
- Problème idéal : $\min_{\boldsymbol{w}} f(\boldsymbol{w}) + \lambda \|\boldsymbol{w}\|_0$, avec $\|\boldsymbol{v}\|_0 := |\{i | [\boldsymbol{v}]_i \neq 0\}|$. Mais la fonction $\| \cdots \|_0$ est non lisse, discontinue et introduit de la combinatoire.

A better approach: LASSO regularization

LASSO=Least Absolute Shrinkage and Selection Operator

$$\min_{\boldsymbol{w} \in \mathbb{R}^d} f(\boldsymbol{w}) + \lambda \|\boldsymbol{w}\|_1, \quad \|\boldsymbol{w}\|_1 = \sum_{i=1}^{d} |w_i|.$$

- $\| \cdot \|_1$ convexe, continue, norme;
- Non lisse mais possède des sous-gradients.

Résolution

Cadre : Optimisation composite

$$\operatorname*{minimiser}_{\boldsymbol{w} \in \mathbb{R}^d} f(\boldsymbol{w}) + \lambda \Omega(\boldsymbol{w}).$$

- $f \in \mathcal{C}^{1,1}$;
- Ω convexe.

Approche proximale

- Classique en optimisation : remplacer un problème par une suite de sous-problèmes plus simples;
- Ici on exploite la "douceur" de f et la structure de Ω pour construire des problèmes que l'on peut résoudre efficacement.

Méthode de gradient proximale

Itération

$$\boldsymbol{w}_{k+1} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^d} \left\{ f(\boldsymbol{w}_k) + \nabla f(\boldsymbol{w}_k)^{\mathrm{T}}(\boldsymbol{w} - \boldsymbol{w}_k) + \frac{1}{2\alpha_k}\|\boldsymbol{w} - \boldsymbol{w}_k\|_2^2 + \lambda\Omega(\boldsymbol{w}) \right\}.$$

- Si $\Omega \equiv 0$, la solution est $\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \alpha_k \nabla f(\boldsymbol{w}_k)$: c'est l'itération de la descente de gradient !
- En général, une itération coûte un calcul de gradient + 1 résolution de sous-problème.

Propriétés

- Complexité/Vitesses de convergence;
- Règles de choix de longueur de pas;
- Variantes accélérées/avec sous-gradients.

Exemple de méthode proximale : ISTA

Contexte

- Résoudre $\text{minimiser}_{\boldsymbol{w}\in\mathbb{R}^d} f(\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|_1$;
- Problème classique en traitement du signal/de l'image;
- La méthode du gradient proximal a une forme explicite.

Exemple de méthode proximale : ISTA

Contexte

- Résoudre $\operatorname{minimiser}_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w}) + \lambda\|\mathbf{w}\|_1$;
- Problème classique en traitement du signal/de l'image;
- La méthode du gradient proximal a une forme explicite.

Itération ISTA : Iterative Soft-Thresholding Algorithm

Définit $\mathbf{w}_{k+1}$ par coordonnées: pour tout $i \in \{1,\dots,d\}$,

$$[\mathbf{w}_{k+1}]_i \;=\; \begin{cases} [\mathbf{w}_k - \alpha_k \nabla f(\mathbf{w}_k)]_i + \alpha_k\lambda & \text{si } [\mathbf{w}_k - \alpha_k \nabla f(\mathbf{w}_k)]_i < -\alpha_k\lambda \\ [\mathbf{w}_k - \alpha_k \nabla f(\mathbf{w}_k)]_i - \alpha_k\lambda & \text{si } [\mathbf{w}_k - \alpha_k \nabla f(\mathbf{w}_k)]_i > \alpha_k\lambda \\ 0 & \text{si } [\mathbf{w}_k - \alpha_k \nabla f(\mathbf{w}_k)]_i \in [-\alpha_k\lambda, \alpha_k\lambda]. \end{cases}$$

Exemple de méthode proximale : ISTA (2)

Mise à jour dans ISTA

- Part du pas de gradient $\boldsymbol{w}_k - \alpha_k \nabla f(\boldsymbol{w}_k)$;
- Applique l'opérateur de *soft-thresholding* $s_{\alpha_k \lambda}(\bullet)$ à chaque coordonnée, avec

$$s_\mu(t) = \begin{cases} t + \mu & \text{si } t < -\mu \\ t - \mu & \text{si } t > \mu \\ 0 & \text{sinon.} \end{cases}$$

- Favorise les composantes nulles.

Variantes de ISTA

- Changement de longueur de pas;
- Ajout de momentum : FISTA (la plus utilisée en pratique).

Révisons ! Algorithmes d'optimisation convexe

1. Peut-on appliquer l'algorithme de descente de gradient à un problème convexe ?
2. Sur quoi repose l'idée de momentum ?
3. Quel est l'intérêt d'une régularisation ℓ_1 ?

Conclusion

Convexité et optimisation

- Convexité = Contexte favorable pour la minimisation;
- Classes de problèmes bien comprises;
- Solveurs efficaces.

Algorithmes d'optimisation convexe

- Un même algorithme possède de meilleures propriétés sur un problème convexe !
- Méthodes accélérées optimales pour les problèmes convexes;
- Stratégies de régularisation : apportent de la structure, "convexifient" souvent le problème.

Références

Ouvrages :

- S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.
- A. Beck, *First-order methods in optimization*, MOS-SIAM Series on Optimization, 2017.

Codes :

- `CVX`, `CVXPY` : logiciels académiques pour l'optimisation convexe (+`IPOPT` pour le cas non convexe).
- `Gurobi`, `CPLEX`, `MOSEK` : logiciels commerciaux pour les programmes linéaires (en variables continues et discrètes) et au-delà (programmes quadratiques, etc).

Fin de la seconde partie

Notebook

- Les effets de l'accélération...
- ...et quelques solveurs convexes.

Demain : Méthodes stochastiques avec Pierre Ablin.

Fin de la seconde partie

Notebook

- Les effets de l'accélération...
- ...et quelques solveurs convexes.

Demain : Méthodes stochastiques avec Pierre Ablin.

Merci beaucoup !