IFT 3515
Fonctions à plusieurs variables
Optimisation sans contraintes
Méthodes de région de confiance

Fabian Bastin
DIRO
Université de Montréal

Contexte

Nous considérons le problème

$$\min_{x \in \mathcal{R}^n} f(x)$$

où $f(x) \in C^1$ et $\nabla f(x)$ est lipschitzienne sur $\mathcal{R}^n$.

Note : ces conditions de continuité sont parfois violées, alors que les approches continuent de donner de bons résultats.

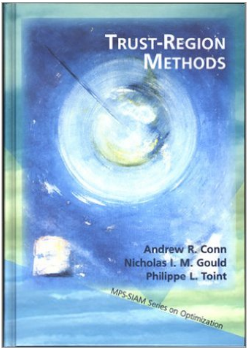
Les algorithmes d'optimisation itératifs travaillent typiquement à chaque itération avec un problème d'optimisation bien plus simple.

- Dans le cas des méthodes de recherche linéaire, le sous-problème est plus simple car il est unidimensionnel.
- En régions de confiance, le sous-problème reste en dimension n, mais la fonction objectif est simplifiée et l'espace de recherche retreint.

Sous-problème

Conceptuellement, l'approche de régions de confiance remplace un problème à n dimensions, sans contraintes, par une approximention à n dimensions, soumis à des contraintes. Avantages :

- le sous-problème ne doit pas être résolu exactement ;
- des algorithmes spécifiques existent pour traiter le sous-problème.



La bible ! A. Conn, M. Gould et Ph. L. Toint, "Trust-region methods", SIAM, 2000.

Presque 1000 pages !

Recherche linéaire vs méthodes de région de confiance

Méthodes de recherche linéaire

1. choisir une direction de descente d_k
2. déterminer un longueur de pas α_k pour réduire suffisamment f le long de d_k : $f(x_k + \alpha_k d_k)$.

Méthodes de région de confiance

1. choisir un pas s_k pour réduire le modèle de $f(x_k + s)$
2. accepter $x_{k+1} = x_k + s_k$ si la réduction prédite par le modèle est vérifiée en $f(x_k + s_k)$
3. autrement poser $x_{k+1} = x_k$ et raffiner le modèle.

Modèle

Nous voudrions un modèle m_k de f en x_k qui satisfait au minimum les hypothèses suivantes :

$$\begin{aligned} &m_k \in C^1 \\ &m_k(0) = f(x_k) \\ &\nabla m_k(0) = \nabla f(x_k) \end{aligned}$$

Généralement, un modèle quadratique est sélectionné :

$$m_k(s) = f(x_k) + \nabla f(x_k)^T s + \frac{1}{2} s^T H_k s$$

avec H_k symétrique (mais pas nécessairement définie positive !).

Région de confiance

Pour de petits s, le modèle devrait ressembler à la fonction f, mais plus s est grand, plus il est probable que m et f soient différents !

Nous allons contrer ce problème en bornant la norme du pas :

$$\|s_k\|_k \leq \Delta_k$$

où Δ_k est le rayon de la région de confiance et $\|\cdot\|_k$ est une norme, dont le choix peut être dépendant de l'itération k. Le plus souvent toutefois, nous travaillerons avec la norme 2 : $\|\cdot\|_2$.

Dès lors, le sous-problème de région de confiance est le problème sous contraintes

$$\begin{aligned} &\min_s \; m_k(s) \\ &\text{t.q. } \|s\|_k \leq \Delta_k. \end{aligned}$$

Région de confiance

La région de confiance à l'itération k est définie comme

$$\mathcal{B}_k = \{x \in \mathcal{R}^n \mid \|x - x_k\|_k \leq \Delta_k\},$$

Outre $\|\cdot\|_2$, des choix usuels pour la norme sont $\|\cdot\|_1$ et $\|\cdot\|_\infty$.

Rappel : si $x = (x_1, x_2, \ldots, x_n)$,

$$\|x\|_2 = \sqrt{x^T x}$$
$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$
$$\|x\|_\infty = \max\{|x_i|,\ i = 1, \ldots, n\}$$

Modèle quadratique

Si $f \in C^2$, on pourra choisir pour H_k la matrice hessienne $\nabla^2 f(x_k)$, mais comme dans le cadre de la recherche linéaire, elle peut être coûteuse à obtenir.

Si $f \in C^1$ et son gradient est une fonction lipschitzienne, on pourra recourir à des approximations, comme l'approche BFGS. Mais comme le caractère défini positif n'est plus requis, d'autres options s'ouvrent, en particulier l'approximation SR1.

Algorithme basique de région de confiance

Étape 0. Initialisation Un point initial x_0 et un rayon initial de région de confiance Δ_0 sont donnés. Les constantes η_1, η_2, γ_1 et γ_2 sont aussi données et satisfont

$$0 < \eta_1 \leq \eta_2 < 1 \text{ et } 0 < \gamma_1 \leq \gamma_2 < 1.$$

Calculer $f(x_0)$ et poser $k = 0$.

Étape 1. Définition du modèle Choisir $\|\cdot\|_k$ et définir un modèle m_k dans $\mathcal{B}_k$.

Étape 2. Calcul du pas Calculer un pas s_k qui "réduit suffisamment le modèle" m_k et tel que $x_k + s_k \in \mathcal{B}_k$.

Étape 3. Acceptation du point candidat Calculer $f(x_k + s_k)$ et définir

$$\rho_k = \frac{f(x_k) - f(x_k + s_k)}{m_k(0) - m_k(s_k)}. \tag{1}$$

Si $\rho_k \geq \eta_1$, définir $x_{k+1} = x_k + s_k$; autrement définir $x_{k+1} = x_k$.

Algorithme basique de région de confiance

Étape 4. Mise à jour de la région de confiance

$$\Delta_{k+1} \in \begin{cases} [\Delta_k, \infty) & \text{si } \rho_k \geq \eta_2, \\ [\gamma_2 \Delta_k, \Delta_k] & \text{si } \rho_k \in [\eta_1, \eta_2), \\ [\gamma_1 \Delta_k, \gamma_2 \Delta_k] & \text{si } \rho_k < \eta_1. \end{cases}$$

Incrémenter k de 1 et retourner à l'étape 1.

Dans cette description, des valeurs raisonnables pour les constantes sont, par exemple,

$$\eta_1 = 0.01, \ \eta_2 = 0.9, \quad \text{et } \gamma_1 = \gamma_2 = 0.5,$$

mais d'autres valeurs peuvent être sélectionnées.

Itérations

Itération réussie Si $\rho_k \geq \eta_1$ à l'étape 1, l'itération k est dite réussie puisque le point candidat $x_k + s_k$ est accepté.

Itération non réussie Sinon, l'itération est déclarée non réussie et le point candidat est rejeté.

Itération très réussie Si $\rho_k \geq \eta_2$, l'accord entre le modèle et la fonction est particulièrement bon, aussi l'itération est dite très réussie. Ceci suggère d'élargir la région de confiance comme décrit à l'étape 4, afin de permettre un plus long pas à la prochaine itération.

Solution approximative du sous-problème

Chaque sous-problème de région de confiance doit être résolu approximativement, et aussi économiquement que possible.

Afin d'être capable de garantir la convergence globale de la méthode, nous voudrions au minimum que la solution approximative atteigne le même niveau de réduction de la valeur du modèle que ne le ferait un pas dans la direction de plus forte pente, contraint par la région de confiance.

Pas de Cauchy

Le pas de Cauchy est défini par $s_k^C := -\alpha_k^C \nabla f(x_k)$, où

$$\begin{aligned}\alpha_k^C :&= \arg\min \{ m_k(-\alpha \nabla f(x_k)) \mid \alpha > 0, \alpha \|\nabla f(x_k)\| \leq \Delta_k \} \\ &= \arg\min \{ m_k(x - \alpha \nabla f(x_k)) \mid 0 < \alpha \leq \Delta_k / \|\nabla f(x_k)\| \}\end{aligned}$$

Calculer le pas de Cauchy est très facile, comme il s'agit de la minimisation d'une fonction quadratique sur un segment de droite.

Habituellement, nous exigerons que la solution approximative du sous-problème de région de confiance entraîne une réduction du modèle au moins équivalente à celle obtenue avec le pas de Cauchy :

$$m_k(s_k) \leq m_k(s_k^C), \;\; \|s_k\|_k \leq \Delta_k.$$

Calcul du pas de Cauchy

Nous devons minimiser

$$m_k(-\alpha\nabla f(x_k)) = f(x_k) - \alpha\|\nabla f(x_k)\|_2^2 + \frac{1}{2}\alpha^2\nabla f(x_k)^T H_k \nabla f(x_k)$$

dont la dérivée en α est

$$\frac{dm_k(-\alpha\nabla f(x_k))}{d\alpha} = -\|\nabla f(x_k)\|_2^2 + \alpha\nabla f(x_k)^T H_k \nabla f(x_k)$$

Supposons $m_k(\cdot)$ strictement convexe en α, i.e. $\nabla f(x_k)^T H_k \nabla f(x_k) > 0$. Le minimum de $m_k(\cdot)$ le long de $-\alpha\nabla f(x_k)$ est obtenu en annulant la dérivée par rapport à α :

$$\alpha^* = \frac{\|\nabla f(x_k)\|_2^2}{\nabla f(x_k)^T H_k \nabla f(x_k)}$$

Mais il se peut que ce minimum soit hors de la région de confiance, i.e.

$$\alpha^*\|\nabla f(x_k)\|_k \geq \Delta_k$$

Calcul du pas de Cauchy (suite)

Dès lors, pour assurer de rester de rester dans la région de confiance, nous prendrons

$$\alpha^* = \min\left\{\frac{\Delta_k}{\|\nabla f(x_k)\|_k}, \frac{\|\nabla f(x_k)\|_2^2}{\nabla f(x_k)^T H_k \nabla f(x_k)}\right\}$$

Si $m_k(\cdot)$ n'est pas strictement convexe en α, le modèle n'est pas borné inférieurement le long de $-\alpha\nabla f(x_k)$), et il faut s'arrêter sur la frontière de la région de confiance, ce qui donne

$$\alpha^* = \frac{\Delta_k}{\|\nabla f(x_k)\|_k}$$

Théorie de convergence

Théorème (Décroissance du modèle atteignable)

Soit $m_k(s)$ un modèle quadratique de f autour de x_k, et soit s_k^C le pas de Cauchy, satisfaisant $\|s_k^C\|_k \leq \Delta_k$. Alors la diminution atteignable du modèle est d'au moins

$$f(x_k) - m_k(s_k^C) \geq \frac{1}{2}\|\nabla f(x_k)\|_2 \min\left\{\frac{\|\nabla f(x_k)\|}{1+\|H_k\|_2}, \kappa_s \Delta_k\right\}$$

Note : κ_s est un facteur d'équivalence de normes. On choisit la norme telle que

$$\kappa_s \|\cdot\|_k \leq \|\cdot\|_2 \leq \kappa_l \|\cdot\|_k$$

avec $\kappa_l \geq \kappa_s > 0$. Les normes usuelles satisfont cette condition :

$$\frac{1}{\sqrt{n}}\|\cdot\|_1 \leq \|\cdot\|_2 \leq \|\cdot\|_1$$

$$\|\cdot\|_\infty \leq \|\cdot\|_2 \leq n\|\cdot\|_\infty$$

Théorie de convergence

Corollaire

Soit $m_k(s)$ un modèle quadratique de f et s_k tel que $m_k(s_k) \leq m_k(s_k^C)$, et $\|s_k\|_k \leq \Delta_k$. Alors

$$f(x_k) - m_k(s_k) \geq \frac{1}{2}\|\nabla f(x_k)\|_2 \min\left\{\frac{\|\nabla f(x_k)\|}{1+\|H_k\|_2}, \kappa_s \Delta_k\right\}$$

De plus, si l'itération k est très réussie,

$$f(x_k) - f(x_{k+1}) \geq \eta_2 \frac{1}{2}\|\nabla f(x_k)\|_2 \min\left\{\frac{\|\nabla f(x_k)\|}{1+\|H_k\|_2}, \kappa_s \Delta_k\right\}$$

Adéquation du modèle

Lemme (Différence entre le modèle et la fonction)

Soit $f \in C^2$, et supposons qu'il existe des constantes $\kappa_H \geq 1$ et $\kappa_m \geq 0$ telles que $\|\nabla^2 f(x_k)\|_2 \leq \kappa_H$ et $\|H_k\|_2 \leq \kappa_m$ pour tout k. Alors

$$|f(x_k + s_k) - m_k(s_k)| \leq \kappa_d \Delta_k^2$$

où $\kappa_d = \frac{1}{2}\kappa_l^2(\kappa_H + \kappa_m)$.

Comportement du rayon de la région de confiance

L'algorithme garantit que le rayon de la région de confiance ne peut être arbitrairement petit à des points non optimums.

Théorème

Soit $f \in C^2$, *et supposons qu'il existe des constantes* $\kappa_H \geq 1$ *et* $\kappa_m \geq 0$ *telles que* $\|\nabla^2 f(x_k)\|_2 \leq \kappa_H$ *et* $\|H_k\|_2 \leq \kappa_m$ *pour tout* k. *Soit* $\kappa_d = \frac{1}{2}\kappa_l^2(\kappa_H + \kappa_m)$. *Si à l'itération* k, *nous avons* $\nabla f(x_k) \neq 0$ *et*

$$\Delta_k \leq \|\nabla f(x_k)\|_2 \min\left\{\frac{1}{\kappa_s(\kappa_H + \kappa_b)}, \frac{\kappa_s(1-\eta_2)}{2\kappa_d}\right\}, \ \forall k,$$

alors l'itération k *est très réussie et* $\Delta_{k+1} \geq \Delta_k$.

Comportement du rayon de la région de confiance

Corollaire

Soit $f \in C^2$, et supposons qu'il existe des constantes $\kappa_H \geq 1$ et $\kappa_m \geq 0$ telles que $\|\nabla^2 f(x_k)\|_2 \leq \kappa_H$ et $\|H_k\|_2 \leq \kappa_m$ pour tout k. Soit $\kappa_d = \frac{1}{2}\kappa_l^2(\kappa_H + \kappa_m)$. S'il existe une constante $\epsilon > 0$ telle que $\|\nabla f(x_k)\|_2 \geq \epsilon$ pour tout k, alors

$$\Delta_k \geq \kappa_\epsilon := \epsilon\gamma_1 \min\left\{\frac{1}{\kappa_s(\kappa_H + \kappa_b)}, \frac{\kappa_s(1-\eta_2)}{2\kappa_d}\right\}, \ \forall k.$$

Corollaire (Possible arrêt en un nombre fini d'itérations)

Soit $f \in C^2$, et supposons que les matrices hessiennes de l'objectif et du modèle uniformément bornées positivement. Si la méthode de région de confiance a seulement un nombre fini d'itérations réussies, alors $x_k = x^*$ et $\nabla f(x^*)$ pour tout k assez grand.

Convergence globale

Théorème (Convergence globale)

Soit $f \in C^2$, et supposons que les matrices hessiennes de l'objectif et du modèle uniformément bornées positivement. Un des trois cas suivants a lieu :

1. $\nabla f(x_k) = 0$ *pour un certain* $k \in \mathcal{N}$
2. $\lim_{k\to\infty} f(x_k) = -\infty$
3. $\lim_{k\to\infty} \nabla f(x_k) = 0$

Méthodes pour résoudre le sous-problème de région de confiance

Examinons plus en détails comment résoudre le sous-problème de région de confiance

$$\min_{s \in \mathcal{R}^n} \; m(s) = s^T \nabla f(x_k) + \frac{1}{2} s^T \nabla^2 f(x_k) s$$
$$\text{t.q. } \|s\| \leq \Delta_k .$$

de telle sorte que la théorie de convergence ci-dessus s'applique.

En d'autres mots, nous cherchons $s^* \in \mathcal{R}^n$ tel que

$$m(s^*) \leq m(s^c) \text{ et } \|s^*\| \leq \Delta.$$

On pourrait résoudre

- exactement $\rightarrow$ méthode de type Newton
- approximativement

Sous-problème

À partir de maintenant, nous utiliserons la norme 2 pour déterminer les régions de confiance, aussi avons-nous à résoudre approximativement

$$(TRS) \ \min_{s \in \mathcal{R}^n} m(s) \text{ s.t. } \|s\|_2 \leq \Delta$$

où $m(s) = \nabla f(x_k)^T s + \frac{1}{2} s^T H s$.

Théorème

N'importe quel minimiseur global s^* de (TRS) doit satisfaire

1. $(H + \lambda^* I)s^* = -\nabla f(x_k)$,
2. $H + \lambda^* I$ *semi-définie positive*
3. $\lambda^* \geq 0$
4. $\lambda^*(\|s^*\|_2 - \Delta) = 0$.

De plus, si $H + \lambda^* I$ est définie positive, alors s^* est unique.

Solutions exactes de (TRS)

Si H est définie positive et la solution de $Hs = -\nabla f(x_k)$ satisfait $\|s\|_2 \leq \Delta$, alors $s^* = s$.

Si H est indéfinie ou si la solution du système $Hs = -\nabla f(x_k)$ satisfait $\|s\|_2 > \Delta$, résoudre le système nonlinéaire

$$(H + \lambda I)s = -\nabla f(x_k), \text{ t.q. } \|s\| = \Delta,$$

pour s et λ en utilisant la méthode de Newton. Des difficultés surviennent

- possiblement quand de multiples solutions locales apparaissent,
- ou quand $\nabla f(x_k)$ est proche de l'orthogonalité avec le(s) vecteur(s) propre(s) correspondant à la valeur propre la plus négative de H.

Quand n est grand, la factorisation pour résoudre $Hs = -g$ peut être impossible.

Solutions approximatives de (TRS)

Comme nous n'avons besoin que d'une solution approximative, il est préférable d'employer une méthode itérative.

1. La méthode de plus forte pente mène au pas de Cauchy s^C.
2. Utiliser la méthode des gradients conjugués pour améliorer le pas à partir de s^C. Point à considérer :
    - il faut rester dans la région de confiance,
    - comment gérer une courbure négative ?