Méthodes de Newton et gradient conjugué avec garanties de complexité pour l'optimisation non convexe

Clément Royer

Université du Wisconsin-Madison, États-Unis

23 octobre 2018

Octobre à Madison

Optimisation lisse non convexe

On considère le problème lisse sans contraintes :

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

Hypothèses sur f

f minorée, de classe $\mathcal{C}^2$, non convexe.

Optimisation lisse non convexe

On considère le problème lisse sans contraintes :

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

Hypothèses sur f

f minorée, de classe $\mathcal{C}^2$, non convexe.

Minimisation d'une fonction lisse non convexe

- Objectif : Satisfaire les *conditions nécessaires du second ordre* pour f :

$$\|\nabla f(x)\| = 0 \quad \text{and} \quad \nabla^2 f(x) \succeq 0.$$

- Si x ne vérifie pas ces conditions, $\exists$ d telle que
  1. $d^\top \nabla f(x) < 0$: direction de descente.
     and/or
  2. $d^\top \nabla^2 f(x) d < 0$: direction de courbure négative

  $\Rightarrow$ **pour problèmes non convexes**.

Motivation

Formulations non convexes de problèmes matriciels de rang faible (Bhojanapalli et al. 2016, Ge et al. 2017)

$$\min_{X \in \mathbb{R}^{n \times m}} f(X) \;\Rightarrow\; \min_{U \in \mathbb{R}^{n \times r}, V \in \mathbb{R}^{m \times r}} f(U V^\top).$$

- Les points nécessaires à l'ordre deux sont des minima globaux (ou proche en valeur d'objectif);

Factorisation orthogonale de tenseur d'ordre 4 (Ge et al. 2015)

$$\min_{\{u_i\} \subset \mathbb{S}^{n-1}} \sum_{i \neq j} T(u_i, u_i, u_j, u_j).$$

- Les solutions **sont** les points nécessaires du second ordre.

Notre but : développer des algorithmes **efficaces** pour trouver des points nécessaires au second ordre pour des problèmes non convexes.

Complexité comme mesure d'efficacité

Complexité d'un algorithme

- Borne sur le **coût** de cet algorithme dans le pire des cas;
- Très courant en informatique théorique,
- Impact majeur en optimisation **convexe** (Nemirovski and Yudin 1983).

Complexité comme mesure d'efficacité

Complexité d'un algorithme

- Borne sur le **coût** de cet algorithme dans le pire des cas;
- Très courant en informatique théorique,
- Impact majeur en optimisation **convexe** (Nemirovski and Yudin 1983).

Complexité en optimisation non convexe

Soit une méthode générant les itérés $\{x_k\}$, deux tolérances $\epsilon_g, \epsilon_H \in (0, 1)$:

- **Coût au pire cas** pour obtenir x_K tel que

$$\|\nabla f(x_K)\| \leq \epsilon_g, \qquad \lambda_{\min}(\nabla^2 f(x_K)) \geq -\epsilon_H.$$

  (x_k est un point-(ϵ_g, ϵ_H).)
- Important : **Dépendence de ce coût selon** ϵ_g, ϵ_H.

Résultats de complexité

Point de vue 1 : optimisation non convexe

- Mesure de coût : Nombre d'itérations (coûteuses);
- Meilleures méthodes : variations sur l'algorithme de Newton.

Résultats de complexité

Point de vue 1 : optimisation non convexe

- Mesure de coût : Nombre d'itérations (coûteuses);
- Meilleures méthodes : variations sur l'algorithme de Newton.

Méthodes	Bornes
Régions de confiance (classique)	$\mathcal{O}\left(\max\{\epsilon_g^{-2}\epsilon_H^{-1}, \epsilon_H^{-3}\}\right)$
Régions de confiance (découplées)	$\mathcal{O}\left(\max\{\epsilon_g^{-2}, \epsilon_H^{-3}\}\right)$
Régularisation cubique Régions de confiance *TRACE*	$\mathcal{O}\left(\max\{\epsilon_g^{-\frac{3}{2}}, \epsilon_H^{-3}\}\right)$

Résultats de complexité (2)

Point de vue 2 : apprentissage

- On prendra $\epsilon_g = \epsilon, \epsilon_H = \mathcal{O}(\sqrt{\epsilon})$.

  Meilleure méthode de type Newton $\mathcal{O}(\epsilon^{-\frac{3}{2}})$.
- Mesure de coût : *évaluations de gradient, produits Hessienne-vecteur*
  $\Rightarrow$ couvre le coût des itérations.

Résultats de complexité (2)

Point de vue 2 : apprentissage

- On prendra $\epsilon_g = \epsilon, \epsilon_H = \mathcal{O}(\sqrt{\epsilon})$.

  Meilleure méthode de type Newton $\mathcal{O}(\epsilon^{-\frac{3}{2}})$.
- Mesure de coût : *évaluations de gradient, produits Hessienne-vecteur*
  $\Rightarrow$ couvre le coût des itérations.

Méthodes	Bornes
Méthodes de descente de gradient avec composante aléatoire	$\tilde{\mathcal{O}}\left(\epsilon^{-2}\right)$
Méthodes de gradient accéléré pour problèmes non convexes	$\tilde{\mathcal{O}}(\epsilon^{-\frac{7}{4}})$

- $\tilde{\mathcal{O}}(\cdot)$: facteurs logarithmiques.
- Résultats avec **forte probabilité**.

But de l'exposé

Aborder toutes les notions de complexité...

- Itérations, évaluations;
- Pour différents choix de ϵ_g, ϵ_H;
- Déterministe, en probabilité.

But de l'exposé

Aborder toutes les notions de complexité...

- Itérations, évaluations;
- Pour différents choix de ϵ_g, ϵ_H;
- Déterministe, en probabilité.

...dans un seul cadre algorithmique

- Itérations de type Newton, avec recherche linéaire;
- Coût principal : calcul de gradient/produit matrice Hessienne-vecteur;
- Les meilleures garanties de complexité possibles.

Outline

1. Méthodes de Newton avec courbure négative
   - Cadre général
   - Variantes inexactes
2. Newton+gradient conjugué plafonné
   - Gradient conjugué et quadratiques non convexes
   - Algorithmes de Newton associés
3. Résultats numériques

Outline

1. Méthodes de Newton avec courbure négative
   - Cadre général
   - Variantes inexactes
2. Newton+gradient conjugué plafonné
3. Résultats numériques

Cadre général : recherche linéaire

Paramètres: $x_0 \in \mathbb{R}^n,\ \theta \in (0,1),\ \eta > 0, \epsilon_g \in (0,1),\ \epsilon_H \in (0,1).$

Pour k=$0, 1, 2, \ldots$

1. Calcul d'une direction $d_k = d_k(\epsilon_g, \epsilon_H)$.
2. Recherche linéaire avec retour arrière Calculer le plus grand $\alpha_k \in \{\theta^j\}_{j\in\mathbb{N}}$ tel que $f(x_k + \alpha_k d_k) < f(x_k) - \frac{\eta}{6}\alpha_k^3 \|d_k\|^3$.
3. Poser $x_{k+1} = x_k + \alpha_k\, d_k$.

- Recherche linéaire : Garantit une décroissance;
- Décroissance cubique pour la complexité.

Itération de Newton

Principe

- Itération k : Calculer d_k comme solution du système linéaire :

$$\nabla^2 f(x_k) d_k \;=\; -\nabla f(x_k);$$

  et poser $x_{k+1} = x_k + d_k$;
- Convergence globale assurée avec une recherche linéaire;
- Solution unique lorsque $\nabla^2 f(x_k) \succ 0$.

Pour des problèmes non convexes

- Utiliser un seuil ϵ_H pour $\lambda_{\min}\left(\nabla^2 f(x_k)\right)$;
- Régulariser pour garantir que $\nabla^2 f(x_k) + \alpha I \succ \epsilon_H I$;
- **Second ordre :** Mettre à profit des directions de courbure négative d telles que $d^\top \nabla^2 f(x_k) d \leq -\epsilon_H \|d\|^2$.

Algorithme de Newton d'ordre deux avec recherche linéaire

Paramètres: $x_0 \in \mathbb{R}^n$, $\theta \in (0,1)$, $\eta > 0$, $\epsilon_H \in (0,1)$.

Pour k=$0, 1, 2, \ldots$

1. Calcul d'une direction de recherche d_k
   - Calculer $\lambda = \lambda_{\min}(\nabla^2 f(x_k))$;
   - Si $\lambda < -\epsilon_H$, choisir une **direction de courbure négative** d_k telle que

$$d_k^\top \nabla f(x_k) \leq 0, \quad d_k^\top \nabla^2 f(x_k) d_k = \lambda \|d_k\|^2;$$

   - Sinon, calculer une **direction de Newton** (possiblement **regularisée**) d_k en résolvant

$$\left(\nabla^2 f(x_k) + 2\epsilon_H I\right) d_k \;=\; -\nabla f(x_k).$$

2. Recherche linéaire avec retour arrière Calculer le plus grand $\alpha_k \in \{\theta^j\}_{j\in\mathbb{N}}$ tel que $f(x_k + \alpha_k d_k) \;<\; f(x_k) - \frac{\eta}{6}\alpha_k^3 \|d_k\|^3$.
3. Poser $x_{k+1} = x_k + \alpha_k\, d_k$.

Complexité de la méthode de Newton d'ordre deux

Théorème (Royer et Wright 2018)

La méthode produit x_k tel que $\|\nabla f(x_k)\| \leq \epsilon_g$ et $\nabla^2 f(x_k) \succeq -\epsilon_H I$ en au plus

$$\mathcal{O}(\max\{\epsilon_g^{-3}\epsilon_H^3, \epsilon_H^{-3}\})$$

itérations.

Complexité de la méthode de Newton d'ordre deux

Théorème (Royer et Wright 2018)

La méthode produit x_k tel que $\|\nabla f(x_k)\| \leq \epsilon_g$ et $\nabla^2 f(x_k) \succeq -\epsilon_H I$ en au plus

$$\mathcal{O}(\max\{\epsilon_g^{-3}\epsilon_H^3, \epsilon_H^{-3}\})$$

itérations.

- $\epsilon_H = \epsilon_g^{1/2}$: borne en $\mathcal{O}(\epsilon_g^{-3/2})$;
- Optimal sur une classe de méthodes d'ordre deux (Cartis, Gould, Toint 2018).

Sommaire

1. Méthodes de Newton avec courbure négative
   - Cadre général
   - Variantes inexactes
2. Newton+gradient conjugué plafonné
3. Résultats numériques

Sources d'inexactitude

Notre méthode utilise des **opérations matricielles** :

- Calcul de valeur/vecteur propre;
- Résolution de système linéaire.

Sources d'inexactitude

Notre méthode utilise des **opérations matricielles** :

- Calcul de valeur/vecteur propre;
- Résolution de système linéaire.

Approches inexactes

- Méthodes d'algèbre linéaire itératives (avec/sans aléatoire) pour les opérations matricielles.
- Coût principal : **Produits matrice Hessienne-vecteur**.

Deux types de direction

En fonction de $\epsilon_H > 0$ (estimation de la plus petite valeur propre):

- **Direction de Newton régularisée**:

$$(\nabla^2 f(x_k) + 2\epsilon_H I)d = -\nabla f(x_k), \quad \nabla^2 f(x_k) + 2\epsilon_H I \succeq \epsilon_H I.$$

- **Direction de courbure négative suffisante**:

$$d^\top \nabla f(x_k) \leq 0, \quad d^\top \nabla^2 f(x_k) d \leq -\epsilon_H \|d\|^2.$$

De l'optimisation vers l'algèbre linéaire

Deux types de direction

En fonction de $\epsilon_H > 0$ (estimation de la plus petite valeur propre):

- **Direction de Newton régularisée**:

$$(\nabla^2 f(x_k) + 2\epsilon_H I)d = -\nabla f(x_k), \quad \nabla^2 f(x_k) + 2\epsilon_H I \succeq \epsilon_H I.$$

- **Direction de courbure négative suffisante**:

$$d^\top \nabla f(x_k) \leq 0, \quad d^\top \nabla^2 f(x_k) d \leq -\epsilon_H \|d\|^2.$$

Problème(s) d'algèbre linéaire associés

Soit $H \in \mathbb{R}^{n\times n}$ symétrique, $g \in \mathbb{R}^n$ et $\epsilon_H > 0$:

- Résoudre $(H + 2\epsilon_H I)d = -g$ où $\lambda_{\min}(H) > -\epsilon_H$;
- Calculer d telle que $d^\top H d \leq -\epsilon_H \|d\|^2$ sinon.

De l'optimisation vers l'algèbre linéaire

Deux types de direction

En fonction de $\epsilon_H > 0$ (estimation de la plus petite valeur propre):

- **Direction de Newton régularisée approchée**:

$$(\nabla^2 f(x_k) + 2\epsilon_H I)d \approx -\nabla f(x_k), \quad \nabla^2 f(x_k) + 2\epsilon_H I \succeq \epsilon_H I.$$

- **Direction de courbure négative suffisante**:

$$d^\top \nabla f(x_k) \leq 0, \quad d^\top \nabla^2 f(x_k) d \leq -\epsilon_H \|d\|^2.$$

Problème(s) d'algèbre linéaire associés

Soit $H \in \mathbb{R}^{n \times n}$ symétrique, $g \in \mathbb{R}^n$ et $\epsilon_H > 0$:

- Approcher la solution de $(H + 2\epsilon_H I)d = -g$ où $\lambda > -\epsilon_H, \lambda \approx \lambda_{\min}(H)$;
- Calculer d telle que $d^\top H d \leq -\epsilon_H \|d\|^2$ sinon.

Gradient conjugué pour systèmes linéaires symétriques

- On considère $Hd = -g$ avec $H = H^\top \succeq \epsilon_H I$.

Gradient conjugué pour systèmes linéaires symétriques

- On considère $Hd = -g$ avec $H = H^\top \succeq \epsilon_H I$.

Gradient Conjugué

- On applique le gradient conjugué avec critère d'arrêt :

$$\|Hd + g\| \leq \frac{\xi}{2} \min\{\|g\|, \epsilon_H \|d\|\}, \qquad \xi \in (0, 1).$$

- Soit $\kappa = \lambda_{\max}(H)/\lambda_{\min}(H)$, l'algorithme termine en au plus

$$\min\left\{n, \tfrac{1}{2}\sqrt{\kappa}\log\left(4\kappa^{\frac{5}{2}}/\xi\right)\right\} = \min\left\{n, \mathcal{O}\left(\sqrt{\kappa}\log(\kappa/\xi)\right)\right\}$$

itérations/**produits matrice-vecteur.**

Lanczos pour le calcul de valeurs propres

- La méthode de Lanczos peut calculer des approximations de valeurs propres;
- Point-clé : utiliser un vecteur initial **aléatoire** uniforme sur la sphère unité (Kuczyński et Woźniakowski 1992).

Lanczos pour le calcul de valeurs propres

- La méthode de Lanczos peut calculer des approximations de valeurs propres;
- Point-clé : utiliser un vecteur initial **aléatoire** uniforme sur la sphère unité (Kuczyński et Woźniakowski 1992).

Itérations de Lanczos

Soit $H \in \mathbb{R}^{n \times n}$ symétrique avec $\|H\| \leq M,\ \epsilon_H > 0,\ \delta \in (0,1)$.
Avec une probabilité d'au moins $1 - \delta$, on peut obtenir du procédé de Lanczos un vecteur unitaire v tel que

$$v^\top H v \ \leq\ \lambda_{\min}(H) + \epsilon_H.$$

en au plus

$$\min\left\{n, \frac{\ln(n/\delta^2)}{2\sqrt{2}}\sqrt{\frac{M}{\epsilon_H}}\right\}$$

itérations/produits matrice-vecteur.

Gradient conjugué pour un problème de valeur propre ?

- Gradient conjugué et Lanczos construisent les mêmes **sous-espaces de Krylov** partant du même vecteur;
- Invariance des sous-espaces par translation.

Gradient conjugué pour un problème de valeur propre ?

- Gradient conjugué et Lanczos construisent les mêmes **sous-espaces de Krylov** partant du même vecteur;
- Invariance des sous-espaces par translation.

Théorème (Royer, O'Neill, Wright 2018)

Soit $H \in \mathbb{R}^{n\times n}$ symétrique avec $\|H\| \leq M$. On applique l'algorithme du gradient conjugué à $\left(H + \frac{\epsilon_H}{2} I\right) d = b$ avec $b \sim \mathcal{U}(\mathbb{S}^{n-1})$. Alors, pour tout $\delta \in [0, 1)$, avec probabilité au moins $1 - \delta$:

1. Si $\lambda_{\min}(H) < -\epsilon_H$, le gradient conjugué calcule une direction de courbure $\leq -\frac{\epsilon_H}{2}$ en au plus $j \leq J$ itérations, où

$$J = \min\left\{n, \left\lceil \frac{\ln(n/\delta^2)}{2}\sqrt{\frac{M}{\epsilon_H}} \right\rceil\right\}.$$

2. Sinon, l'algorithme effectue J itérations et certifie que $H \succeq -\epsilon_H I$.

Oracles de plus petite valeur propre

Corollaire

Pour une matrice $\nabla^2 f(x_k)$, on considère les deux algorithmes suivants :

1. Gradient conjugué appliqué à $\left(\nabla^2 f(x_k) + \frac{\epsilon_H}{2} I\right) d = b$, avec $b \sim \mathbb{S}^{n-1}$;
2. Lanczos appliqué à $\nabla^2 f(x_k)$ partant de $b \sim \mathbb{S}^{n-1}$.

Alors, pour tout $\delta \in [0, 1)$, on obtient

- une direction de courbure négative de courbure $\leq -\epsilon_H/2$,
- OU un certificat que $\nabla^2 f(x_k) \succeq -\epsilon_H I$,

en au plus $\tilde{\mathcal{O}}\left(\min\{n, \epsilon_H^{-1/2}\}\right)$ appels de gradients/produits Hessienne-vecteur, le tout en probabilité au moins $1 - \delta$.

Oracles de plus petite valeur propre

Corollaire

Pour une matrice $\nabla^2 f(x_k)$, on considère les deux algorithmes suivants :

1. Gradient conjugué appliqué à $\left(\nabla^2 f(x_k) + \frac{\epsilon_H}{2} I\right) d = b$, avec $b \sim \mathbb{S}^{n-1}$;
2. Lanczos appliqué à $\nabla^2 f(x_k)$ partant de $b \sim \mathbb{S}^{n-1}$.

Alors, pour tout $\delta \in [0, 1)$, on obtient

- une direction de courbure négative de courbure $\leq -\epsilon_H/2$,
- OU un certificat que $\nabla^2 f(x_k) \succeq -\epsilon_H I$,

en au plus $\tilde{\mathcal{O}}\left(\min\{n, \epsilon_H^{-1/2}\}\right)$ appels de gradients/produits Hessienne-vecteur, le tout en probabilité au moins $1 - \delta$.

On dira que l'on dispose d'un oracle de (plus petite) valeur propre.

Variante inexacte de la méthode de Newton

Paramètres: $x_0 \in \mathbb{R}^n$, $\theta \in (0,1)$, $\eta > 0$, $\epsilon_H \in (0,1)$.

Pour k=$0, 1, 2, \ldots$

1. Calcul d'une direction de recherche d_k
   - **Calculer $\lambda \approx \lambda_{\min}(\nabla^2 f(x_k))$ via un oracle de valeur propre**;
   - Si $\lambda < -\epsilon_H$, choisir une **direction de courbure négative** d_k telle que

$$d_k^\top \nabla f(x_k) \leq 0, \quad d_k^\top \nabla^2 f(x_k) d_k = \lambda \|d_k\|^2;$$

   - Sinon, calculer une **direction de Newton** (possiblement **regularisée**) d_k **via le gradient conjugué de sorte que**

$$\left\| \left(\nabla^2 f(x_k) + 2\epsilon_H I \right) d_k + \nabla f(x_k) \right\| \leq \tfrac{\xi}{2} \min \left\{ \|\nabla f(x_k)\|, \epsilon_H \|d_k\| \right\}$$

2. Recherche linéaire avec retour arrière (inchangée)
3. Poser $x_{k+1} = x_k + \alpha_k d_k$.

Résultat de complexité pour les variantes inexactes

On pose $\epsilon_g = \epsilon, \epsilon_H = \sqrt{\epsilon}$, et on suppose que $n >> \epsilon^{-1/2}$.

Un point-$(\epsilon, \sqrt{\epsilon})$ est obtenu en au plus

- $\mathcal{O}(\epsilon^{-\frac{3}{2}})$ itérations externes,
- $\tilde{\mathcal{O}}\left(\min\left\{n\epsilon^{-\frac{3}{2}}, \epsilon^{-\frac{7}{4}}\right\}\right)$ produits Hessienne-vecteur/appels de gradients,

en probabilité au moins $1 - \mathcal{O}(\epsilon^{-\frac{3}{2}}\delta)$.

Résultat de complexité pour les variantes inexactes

On pose $\epsilon_g = \epsilon, \epsilon_H = \sqrt{\epsilon}$, et on suppose que $n >> \epsilon^{-1/2}$.

Un point-$(\epsilon, \sqrt{\epsilon})$ est obtenu en au plus

- $\mathcal{O}(\epsilon^{-\frac{3}{2}})$ itérations externes,
- $\tilde{\mathcal{O}}\left(\min\left\{n\epsilon^{-\frac{3}{2}}, \epsilon^{-\frac{7}{4}}\right\}\right)$ produits Hessienne-vecteur/appels de gradients,

en probabilité au moins $1 - \mathcal{O}(\epsilon^{-\frac{3}{2}}\delta)$.

Poser $\delta = 0$ conduit à des résultats *presque sûrs*:

- Itérations: $\mathcal{O}(\epsilon^{-\frac{3}{2}})$.
- Hessienne-vecteur/gradients: $\mathcal{O}\left(n\,\epsilon^{-\frac{3}{2}}\right)$.

Sommaire

1. Méthodes de Newton avec courbure négative
2. Newton+gradient conjugué plafonné
   - Gradient conjugué et quadratiques non convexes
   - Algorithmes de Newton associés
3. Résultats numériques

Revisiter le gradient conjugué

Idées

- Appliquer le gradient conjugué à un système linéaire $\bar{H}d = -g$; **potentiellement non défini positif**.
- Équivaut à essayer de résoudre

$$\min_d \frac{1}{2} d^\top \bar{H} d + g^\top d$$

sans savoir si cette fonction quadratique est bornée.

Pourquoi ?

- Théorie de convergence dans le cas défini positif;
- Appliqué à une quadratique indéfinie:
  - **Peut** ne pas fonctionner...
  - ...en présence de courbure négative.

Gradient conjugué pour $\bar{H}y = -g$

Algorithme

Init: Poser $y_0 = 0_{\mathbb{R}^n}$, $r_0 = g$, $p_0 = -g$, $j = 0$.

Tant que $p_j^\top \bar{H} p_j > 0$

- Calculer $y_{j+1} = y_j + \alpha_j p_j$, $r_{j+1} = \bar{H} y_{j+1} + g$ et p_{j+1}.
- Poser $j = j + 1$; terminer si $\|r_j\| \leq \zeta \|r_0\|$.

Gradient conjugué pour $\bar{H}y = -g$

Algorithme

Init: Poser $y_0 = 0_{\mathbb{R}^n}$, $r_0 = g$, $p_0 = -g$, $j = 0$.
Tant que $p_j^\top \bar{H} p_j > 0$

- Calculer $y_{j+1} = y_j + \alpha_j p_j$, $r_{j+1} = \bar{H} y_{j+1} + g$ et p_{j+1}.
- Poser $j = j + 1$; terminer si $\|r_j\| \leq \zeta \|r_0\|$.

- Si $\bar{H} \succ 0$, $\|r_n\| = 0$;
- Si $\epsilon_H I \preceq \bar{H} \preceq MI$,

$$\|r_j\|^2 \leq 4\kappa \left(1 - \frac{2}{\sqrt{\kappa} + 1}\right)^{2j} \|r_0\|^2, \quad \kappa = \frac{M}{\epsilon_H}.$$

Gradient conjugué pour $\bar{H}y = -g$

Algorithme pour $\epsilon_H I \preceq \bar{H} \preceq MI$

Init: Poser $y_0 = 0_{\mathbb{R}^n}$, $r_0 = g$, $p_0 = -g$, $j = 0$.

Tant que $p_j^\top \bar{H} p_j > \epsilon_H \|p_j\|^2$ **et** $\|r_j\|^2 \leq 4\kappa \left(1 - \frac{2}{\sqrt{\kappa}+1}\right)^{2j} \|r_0\|^2$

- Calculer $y_{j+1} = y_j + \alpha_j p_j$, $r_{j+1} = \bar{H} y_{j+1} + g$ et p_{j+1}.
- Poser $j = j + 1$; terminer si $\|r_j\| \leq \zeta \|r_0\|$.

- Si $\bar{H} \succ 0$, $\|r_n\| = 0$;
- Si $\epsilon_H I \preceq \bar{H} \preceq MI$,

$$\|r_j\|^2 \leq 4\kappa \left(1 - \frac{2}{\sqrt{\kappa}+1}\right)^{2j} \|r_0\|^2, \quad \kappa = \frac{M}{\epsilon_H}.$$

Gradient conjugué pour $\bar{H}y = -g$

Algorithme pour $\epsilon_H I \preceq \bar{H} \preceq MI$

Init: Poser $y_0 = 0_{\mathbb{R}^n}$, $r_0 = g$, $p_0 = -g$, $j = 0$.

Tant que $p_j^\top \bar{H} p_j > \epsilon_H \|p_j\|^2$ **et** $\|r_j\|^2 \leq 4\kappa \left(1 - \frac{2}{\sqrt{\kappa}+1}\right)^{2j} \|r_0\|^2$

- Calculer $y_{j+1} = y_j + \alpha_j p_j$, $r_{j+1} = \bar{H} y_{j+1} + g$ et p_{j+1}.
- Poser $j = j + 1$; terminer si $\|r_j\| \leq \zeta \|r_0\|$.

- Si $\bar{H} \succ 0$, $\|r_n\| = 0$;
- Si $\epsilon_H I \preceq \bar{H} \preceq MI$,

$$\|r_j\|^2 \leq 4\kappa \left(1 - \frac{2}{\sqrt{\kappa}+1}\right)^{2j} \|r_0\|^2, \quad \kappa = \frac{M}{\epsilon_H}.$$

Et si $\bar{H}$ est indéfinie ?

Gradient conjugué pour un système indéfini

Gradient conjugué plafonné

Init : Poser $y_0 = 0_{\mathbb{R}^n}$, $r_0 = g$, $p_0 = -g$, $j = 0$.

Tant que $p_j^\top \bar{H} p_j > \epsilon_H \|p_j\|^2$ et $\|r_j\|^2 \leq T\tau^j \|r_0\|^2$

- Calculer y_{j+1}, $r_{j+1} = \bar{H} y_{j+1} + g$ et p_{j+1}.
- Poser $j = j + 1$; terminer si $\|r_j\| \leq \zeta \|r_0\|$.

Gradient conjugué pour un système indéfini

Gradient conjugué plafonné

Init : *Poser* $y_0 = 0_{\mathbb{R}^n}$, $r_0 = g$, $p_0 = -g$, $j = 0$.

Tant que $p_j^\top \bar{H} p_j > \epsilon_H \|p_j\|^2$ et $\|r_j\|^2 \leq T\tau^j \|r_0\|^2$

- Calculer y_{j+1}, $r_{j+1} = \bar{H} y_{j+1} + g$ et p_{j+1}.
- Poser $j = j + 1$; terminer si $\|r_j\| \leq \zeta \|r_0\|$.

Propriétés pour toute matrice $\bar{H} \preceq MI$

- Si r_j est calculé, alors

$$\|r_j\|^2 \leq 16\kappa^5 \left(1 - \frac{1}{\sqrt{\kappa}+1}\right)^j \|r_0\|^2, \quad \kappa = \frac{M}{\epsilon_H}.$$

  $(T = 16\kappa^5,\ \tau = \frac{\kappa}{\sqrt{\kappa}+1})$;

- Au plus $j_{pla} = \min\left\{n, \tilde{\mathcal{O}}\left(\sqrt{\frac{M}{\epsilon_H}}\right)\right\}$ itérations ("plafond") avant terminaison ou violation d'une des conditions.

Résultat principal - Sortie de boucle du gradient conjugué

Théorème (Royer, O'Neill, Wright 2018)

Si la méthode est appliquée à $\bar{H}d = -g$ pendant J iterations et que $\|r_J\| > \zeta\|r_0\|$, alors

- Soit $p_J^\top \bar{H} p_J \leq \epsilon_H \|p_J\|^2$;
- Soit $\|r_J\| > T\tau^J\|r_0\|$, on peut calculer y_{J+1} et il existe $j = 0, \ldots, J$ tel que

$$(y_{J+1} - y_j)^\top \bar{H}(y_{J+1} - y_j) \leq \epsilon_H \|y_{J+1} - y_j\|^2.$$

Résultat principal - Sortie de boucle du gradient conjugué

Théorème (Royer, O'Neill, Wright 2018)

Si la méthode est appliquée à $(H + 2\epsilon_H I)d = -g$ pendant J iterations et que $\|r_J\| > \zeta\|r_0\|$, alors

- Soit $p_J^\top H p_J \leq -\epsilon_H \|p_J\|^2$;
- Soit $\|r_J\| > T\tau^J\|r_0\|$, on peut calculer y_{J+1} et il existe $j = 0, \ldots, J$ tel que

$$(y_{J+1} - y_j)^\top H(y_{J+1} - y_j) \leq -\epsilon_H \|y_{J+1} - y_j\|^2.$$

Résultat principal - Sortie de boucle du gradient conjugué

Théorème (Royer, O'Neill, Wright 2018)

Si la méthode est appliquée à $(H + 2\epsilon_H I)d = -g$ pendant J iterations et que $\|r_J\| > \zeta\|r_0\|$, alors

- Soit $p_J^\top H p_J \leq -\epsilon_H \|p_J\|^2$;
- Soit $\|r_J\| > T\tau^J\|r_0\|$, on peut calculer y_{J+1} et il existe $j = 0, \ldots, J$ tel que

$$(y_{J+1} - y_j)^\top H(y_{J+1} - y_j) \leq -\epsilon_H \|y_{J+1} - y_j\|^2.$$

- Démonstration : utilise uniquement des propriétés intrinsèques au gradient conjugué !
- Suit le raisonnement de Bubeck (2014) et sa déclinaison pour le gradient acceleré (Carmon et al 2017), mais appliqué ici à des fonctions quadratiques $\Rightarrow$ Obtention directe de directions de courbure négative !

Gradient Conjugué plafonné - en bref

Exécution

Le gradient conjugué plafonné appliqué à

$$\left(\nabla^2 f(x_k) + 2\epsilon_H I\right) d \;=\; -\nabla f(x_k)$$

renvoie

- un pas de Newton régularisé approché d_k avec

$$\left\|(\nabla^2 f(x_k) + 2\epsilon_H I)d_k + \nabla f(x_k)\right\| \;\leq\; \zeta\|r_0\|.$$

- ou une direction de courbure négative $\leq -\epsilon_H/2$.

en au plus $\tilde{\mathcal{O}}(\min\{n, \epsilon_H^{-1/2}\})$ itérations/produits Hessienne-vecteur.

Sommaire

1. Méthodes de Newton avec courbure négative
2. Newton+gradient conjugué plafonné
   - Gradient conjugué et quadratiques non convexes
   - Algorithmes de Newton associés
3. Résultats numériques

Incorporer le gradient conjugué plafonné

Deux nouvelles instances de notre algorithme générique

- Première phase : tant que la norme du gradient est large, calculer les directions uniquement via le gradient conjugué plafonné;
- Seconde phase : itiliser le gradient conjugué standard pour estimer la plus petite valeur propre lorsque la norme du gradient est faible.

On ne calcule plus $\lambda_{\min}(\nabla^2 f)$ avant de décider quelle direction prendre.

Algoritme Newton+Gradient Conjugué plafonné

Paramètres: $x_0 \in \mathbb{R}^n$, $\theta \in (0,1)$, $\eta > 0, \epsilon_g \in (0,1)$, $\epsilon_H \in (0,1), \delta \in (0,1]$.

Pour k=$0, 1, 2, \ldots$

1. Si $\|\nabla f(x_k)\| > \epsilon_g$, calculer d_k via le gradient conjugué plafonné.
2. Sinon, appliquer le gradient conjugué comme oracle de valeur propre. Terminer si cet oracle certifie que $\nabla^2 f(x_k) \succeq -\epsilon_H I$, sinon utiliser sa sortie comme d_k.
3. Recherche linéaire avec retour arrière (inchangée) Calculer le plus grand $\alpha_k \in \{\theta^j\}_{j\in\mathbb{N}}$ tel que $f(x_k + \alpha_k d_k) < f(x_k) - \frac{\eta}{6}\alpha_k^3 \|d_k\|^3$.
4. Poser $x_{k+1} = x_k + \alpha_k\, d_k$.

Analyse probabiliste

- On peut mal terminer.
- Mais l'algorithme est toujours bien défini.

Résultats de complexité à l'ordre un

Théorème - Nombre d'itérations

- On atteint x_k tel que $\|\nabla f(x_k)\| \leq \epsilon_g$ (point-ϵ_g) en au plus $\mathcal{O}\left(\max\left\{\epsilon_g^{-3}\epsilon_H^3, \epsilon_H^{-3}\right\}\right)$ itérations;
- Chaque itération coûte au plus $\tilde{\mathcal{O}}(\min\{n, \epsilon_H^{-1/2}\})$ gradients/produits Hessienne-vecteur.

Théorème - Complexité de calcul

On atteint un point-ϵ_g en au plus

$$\tilde{\mathcal{O}}\left(\min\{n, \epsilon_H^{-1/2}\} \times \max\left\{\epsilon_g^{-3}\epsilon_H^3, \epsilon_H^{-3}\right\}\right).$$

gradients/produits Hessienne-vecteur.

- $\epsilon_H = \epsilon_g^{1/2} \Rightarrow \tilde{\mathcal{O}}\left(\max\{n\epsilon_g^{-3/2}, \epsilon_g^{-7/4}\}\right)$.
- **Meilleure borne connue sans calcul direct de Hessienne**.

Complexité à l'ordre deux (générale)

- But : Atteindre un point-(ϵ_g, ϵ_H) x_k tel que

$$\|\nabla f(x_k)\| \leq \epsilon_g, \qquad \lambda_k = \lambda_{\min}(\nabla^2 f(x_k)) \geq -\epsilon_H.$$

- Utiliser le gradient conjugué comme oracle avec $\delta \in [0, 1)$.

Théorème

Un point-(ϵ_g, ϵ_H) est atteint en au plus

- $\mathcal{O}(\max\{\epsilon_g^{-3}\epsilon_H^3, \epsilon_H^{-3}\})$ itérations,
- $\tilde{\mathcal{O}}\left(\min\{n, \epsilon^{-1/2}\} \times \max\{\epsilon_g^{-3}\epsilon_H^3, \epsilon_H^{-3}\}\right)$ gradients/produits Hessienne-vecteur,

en probabilité au moins $(1-\delta)^{\mathcal{O}(\min\{\epsilon_g^3\epsilon_H^{-3}, \epsilon_H^3\})}$.

Complexité à l'ordre deux (spécifique)

- But : Atteindre un point-$(\epsilon, \sqrt{\epsilon})$ x_k tel que

$$\|\nabla f(x_k)\| \leq \epsilon, \qquad \lambda_k = \lambda_{\min}(\nabla^2 f(x_k)) \geq -\sqrt{\epsilon}.$$

- Utiliser le gradient conjugué comme oracle avec $\delta \in [0, 1)$.

Théorème

Un point-$(\epsilon, \sqrt{\epsilon})$ est atteint en au plus

- $\mathcal{O}(\epsilon^{-3/2})$ itérations,
- $\tilde{\mathcal{O}}\left(\min\{n\epsilon^{-3/2}, \epsilon^{-7/4}\}\right)$ gradients/produits Hessienne-vecteur,

en probabilité au moins $(1-\delta)^{\mathcal{O}(\epsilon^{-3/2})}$.

Sommaire

1 Méthodes de Newton avec courbure négative

2 Newton+gradient conjugué plafonné

3 Résultats numériques

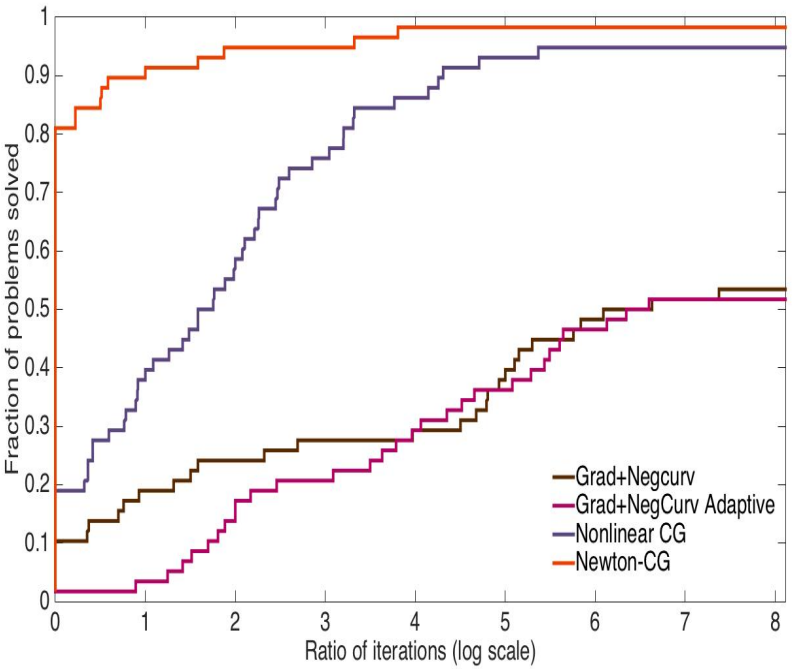
Nos tests

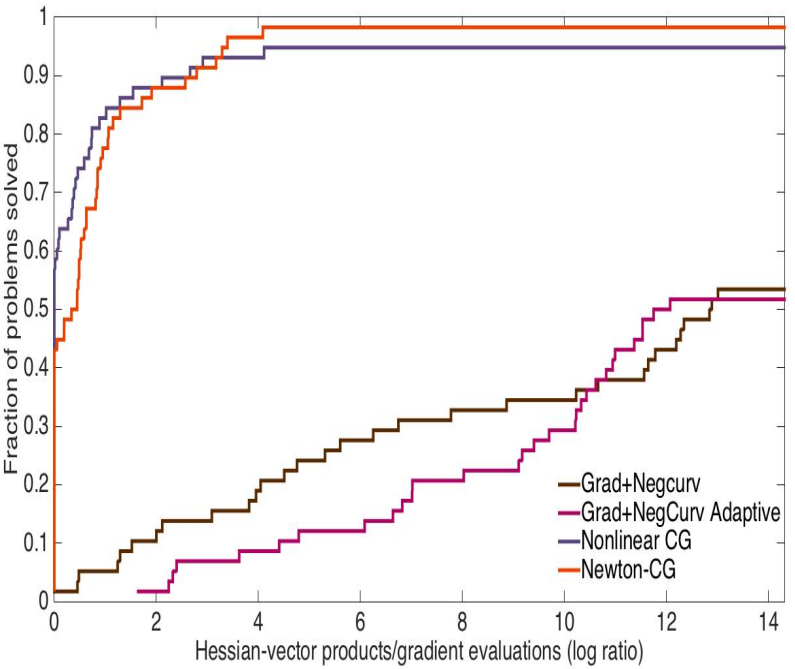
- Extraits d'un processus exhaustif de tests en cours;
- Focus : méthode de Newton+gradient conjugué plafonné (donne les meilleurs résultats parmi nos variantes);
- Comparaison avec d'autres algorithmes pour l'optimisation de grande taille (gradient conjugué non linéaire) et d'autres populaires en sciences des données (gradient accéléré, méthodes *ad hoc*).

Tests d'optimisation classiques

Contexte

- 61 problèmes non convexes de CUTEst, dimensions entre 2 et 500;
- $\epsilon_g = 10^{-5}$, $\epsilon_H = \sqrt{\epsilon_g}$;





Un problème d'estimation non convexe

Fonction de perte de Tukey (Carmon et al, 2017)

$$\min_{x\in\mathbb{R}^n} f(x) = \sum_{i=1}^{30} h(a_i^\top x - b_i) \quad \text{où } h(\theta) = \theta^2/(1+\theta^2),$$

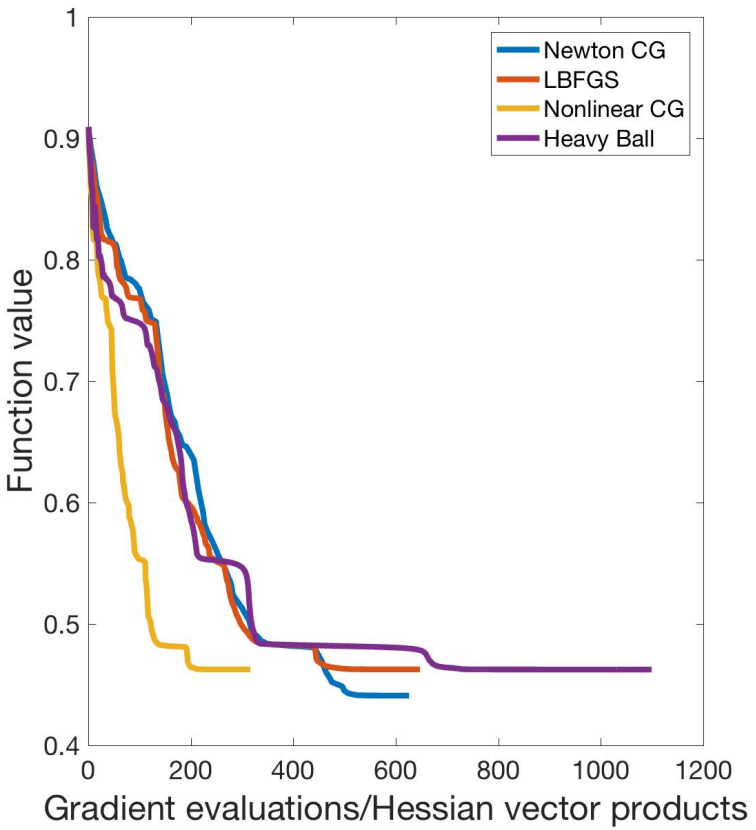
avec $a_i \equiv \mathcal{N}(0, I_n)$ et $b_i = a_i^\top x +$ bruit non Gaussien.

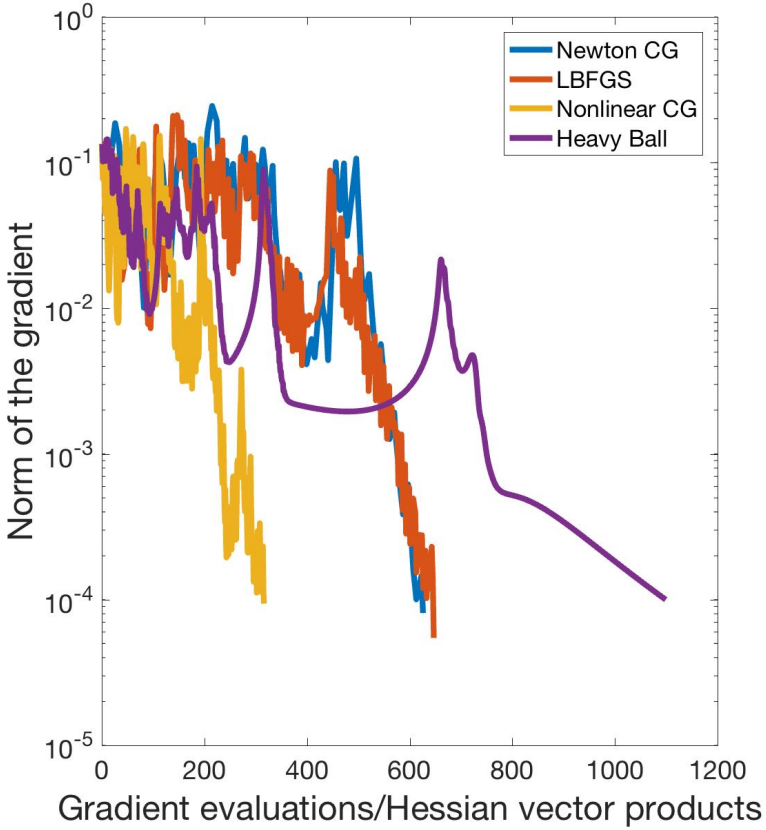
- **Critère d'arrêt:** $\|\nabla f(x)\| \leq \epsilon_g = 10^{-4}$.

Quatre algorithmes

- Newton+Gradient Conjugué;
- Gradient Conjugué non linéaire (Polak-Ribière);
- LBFGS;
- Méthode de la boule "pesante" *(Heavy ball)*.

Problème d'estimation non convexe : résultats





Un problème matriciel

Complétion de matrice

$$\min_{U,V} \frac{1}{2} \left\| P_\Omega(UV^\top - M) \right\|_F^2,$$

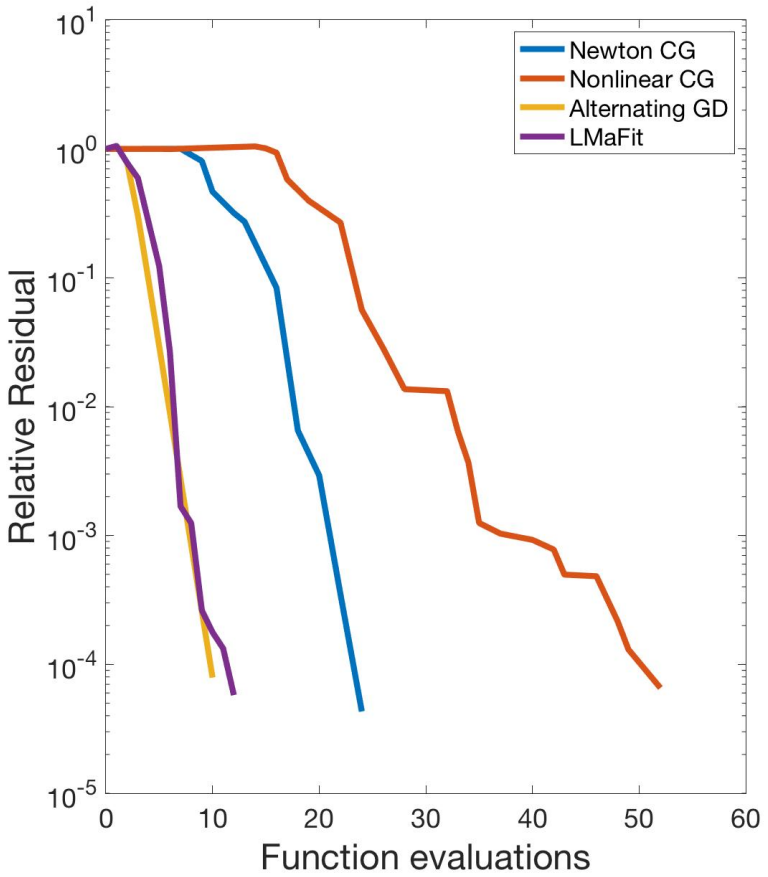
avec $M \in \mathbb{R}^{m \times n}$, $U \in \mathbb{R}^{m \times r}$, $V \in \mathbb{R}^{n \times r}$, $|\Omega| = 15\% \times mn$.

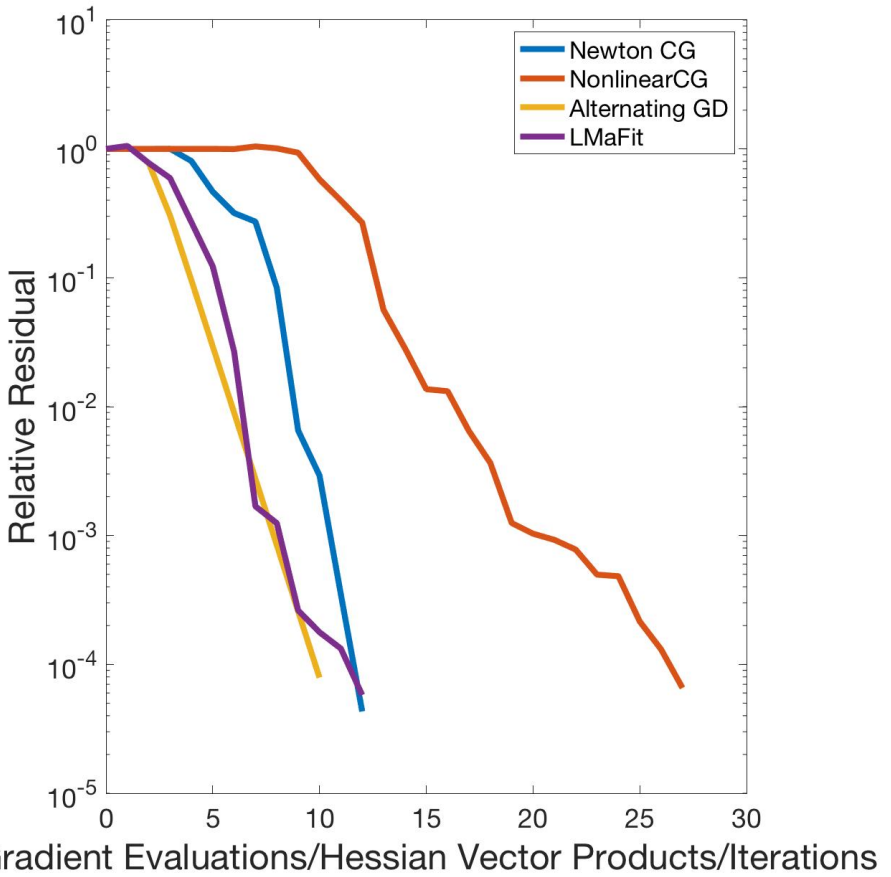
- Jeu de données MNIST (Chiffres 0-1) : trouver la composante principale ($r = 1$).

Comparaison

- Méthodes génériques :
  - Newton+Gradient conjugué plafonné;
  - Gradient conjugué non linéaire Polak-Ribière;
- Méthodes spécifiques :
  - Plus forte pente alternée (Tanner et Wei 2016);
  - LMaFit (Wen et al. 2012).

Problème de complétion de matrice : résultats





Conclusion

Newton + Gradient Conjugué : la vision classique

- Utilisée pour les problèmes de grande taille;
- Pas de garanties de complexité spécifiques...
- ...ni de justification de ses propriétés de détection de courbure négative.

Newton + Gradient Conjugué : un certain regard

- Gradient conjugué :
  - Analysé sur des quadratiques indéfinies;
  - Étudié en tant qu'oracle de valeur propre;
- Newton + Gradient Conjugué Plafonné :
  - Complexité en itérations optimales
  - Premier ordre : complexité déterministe en $\tilde{\mathcal{O}}(\epsilon_g^{-7/4})$;
  - Résultats probabilistes au second ordre.

À suivre

Pour plus de détails...

- **Complexity analysis of second-order line-search algorithms for smooth nonconvex optimization**, C. W. Royer et S. J. Wright, *SIAM J. Optim.* 28(2):1448-1477, 2018.
- **A Newton-CG algorithm with complexity guarantees for smooth unconstrained optimization**, C. W. Royer, M. O'Neill et S. J. Wright, arXiv:1803.02924.

Travaux en cours

- **Synthèse de nos tests numériques**;
- Problèmes avec structure : variable matricielles, somme finie;
- Optimisation avec contraintes.

À suivre

Pour plus de détails...

- **Complexity analysis of second-order line-search algorithms for smooth nonconvex optimization**, C. W. Royer et S. J. Wright, *SIAM J. Optim.* 28(2):1448-1477, 2018.
- **A Newton-CG algorithm with complexity guarantees for smooth unconstrained optimization**, C. W. Royer, M. O'Neill et S. J. Wright, arXiv:1803.02924.

Travaux en cours

- **Synthèse de nos tests numériques**;
- Problèmes avec structure : variable matricielles, somme finie;
- Optimisation avec contraintes.

Merci de votre attention!

croyer2@wisc.edu