Análise de complexidade para otimização não linear com restrições usando condições KKT não escaladas e modelos de ordem superior

Sandra Augusta Santos

em colaboração com

E. G. Birgin, J. L. Gardenghi, J. M. Martínez e Ph. L. Toint

Departamento de Matemática Aplicada

IMECC – Unicamp

Campinas – SP

8 de novembro de 2016

Apoios parciais CNPq, FAPESP (Brasil) & FNRS (Bélgica)

Sumário

1. Nosso objetivo
2. Fundamentos
3. `FTarget`
4. Análise de convergência
5. Perspectivas

Complexidade (pior caso) para PNL

O problema

$$\begin{array}{ll} \text{Minimizar} & f(x) \\ \text{sujeito a} & h(x) = 0 \\ & g(x) \leq 0, \end{array}$$

em que $f : \mathbb{R}^n \to \mathbb{R}$, $h : \mathbb{R}^n \to \mathbb{R}^m$, e $g : \mathbb{R}^n \to \mathbb{R}^p$ são suficientemente suaves.

Complexidade (pior caso) para PNL

O problema

$$\begin{array}{ll} \text{Minimizar} & f(x) \\ \text{sujeito a} & h(x) = 0 \\ & g(x) \leq 0, \end{array}$$

em que $f : \mathbb{R}^n \to \mathbb{R}$, $h : \mathbb{R}^n \to \mathbb{R}^m$, e $g : \mathbb{R}^n \to \mathbb{R}^p$ são suficientemente suaves.

Questão: Dado um ponto inicial x_0 e uma tolerância $\varepsilon > 0$, quanto custa (em termos do número de avaliações de f, g, and h) encontrar um ponto $x^*(\varepsilon)$ que satisfaça as condições KKT com tolerancia ε ou declarar que o algoritmo falhou nessa tarefa?

Complexidade (pior caso) para PNL

O problema

$$\begin{array}{ll} \text{Minimizar} & f(x) \\ \text{sujeito a} & h(x) = 0 \\ & g(x) \leq 0, \end{array}$$

em que $f : \mathbb{R}^n \to \mathbb{R}$, $h : \mathbb{R}^n \to \mathbb{R}^m$, e $g : \mathbb{R}^n \to \mathbb{R}^p$ são suficientemente suaves.

Questão: Dado um ponto inicial x_0 e uma tolerância $\varepsilon > 0$, quanto custa (em termos do número de avaliações de f, g, and h) encontrar um ponto $x^*(\varepsilon)$ que satisfaça as condições KKT com tolerancia ε ou declarar que o algoritmo falhou nessa tarefa?

Observações: (i) A *análise de pior caso*, por ser muito pessimista, pode não ter implicações práticas. (ii) Na *complexidade de avaliação funcional*, os demais custos são ignorados (álgebra linear, solução de subproblemas, etc.)

Condições KKT não escaladas $\times$ escaladas

Neste trabalho, nosso objetivo é analisar a complexidade de pior caso em termos de avaliações funcionais, para a tarefa de computar um ponto ε-KKT, i.e. $(x^*, \lambda^*, \mu^*) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p_+$ que satisfaça:

$$\begin{aligned} \|h(x^*)\| + \|g(x^*)_+\| &\leq \varepsilon, \\ \|\nabla f(x^*) + \sum_{j=1}^m \lambda_j^* \nabla h_j(x^*) + \sum_{j=1}^p \mu_j^* \nabla g_j(x^*)\| &\leq \varepsilon, \\ \|\min\{-g(x^*), \mu^*\}\| &\leq \varepsilon. \end{aligned}$$

Condições KKT não escaladas $\times$ escaladas

Neste trabalho, nosso objetivo é analisar a complexidade de pior caso em termos de avaliações funcionais, para a tarefa de computar um ponto ε-KKT, i.e. $(x^*, \lambda^*, \mu^*) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p_+$ que satisfaça:

$$\begin{aligned} \|h(x^*)\| + \|g(x^*)_+\| &\leq \varepsilon, \\ \|\nabla f(x^*) + \sum_{j=1}^m \lambda_j^* \nabla h_j(x^*) + \sum_{j=1}^p \mu_j^* \nabla g_j(x^*)\| &\leq \varepsilon, \\ \|\min\{-g(x^*), \mu^*\}\| &\leq \varepsilon. \end{aligned}$$

Trabalhos anteriores consideraram uma versão *escalada* das condições ε-KKT:

$$\begin{aligned} \|h(x^*)\| + \|g(x^*)_+\| &\leq \varepsilon, \\ \|\nabla f(x^*) + \sum_{j=1}^m \lambda_j^* \nabla h_j(x^*) + \sum_{j=1}^p \mu_j^* \nabla g_j(x^*)\| &\leq \varepsilon \max\{1, \|\lambda^*\|, \|\mu^*\|\}, \\ \|\min\{-g(x^*), \mu^*\}\| &\leq \varepsilon \max\{1, \|\lambda^*\|, \|\mu^*\|\}. \end{aligned}$$

Condições KKT não escaladas $\times$ escaladas

Neste trabalho, nosso objetivo é analisar a complexidade de pior caso em termos de avaliações funcionais, para a tarefa de computar um ponto ε-KKT, i.e. $(x^*, \lambda^*, \mu^*) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p_+$ que satisfaça:

$$\begin{aligned} \|h(x^*)\| + \|g(x^*)_+\| &\leq \varepsilon, \\ \|\nabla f(x^*) + \sum_{j=1}^m \lambda_j^* \nabla h_j(x^*) + \sum_{j=1}^p \mu_j^* \nabla g_j(x^*)\| &\leq \varepsilon, \\ \|\min\{-g(x^*), \mu^*\}\| &\leq \varepsilon. \end{aligned}$$

Trabalhos anteriores consideraram uma versão *escalada* das condições ε-KKT:

$$\begin{aligned} \|h(x^*)\| + \|g(x^*)_+\| &\leq \varepsilon, \\ \|\nabla f(x^*) + \sum_{j=1}^m \lambda_j^* \nabla h_j(x^*) + \sum_{j=1}^p \mu_j^* \nabla g_j(x^*)\| &\leq \varepsilon \max\{1, \|\lambda^*\|, \|\mu^*\|\}, \\ \|\min\{-g(x^*), \mu^*\}\| &\leq \varepsilon \max\{1, \|\lambda^*\|, \|\mu^*\|\}. \end{aligned}$$

Não se trata de um simples escalamento: todo ponto *viável* x para o qual existam $\lambda \in \mathbb{R}^m$ e $\mu \in \mathbb{R}^p_+$ t.q $\sum_{j=1}^m \lambda_j \nabla h_j(x) + \sum_{j \in \mathcal{A}(x)} \mu_j \nabla g_j(x) = 0$ cumpre tal condição, indep. de x satisfazer, ou não, uma propriedade minimizadora.

Minimizar $0.5\, x^2$ s.a $0\, x = 0$: dados quaisquer $x^* \in \mathbb{R}$ e $\varepsilon > 0$, tome $\lambda^* = x^*/\varepsilon$.

Situações em que o método pode falhar

Um método $\mathcal{O}(1)$:

> Dados x_0 e $\varepsilon > 0$, verifique se x_0 é um ponto ε-KKT.
> Em caso afirmativo, pare e retorne x_0.
> Caso contrário, pare e declare que o método falhou.

Situações em que o método pode falhar

Um método $\mathcal{O}(1)$:

> Dados x_0 e $\varepsilon > 0$, verifique se x_0 é um ponto ε-KKT.
> Em caso afirmativo, pare e retorne x_0.
> Caso contrário, pare e declare que o método falhou.

Com isso, queremos enfatizar que toda análise deve claramente estabelecer as condições específicas em que o método pode falhar.

Situações em que o método pode falhar

Um método $\mathcal{O}(1)$:

> Dados x_0 e $\varepsilon > 0$, verifique se x_0 é um ponto ε-KKT.
> Em caso afirmativo, pare e retorne x_0.
> Caso contrário, pare e declare que o método falhou.

Com isso, queremos enfatizar que toda análise deve claramente estabelecer as condições específicas em que o método pode falhar.

Introduziremos uma família de métodos que depende:

- da estratégia usada para resolver os *subproblemas irrestritos*;
- da ordem das derivadas consideradas;
- das situações em que o método falha em obter um ponto ε-KKT.

Como esperado, o método mais barato é o que desiste mais facilmente.

Preliminares

Medida da inviabilidade: $V(x) = \|h(x)\|_2^2 + \|g(x)_+\|_2^2$.

Função de mérito: $\Phi(x,t) = V(x) + [f(x) - t]_+^2$, t é o *target*.

Sejam $\eta > 0$ e $\varepsilon > 0$ tolerâncias dadas para viabilidade e otimalidade, resp.

Preliminares

Medida da inviabilidade: $V(x) = \|h(x)\|_2^2 + \|g(x)_+\|_2^2$.

Função de mérito: $\Phi(x,t) = V(x) + [f(x) - t]_+^2$, t é o *target*.

Sejam $\eta > 0$ e $\varepsilon > 0$ tolerâncias dadas para viabilidade e otimalidade, resp.

O método proposto tem **duas fases**.

Na **Fase 1**, ao minimizar $V(\cdot)$, o método tenta obter um ponto x_0 (inicial para a segunda fase) satisfazendo $V(x_0) \leq 0.99\eta$.

Na **Fase 2**, η-viabilidade é sempre preservada.
Fase 2 possui iterações internas e externas. Na iteração externa k, um *target* t_k é determinado. Ao minimizar $\Phi(\cdot, t_k)$, o *solver* interno tenta atingir o *target*, mantendo a 0.99η-viabilidade, ou encontra um ponto ε-KKT para o problema original.

FTarget (do inglês *Feasibility and Target following*)

Entrada: Dados $\eta > 0$, $\varepsilon > 0$, $\rho \in (0,1)$ e $x_{-1} \in \mathbb{R}^n$. Seja $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ uma função contínua e não decrescente tal que $\psi(0) = 0$.

FASE 1: *Cálculo de uma aproximação inicial suficientemente viável*

Passo F1. Utilizando um algoritmo para Minimização Irrestrita (MI) aplicado a $V(x)$ começando em $x_{-1,0} = x_{-1}$, calcule um iterado $x_{-1,j}$, $j \in \{0, 1, 2, \dots\}$, tal que $V(x_{-1,j}) \leq V(x_{-1,0})$ e tal que $x_{-1,j}$ satisfaça pelo menos uma das seguintes condições:

$$V(x_{-1,j}) \leq 0.99\eta, \tag{1}$$

$$V(x_{-1,j}) > 0.99\eta \quad \text{e} \quad \|\nabla V(x_{-1,j})\| \leq \psi(\eta). \tag{2}$$

Defina $x_0 = x_{-1,j}$.

Passo F2. Se vale (2) então **pare** e retorne x_0.

▸ proposição

▸ teorema

FTarget (do inglês *Feasibility and Target following*) – (*cont.*)

FASE 2: *Targeting*

Passo T0. Inicialize $k \leftarrow 0$.

Passo T1. Calcule $t_k = f(x_k) - \sqrt{\eta - V(x_k)}$. ▸ Proposição

Passo T2. Utilizando um algoritmo (MI) aplicado a $\Phi(x, t_k)$ começando em $x_{k,0} = x_k$, calcule um iterado $x_{k,j}$, $j \in \{0, 1, 2, \dots\}$, tal que $\Phi(x_{k,j}, t_k) \leq \Phi(x_{k,0}, t_k) = \eta$ e tal que $x_{k,j}$ satisfaça pelo menos uma das seguintes condições:

$$f(x_{k,j}) \leq t_k + \rho(f(x_k) - t_k) \quad \text{e} \quad V(x_{k,j}) \leq 0.99\eta, \tag{3}$$

$$f(x_{k,j}) > t_k \quad \text{e} \quad \|\nabla\Phi(x_{k,j}, t_k)\| \leq 2\varepsilon[f(x_{k,j}) - t_k]_+, \tag{4}$$

$$V(x_{k,j}) > 0.99\eta \quad \text{e} \quad \|\nabla V(x_{k,j})\| \leq \psi(\eta). \tag{5}$$

Defina $x_{k+1} = x_{k,j}$.

Passo T3. Se (4) ou (5) se verificam **pare** e retorne x_{k+1}.

Passo T4. Faça $k \leftarrow k + 1$, e vá para o Passo T1. ▸ Caso 1

Significado dos critérios de parada

Caso 1 (o caso bom): O método parou porque x_{k+1} satisfez (4). ▸ (4)

Neste caso, basta (a) verificar a relação entre o gradiente do Lagrangiano e o gradiente de $\Phi(\cdot,\cdot)$, (b) mostrar que a tolerância obtida no primeiro implica na tolerância requerida no segundo, e (c) mostrar que se cumpre a complementaridade.

De fato, definindo

$$\lambda_j^* = \frac{h_j(x_{k+1})}{f(x_{k+1}) - t_k},\ j = 1,\ldots,m, \quad \text{e} \quad \mu_j^* = \frac{g_j(x_{k+1})_+}{f(x_{k+1}) - t_k},\ j = 1,\ldots,p,$$

e considerando $\eta = \varepsilon^2$, temos que $(x^* \equiv x_{k+1}, \lambda^*, \mu^*) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p_+$ satisfaz

$$\begin{aligned} \sqrt{\|h(x^*)\|_2^2 + \|g(x^*)_+\|_2^2} &\leq \varepsilon, \\ \|\nabla f(x^*) + \textstyle\sum_{j=1}^m \lambda_j^* \nabla h_j(x^*) + \sum_{j=1}^p \mu_j^* \nabla g_j(x^*)\| &\leq \varepsilon, \\ \|\min\{\mu^*, -g(x^*)\}\| &\leq \varepsilon. \end{aligned}$$

Significado dos critérios de parada (*cont.*)

Caso 2: `FTarget` parou na Fase 1 com x_0 satisfazendo (2), ou na Fase 2 com x_{k+1} satisfazendo (5), i.e. $\boxed{V(x) > 0.99\eta \;\text{ e }\; \|\nabla V(x)\| \leq \psi(\eta)}$

Note que ao parar por (5) tambem se cumpre $V(x_{k+1}) \leq \eta$;
no entanto, ao parar por (2), $V(x_0)$ pode ser grande.
$\rightarrow$ Esta última situação é a que caracteriza a existência, no limite, de um ponto inviável e estacionário para a medida de inviabilidade.

Significado dos critérios de parada (*cont.*)

Caso 2: `FTarget` parou na Fase 1 com x_0 satisfazendo (2), ou na Fase 2 com x_{k+1} satisfazendo (5), i.e. $\boxed{V(x) > 0.99\eta \;\text{ e }\; \|\nabla V(x)\| \leq \psi(\eta)}$

Note que ao parar por (5) tambem se cumpre $V(x_{k+1}) \leq \eta$;
no entanto, ao parar por (2), $V(x_0)$ pode ser grande.
$\rightarrow$ Esta última situação é a que caracteriza a existência, no limite, de um ponto inviável e estacionário para a medida de inviabilidade.

Impacto da escolha de $\boxed{\psi(\eta) = \sigma_1 \eta^{\sigma_2}}$, $\sigma_1 > 0, \sigma_2 \in [\frac{1}{2}, 1]$:

- $\sigma_2 = \frac{1}{2} \rightarrow$ um ponto η-viável que não possui gradientes das restrições ativas ξ-uniformemente PLI (com $\xi = \sigma_1/(2\sqrt{0.99})$) foi encontrado;
- $\sigma_2 \in (\frac{1}{2}, 1) \rightarrow$ um ponto η-viável que não satisfaz a cond. de qualificação de Mangasarian-Fromowitz existe no limite;
- $\sigma_2 = 1 \rightarrow$ um ponto η-viável para o qual $V(\cdot)$ não satisfaz a desigualdade de Łojasiewicz existe no limite.

Terminação finita, convergência e complexidade

Seja $S_\eta = \{x \in \mathbb{R}^n \mid V(x) \leq \eta\}$.
Assumiremos que a função f é limitada superiorm. (por f^{up}) e inferiorm. (por f_{low}) no conjunto S_η e que $\|\nabla f\|$ é limitada (por Γ) em S_η.

Proposição: `FTarget` para na Fase 1 ou executa, no máximo, ▸ Fase 1

$$\left\lfloor \frac{f^{\text{up}} - f_{\text{low}}}{0.1(1-\rho)\sqrt{\eta}} \right\rfloor + 1$$

iterações externas na Fase 2. (Sobre usar $f(x_0)$ ao invés de f^{up}.) ▸ Fase 2
$\rightarrow$ Basta observar que $f(x_{k+1}) \leq f(x_k) - 0.1(1-\rho)\sqrt{\eta}$ para todo k.

Precisamos agora medir o esforço do algoritmo MI escolhido para resolver o problema irrestrito da Fase 1 e o problema irrestrito de cada iteração externa da Fase 2.

Terminação finita, convergência e complexidade (*cont.*)

Assumimos agora que, dado $\bar{\varepsilon} > 0$, o algoritmo MI, quando aplicado à minimização de $V(\cdot)$ a partir de z_0, encontra z_k tal que $\|\nabla V(z_k)\| \leq \bar{\varepsilon}$ usando no máximo

$$c_V \times \left(\frac{V(z_0) - V_{\text{low}}}{\bar{\varepsilon}^{\alpha}} \right)$$

avaliações de V e de suas derivadas. Tal hipótese nos dá imediatamente o resultado desejado para o custo da Fase 1.

Também assumimos uma relação similar com relação a $\Phi(\cdot, t)$.
O resultado para cada subproblema da Fase 2 vem do seguinte fato: sempre que o método não para em um ponto $x_{k,j}$, temos

$$\|\nabla \Phi(x_{k,j}, t_k)\| \geq \min \left\{ 0.2\rho\varepsilon\sqrt{\eta},\ \frac{\psi(\eta)}{2},\ \frac{\psi(\eta)\varepsilon}{2\Gamma} \right\}.$$

Terminação finita, convergência e complexidade (*cont.*)

▸ F2

Teorema: `FTarget` para no Passo F2 da Fase 1, ou existe $k \in \{0, 1, 2, \dots\}$ tal que `FTarget` para na iteração k da Fase 2 e retorna x_{k+1}. Além disso, na Fase 1, `FTarget` utiliza, no máximo,

$$c_V \times \left(\frac{V(x_{-1})}{\psi(\eta)^\alpha} \right)$$

avaliações de h, g, e suas derivadas, e, na Fase 2, no máximo

$$c_\Phi \times \left(\frac{\eta}{\min\left\{ 0.2\rho\varepsilon\sqrt{\eta},\ \frac{\psi(\eta)}{2},\ \frac{\psi(\eta)\varepsilon}{2\Gamma} \right\}^\alpha} \right) \times \left(\left\lfloor \frac{f^{\text{up}} - f_{\text{low}}}{0.1(1-\rho)\sqrt{\eta}} \right\rfloor + 1 \right),$$

avaliações de f, h, g, e suas derivadas.

A constante c_V depende apenas de h, g, e parâmetros do algoritmo MI, enquanto a constante c_Φ depende apenas de f, h, g, e parâmetros do MI.

Resumo dos resultados de complexidade (com $\eta = \varepsilon^2$)

		$\psi(\eta) = \sigma_1 \eta^{\sigma_2}$		
		$\sigma_2 = \frac{1}{2}$	$\sigma_2 \in (\frac{1}{2}, 1)$	$\sigma_2 = 1$
Resultados de convergencia		ε-KKT ou gradientes das restr. ativas não ξ-uniform. PLI com $\xi = \sigma_1/(2\sqrt{0.99})$	ε-KKT ou não MFCQ	ε-KKT ou não Łojasiewicz
Complexidade	$\alpha = 1.5$	Fase 1: $\mathcal{O}(\varepsilon^{-1.5})$ Fase 2: $\mathcal{O}(\varepsilon^{-2})$	Fase 1: $\mathcal{O}(\varepsilon^{-1.5-3\sigma_2})$ Fase 2: $\mathcal{O}(\varepsilon^{-2-3\sigma_2})$	Fase 1: $\mathcal{O}(\varepsilon^{-3})$ Fase 2: $\mathcal{O}(\varepsilon^{-3.5})$
	$\alpha = 2$	Fase 1: $\mathcal{O}(\varepsilon^{-2})$ Fase 2: $\mathcal{O}(\varepsilon^{-3})$	Fase 1: $\mathcal{O}(\varepsilon^{-2-4\sigma_2})$ Fase 2: $\mathcal{O}(\varepsilon^{-3-4\sigma_2})$	Fase 1: $\mathcal{O}(\varepsilon^{-4})$ Fase 2: $\mathcal{O}(\varepsilon^{-5})$

Conclusões

- Introduzimos um algoritmo com resultado de complexidade no qual dado $\varepsilon > 0$ possivelmente encontra um ponto ε-KKT *não escalado*.

Conclusões

- Introduzimos um algoritmo com resultado de complexidade no qual dado $\varepsilon > 0$ possivelmente encontra um ponto ε-KKT *não escalado*.
- Três diferentes hipóteses que implicam no sucesso do método foram claramente estabelecidas.

Conclusões

- Introduzimos um algoritmo com resultado de complexidade no qual dado $\varepsilon > 0$ possivelmente encontra um ponto ε-KKT *não escalado*.
- Três diferentes hipóteses que implicam no sucesso do método foram claramente estabelecidas.
- Hipóteses mais fracas implicam em resultados de convergência mais fortes, e métodos *mais caros*.

Conclusões

- Introduzimos um algoritmo com resultado de complexidade no qual dado $\varepsilon > 0$ possivelmente encontra um ponto ε-KKT *não escalado*.
- Três diferentes hipóteses que implicam no sucesso do método foram claramente estabelecidas.
- Hipóteses mais fracas implicam em resultados de convergência mais fortes, e métodos *mais caros*.
- Se o algoritmo MI para resolver os subproblemas utiliza um modelo baseado em derivadas de ordem q, sua complexidade é $\mathcal{O}(\varepsilon^{-(q+1)/q})$ e o método introduzido para PNL é

$$\mathcal{O}(\varepsilon^{1-2(q+1)/q}), \quad \mathcal{O}(\varepsilon^{1-(1+2\sigma_2)(q+1)/q}) \quad \text{ou} \quad \mathcal{O}(\varepsilon^{1-3(q+1)/q}),$$

dependendo das situações permitidas que o método falhe.

Referências

- E.G. Birgin, J.L. Gardenghi, J.M. Martínez, S.A. Santos, e Ph.L. Toint, Evaluation complexity for nonlinear constrained optimization using unscaled KKT conditions and high-order models, *SIAM Journal on Optimization* 26, pp. 951–967, 2016.

Referências

- E.G. Birgin, J.L. Gardenghi, J.M. Martínez, S.A. Santos, e Ph.L. Toint, Evaluation complexity for nonlinear constrained optimization using unscaled KKT conditions and high-order models, *SIAM Journal on Optimization* 26, pp. 951–967, 2016.

- E.G. Birgin, J.L. Gardenghi, J.M. Martínez, S.A. Santos, e Ph.L. Toint, Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models, *Mathematical Programming* DOI:10.1007/s10107-016-1065-8

Referências

- E.G. Birgin, J.L. Gardenghi, J.M. Martínez, S.A. Santos, e Ph.L. Toint, Evaluation complexity for nonlinear constrained optimization using unscaled KKT conditions and high-order models, *SIAM Journal on Optimization* 26, pp. 951–967, 2016.

- E.G. Birgin, J.L. Gardenghi, J.M. Martínez, S.A. Santos, e Ph.L. Toint, Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models, *Mathematical Programming* DOI:10.1007/s10107-016-1065-8

Muito obrigada!

sandra@ime.unicamp.br