Dualidade e Métodos iterativos para otimização sem restrições de 1º ordem

Departamento de Matemática
Universidade do Minho

Outline

Dualidade

Ideias gerais:

- A teoria da dualidade mostra como podemos construir um problema alternativo (problema dual) a partir do problema de otimização original (problema primal).
- Em alguns casos, o problema dual é computacionalmente mais fácil de resolver do que o problema primal.
- Noutros casos, o problema dual pode ser usado para obter facilmente um limite inferior para o valor ótimo F^* da função objetivo do problema primal.
- A dualidade tem sido também usada para desenvolver algoritmos para resolver o problema primal.

Função dual de Lagrange

Para simplificar a exposição, considera-se o caso especial do problema de otimização com restrições de desigualdade:

$$\begin{aligned} \underset{w \in \mathbb{R}^d}{\text{minimizar}} \quad & F(w) \\ \text{sujeito a} \quad & c_n(w) \geq 0, \quad n = 1, \ldots, N \end{aligned} \tag{1}$$

A Função Lagrangiana associada a este problema é

$$L(w, \lambda) = F(w) - \sum_{n=1}^{N} \lambda_n c_n(w)$$

▷ $\lambda = (\lambda_1, \ldots, \lambda_N)^T$ é vetor dos multiplicadores de Lagrange associadas às restrições $c_n(w) \geq 0$.

Função dual de Lagrange

A função dual $F_D : \mathbb{R}^N \to \mathbb{R}$ é definida pelo ínfimo (valor mínimo) da função Lagrangiana sobre w: para $\lambda \in \mathbb{R}^N$

$$F_D(\lambda) = \inf_{w \in \mathbb{R}^d} L(w, \lambda) \tag{2}$$

- Se a função Lagrangeana é ilimitada inferiormente em w, para alguns valores de λ, então a função dual toma o valor $-\infty$.
- Considera-se para domínio de F_D o conjunto dos valores de $\lambda \in \mathbb{R}^N$ para os quais F_D é finita, ou seja,

$$dom\, F_D = \{\lambda \in \mathbb{R}^N : F_D(\lambda) > -\infty\}$$

Teorema 1

Para qualquer ponto $\tilde{w}$ admissível no problema (1) e para qualquer $\tilde{\lambda} \geq 0$. tem-se que:

$$F_D(\tilde{\lambda}) \leq F(\tilde{w}).$$

Demonstração.

$$\begin{aligned} F_D(\tilde{\lambda}) &= \inf_{w\in\mathbb{R}^d} L(w,\tilde{\lambda}) = \inf_{w\in\mathbb{R}^d} F(w) - \textstyle\sum_{n=1}^{N} \tilde{\lambda}_n c_n(w) \\ &\leq F(\tilde{w}) - \textstyle\sum_{n=1}^{N} \tilde{\lambda}_n c_n(\tilde{w}) \\ &\leq F(\tilde{w}), \end{aligned}$$

onde a desigualdade final segue-se por $\tilde{\lambda} \geq 0$ e $c_n(\tilde{w}) \geq 0$, para todo $n = 1, \ldots, N$. □

Nota: A função dual produz limites inferiores no valor ótimo $F(w^*)$ do problema (1).

Definição 2

O problema dual para o problema (1) é definido da forma:

$$\begin{array}{ll} \underset{\lambda \in \mathbb{R}^{\mathrm{N}}}{\text{maximizar}} & F_D(\lambda) \equiv \inf_{w \in \mathbb{R}^d} L(w, \lambda) \\ \text{sujeito a} & \lambda \geq 0 \end{array} \tag{3}$$

Calcular o ínfimo de (2) implica encontrar o minimizante global da função $L(., \lambda)$ para um λ dado, o que pode ser extremamente difícil na prática.

Porém, quando F e $-c_n$ são funções convexas e $\lambda \geq 0$, a função Lagrangiana $L(., \lambda)$ é também convexa, todos os minimizantes locais são minimizantes globais.

Quando é garantida a convexidade de um problema primal e a regularidade do ponto ótimo, é possível concluir que a solução do problema dual é igual à solução do problema primal.

Teorema 3

Seja $F, -c_i, \ i = 1, 2, \ldots N$ funções convexas e continuamente diferenciáveis em $\mathbb{R}^d$. Seja w^* um ponto regular e uma solução para o problema (1). Seja $\hat{\lambda}$ uma solução para o problema dual (3) e que o seu ínfimo é $L(w, \hat{\lambda})$ é alcançado em $\hat{w}$. Assume-se que $L(\cdot, \hat{\lambda})$ é uma função estritamente convexa. Então $w^* = \hat{w}$ (única solução) e $F(w*) = L(\hat{w}, \hat{\lambda})$.

Um pequena alteração na formulação da dualidade é necessária para facilitar a computação, esta nova formulação é conhecida por Dual de Wolfe.

Teorema 4 (Dual de Wolfe)

Se (w^*, λ^*) é um par solução do problema (1), se F e $-c_n$, $n = 1, \ldots, N$, são funções convexas e continuamente diferenciáveis, e se w^* é ponto regular, então (w^*, λ^*) é solução do problema dual

$$\begin{array}{ll} \underset{w \in \mathbb{R}^d, \alpha \in \mathbb{R}^N}{\text{maximizar}} & L(w, \lambda) \\ \text{sujeito a} & \nabla_w L(w, \lambda) = 0, \lambda \geq 0. \end{array}$$

Além disso, $F(w^*) = L(w^*, \lambda^*)$.

Demonstração.

Das condições de KKT, temos que o par (w^*, λ^*) satisfaz: $\nabla_w L(w, \lambda) = 0, \lambda \geq 0$ e $L(w^*, \lambda^*) = F(w^*)$. □

Demonstração.

Por tanto, para qualquer par (w^*, λ^*), temos que:

$$\begin{aligned} L(w^*, \lambda^*) &= F(w^*) \\ &\geq F(w^*) - \textstyle\sum_{n=1}^{N} \lambda_n c_n(w^*) \\ &= L(w^*, \lambda) \\ &\geq L(w, \lambda) + \nabla_w L(w, \lambda)^T (w^* - w) \\ &= L(w, \lambda), \end{aligned}$$

onde a segunda desigualdade vem da convexidade $L(\cdot, \lambda)$. Portanto, mostramos que (w^*, λ^*) maximiza L sob a restrição $\nabla_w L(w, \lambda) = 0, \lambda \geq 0$, logo, é solução para o problema dual. □

Exercício: Utilize o teorema Dual de Wolfe para determinar a solução para o problema

$$\begin{aligned} \underset{w \in \mathbb{R}^2}{\text{minimizar}} \quad & 0.5(w_1^2 + w_2^2) \\ \text{sujeito a} \quad & w_1 - 1 \geq 0. \end{aligned}$$

Exercício: Reescrever o problema de programação linear aplicando o teorema Dual de Wolfe

$$\begin{array}{rc} \underset{w\in\mathbb{R}^d}{\text{minimizar}} & c^T w \\ \text{sujeito a} & Aw - b \geq 0. \end{array}$$

com $A \neq 0$.

Exercício: Mostre que problema de programação convexa quadrática

$$\begin{array}{rc} \underset{w\in\mathbb{R}^d}{\text{minimizar}} & \frac{1}{2}w^T Gw + c^T w \\ \text{sujeito a} & Aw - b \geq 0. \end{array}$$

onde G é uma matriz simétrica definida positiva e $A \neq 0$ é equivalente ao problema:

$$\begin{array}{rr} \underset{w\in\mathbb{R}^d,\lambda\in\mathbb{R}^N}{\text{maximizar}} & -\frac{1}{2}w^T Gw + \lambda^T b \\ \text{sujeito a} & Gw + c - A^T\lambda = 0. \\ & \lambda \geq 0. \end{array}$$

Em geral, não se consegue resolver

$$\underset{w \in \mathbb{R}^d}{\text{minimizar}} \; F(w)$$

analiticamente ... recorre-se a métodos iterativos.

Métodos iterativos para Otimização

- Começam a partir de uma aproximação inicial à solução, $w^{(1)}$
- Dado $w^{(k)}$, calculam um novo (melhor) ponto $w^{(k+1)}$, e este processo repete-se ($k = 1, 2, \dots$)
- Geram uma sucessão $\{w^{(k)}\}$ de aproximações na qual a função F decresce, que espera-se que convirja para a solução ótima w^*.

Critério 1 (proposto por Wolfe). Parar o algoritmo para otimização sem restrições se:

$$\underbrace{\left\|\nabla F(w^{(k)})\right\|}_{\text{medida de estacionaridade}} \leq \varepsilon_1$$

e

$$\underbrace{\frac{\left\|w^{(k)} - w^{(k-1)}\right\|}{\left\|w^{(k)}\right\|}}_{\text{erro relativo da aproximação}} \leq \varepsilon_2$$

e

$$\frac{\left|F(w^{(k)}) - F(w^{(k-1)})\right|}{\left|F(w^{(k)})\right|} \leq \varepsilon_3$$

$\varepsilon_1, \varepsilon_2, \varepsilon_3$ constantes positivas próximas de zero.

Critérios de paragem

Critério 2 (proposto por Gill e Murray). Parar o algoritmo para otimização sem restrições se:

$$\left\| \nabla F(w^{(k)}) \right\| \leq \varepsilon^{\frac{1}{3}} \left(1 + \left| F(w^{(k)}) \right| \right)$$

e

$$\left\| w^{(k)} - w^{(k-1)} \right\| \leq \varepsilon \left(1 + \left\| w^{(k)} \right\| \right)$$

e

$$\left| F(w^{(k)}) - F(w^{(k-1)}) \right| \leq \varepsilon^2 \left(1 + \left| F(w^{(k)}) \right| \right)$$

ε é uma constante positiva próxima de zero.

Observação: Este critério é de aplicação mais geral! Pode ser aplicado a problemas em que a solução ótima é o vetor nulo ou o valor ótimo da função objetivo é zero.

Diz-se que o método de otimização tem convergência de primeira ordem se a sucessão $\{w^{(k)}\}$ de aproximações à solução converge para um ponto estacionário de F.

- Usa-se o termo convergência global se o método tem convergência de primeira ordem qualquer que seja a aproximação inicial $w^{(1)}$ do processo iterativo.
- Usa-se o termo convergência local se o método tem convergência de primeira ordem apenas quando a aproximação inicial $w^{(1)}$ estiver suficientemente perto da solução.

Métodos de Direção de Descida

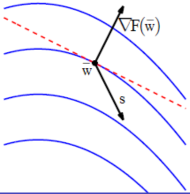
Definição 5 (Direção de descida)

Considere uma função $F : \mathbb{R}^d \to \mathbb{R}$ e um ponto $\overline{w} \in \mathbb{R}^d$. Uma direção $s \in \mathbb{R}^d \setminus \{0\}$ é uma direção de descida para F a partir de $\overline{w}$, se existe $\overline{\eta} > 0$ tal que:

$$F(\overline{w} + \eta s) < F(\overline{w}), \text{ para todo } \eta \in (0, \overline{\eta}).$$

Teorema 6

Se $\nabla F(\overline{w})^T s < 0$, então s é uma direção de descida para F a partir de $\overline{w}$.



Recordar: $\nabla F(\overline{w})^T s < 0 \Rightarrow$ o declive de F em $\overline{w}$ na direção de s é negativo.

Ideia:

- calcular uma direção de descida, $s^{(k)}$
- procurar uma redução no valor de F ao longo da direção $s^{(k)}$

Algoritmo: Método de direção de descida geral

1. Dar: $w^{(1)}$
2. Fazer $k = 1$
3. **Enquanto** ($w^{(k)}$ não satisfaz o critério de paragem)
4. Calcular uma direção de procura $s^{(k)}$ tal que: $\nabla F(w^{(k)})^T s^{(k)} < 0$
5. Encontrar o comprimento do passo η_k tal que:

$$F(w^{(k)} + \eta_k s^{(k)}) < F(w^{(k)})$$

6. Fazer $w^{(k+1)} = w^{(k)} + \eta_k s^{(k)}$
7. Fazer $k = k + 1$
8. **Fim enquanto**

Observações sobre o método de direções de descida geral:

- existem várias escolhas possíveis para $s^{(k)}$;
- procura exata do η_k:

$$\underset{\eta\in\mathbb{R}}{\text{minimizar}}\ F(w^{(k)} + \eta s^{(k)})$$

... é, em geral, impraticável, devido ao elevado custo computacional e, em geral, não é necessária. Na prática, recorre-se a técnicas de procura não exata.

- A condição de redução simples $F(w^{(k)} + \eta_k s^{(k)}) < F(w^{(k)})$, não garante convergência para um ponto estacionário (ver exemplo).

Exercício: Considere $F : \mathbb{R} \to \mathbb{R}$ dada por $F(w) = w^2$, utilize o algoritmo de direção de descida geral para determinar o minimizante com $w^{(1)} = \frac{3}{4}$, $s^{(k)} = -1$ e $\eta_k = 2^{(-k-2)}$. Que pode concluir?

Nota: Tem se que:

$$w^{(k+1)} = w^{(k)} + \eta_k s^{(k)} = w^{(k)} - \eta_k = w^{(1)} - \sum_{i=1}^{k} \eta_k$$

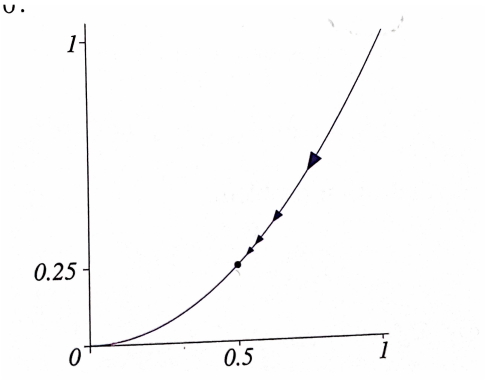
e então

$$w^{(k+1)} = \frac{3}{4} - \sum_{i=1}^{k} 2^{(-k-2)} = \frac{3}{4} - \frac{1}{8} \sum_{i=1}^{k} \left(\frac{1}{2}\right)^{k-1} = \frac{1}{2} + \frac{1}{4}\left(\frac{1}{2}\right)^{k} = \frac{1}{2} + \left(\frac{1}{2}\right)^{k+2}$$

Observa-se que

$$0 < w^{(k+1)} < w^{(k)}, \text{ logo } F\left(w^{(k+1)}\right) < F\left(w^{(k)}\right),$$

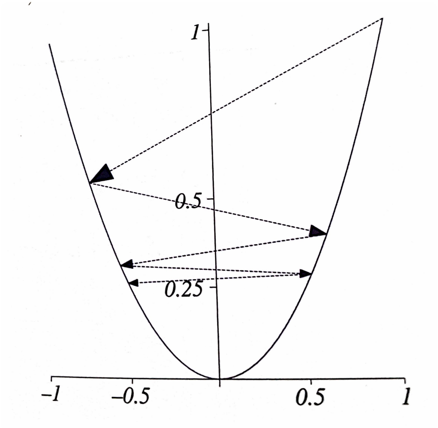
com $w^{(k)} \to \frac{1}{2}$ e $F(w^{(k)}) \to \frac{1}{4}$. Contudo, o minimizante é 0 e o mínimo também é 0.



passos muito curtos

Exercício: Resolva o exercício anterior com $w^{(1)} = \frac{3}{4}$, $s^{(k)} = (-1)^k$ e $\eta_k = 1 + \frac{3}{2^{k+2}}$. Que pode concluir?

Nota: $w^{(k+1)} = \frac{1}{2}(-1)^k \left(1 + \frac{1}{2^{k+1}}\right).$



passos muito longos

Método do Gradiente

Pode-se utilizar várias técnicas de otimização no algoritmo de descida mais rápida, aqui vamos dar especial atenção ao método do gradiente.

No **método do gradiente**, a direção de procura é dada pela direção de descida máxima:

$$s^{(k)} = -\nabla F(w^{(k)}).$$

Algoritmo: Método do Gradiente

1. Dar: $w^{(1)}$
2. Fazer $k = 1$
3. **Enquanto** ($w^{(k)}$ não satisfaz o critério de paragem)
4. Calcular a direção de descida máxima $s^{(k)} = -\nabla F(w^{(k)})$
5. Calcular o comprimento do passo η_k tal que
$$F(w^{(k)} + \eta_k s^{(k)}) < F(w^{(k)})$$
6. Fazer $w^{(k+1)} = w^{(k)} + \eta_k s^{(k)}$
7. Fazer $k = k + 1$
8. **Fim enquanto**

Exercício: Considerando a direção do método do gradiente, determine a procura exata do η_k para a função quadrática definida por

$$F(w) = \frac{1}{2} w^T Q w + b^T w$$

onde Q é uma matriz definida positiva simétrica e b é um vetor em $\mathbb{R}^d$.

Exercício: Utilize o Método do Gradiente para determinar a solução do problema

$$\underset{w \in \mathbb{R}^d}{\text{minimizar}}\, w_1^2 + 2w_2^2$$

.

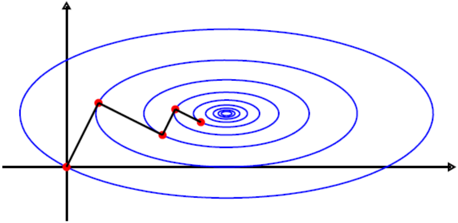
Note-se que: $F(w) = w_1^2 + 2w_2^2$ pode ser escrita na forma $\frac{1}{2} w^T Q w + b^T w$ com $Q = \nabla^2 F(w)$ e $b = \nabla F(0)$.

Propriedade do Método do gradiente

- Se no método do gradiente o comprimento do passo, η_k, é obtido por procura exata, então as sucessivas direções de descida máxima definem ângulos retos:

$$s^{(k+1)\ T} s^{(k)} = 0$$

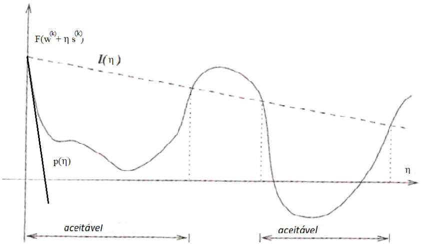
e o método apresenta um comportamento em ziguezague que se traduz num processo muito lento quando já está perto do mínimo.



Procura não exata - Condição de Armijo

Dados $w^{(k)}, s^{(k)} \in \mathbb{R}^{\mathrm{I}}$ e $\delta \in (0,1)$. A **Condição de Armijo** encontra o η_k que origina uma redução significativa no valor de F, dada por:

$$F(w^{(k)} + \eta s^{(k)}) \leq \underbrace{F(w^{(k)}) + \delta\eta\nabla F^T \left(w^{(k)}\right) s^{(k)}}_{l(\eta)}$$



Condição de redução significativa

$\Rightarrow$ impede comprimentos de passos muito longos

Algoritmo - condição de Armijo com backtracking

Na prática, para prevenir que a condição Armijo seja verificada por comprimentos do passo muito pequenos (ver figura), aplica-se uma estratégia de bracktracking.

Algoritmo: condição de Armijo com backtracking

1. Dar $\overline{\eta} > 0$, $\delta \in (0, 1)$
2. Fazer $\eta \leftarrow \overline{\eta}$
3. **Enquanto** $F(w^{(k)} + \eta s^{(k)}) > F(w^{(k)}) + \delta\eta\nabla^T F\left(w^{(k)}\right) s^{(k)}$
4. Fazer $\eta \leftarrow \eta/2$
5. **Fim enquanto**
6. Fazer $\eta_k \leftarrow \eta$

- Se no método do gradiente o comprimento do passo, η_k, é obtido por procura exata (ou por procura não exata: condição de Armijo) então o método é globalmente convergente.

Exercício

Resolva o problema

$$\underset{w \in \mathbb{R}^2}{\text{minimizar}} \; F(w) = w_1^2 + 2w_2^2$$

usando o método do gradiente. O processo iterativo deve ser iniciado com o ponto $(1, 1)$ e deve terminar quando o critério de paragem baseado na condição $\|\nabla F(w)\| \leq \varepsilon$ for verificado para $\varepsilon = 0.1$. Usar o algoritmo de procura de Armijo com backtracking para calcular o η_k, em cada iteração. Considere $\delta = 0.1$. e $\overline{\eta} = 1$

Os apontamentos foram baseados na seguinte bibliografia:[?] e [?].

W. Forst and D. Hoffmann.
Optimization—theory and practice.
Springer Science & Business Media, 2010.

J. Nocedal and S. J. Wright.
Numerical optimization.
Springer, 1999.