Métodos quase-Newton

Marina Andretta

ICMC-USP

21 de setembro de 2010

Métodos quase-Newton

Métodos quase-Newton, assim como o método de máxima de descida, necessitam que apenas o gradiente da função objetivo esteja disponível em cada iteração.

Ao medir as mudanças no gradiente de uma iteração para outra, eles tentam construir um modelo para a função objetivo bom o bastante para produzir convergência superlinear.

A melhora em relação ao método de máxima descida é dramática, principalmente em problemas difíceis.

Por não necessitar de segundas derivadas, métodos quase-Newton podem ser até mais eficientes do que métodos de Newton em alguns casos.

Métodos DFP e BFGS

O método quase-Newton mais popular é o método BFGS, criado por Broyden, Fletcher, Goldfarb e Shanno. Para entender como foi desenvolvido o método BFGS, veremos primeiro o desenvolvimento do método DFP.

Primeiramente, considere o seguinte modelo quadrático para a função objetivo no ponto x_k

$$m_k(p) = f_k + \nabla f_k^T p + \frac{1}{2} p^T B_k p,$$

com B_k simétrica e definida positiva. Esta matriz será atualizada a cada iteração.

Método DFP

Note que, no ponto $p = 0$, o valor do modelo e seu gradiente coincidem com o valor da função objetivo e de seu gradiente.

O minimizador p_k deste modelo, dado por

$$\boxed{p_k = -B_k^{-1} \nabla f_k,} \tag{1}$$

é usado como direção de busca.

Método DFP

O novo iterando é dado por

$$x_{k+1} = x_k + \alpha_k p_k,$$

com α_k satisfazendo as condições de Wolfe.

Esta iteração é muito similar a uma iteração do método de Newton. A diferença está no uso da aproximação B_k no lugar a Hessiana verdadeira.

Método DFP

Em vez de recalcular B_k a cada iteração, a ideia é atualizar B_k de maneira simples a cada iteração, usando informação de curvatura obtida na iteração atual e na anterior.

Suponha que tenhamos calculado um novo iterando x_{k+1} e desejamos construir um novo modelo quadrático

$$m_{k+1}(p) = f_{k+1} + \nabla f_{k+1}^T p + \frac{1}{2} p^T B_{k+1} p.$$

Método DFP

Se nos basearmos nas informações obtidas durante a última iteração, que condições B_{k+1} deve satisfazer?

Uma condição razoável seria exigir que o gradiente do novo modelo m_{k+1} coincidisse com o gradiente da função f nos últimos dois iterandos x_k e x_{k+1}.

Como $\nabla m_{k+1}(0) = \nabla f_{k+1}$, precisamos apenas exigir que

$$\nabla m_{k+1}(-\alpha_k p_k) = \nabla f_{k+1} - \alpha_k B_{k+1} p_k = \nabla f_k.$$

Método DFP

Reordenando a equação, temos que

$$B_{k+1}\alpha_k p_k = \nabla f_{k+1} - \nabla f_k.$$

Para simplificar a notação, vamos definir os vetores

$$s_k = x_{k+1} - x_k, \quad y_k = \nabla f_{k+1} - \nabla f_k.$$

Método DFP

Assim, queremos satisfazer a seguinte equação:

$$\boxed{B_{k+1}s_k = y_k,} \tag{2}$$

chamada de equação secante.

Método DFP

Dadas as mudanças nos iterandos (s_k) e em seus gradientes (y_k), a equação secante exige que a matriz simétrica definida positiva B_{k+1} mapeie s_k em y_k.

Isso será possível apenas se s_k e y_k satisfizerem a condição de curvatura

$$\boxed{s_k^T y_k > 0,} \tag{3}$$

o que pode ser visto facilmente premultiplicando (2) por s_k^T.

Método DFP

Quando f é fortemente convexa, a desigualdade $s_k^T y_k > 0$ é satisfeita para quaisquer dois pontos x_k e x_{k+1}.

No entanto, esta condição nem sempre valerá para funções não-convexas. Neste caso, teremos de forçar que ela valha explicitamente, impondo restrições à busca linear que calcula α_k.

Quando α_k satisfaz as condições de Wolfe ou as condições fortes de Wolfe, temos que $s_k^T y_k > 0$.

Método DFP

Para verificar este fato, basta notar que, nas condições de Wolfe, temos que

$$\nabla f_{k+1}^T p_k \geq c_2 \nabla f_k^T p_k.$$

Ou seja,

$$\nabla f_{k+1}^T s_k \geq c_2 \nabla f_k^T s_k.$$

Método DFP

Daí temos que

$$s_k^T y_k \geq (c_2 - 1)\alpha_k \nabla f_k^T p_k.$$

Como p_k é uma direção de descida em relação a x_k, $c_2 < 1$ e $\alpha_k > 0$, temos que a condição da curvatura (3) vale.

Método DFP

Quando a condição de curvatura é satisfeita, a equação secante (2) tem solução B_{k+1}.

De fato, infinitas soluções são possíveis, já que o grau de liberdade em uma matriz simétrica é $n(n+1)$ e a equação secante representa apenas n restrições.

A condição de que B_{k+1} deve ser definida positiva impõe mais n restrições (todos os menores principais devem ser positivos), mas estas condições não absorvem todo o grau de liberdade.

Método DFP

Para determinar B_{k+1} de maneira única, impomos mais uma condição: dentre todas as matrizes simétricas que satisfazem a equação secante, B_{k+1} deve ser, em algum sentido, a mais próxima da matriz atual B_k.

Ou seja, queremos encontrar a matriz solução para o seguinte problema:

$$\begin{array}{ll} \text{Minimizar}_B & \|B - B_k\| \\ \text{Sujeita a} & B = B^T, \quad Bs_k = y_k, \end{array} \tag{4}$$

com B_k simétrica definida positiva e $y_k^T s_k > 0$.

Método DFP

Muitas normas de matriz podem ser usadas no problema (4). Cada uma delas gera um método quase-Newton diferente.

Uma norma que faz com que o problema (4) seja facilmente resolvido e seja invariante quanto ao escalamento é a norma de Frobenius com peso

$$\|A\|_W \equiv \|W^{1/2} A W^{1/2}\|_F. \tag{5}$$

Método DFP

A matriz W pode ser qualquer matriz que satisfaça a relação $Wy_k = s_k$.

Para sermos concretos, podemos supor que $W = \bar{G}_k^{-1}$, com $\bar{G}_k$ a Hessiana média definida por

$$\bar{G}_k = \left[\int_0^1 \nabla^2 f(x_k + \tau \alpha_k p_k) d\tau \right].$$

Método DFP

A propriedade

$$y_k = \bar{G}_k \alpha_k p_k = \bar{G}_k s_k.$$

segue do teorema de Taylor.

Com esta escolha de W, a norma (5) é adimensional. Isto é desejável, já que não queremos que a matriz B_{k+1}, solução de (4), dependa das unidades do problema.

Método DFP

Com esta matriz W de peso e a norma de Frobenius com peso, a solução única para o problema (4) é dada por

$$B_{k+1} = (I - \gamma_k y_k s_k^T) B_k (I - \gamma_k s_k y_k^T) + \gamma_k y_k y_k^T,$$

com

$$\gamma_k = \frac{1}{y_k^T s_k}.$$

Método DFP

Esta fórmula é chamada de fórmula de atualização DFP, já que foi originalmente proposta por Davidson e depois estudada, implementada e popularizada por Fletcher e Powell.

A inversa de B_K, que chamaremos de $H_k = B_k^{-1}$, é bastante útil na implementação do método, já que permite que a direção p_k (1) seja calculada por um produto matriz-vetor.

Método DFP

Usando a fórmula de Sherman-Morrison-Woodbury, podemos definir a seguinte fórmula de atualização da inversa da aproximação da Hessiana H_k que corresponde à atualização DFP de B_k:

$$H_{k+1} = H_k - \frac{H_k y_k y_k^T H_k}{y_k^T H_k y_k} + \frac{s_k s_k^T}{y_k^T s_k}.$$

Método BFGS

A fórmula de atualização DFP é muito eficaz, mas foi rapidamente substituída pela fórmula BFGS.

Para chegar à formula de atualização BFGS, basta fazer uma pequena mudança nos argumentos que levaram à fórmula DFP: em vez de impor condições à aproximação da Hessiana B_k, são impostas condições similares à sua inversa H_k.

A aproximação atualizada H_{k+1} deve ser simétrica definida positiva e deve satisfazer a equação secante

$$H_{k+1} y_k = s_k.$$

Método BFGS

A condição de proximidade a H_k é especificada de forma análoga a (4):

$$\begin{array}{ll} \text{Minimizar}_H & \|H - H_k\| \\ \text{Sujeita a} & H = H^T, \quad Hy_k = s_k. \end{array} \tag{6}$$

Método BFGS

A norma usada é, novamente, a de Frobenius com peso, com a matriz W, agora, satisfazendo $Ws_k = y_k$. Novamente, usaremos $W = \bar{G}_k^{-1}$.

A solução única H_{k+1} de (6) é dada por

$$H_{k+1} = (I - \rho_k s_k y_k^T) H_k (I - \rho_k y_k s_k^T) + \rho_k s_k s_k^T, \tag{7}$$

com

$$\rho_k = \frac{1}{y_k^T s_k}.$$

Método BFGS

Antes que possamos definir completamente o método BFGS, há apenas uma questão em aberto: como definir a aproximação inicial H_0?

Infelizmente, não há uma aproximação inicial que funcione bem em todos os casos. Algumas alternativas são:

- Utilizar informação do problema. Por exemplo, H_0 pode ser definida como a inversa da aproximação por diferenças finitas da Hessiana de f em x_0.
- Utilizar a matriz identidade ou um múltiplo da matriz identidade, para refletir o escalamento das variáveis.

Método BFGS

Método BFGS: Dados um ponto inicial $x^{(0)}$, uma aproximação da inversa da Hessiana H_0 e um escalar $\epsilon > 0$.

Passo 1: Faça $k \leftarrow 0$.

Passo 2: Se $\|\nabla f(x_k)\| \leq \epsilon$, pare com x_k como solução.

Passo 3: Calcule uma direção de busca $p_k = -H_k \nabla f_k$.

Passo 4: Faça $x_{k+1} \leftarrow x_k + \alpha_k p_k$, α_k calculado com busca linear satisfazendo condições de Wolfe.

Passo 5: Faça $s_k \leftarrow x_{k+1} - x_k$ e $y_k \leftarrow \nabla f_{k+1} - \nabla f_k$.

Passo 6: Calcule H_{k+1} usando a atualização BFGS (7).

Passo 7: Faça $k \leftarrow k + 1$ e volte para o Passo 2.

Método BFGS

O custo de cada iteração do método BFGS é de $O(n^2)$ operações aritméticas.

Uma vantagem deste método em relação ao método de Newton é o uso apenas das primeiras derivadas. Quando as Hessianas não estão disponíveis ou suas fatorações têm custo computacional proibitivo, o método BFGS é uma boa alternativa ao método de Newton.

Convergência do método BFGS

Apesar do método BFGS ser muito robusto na prática, não podemos estabelecer resultados de convergência verdadeiramente globais para funções não-lineares gerais. Ou seja, não podemos provar que os iterandos gerados pelo método BFGS convergem a um ponto estacionário da função objetivo partindo de qualquer ponto inicial e qualquer aproximação inicial (razoável) da Hessiana.

De fato, não se sabe se este método possui esta propriedade.

Na análise a seguir, iremos supor que ou a função objetivo é convexa ou os iterandos gerados pelo método satisfazem algumas propriedades.

Por outro lado, sob algumas hipóteses razoáveis, há resultados conhecidos de convergência local superlinear.

Convergência do método BFGS

Teorema 1: *Seja B_0 uma matriz inicial simétrica definida positiva. Suponha que f tenha segunda derivada contínua. Seja x_0 um ponto inicial tal que o conjunto de nível $\Omega = \{x \in \boldsymbol{R}^n | f(x) \leq f(x_0)\}$ seja convexo e existam constantes m e M tais que*

$$m\|z\|^2 \leq z^T \nabla^2 f(x) z \leq M\|z\|^2$$

para todo $z \in \boldsymbol{R}^n$ e $x \in \Omega$.
Então, a sequência $\{x_k\}$ gerada pelo método BFGS converge ao minimizador x^* de f.

Convergência do método BFGS

Teorema 2: *Suponha que f tenha segunda derivada contínua. Suponha que os iterandos gerados pelo método BFGS convirjam a um minimizador x^* para o qual a Hessiana é Lipschitz contínua. Suponha também que $\sum_{k=1}^{\infty} \|x_k - x^*\| < \infty$.*
Então, a sequência $\{x_k\}$ converge para x^* superlinearmente.