FGV EMAp

João Pedro Jerônimo

Otimização para Ciência de Dados

Revisão para A2

Rio de Janeiro

2025

Conteúdo

Método do Gradiente

É o método de otimização mais clássico que existe! Vamos supor que queremos resolver o problema:

$$\min_{x \in \mathbb{R}^n} f(x) \tag{1}$$

Lembra do que vimos em cálculo? Que $\nabla f(x)$ é o vetor que aponta pra direção em que $f(x)$ aumenta? Então que tal a gente seguir na direção contrária a $f(x)$? Isso faz bastante sentido, e funciona! Mas deve ter um motivo mais matemático por trás, não é? Vamos primeiro mostrar o algoritmo:

```
1 func GradientDescent(f) {
2 |  x^(0) ∈ ℝ^n
3 |  α > 0
4 |  for t ∈ [T] do {
5 |  |  x^(t+1) = x^(t) − α∇f(x^(t))
6 |  }
7 |  return x^(T)
8 }
```

Algoritmo 1: Gradient Descent

Antes de entender um motivo mais matemático por trás do algoritmo, vamos ver algumas definições

Definição 1.1 (Funções M-Lipschitz): Dizemos que uma função $f : \mathbb{R}^n \to \mathbb{R}^m$ é M-Lipschitz quando:

$$\|f(x) - f(y)\| \leq M \ \|x - y\| \qquad \forall x, y \in \mathbb{R}^n \tag{2}$$

Definição 1.2 (Função L-Suave): Uma função $f : \mathbb{R}^n \to \mathbb{R}$ é L-suave quando seu gradiente é L-Lipschitz:

$$\|\nabla f(x) - \nabla f(y)\| \leq L \ \|x - y\| \qquad \forall x, y \in \mathbb{R}^n \tag{3}$$

Essa definição de suavidade tem uma interpretação, imagine que, se eu estou na posição x e vou pra posição y, a variação que eu vou ter na função, dentro dessa passada, não ultrapassa o quanto eu andei vezes uma constante L. Então funções muito onduladas, e com ondulações

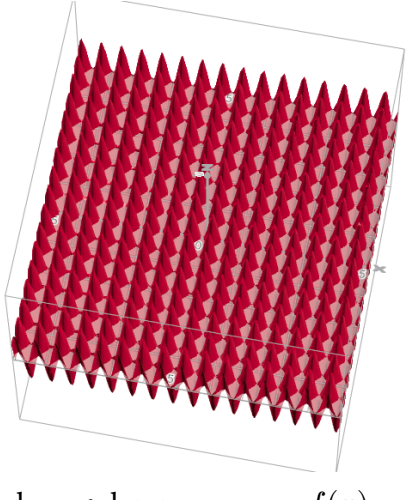

Figura 2: Função bem desregular, mas suave, $f(x) = \sin(10x) + \cos(10y)$

Teorema 1.1 (Aproximação linear de funções suaves): Seja $f : \mathbb{R}^n \to \mathbb{R}$ uma função diferenciável. Então f é L-suave se, e somente se, $\forall x, y \in \mathbb{R}^n$:

$$|f(y) - f(x) + \nabla f(x)^T (y - x)| \leq \frac{L}{2} \|y - x\|_2^2 \tag{4}$$

Usando esse teorema, a gente pode escrever isso:

$$\begin{aligned} f\big(x^{(t+1)}\big) &\leq f\big(x^{(t)}\big) + \nabla f\big(x^{(t)}\big)^T \big(x^{(t+1)} - x^{(t)}\big) + \frac{L}{2} \|x^{(t+1)} - x^{(t)}\|^2 \\ &\leq f\big(x^{(t)}\big) - \alpha \|\nabla f\big(x^{(t)}\big)\|^2 + \frac{\alpha^2 L}{2} \|\nabla f\big(x^{(t)}\big)\|^2 \\ &\leq f\big(x^{(t)}\big) - \alpha\left(1 - \frac{\alpha L}{2}\right) \|\nabla f\big(x^{(t)}\big)\|^2 \end{aligned} \tag{5}$$

E isso vale quando $\alpha \in \big(0, \frac{2}{L}\big)$, ou seja, se o ponto atual $x^{(t)}$ **NÃO É ESTACIONÁRIO**, o valor da função no próximo ponto será **menor** que o valor do ponto atual menos o tamanho do gradiente ao quadrado vezes um termo de regulação. Parece complicado, mas o que isso quer dizer? Eu vou usar esse fato para mostrar que, independente do ponto que eu iniciar o método do gradiente, eu **sempre vou encontrar um ponto mínimo local utilizando o método do gradiente**

Teorema 1.2: Suponha que $f : \mathbb{R}^n \to \mathbb{R}$ é L-suave. Tome qualquer passo:

$$\alpha = \frac{\beta}{L} \tag{6}$$

para algum $\beta \in (0, 2)$. Então:

$$\min_{t \in [T]} \|\nabla f\big(x^{(t)}\big)\|_2^2 \leq \frac{1}{T} \sum_{t=1}^{T} \|\nabla f\big(x^{(t)}\big)\|_2^2 \leq \left(\frac{2/\beta}{2 - \beta}\right) \frac{L\big(f\big(x^{(1)}\big) - f^*\big)}{T} \tag{7}$$

Demonstração: Sabemos que, dado n pontos x_i, a média $\frac{1}{n}\sum_{i=1}^{n} x_i \in [\min(x_i), \max(x_i)]$, então a desigualdade inicial já está provada. Vamos provar a segunda. Usando a equação (5), temos que:

$$\alpha\left(1 - \frac{\alpha L}{2}\right) \|\nabla f\big(x^{(t)}\big)\|^2 \leq f\big(x^{(t)}\big) - f\big(x^{(t+1)}\big) \tag{8}$$

isso para todo $t \in [T]$, então vamos somar todos os termos para obter:

$$\begin{aligned} \alpha\left(1 - \frac{\alpha L}{2}\right) \sum_{t=1}^{T} \|\nabla f\big(x^{(t)}\big)\|_2^2 &\leq \sum_{t=1}^{T} \big(f\big(x^{(t)}\big) - f\big(x^{(t+1)}\big)\big) \\ &\leq f\big(x^{(1)}\big) - f\big(x^{(T+1)}\big) \\ &= f\big(x^{(1)}\big) - f^* + f^* + f\big(x^{(T+1)}\big) \\ &\leq f\big(x^{(1)}\big) - f^* \end{aligned} \tag{9}$$

A primeria desigualdade eu fiz uma soma telescópica, depois eu somei 0 ($f^* - f^*$) e, como f^* é o valor mínimo da função, com certeza subtrair a parte que eu somei f^* vai dar um valor maior, então eu obtenho o resultado do enunciado do teorema dividindo tudo por $\alpha\left(1 - \frac{\alpha L}{2}\right)T$ $\square$

Por que esse teorema mostra que, independentemente do lugar, o algoritmo converge para um ponto estacionário? Ele ta me dizendo isso daqui:

$$\min_{t \in [T]} \|\nabla f\left(x^{(t)}\right)\|_2^2 = O\left(\frac{1}{T}\right) \tag{10}$$

Ou seja, o mínimo **converge para 0** conforme $T \to \infty$ **independentemente do ponto inicial** $\boldsymbol{x^{(0)}}$