

T319 - Introdução ao Aprendizado de Máquina: *Regressão Linear (Parte III)*



Inatel

Felipe Augusto Pereira de Figueiredo
felipe.figueiredo@inatel.br

Recapitulando

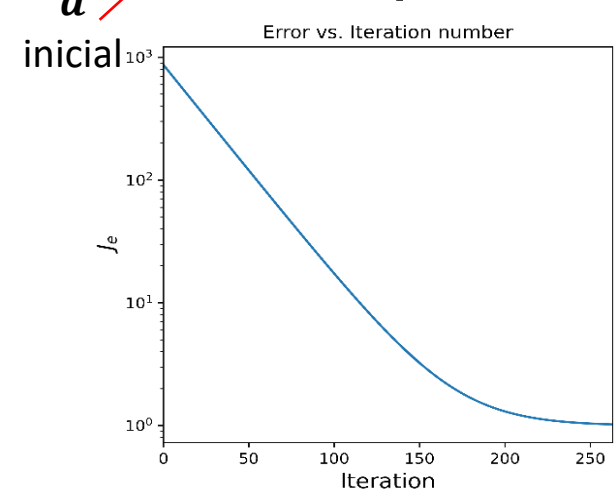
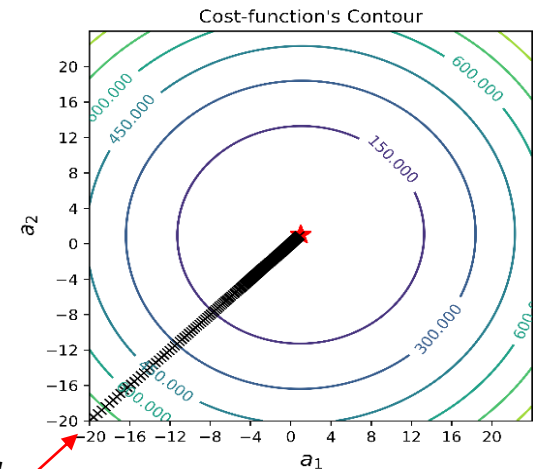
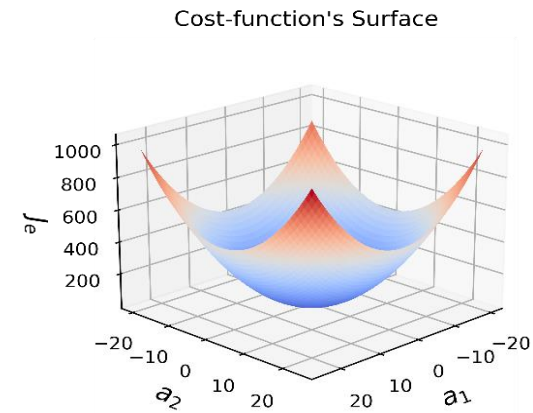
- Discutimos sobre o vetor gradiente.
- Aprendemos dois algoritmos que usam o vetor gradiente para a resolução de problemas de otimização.
- Vimos as três versões do gradiente descendente e as comparamos.
- Nesta parte, discutiremos o quão importante é o ajuste do passo de aprendizagem, α .

Escolha do Passo de Aprendizagem

- Conforme nós já aprendemos, enquanto a **direção** para o ponto de mínimo é determinada pelo **vetor gradiente** da **função de erro**, o **passo de aprendizagem** determina o quão grande esse passo é dado naquela direção.

$$\mathbf{a} \leftarrow \mathbf{a} - \alpha \frac{\partial J_e(\mathbf{a})}{\partial \mathbf{a}}$$

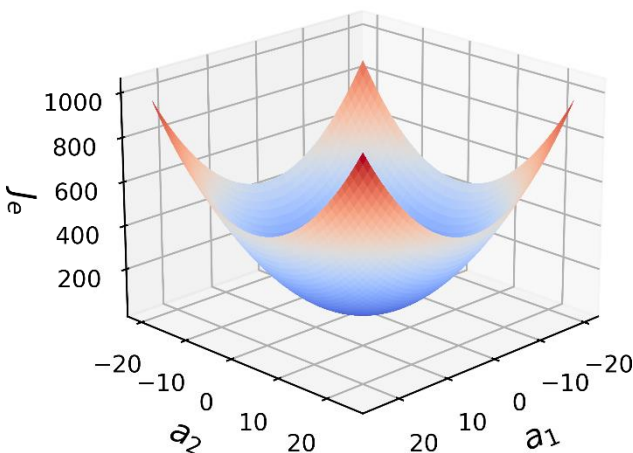
- Portanto, a **escolha do passo de aprendizagem (hiperparâmetro) é muito importante**:
 - Caso ele seja muito pequeno, a convergência do algoritmo levará muito tempo.
 - Exemplo**: com $\alpha = 0.01$, o algoritmo atinge o valor ótimo após mais de 250 épocas.
 - Passos muito curtos, fazem com que o algoritmo caminhe vagarosamente em direção ao **mínimo global** da **função de erro**.



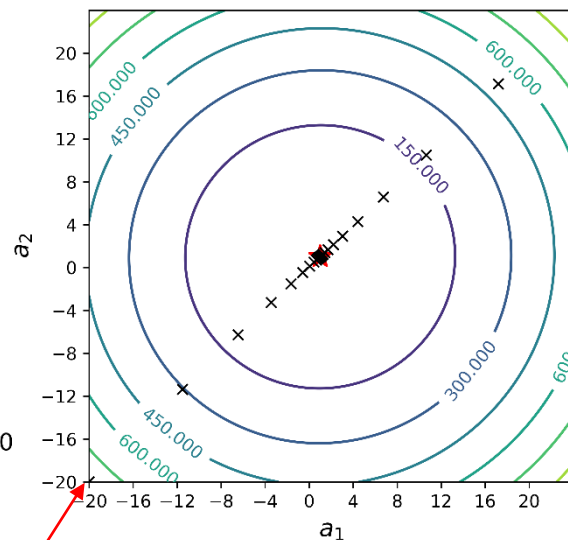
Escolha do Passo de Aprendizagem

- Caso o **passo de aprendizagem** seja muito grande, o algoritmo pode nunca convergir.
- Se α for grande, mas não tão grande assim, o algoritmo fica “pulando” ou “oscilando” de um lado para o outro da superfície até que converge, por sorte (veja exemplo #1).
- Em outros casos, quando α é bem grande, a cada iteração o algoritmo “pula” para um valor mais alto que antes, e assim, divergindo (veja exemplo #2).

Cost-function's Surface



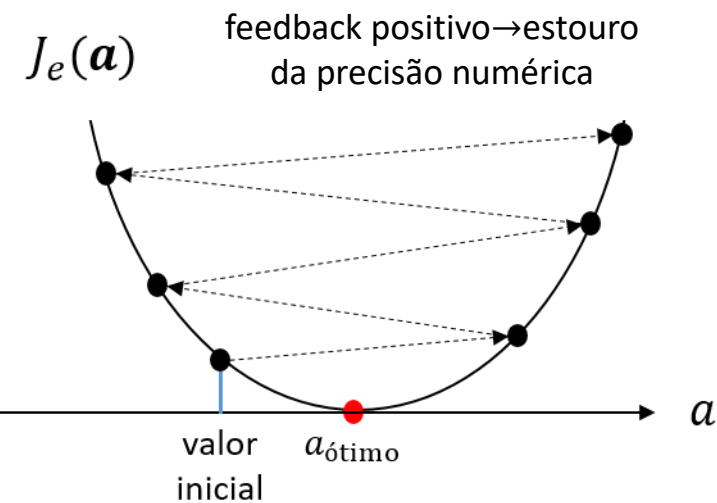
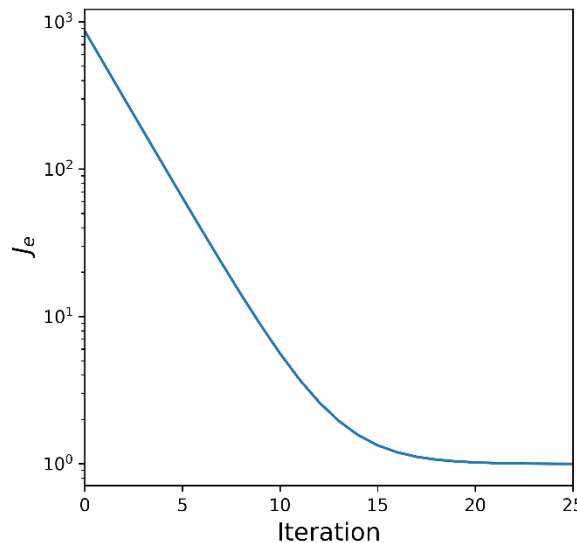
Cost-function's Contour



a inicial

Exemplo #1

Error vs. Iteration number



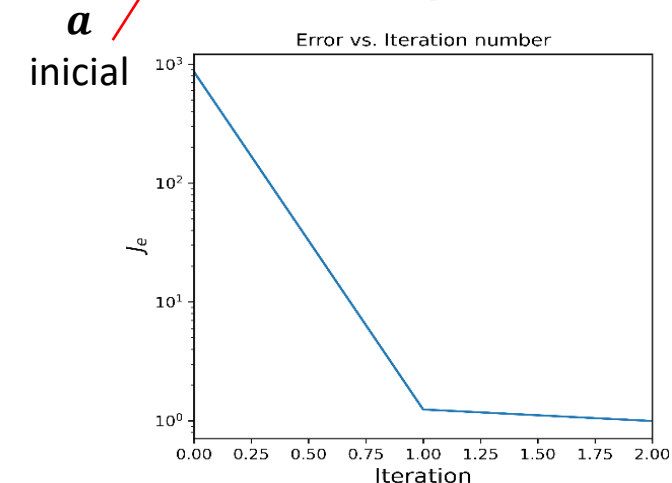
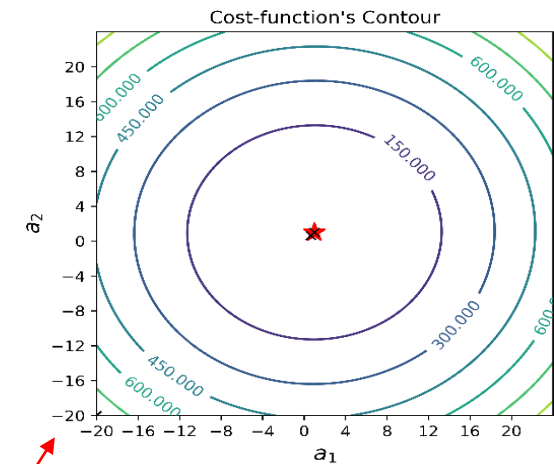
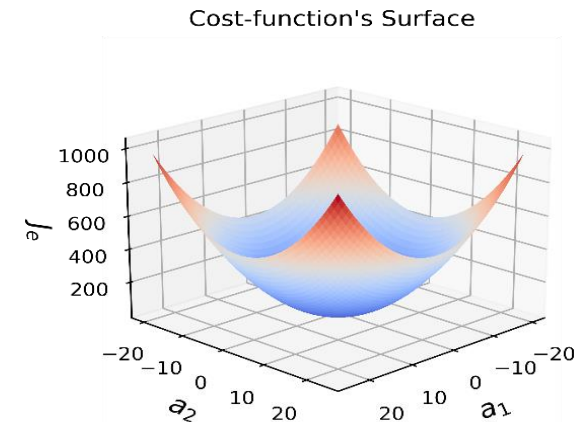
Exemplo #2

Escolha do Passo de Aprendizagem

- Portanto, o valor ***passo de aprendizagem*** deve ser ***explorado*** para se encontrar um ***valor ideal*** que acelere a ***descida do gradiente*** de forma ***estável*** (ou seja, acelere a convergência).
 - O exemplo ao lado, converge para o ***mínimo global*** em apenas 2 iterações.
- Portanto, a escolha do ***passo de aprendizagem*** pode ser bastante demorada.
- Uma regra empírica para exploração do passo de aprendizagem é usar a seguinte sequência (***ajuste manual***):

..., 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1.0, ...

3 × ≈ 3 × 3 × ≈ 3 × 3 × ≈ 3 ×



Como depurar o algoritmo do GD?

- Uma maneira de se **depurar** (principalmente quando não é possível se plotar o gráfico da superfície de contorno) o algoritmo do **gradiente descendente** é plotar o gráfico do erro (EQM) em função do número de iterações ou épocas.
 - Figura A \Rightarrow Passo ideal: converge rapidamente
 - Erro diminui rapidamente nas primeiras épocas e depois diminui quase que a uma taxa constante.
 - Convergência pode ser declarada quando o erro entre duas iterações subsequentes for menor do que um limiar pré-definido (e.g., $1e-3$).
 - Figura B \Rightarrow Passo pequeno demais: convergência lenta.
 - Figuras C e D \Rightarrow Passo grande demais: divergência.

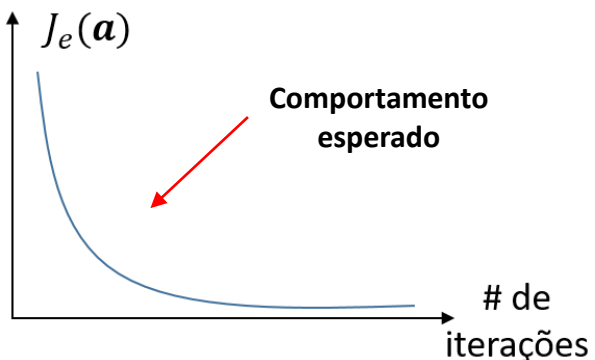


Figura A

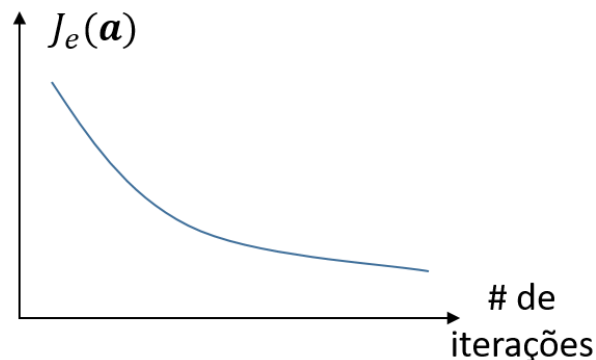


Figura B

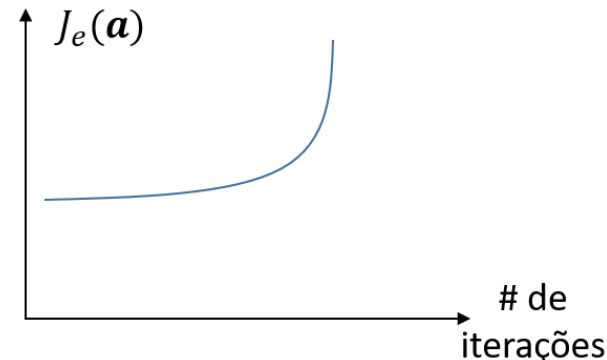


Figura C

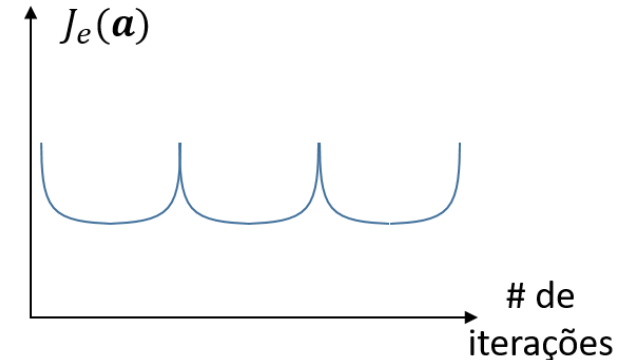
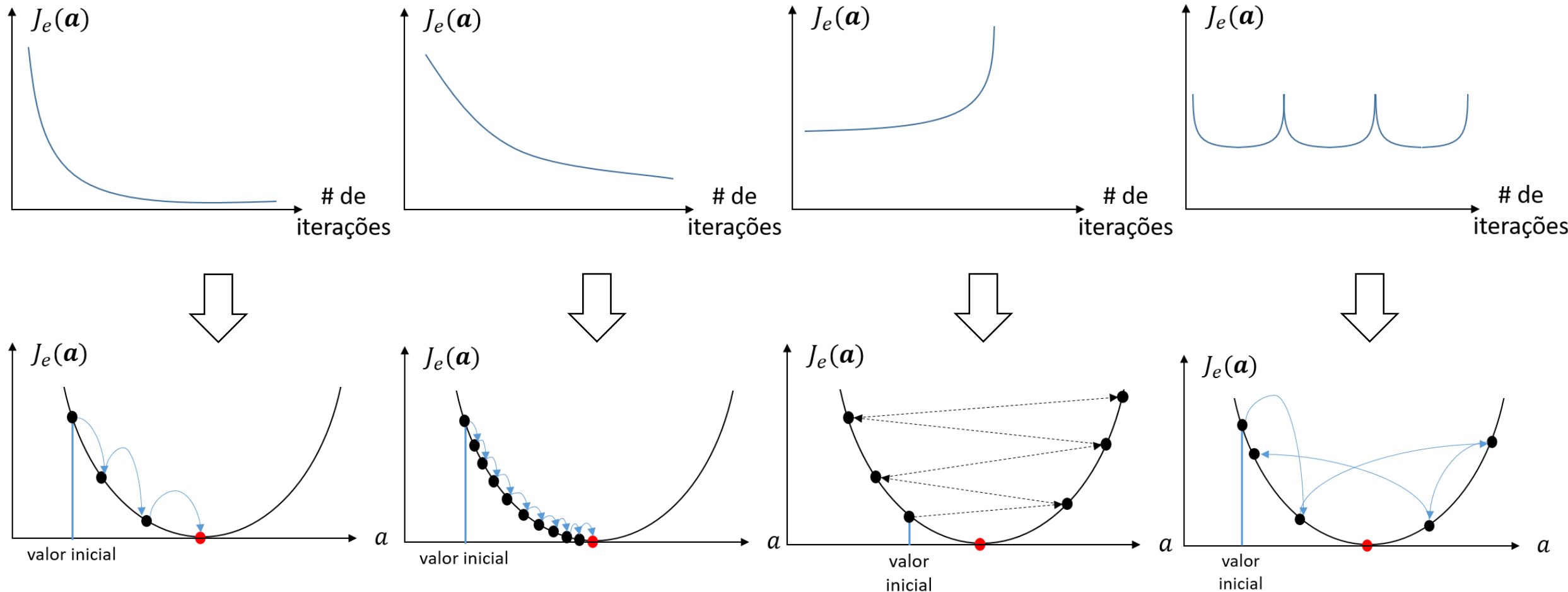


Figura D

Como depurar o algoritmo do GD?



Como configurar o passo de aprendizagem?

Além do ***ajuste manual*** (escolha de α por tentativa e erro), podemos também usar as seguintes abordagens para configurar α :

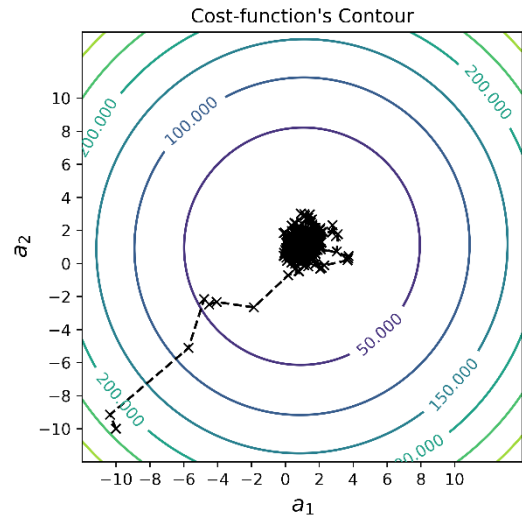
- **Redução programada:** redução do passo de aprendizagem ao longo do processo de treinamento.
 - A forma mais simples é diminuir o passo de aprendizagem linearmente de um valor inicial grande até um valor pequeno.
 - Abordagem muito usada com GD estocástico e mini-batch para garantir a convergência para o ponto de mínimo.
- **Variação adaptativa:** α é adaptativamente ajustado de acordo com a performance do modelo além disso, pode ter passos diferentes para cada peso do modelo e os atualiza independentemente.
 - **Vantagem:** na maioria dos casos, não é necessário se ajustar manualmente nenhum ***hiperparâmetro*** como no caso dos esquemas de redução programada.

Redução Programada do Passo de Aprendizagem

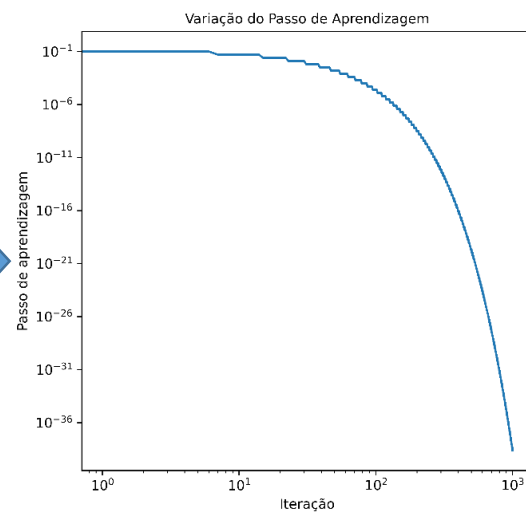
- Os três tipos mais comuns de implementação da **redução programada** do passo de aprendizagem são:
 - **Decaimento gradual**: também conhecido como **decaimento por etapas** ou **por degraus**. Ele reduz a taxa de aprendizagem de um fator α inicial, α_0 , a cada número pré-definido de iterações ou épocas, β . Um valor típico para reduzir a taxa de aprendizado é de $\alpha = 0.5$ a cada número pré-definido de épocas.
 - **Decaimento exponencial**: tem a forma matemática $\alpha = \alpha_0 e^{-kt}$, onde α_0 e k são hiperparâmetros e t é o número da iteração atual (pode-se se usar também o número de épocas).
 - **Decaimento temporal**: tem a forma matemática $\alpha = \alpha_0 / (1+kt)$ onde α_0 e k são hiperparâmetros e t é o número da iteração.
- Na prática, o **decaimento gradual** é o mais utilizado entre os 3, pois seus **hiperparâmetros** (a fração de decaimento e os intervalos de tempo para redução) são mais interpretáveis do que o hiperparâmetro k , que dita a taxa de decaimento do passo de aprendizagem.

Exemplo: GDE com Redução Programada de α

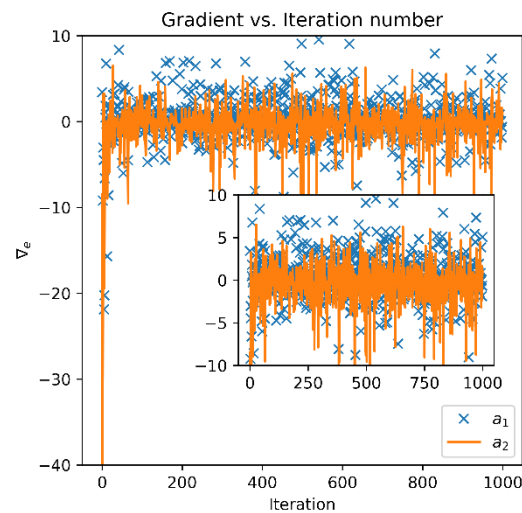
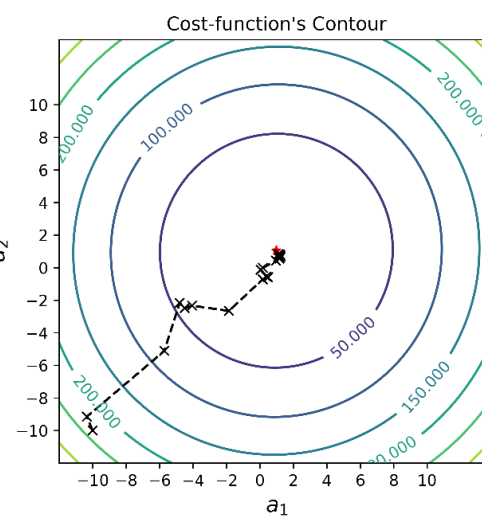
GDE sem redução programada



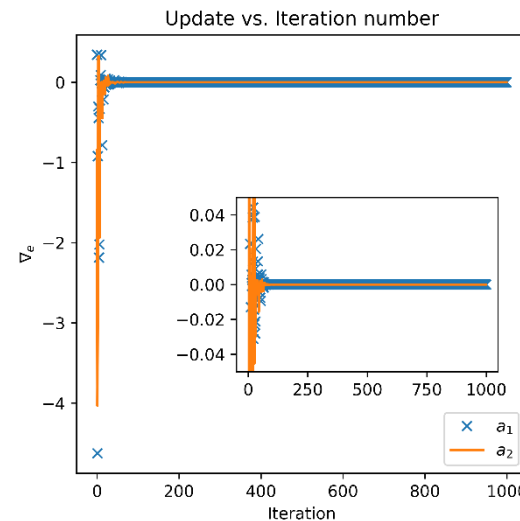
Redução programada



GDE com redução programada



```
# Learning schedule function.  
def stepDecay(alpha_init, t, epochs_drop=8):  
    drop = 0.5  
    epochs_drop = 4.0  
    alpha = alpha_init*pow(drop, floor((1+t)/epochs_drop))  
    return alpha
```

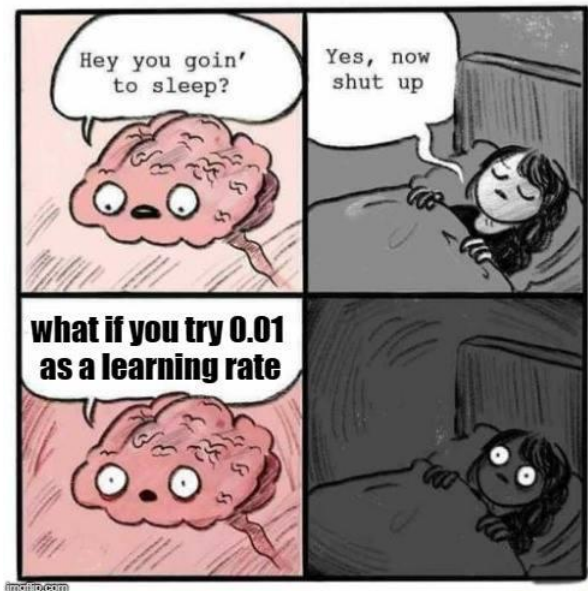


- Exemplo usando GDE com **decaimento gradual**.
- O caminho com **decaimento gradual** também não é regular para o ponto de mínimo.
- Apresenta algumas mudanças de direção ao longo do caminho.
- Porém, a oscilação em torno do mínimo é bastante minimizada devido à **diminuição gradual** de α .
- O passo inicial com valor grande e diminui ao longo das iterações, permitindo que o algoritmo se estabilize próximo ao ponto de mínimo global.
- Conseguimos visualizar melhor o efeito da redução de α nas figuras que mostram o gradiente.

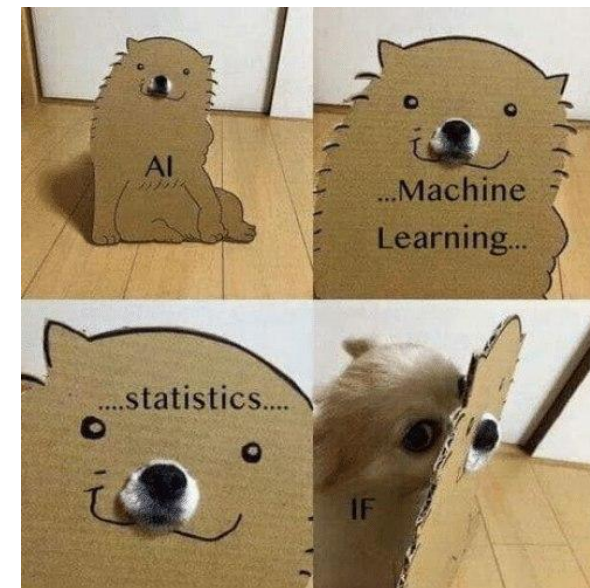
Tarefas

- **Quiz:** “*T319 - Quiz - Regressão: Parte III*” que se encontra no MS Teams.
- **Exercício Prático:** [Laboratório #4](#).
 - Pode ser baixado do MS Teams ou do GitHub.
 - Pode ser respondido através do link acima (na nuvem) ou localmente.
 - [Instruções para resolução e entrega dos laboratórios](#).
 - **Laboratórios podem ser feitos em grupo, mas as entregas devem ser individuais.**

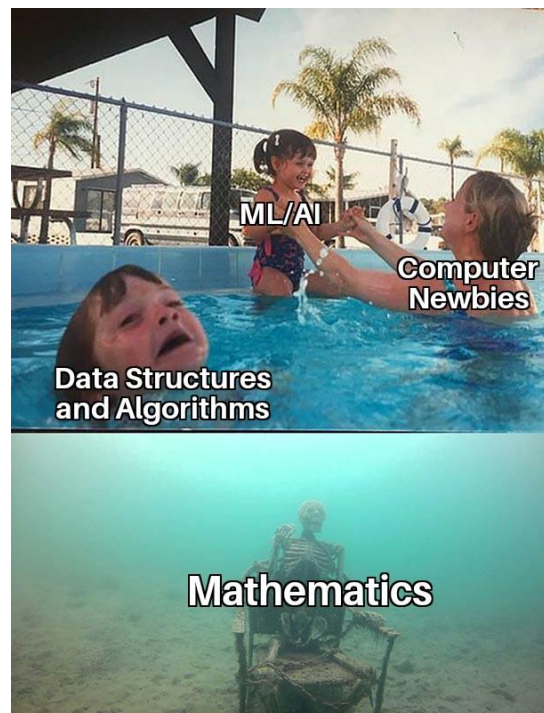
Obrigado!



When someone asks why you never stops talking about machine learning

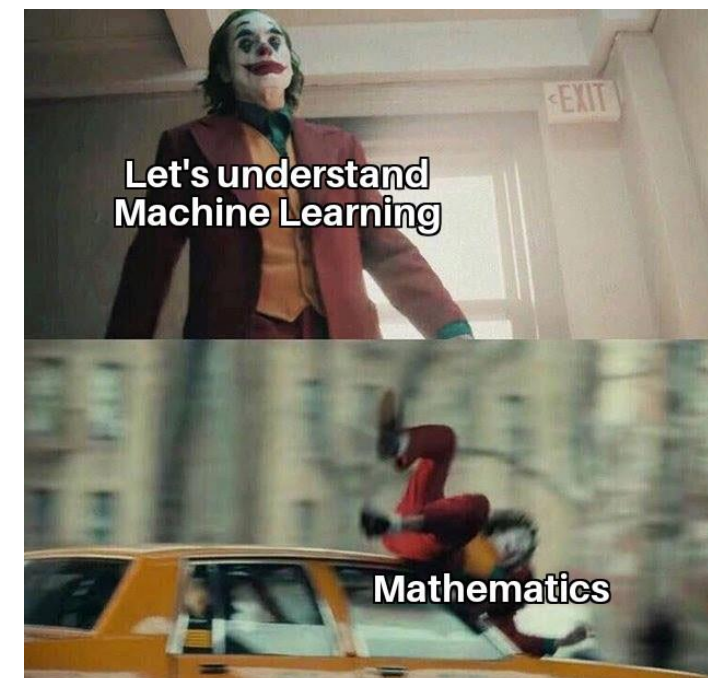


IF IF IF IF IF IF IF IF IF IF WE!

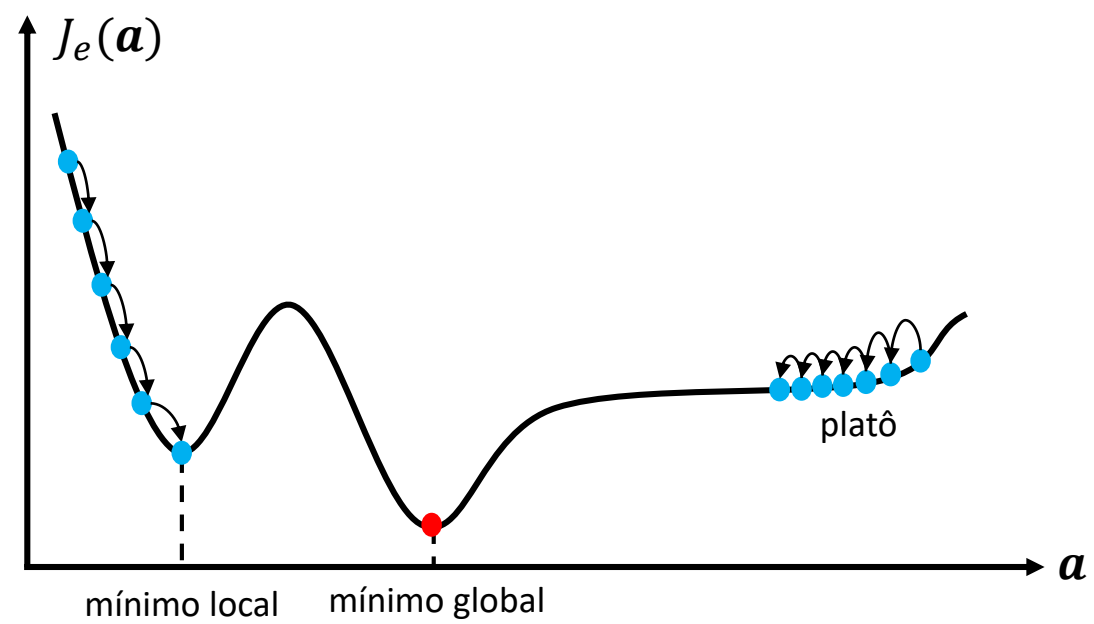


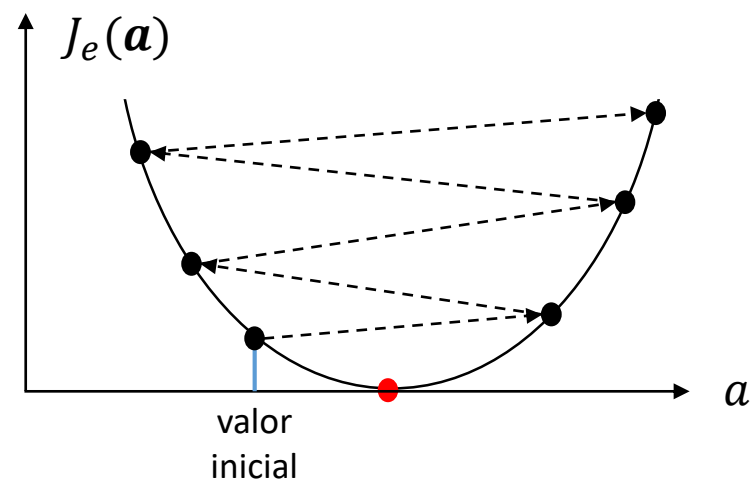
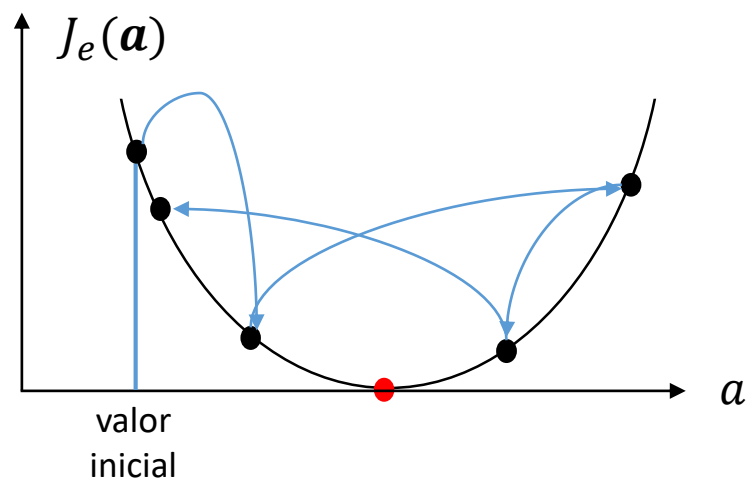
Albert Einstein: Insanity Is Doing the Same Thing Over and Over Again and Expecting Different Results

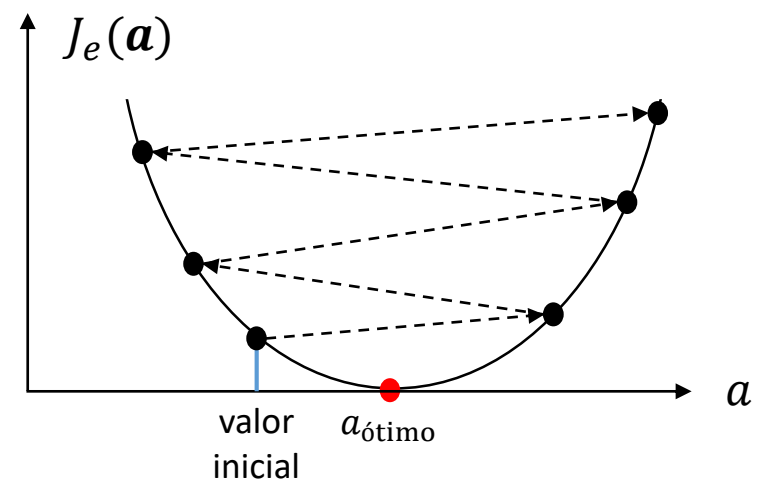
Machine learning:

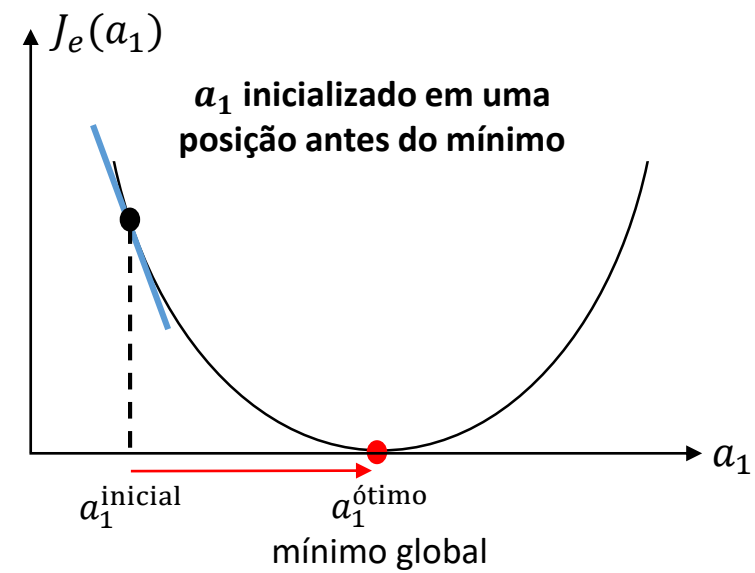


FIGURAS

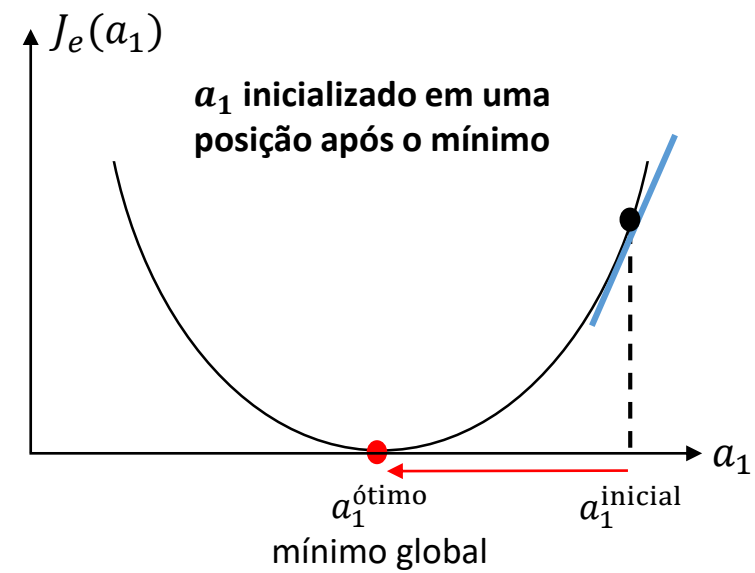




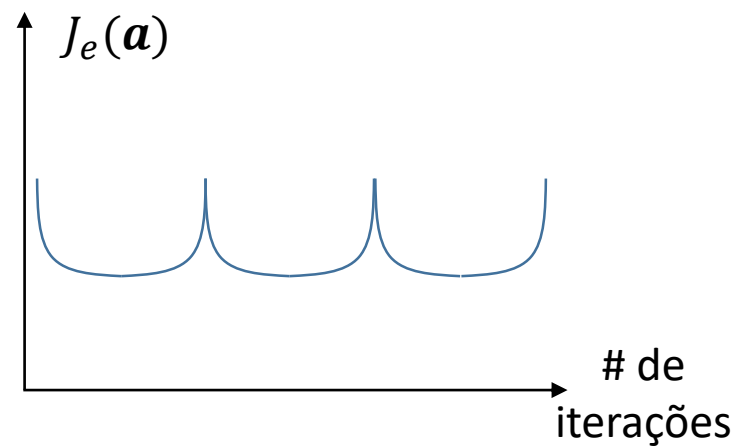
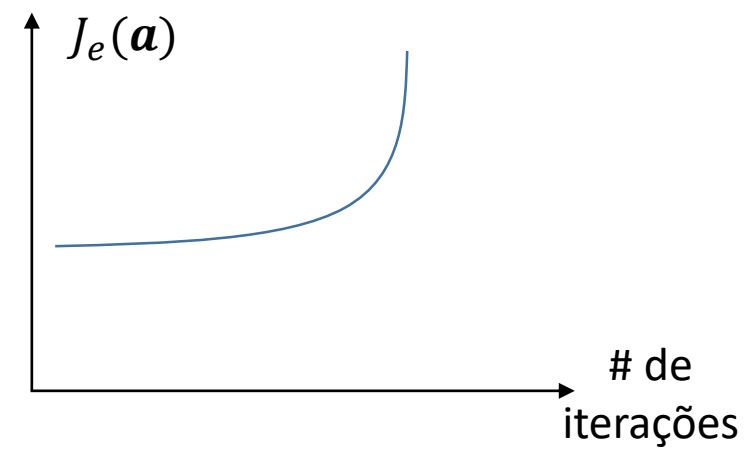
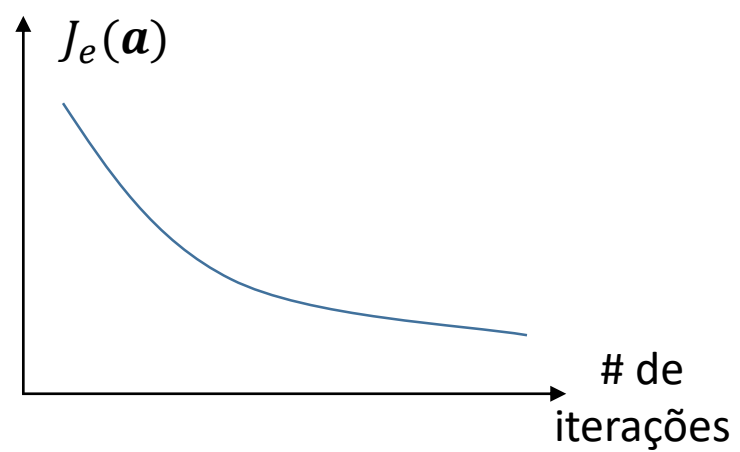
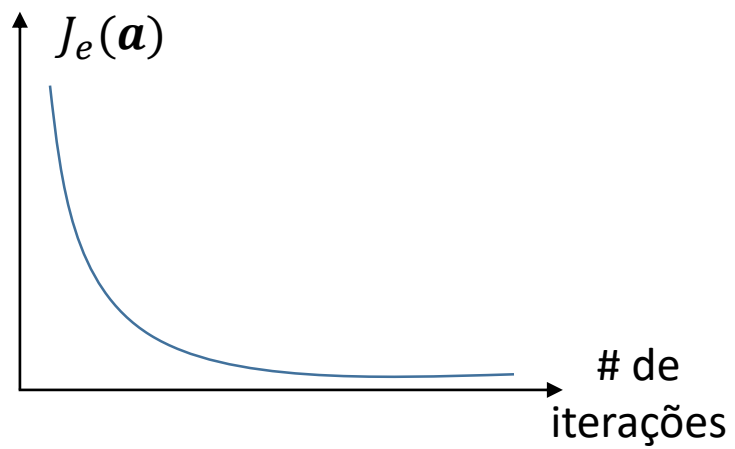


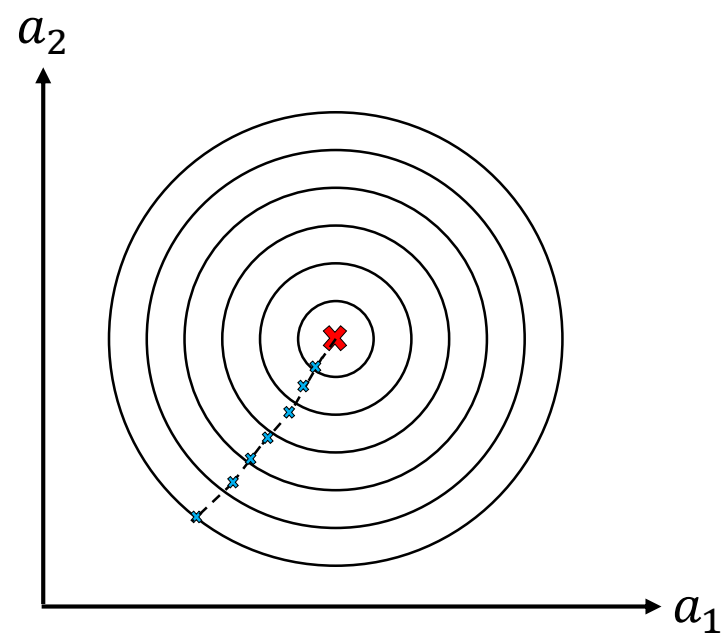
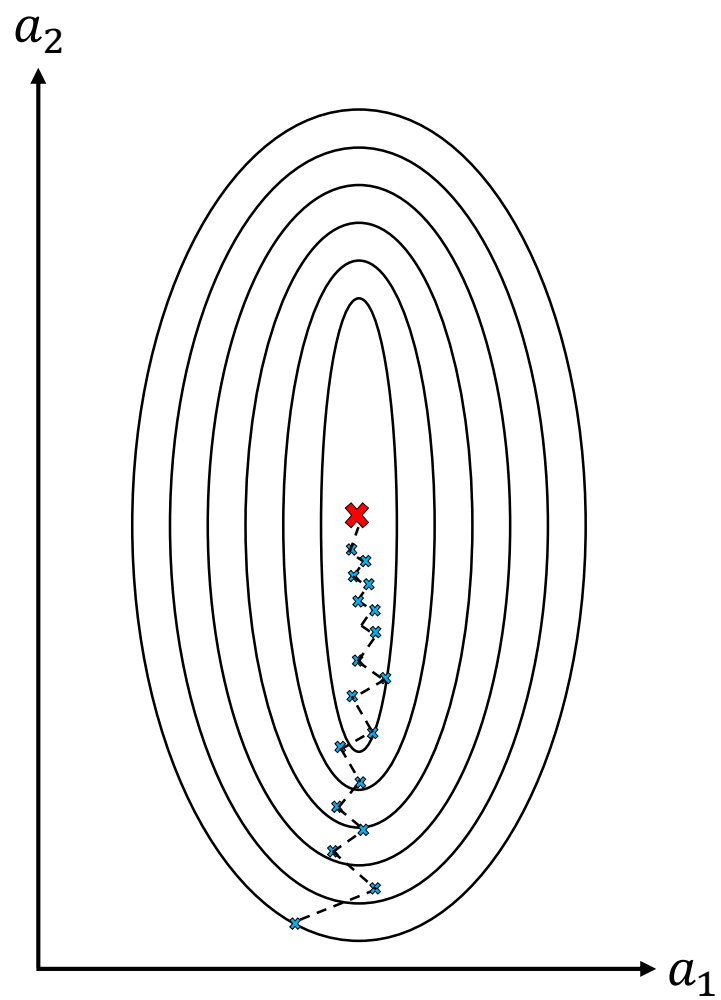


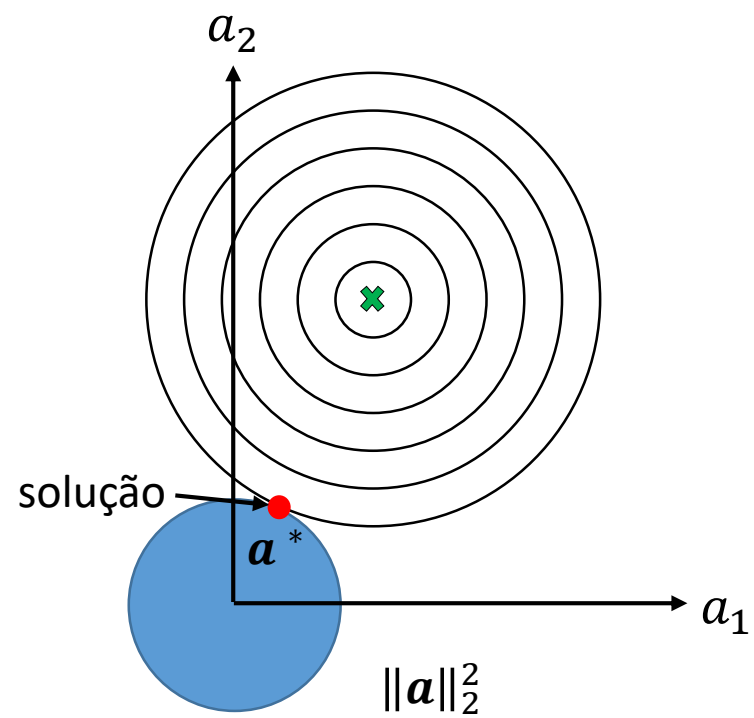
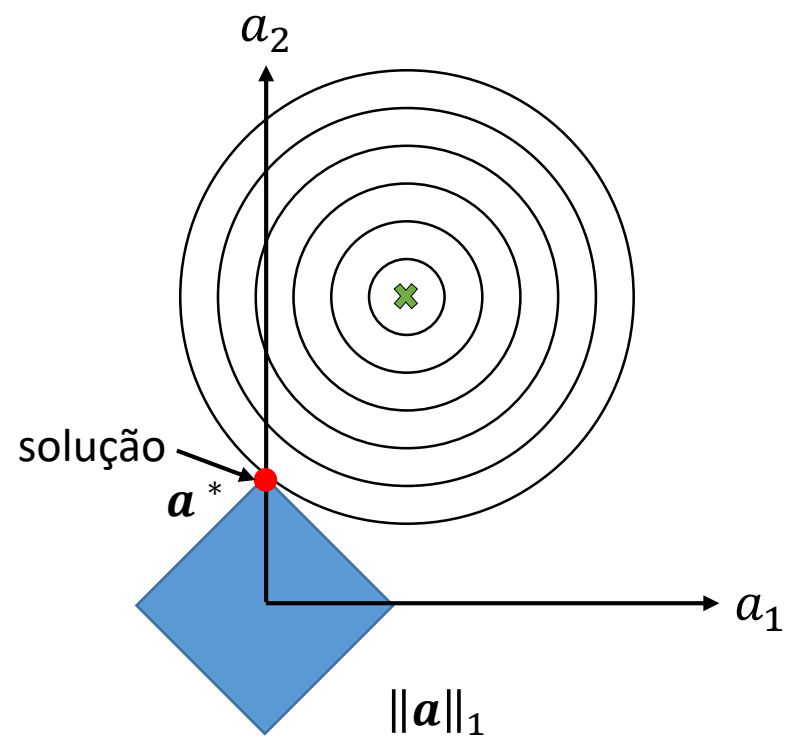
gradiente negativo: $a_1 = a_1^{\text{inicial}} + \alpha \nabla J_e(a_1)$
 a_1 aumenta e se aproxima do mínimo



gradiente positivo: $a_1 = a_1^{\text{inicial}} - \alpha \nabla J_e(a_1)$
 a_1 diminuiu e se aproxima do mínimo







Gradiente Descendente Estocástico

