Ficha - Métodos de Otimização Sem Restrições

Métodos de descida acelerada: Momentum, NAG, Adagrad, RMSProp, Adam

1. Considere o problema (Rosenbrock)

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

com $a = 1$ e $b = 100$. (Solução: minimizante global $w^* = (a, a^2)$)
Considere para ponto inicial $w^{(0)} = (0, 0)^T$, para critério de paragem considere a condição $\|\nabla F(w)\| \leq 10^{-6}$, e faça $\eta =$ constante. Em cada iteração, grave para um ficheiro a informação:

k	${w^{(k)}}^T$	$\nabla F(w^{(k)})$	$\|\nabla F(w^{(k)}\|_2$	$F(w(k))$
...	...	...	...	...

   (a) Resolva o problema usando o método NAG.
   (b) Resolva o problema usando o método RMSProp.
   (c) Resolva o problema usando o método Adam.
   (d) Faça uma análise comparativa dos resultados obtidos.

2. Considere novamente o problema de *machine learning*. Dado o data set $D = (x^n, y^n)_{n=1}^N$ pretende-se determinar os coeficientes de um polinómio de grau I

$$\begin{aligned}\phi(w; x) = &\quad w_0 + w_1 x + w_2 x^2 + \cdots + w_{\mathrm{I}} x^{\mathrm{I}} \\ = &\quad w^T p(x)\end{aligned}$$

onde $p(x) = (1, x, x^2, \ldots, x^{\mathrm{I}})^T$ que melhor ajustam o polinómio aos dados D no sentido da minimização da função MSE (*Mean Squared error*):

$$MSE(w; D) = \frac{1}{N}\sum_{n=1}^{N}(\phi(w; x^n) - y^n)^2.$$

Nota: O gradiente da função $MSE(w; D)$ é dado por $\nabla MSE(w; D) = \dfrac{2}{N}\sum_{n=1}^{N}(\phi(w; x^n) - y^n)p(x^n)$.

   (a) Resolver o Problema apresentado com o data set data1.csv (N=100).
   Dividir o data set em duas partes: 80% para treino D_t e 20% para validação D_v. Esta seleção deverá ser aleatória. Resolva o problema com o método Adam (batch) e como aproximações iniciais aos parâmetros considere $w^{(0)} = (0, \ldots, 0)$, e para critério de paragem considere

   $$\|\nabla MSE(w)\| \leq 10^{-4} \ \text{ e } \ k \leq 10N_t.$$

   Resolva o problema considerando polinómios de grau $\mathrm{I} = 2, \ldots, 7$. Para cada um dos polinómios calcule: w^*, o erro de treino (*in-sample error*) $MSE(w^*; D_t)$, e o erro de validação (*out-sample error*) $MSE(w^*; D_v)$. Fazer o gráfico dos erros e indicar qual o grau I que fornece a melhor aproximação.
   (b) Faça o exercício 2(a) mas implemente a versão Adam estocástico.
   (c) Faça o exercício 2(a) mas implemente a versão Adam estocástico mini-batch. Considere para mini-batch 5% dos dados do D_t. Esta seleção deverá ser aleatória em cada iteração.