2019年05月18日

CSC421

2 a) Adam

$v_t \leftarrow B_2 v_{t-1} + (1-B_2) g_t^2$

$\theta_t \leftarrow \theta_{t-1} - \alpha_A m_t / (\sqrt{v_t} + \epsilon_A)$

$\hookrightarrow m_t \leftarrow B_1 m_{t-1} + (1-B_1) g_t$

RMSProp

$v_t \leftarrow \gamma v_{t-1} + (1-\gamma) g_t^2$

$\theta_t \leftarrow \theta_{t-1} - \alpha_R g_t / (\sqrt{v_t} + \epsilon_R)$

$\therefore$ Set $B_2 = \gamma$.

$B_1 = 0 \Rightarrow m_t = g_t$

$\alpha_A = \alpha_R$

$\epsilon_A = \epsilon_R$

b) SGD w/momentum

$p_t \leftarrow \mu p_{t-1} - (1-\mu) \nabla F(\theta_{t-1})$

$\theta_t \leftarrow \theta_{t-1} + \alpha_S p_t$

Want: $\alpha_S p_t \approx \alpha_A m_t / (\sqrt{v_t} + \epsilon_A)$

Setting $B_2 = 1 \Rightarrow v_t \leftarrow v_{t-1} = 0 \quad \forall t$

Setting $\epsilon_A = 1$, $\alpha_A = -\alpha_S$

$\alpha_S p_t = -\alpha_S m_t$

Want: $p_t = -m_t$.

Setting $B_1 = \mu \Rightarrow m_t = \mu m_{t-1} + (1-\mu) \nabla F(\theta_{t-1})$

$\rightarrow$ update eq^n's almost the same. m_t sums the gradient, p_t sums the neg.

Noting that $p_0 = m_0 = 0$, then:

$p_1 = \mu p_0 - (1-\mu) \nabla F(\theta_0)$

$p_2 = -\mu(1-\mu) \nabla F(\theta_0) - (1-\mu) \nabla F(\theta_1)$

$p_3 = -\mu^2(1-\mu) \nabla F(\theta_0) - \mu(1-\mu) \nabla F(\theta_1) - (1-\mu) \nabla F(\theta_2)$

$p_T = \sum_{t=1}^{T} - \mu^{T-t}(1-\mu) \nabla F(\theta_{t-1})$

similarily: $m_T = \Sigma + \mu^{T-t}(1-\mu) \nabla F(\theta_{t-1})$

as required. $p_t = -m_t$

c) Suppose $\tilde{F} = C \cdot F$

$\tilde{g}_t \leftarrow \nabla C \cdot F = C \cdot \nabla F = C \cdot g_t$

$\tilde{m}_t \leftarrow B_1 \tilde{m}_{t-1} + (1-B_1) C g_t$

$\tilde{V}_t \leftarrow B_2 \tilde{V}_{t-1} + (1+B_2) C^2 g_t^2$

$\tilde{\Theta}_t \leftarrow \tilde{\Theta}_{t-1} - \alpha_A \tilde{m}_t / \sqrt{\tilde{V}_t}$

Recall: $m_T = \sum_{t=1}^{T} B_1^{T-t} (1-B_1) \nabla F(\Theta_{t-1})$

Clearly $\tilde{m}_T = C m_T$ since $\tilde{\nabla F} = C \nabla F$.

By a very similar argument (e.g. let $\hat{g}_t = g_t^2$)

$V_T = \sum B_2^{T-t} (1-B_2) (\nabla F(\Theta_{t-1}))^2$

Again, noting $\nabla \tilde{F} = C \nabla F$, $\tilde{V}_T = C^2 V_T$

Then: $\tilde{\Theta}_t \leftarrow \tilde{\Theta}_{t-1} - \alpha_A C m_t / \sqrt{C^2 V_T}$

$\tilde{\Theta}_t \leftarrow \tilde{\Theta}_{t-1} - \alpha_A m_t / \sqrt{V_T}$

Assuming $\tilde{\Theta}_{t-1} = \Theta_{t-1}$, this is the same recurrence.

Base case: $\Theta_0 = 0 = \tilde{\Theta}_0$

$\therefore$ If $\epsilon_A = 0$, then the trejectory is invariant to the scale of the loss fct.