22/9/23

Deep Learning

- Recap: SGD, Momentum, Nesterov Momentum ✓
- Adaptive Learning Rate Algorithms
  - Adaptive Gradient (AdaGrad) ✓
  - RMS Prop ✓
  - Adaptive Moment (Adam) ✓

Reference: Chapter 8 in DL book by Goodfellow et al.

- Recap - SGD: $R_{SGD}(\theta) = \sum_{i=1}^{m} d(y^{(i)}, \hat{y}^{(i)})$; m samples out of n training samples
  - Momentum: $v^{(t)} = \alpha v^{(t-1)} - \eta^{(t)} \nabla_{\theta^{(t-1)}} R_{SGD}(\theta)$ $(0 < \alpha < 1)$
    $\theta^{(t)} = \theta^{(t-1)} + v^{(t)}$
  - Nesterov momentum: $v^{(t)} = \alpha v^{(t-1)} - \eta^{(t)} \nabla_{\theta^{(t-1)}} R_{SGD}(\theta^{(t-1)} + \alpha v^{(t-1)})$
    $\theta^{(t)} = \theta^{(t-1)} + v^{(t)}$
- Adaptive Learning Rate Algorithms:
  - Adaptive Gradient (AdaGrad): Smaller steps along steeper directions & bigger steps along flatter directions
    - $h^{(0)} = 0$
    - $h^{(t)} = h^{(t-1)} + \nabla_{\theta^{(t-1)}} \odot \nabla_{\theta^{(t-1)}}$ → cumulative gradient square; $\odot$: element-wise product
    - $\Delta\theta^{(t)} = \left(\frac{-\eta^{(t)}}{\delta + \sqrt{h^{(t)}}} \right) \odot \nabla_{\theta^{(t-1)}}$ → adaptive learning rate; δ: small stabilizing constant; Element-wise division
    - $\theta^{(t)} = \theta^{(t-1)} + \Delta\theta^{(t)}$

  Drawback: Since $h^{(t)}$ is cumulative, older noisy gradients could adversely impact the learning rate

  - RMS Prop: To address this issue, RMS Prop proposes a weighting strategy
    - $h^{(0)} = 0$; $0 \leq \rho < 1$
    - $h^{(t)} = \rho \cdot h^{(t-1)} + (1-\rho) \cdot \nabla_{\theta^{(t-1)}} \odot \nabla_{\theta^{(t-1)}}$ ← damps older gradients more.
    - $\Delta\theta^{(t)} = \frac{-\eta^{(t)}}{\sqrt{\delta + h^{(t)}}} \odot \nabla_{\theta^{(t-1)}}$
    - $\theta^{(t)} = \theta^{(t-1)} + \Delta\theta^{(t)}$

- Adam: Combines momentum idea with RMSProp to further improve performance

  - $g^{(0)} = h^{(0)} = 0$; $\rho_1, \rho_2 \in [0, 1)$
  - $g^{(r)} = \rho_1 g^{(r-1)} + (1-\rho_1) \cdot \nabla_{\theta^{(r-1)}}$ (biased estimate of "first moment")
  - $h^{(r)} = \rho_2 h^{(r-1)} + (1-\rho_2) \cdot \nabla_{\theta^{(r-1)}} \odot \nabla_{\theta^{(r-1)}}$ (biased "second moment")
  - $\hat{g}^{(r)} = \dfrac{g^{(r)}}{(1-\rho_1^r)}$ (bias correction) → ρ_1 raised to the power r
  - $\hat{h}^{(r)} = \dfrac{h^{(r)}}{(1-\rho_2^r)}$ → ρ_2 raised to the power r
  - $\Delta\theta^{(r)} = \left(\dfrac{-\eta^{(r)}}{\delta + \sqrt{\hat{h}^{(r)}}} \right) \odot \hat{g}^{(r)}$ → adaptive learning rate
  - $\theta^{(r)} = \theta^{(r-1)} + \Delta\theta^{(r)}$

- Second-order Taylor expansion of $R(\theta)$ about $(\theta + \Delta\theta)$ for univariate case

$$R(\theta+\Delta\theta) = R(\theta) + \Delta\theta \cdot R'(\theta) + \frac{(\Delta\theta)^2}{2} R''(\theta) + \dots$$