Lecture 14. Advanced opt. Algos. II.

RMSprop

$$w^{i+1} = w^i + \Delta w^i$$

$$\Delta w^i = \frac{-\varepsilon}{\sqrt{r^i + \delta}} (\nabla_{w^i} L)$$

$$r^i = \rho \cdot r^{i-1} + (1-\rho)(\nabla_{w^i} L)^2 \quad 0 < \rho < 1$$

δ: small positive constant.

i is large $\rightarrow$ $\rho^{i-1} \rightarrow 0$

$$r^i = \rho^{i-1} r^1 + \rho^{i-2} \cdot r^2$$

1. RMSprop with Nesterov momentum.

$$\Delta w^i = \alpha \cdot \Delta w^{i-1} + \varepsilon/\sqrt{r^i}(-\nabla_{w^i} L)$$

$$r^i = \rho \cdot r^{i-1} + (1-\rho)(\nabla L)^2 .$$

2. Adaptive moments (Adam, Kingma and Ba, 2014)

ideal is to combine RMSprop + Momentum.

$$\Delta w^i = \frac{-\varepsilon}{\sqrt{\delta + \hat{r}^i}} \cdot \hat{s}^i \qquad 0 < \rho_1 < 1,\ 0 < \rho_2 < 1$$

$r^i = \rho_2 \cdot r^{i-1} + (1-\rho_2)\cdot (\nabla_w L)^2$: 2nd moment variable.

$s^i = \rho_1 \cdot s^{i-1} + (1-\rho_1)\cdot (\nabla_w L)$: 1st moment variable.

$\hat{r}^i = \frac{r^i}{1-(\rho_2)^i}$ → power. $\quad \hat{s}^i = \frac{s^i}{1-(\rho_1)^i}$

$r^0 = s^0 = 0$

$i \to$ large $\to +\infty \to (\rho_2)^i = 0,\ (\rho_1)^i = 0$

$\hat{r}^i = r^i \quad \hat{s}^i = s^i$

$\hat{r}^i$ & $\hat{s}^i$ introduced to correct bias at the begining of training with small initial w.

3. $w^i = w^{i-1} + \Delta w^i$ → combined step size

Algorithm	Δw^i	pros	Cons
SGD	$\Delta w^i = -\varepsilon \cdot (\nabla L)$	Computational efficient	fixed ε.
SGD w/ Momentum	$\Delta w^i = \partial \cdot \Delta w^{i-1} + (1-\partial) \cdot (\nabla_{w^i} L)$	accelerate training	need hyperparameter ∂.
AdaGrad	$\Delta w^i = \frac{-\varepsilon}{\delta + \sqrt{r_i}} (\nabla_{w^i} L)$ $r^i = r^{i-1} + (\nabla L)^2$	No hyperp. Rescale ε adapt.	excessively decrease ε.
RMS prop	$\Delta w^i = \frac{-\varepsilon}{\sqrt{\delta + r_i}} (\nabla L)$ $r^i = \rho \cdot r^{i-1} + (1-\rho)(\nabla L)^2$	Discard extremy past gradients ↓ faster.	one additional ρ.
RMSprop with Nestrov M.	$\Delta w^i = \partial \Delta w^{i-1} + \frac{\varepsilon}{\sqrt{r_i}} (-\nabla L)$		∂. ρ.
Adam.	$\Delta w^i = \frac{-\varepsilon}{\sqrt{\delta + \hat{r}_i}} \cdot \hat{s}^i$ $s^i = \rho_1 \cdot s^{i-1} + (1-\rho_1)(\nabla L)$ $r^i = \rho_2 \cdot r^{i-1} + (1-\rho_2)(\nabla L)^2$	momentum bias correction	ρ_1, ρ_2