CSC2516 - Homework II

Sergio E. Betancourt (998548585)

2019-02-11

Part 2

(a) Goal: Specify hyperparameters $(\alpha_A, \beta_1, \beta_2, \epsilon_A)$ that make Adam equivalent to RMSprop with $(\alpha_R, \gamma, \epsilon_R)$.

Note that RMSprop divides the learning rate by a weighted average of the squared gradient plus a dampener. To adapt Adam to perform RMSprop, we simply let the Adam hyperparameters be:

$$\beta_{\mathbf{1}} = 0,\ \beta_{\mathbf{2}} = \gamma,\ \alpha_{\mathbf{A}} = \alpha_{\mathbf{R}}, \text{ and } \epsilon_{\mathbf{A}} = \epsilon_{\mathbf{R}}.$$

The above means that we eliminate the first time scale and let $\mathbf{m}_t \leftarrow \mathbf{g}_t$.

(b) Goal: Specify hyperparameters $(\alpha_A, \beta_1, \beta_2, \epsilon_A)$ that make Adam equivalent to Momentum SGD with (μ, α_S).

Note that momentum SGD does not make use of a second timescale with respect to the squared gradient term $\mathbf{g}_t^2$. Therefore, Adam is equivalent to momentum SGD when we take away Adam's $\mathbf{v_t}$ update rule, leaving it as initialized $\forall t, \mathbf{v_t} = 0$ and elevate the dampener to 1 to prevent division by 0 and undesired scaling.

Thus, the right parameters are:

$$\alpha_{\mathbf{A}} = \alpha_{\mathbf{S}},\ \beta_{\mathbf{1}} = \mu,\ \beta_{\mathbf{2}} = 1, \text{ and } \epsilon_{\mathbf{A}} = 1.$$

(c) Goal: Show that $\epsilon_A = 0 \implies$ Adam algorithm is invariant to re-scaling.

First, denote $\tilde{\mathcal{J}}(\theta_t){=}C \cdot \mathcal{J}(\theta_t)$, assuming $C > 0$, and let $\epsilon_A = 0$ and $\tilde{\theta}_0 = \theta_0$. Note that $\nabla\tilde{\mathcal{J}}(\theta_t) = C \cdot \nabla\mathcal{J}(\theta_t)$.

Then, for t=1,

$$\tilde{\mathbf{g}}_1 \leftarrow \nabla\tilde{\mathcal{J}}(\theta_0)$$

$$\tilde{\mathbf{m}}_1 \leftarrow (1-\beta_1)\tilde{\mathbf{g}}_1\ ,\ \tilde{\mathbf{v}}_1 \leftarrow (1-\beta_2)\tilde{\mathbf{g}}_1^2$$

$$\tilde{\theta}_1 \leftarrow \tilde{\theta}_0 + \alpha_A \tilde{\mathbf{m}}_1/\sqrt{\tilde{\mathbf{v}}_1}$$

but $\tilde{\mathbf{m}}_1/\sqrt{\tilde{\mathbf{v}}_1} = \frac{(1-\beta_1)\tilde{\mathbf{g}}_1}{\sqrt{(1-\beta_2)\tilde{\mathbf{g}}_1^2}} = \frac{C}{|C|} \cdot \frac{(1-\beta_1)\mathbf{g}_1}{\sqrt{(1-\beta_2)\mathbf{g}_1^2}} = \mathbf{m}_1/\sqrt{\mathbf{v}_1}$, given $\frac{C}{|C|} = 1$. Then $\tilde{\mathbf{m}}_1 = C\mathbf{m}_1$ and $\tilde{\mathbf{v}}_1 = C^2\mathbf{v}_1$, and thus $\tilde{\theta}_1 = \theta_1$.

Now, assume the above holds for $t = k$ for induction. Explicitly, assume

$$\tilde{\theta}_k = \theta_k, \text{ given } \tilde{\theta}_{k-1} = \theta_{k-1} \text{ and } \tilde{\mathbf{m}}_k/\sqrt{\tilde{\mathbf{v}}_k} = \mathbf{m}_k/\sqrt{\mathbf{v}_k}$$

$$\text{where } \tilde{\mathbf{m}}_k = C\mathbf{m}_k \text{ and } \tilde{\mathbf{v}}_k = C^2\mathbf{v}_k$$

Consider $t = k + 1$:

$$\tilde{\mathbf{g}}_{k+1} \leftarrow \nabla\tilde{\mathcal{J}}(\theta_k)$$

$$\tilde{\mathbf{m}}_{k+1} \leftarrow \beta_1\tilde{\mathbf{m}}_k + (1-\beta_1)\tilde{\mathbf{g}}_{k+1}\ ,\ \tilde{\mathbf{v}}_{k+1} \leftarrow \beta_2\tilde{\mathbf{v}}_k + (1-\beta_2)\tilde{\mathbf{g}}_{k+1}^2$$

$$\tilde{\theta}_{k+1} \leftarrow \tilde{\theta}_k + \alpha_A\tilde{\mathbf{m}}_{k+1}/\sqrt{\tilde{\mathbf{v}}_{k+1}}$$

We have:

$$\begin{aligned}\tilde{\mathbf{m}}_{k+1} &\leftarrow \beta_1\tilde{\mathbf{m}}_k + (1-\beta_1)\tilde{\mathbf{g}}_{k+1} \\ &= C\beta_1\mathbf{m}_k + (1-\beta_1)C\mathbf{g}_{k+1} \\ &= C(\beta_1\mathbf{m}_k + (1-\beta_1)\mathbf{g}_{k+1}) \\ &= C\mathbf{m}_{k+1}\end{aligned} \qquad \begin{aligned}\tilde{\mathbf{v}}_{k+1} &\leftarrow \beta_2\tilde{\mathbf{v}}_k + (1-\beta_2)\tilde{\mathbf{g}}_{k+1}^2 \\ &= C^2\beta_2\mathbf{v}_k + (1-\beta_2)C^2\mathbf{g}_{k+1} \\ &= C^2(\beta_2\mathbf{v}_k + (1-\beta_2)\mathbf{g}_{k+1}) \\ &= C^2\mathbf{v}_{k+1}\end{aligned}$$

and thus,

$$\tilde{\mathbf{m}}_{k+1}/\sqrt{\tilde{\mathbf{v}}_{k+1}} = \mathbf{m}_{k+1}/\sqrt{\mathbf{v}_{k+1}} \implies \tilde{\theta}_{k+1} \leftarrow \theta_k + \alpha_A\mathbf{m}_{k+1}/\sqrt{\mathbf{v}_{k+1}}$$

resulting in $\tilde{\theta}_{k+1} = \theta_{k+1}$. Therefore, Adam with $\epsilon_A = 0$ is invariant to re-scaling.