# CSC2516: Homework #2

Due on February 11

*Roger Grosse, Jimmy Ba*

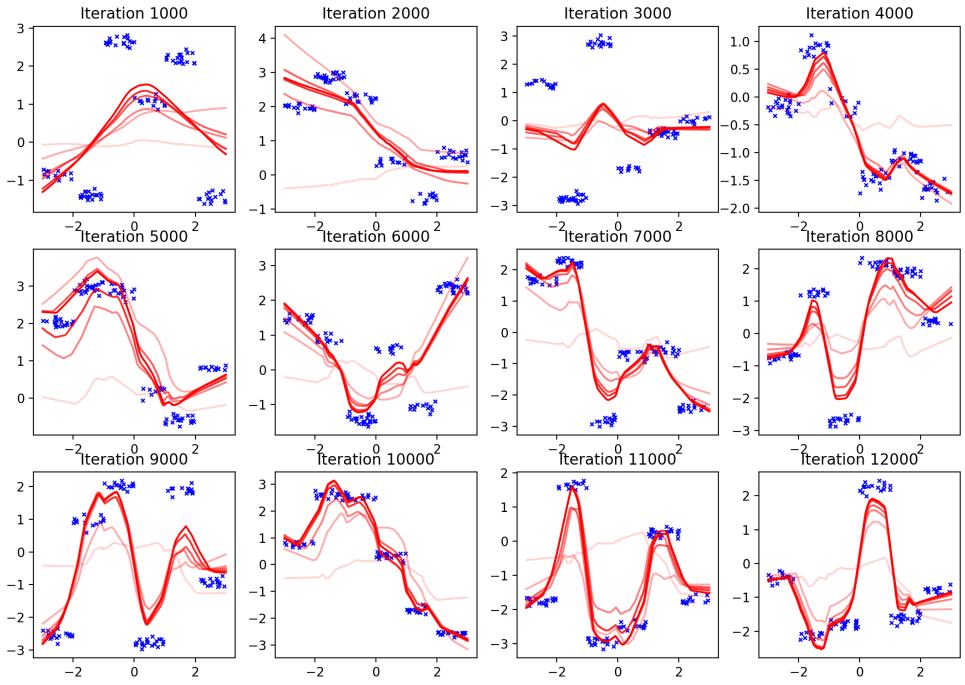
**Thomas Hollis**

# Problem 1

Submit your code solution as maml.py. You dont need to submit anything else for this question.

### Solution

See `maml.py` attached. This code produces the following output:



# Problem 2

Specify Adam hyperparameters ($\alpha_A$, $\beta_1$, $\beta_2$, $\epsilon_A$) which make Adam equivalent to RMSprop with hyperparameters ($\alpha_R$, $\gamma$, $\epsilon_R$). You should explain your answer, though a full derivation isnt required.

### Solution

We know that Adam is, in essence, RMSprop with momentum. Thus the goal here would be to adjust the parameters such that the momentum of Adam no longer has an impact. Hence, the Adam hyperparameters that correspond to RMSprop with the above hyperparameters are:

$$\begin{gathered}\alpha_A = \alpha_R \\ \beta_1 = 0 \\ \beta_2 = \gamma \\ \epsilon_A = \epsilon_R\end{gathered}$$

Indeed, this means the learning rate of Adam becomes the learning rate of RMSprop, the moment timescale $\beta_1$ gets set to 0 and the other moment timescale becomes the $\gamma$ parameter of RMSprop while the Adam damping term $\epsilon_A$ becomes the damping term $\epsilon_R$ of RMSprop.

This is because if we plug these parameters above into the equations for Adam given, we end up with the exact same equations as those given for RMSprop:

$$\begin{gathered}\mathbf{g}_t \longleftarrow \nabla J(\boldsymbol{\theta}_{t-1}) \\ \mathbf{v}_t \longleftarrow \gamma \mathbf{v}_{t-1} + (1-\gamma)\mathbf{g}_t^2 \\ \boldsymbol{\theta}_t \longleftarrow \boldsymbol{\theta}_{t-1} - \alpha_R \mathbf{g}_t / (\sqrt{\mathbf{v}_t} + \epsilon_R)\end{gathered}$$

# Problem 3

Specify Adam hyperparameters ($\alpha_A$, $\beta_1$, $\beta_2$, $\epsilon_A$) which make Adam approximately equivalent to momentum SGD with parameters ($\alpha_S$, $\mu$). Explain your answer.

**Solution**

We know that Adam is, in essence, RMSprop with momentum. Thus the goal here would be to adjust the parameters such that the RMSprop part of Adam no longer has an impact.

We can start off by eliminating $\mathbf{g}_t^2$ from Adam. This can be done by setting $\beta_2$ to 1. Since $\mathbf{v}_t$ is initialised to 0 and no longer updated thus Adam reduces to:

$$\begin{aligned}\mathbf{m}_t &\longleftarrow \beta_1 \mathbf{m}_{t-1} + (1-\beta_1)\mathbf{g}_t \\ \boldsymbol{\theta}_t &\longleftarrow \boldsymbol{\theta}_{t-1} - \alpha_A \mathbf{m}_t / \epsilon_R\end{aligned}$$

We can then set $\epsilon_R$ to 1 reducing Adam to:

$$\begin{aligned}\mathbf{m}_t &\longleftarrow \beta_1 \mathbf{m}_{t-1} + (1-\beta_1)\mathbf{g}_t \\ \boldsymbol{\theta}_t &\longleftarrow \boldsymbol{\theta}_{t-1} - \alpha_A \mathbf{m}_t\end{aligned}$$

To proceed further we must now find the general form for $\mathbf{m}_t$ (Adam) expanded out as a sum rather than as updates, and its corresponding general form for $\mathbf{p}_t$ (Momentum SGD). By expanding out the updates we get:

$$\begin{aligned}\mathbf{m}_t &= \textstyle\sum_{i=1}^{t} \beta_1^{(t-i)}(1-\beta_1)\mathbf{g}_i \\ \mathbf{p}_t &= \textstyle\sum_{i=1}^{t} \mu^{(t-i)}(\mu-1)\mathbf{g}_i\end{aligned}$$

Hence, plugging the above expressions into their respective equations for $\boldsymbol{\theta}_t$ yields:

$$\begin{aligned}\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \alpha_A \textstyle\sum_{i=1}^{t} \beta_1^{(t-i)}(1-\beta_1)\mathbf{g}_i = \boldsymbol{\theta}_{t-1} + \alpha_A(\beta_1 - 1)\textstyle\sum_{i=1}^{t} \beta_1^{(t-i)}\mathbf{g}_i \\ \boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \alpha_A \textstyle\sum_{i=1}^{t} \mu^{(t-i)}(\mu-1)\mathbf{g}_i = \boldsymbol{\theta}_{t-1} + \alpha_A(\mu - 1)\textstyle\sum_{i=1}^{t} \mu^{(t-i)}\mathbf{g}_i\end{aligned}$$

Thus from the above, we can set by inspection $\alpha_A = \alpha_S$ and $\beta_1 = \mu$.

Hence, the Adam hyperparameters that correspond to momentum SGD with the hyperparameters ($\alpha_S$, $\mu$) are:

$$\begin{aligned}\alpha_A &= \alpha_S \\ \beta_1 &= \mu \\ \beta_2 &= 1 \\ \epsilon_A &= 1\end{aligned}$$

# Problem 4

An important fact about Adam is that it is invariant to rescaling of the loss function. I.e., suppose we have a loss function $L(y,t)$, and we define a modified loss function as $\tilde{L}(y,t) = C \cdot L(y,t)$ for some positive constant $C$. Show that for $\epsilon_A = 0$, Adam is invariant to this rescaling, i.e. it passes through the same sequence of iterates $\boldsymbol{\theta}_0, \cdots, \boldsymbol{\theta}_T$. Hint: Denote the quantities computed by Adam on the modified loss function as $\tilde{\mathbf{g}}_t, \tilde{\mathbf{m}}_t$, etc. Use induction to find relationships between these and the original $\mathbf{g}_t, \mathbf{m}_t$, etc.)

**Solution**

The reason why Adam is invariant to rescaling of the loss function can be proved by mathematical induction as follows.

For the base case where $t = 1$, setting $\epsilon_A = 0$, we have:

$$\begin{aligned} g_1 &= \nabla J(\theta_0) \\ m_1 &= (1-\beta_1)\nabla J(\theta_0) \\ v_1 &= (1-\beta_2)(\nabla J(\theta_0))^2 \\ \theta_1 &= \theta_0 - \tfrac{\alpha_A(1-\beta_1)\nabla J(\theta_0)}{\sqrt{(1-\beta_2)\nabla J(\theta_0)}} \end{aligned} \qquad \begin{aligned} \tilde{g}_1 &= \nabla J(\theta_0) \cdot C \\ \tilde{m}_1 &= (1-\beta_1)J(\theta_0) \cdot C \\ \tilde{v}_1 &= (1-\beta_2)(\nabla J(\theta_0) \cdot C)^2 \\ \tilde{\theta}_1 &= \tilde{\theta}_0 - \tfrac{\alpha_A(1-\beta_1)\nabla J(\theta_0)\cdot C}{\sqrt{(1-\beta_2)\nabla J(\theta_0)\cdot C}} \end{aligned}$$

$$\theta_0 - \tfrac{\alpha_A(1-\beta_1)\nabla J(\theta_0)}{\sqrt{(1-\beta_2)\nabla J(\theta_0)}} = \tilde{\theta}_0 - \tfrac{\alpha_A(1-\beta_1)\nabla J(\theta_0)\cdot C}{\sqrt{(1-\beta_2)\nabla J(\theta_0)\cdot C}}$$
$$\therefore \tilde{\theta}_1 = \theta_1$$
$$\therefore \tilde{v}_1 = v_1 \cdot C^2$$
$$\therefore \tilde{m}_1 = m_1 \cdot C$$

Hence, $\tilde{\boldsymbol{\theta}}_t = \boldsymbol{\theta}_t$ proved for $t = 1$.

Assuming the case for $t = n$ is true, we have:

$$\tilde{\theta}_n = \theta_n$$
$$\tilde{v}_n = v_n \cdot C^2$$
$$\tilde{m}_n = m_n \cdot C$$

Thus for $t = n+1$, we have:

$$\tilde{v}_{n+1} = \beta_2\tilde{v}_n + (1-\beta_2)(\nabla J(\theta_n)\cdot C)^2 = \beta_2 v_n \cdot C^2 + (1-\beta_2)(\nabla J(\theta_n))^2 \cdot C^2 = (\beta_2 v_n + (1-\beta_2)(\nabla J(\theta_n))^2) \cdot C^2$$
$$\therefore \tilde{v}_{n+1} = v_{n+1} \cdot C^2$$

$$\tilde{m}_{n+1} = \beta_1\tilde{m}_n + (1-\beta_1)\nabla J(\theta_n)\cdot C = \beta_1 m_n \cdot C + (1-\beta_1)\nabla J(\theta_n)\cdot C = (\beta_1 m_n + (1-\beta_1)\nabla J(\theta_n)) \cdot C$$
$$\therefore \tilde{m}_{n+1} = m_{n+1} \cdot C$$

$$\tilde{\theta}_{n+1} = \tilde{\theta}_n - \tfrac{\alpha_A\tilde{m}_n}{\sqrt{\tilde{v}_n}} = \theta_n - \tfrac{\alpha_A m_n \cdot C}{\sqrt{\tilde{v}_n \cdot C^2}} = \theta_n - \tfrac{\alpha_A m_n}{\sqrt{\tilde{v}_n}}$$
$$\therefore \tilde{\theta}_{n+1} = \theta_{n+1}$$

Hence, $\tilde{\boldsymbol{\theta}}_t = \boldsymbol{\theta}_t$ proved for $t = n+1$.

Hence, since $\tilde{\boldsymbol{\theta}}_t = \boldsymbol{\theta}_t$ proved for $t = 1$ and $t = n+1$ (assuming true for $t = n$), therefore by principle of mathematical induction the statement is true for all natural numbers:

$$\therefore \tilde{\boldsymbol{\theta}}_t = \boldsymbol{\theta}_t \quad \forall \quad n \in \mathbb{N}$$