MARS is generalised Nesterov with longer lookahead

Anonymous Author(s)
Affiliation
Address
`email`

MARS (Yuan et al., 2024) is clearly great. But the justification given in the paper makes no sense. Specifically, the justification is given in terms of reducing the minibatch variance by computing the gradient for the same minibatch at different values of the parameters. This would indeed reduce variance. However, this doubles the compute requirements, because you now need to compute the gradient for each datapoint at two setting of the parameters. Instead, they just use the current and previous gradients, evaluated on *different* minibatches. There is no sense in which this can reduce minibatch gradients. Instead, here we argue that MARS should be understood as a generalisation of Nesterov, where you lookahead further than usual.

Nesterov is (Sutskever et al., 2013, from)

$$v_{t+1} = \beta_1 v_t - \varepsilon f(\theta_t + \mu v_t) \tag{1a}$$
$$\theta_{t+1} = \theta_t + v_{t+1} \tag{1b}$$

The usual interpretation of Nesterov is that you compute the gradient, $f(\theta_t + \mu v_t)$, at a "lookahead" location, $\theta_t + \mu v_t$. Note that in standard Nesterov, the "lookahead" is one momentum step (i.e. $\beta_1 = \mu$). We have generalised Nesterov slightly, by allowing an arbitrary lookahead of size μ.

Now, to reframe this to look more like MARS, we start by noting that the gradient steps are performed at $\theta_t + \mu v_t$. We therefore write the v update as,

$$v_{t+1} = \beta_1 v_t - \varepsilon f(\theta'_t) \tag{2}$$

which is equivalent to the Nesterov momentum update (Eq. 1a) if we define,

$$\theta'_t = \theta_t + \mu v_t. \tag{3}$$

To write the updates for θ', we take Eq. (3) for timestep $t+1$ and substitute Eq. (1b),

$$\theta'_{t+1} = \theta_{t+1} + \mu v_{t+1} = \theta_t + v_{t+1} + \mu v_{t+1} \tag{4}$$

Adding and subtracting μv_t,

$$\theta'_{t+1} = \theta_t + \mu v_t - \mu v_t + v_{t+1} + \mu v_{t+1} \tag{5}$$

Noticing that $\theta'_t = \theta_t + \mu v_t$ (Eq. 3),

$$\theta'_{t+1} = \theta'_t + v_{t+1} + \mu \left(v_{t+1} - v_t\right) \tag{6}$$

Thus, the overall updates become,

$$v_{t+1} = \beta_1 v_t - \varepsilon g_t \tag{7a}$$
$$\theta'_{t+1} = \theta'_t + v_{t+1} + \mu \left(v_{t+1} - v_t\right) \tag{7b}$$

where $g_t = f(\theta'_t)$ is the gradient evaluated at θ'_t.

Here, the $v_{t+1} - v_t$ correction term arises in the parameter update, not in the parameter update, as in MARS. We can push the correction into the momentum by writing the parameter update as,

$$\theta'_{t+1} = \theta'_t + v'_{t+1} \tag{8}$$

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

which is equivalent to the Nesterov parameter update (Eq. 7b) if we define,

$$v'_{t+1} = v_{t+1} + \mu\left(v_{t+1} - v_t\right). \tag{9}$$

To write the updates for v', we substitute the update for v (Eq. (7a), into Eq. 9,

$$v'_{t+1} = (\beta_1 v_t - \varepsilon g_t) + \mu\left((\beta_1 v_t - \varepsilon g_t)) - (\beta_1 v_{t-1} - \varepsilon g_{t-1})\right)) \tag{10}$$

Rearranging,

$$v'_{t+1} = \beta_1\left(v_t + \mu\left(v_t - v_{t-1}\right)\right) - \varepsilon\left(g_t + \mu\left(g_t - g_{t-1}\right)\right)) \tag{11}$$

Identifying $v'_t = v_t + \mu\left(v_t - v_{t-1}\right)$ (Eq. 9),

$$v'_{t+1} = \beta_1 v'_t - \varepsilon\left(g_t + \mu\left(g_t - g_{t-1}\right)\right). \tag{12}$$

Overall, the updates become,

$$\theta'_{t+1} = \theta'_t + v'_{t+1} \tag{13a}$$

$$v'_{t+1} = \beta_1 v'_t - \varepsilon\left(g_t + \mu\left(g_t - g_{t-1}\right)\right). \tag{13b}$$

This is starting to resemble MARS! Critically, in MARS, there is a learning rate that applies to the gradient in the parameter update. To get something like this in our setup, we define v'' as a scaled version of v',

$$\theta'_{t+1} = \theta'_t + \eta v''_{t+1} \tag{14}$$

This is equivalent to Eq. (13b) if we set,

$$v''_{t+1} = \tfrac{1}{\eta} v'_{t+1} \tag{15}$$

To get updates for v'', we susbtitute Eq. 13b,

$$v''_{t+1} = \tfrac{1}{\eta}\beta_1 v'_t - \tfrac{\varepsilon}{\eta}\left(g_t + \mu\left(g_t - g_{t-1}\right)\right). \tag{16}$$

Recognising that $v'_t/\eta = v''_t$ (Eq. 15), we get,

$$v''_{t+1} = \beta_1 v''_t - \tfrac{\varepsilon}{\eta}\left(g_t + \mu\left(g_t - g_{t-1}\right)\right). \tag{17}$$

Overall, this gives updates,

$$\theta'_{t+1} = \theta'_t + \eta v''_{t+1} \tag{18a}$$

$$v''_{t+1} = \beta_1 v''_t - \tfrac{\varepsilon}{\eta}\left(g_t + \mu\left(g_t - g_{t-1}\right)\right). \tag{18b}$$

which has exactly the same form as MARS, except for the constants. Specifically, the MARS updates are the same for θ', but have different constants for v'',

$$v''_{t+1} = \beta_1 v''_t - (1-\beta_1) g_t + \beta_1 \mu\left(g_t - g_{t-1}\right). \tag{19}$$

Thus, MARS can be connected to generalised Nesterov by taking the constants in Eq. 18b and Eq. 19 to be equal. Specifically, taking the constants for the g_t terms to be equal,

$$\tfrac{\varepsilon}{\eta} = 1 - \beta \tag{20}$$

implies that to make generalised Nesterov equivalent to MARS, we need to set the generalised Nesterov parameter, ε, to,

$$\varepsilon = \eta\left(1 - \beta_1\right) \tag{21}$$

And, taking the constants for the $\mu(g_t - g_{t-1})$ terms to be equal,

$$\tfrac{\varepsilon}{\eta}\mu = \beta_1 \tag{22}$$

implies that to make generalized Nesterov equivalent to MARS, we need to set the genearlised Nesterov parameter, μ, to,

$$\mu = \frac{\beta_1}{\frac{\varepsilon}{\eta}} = \frac{\beta_1}{1-\beta_1} \tag{23}$$

What does this mean for the interpretation of MARS? It means that MARS is generalised Nesterov, where we lookahead a long way: specifically, the look-ahead is set by μ. In standard Nesterov, we would have $\mu = \beta_1$, and remember that β_1 is typically close to 1, e.g. $\mu = \beta_1 = 0.9$. However, in MARS, with $\beta_1 = 0.9$, the lookahead is given by,

$$\mu(\beta_1 = 0.9) = \frac{0.9}{1-0.1} = \frac{0.9}{1-0.1} = \frac{0.9}{0.1} = 9. \tag{24}$$

Implying that if $\beta_1 = 0.9$, MARS is generalised Nesterov which looks ahead ten times further than standard Nesterov.

References

Sutskever, I., Martens, J., Dahl, G., and Hinton, G. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pp. 1139–1147. PMLR, 2013.

Yuan, H., Liu, Y., Wu, S., Zhou, X., and Gu, Q. Mars: Unleashing the power of variance reduction for training large models. *arXiv:2411.10438*, 2024.