Learning values across many orders of magnitude [1]

Guannan Hu

April 2, 2018

Algorithm 1 SGD on squared loss with Pop-Art

For a given differentiable function $h_{\boldsymbol{\theta}}$, initialize $\boldsymbol{\theta}$.
Initialize $\boldsymbol{\theta} = I, \boldsymbol{b} = 0, \boldsymbol{\Sigma} = I$ and $\boldsymbol{\mu} = \mathbf{0}$.
while learning **do**
  Observe input X and target Y
  Use Y to compute new scale $\boldsymbol{\Sigma}_{\text{new}}$ and new shift $\boldsymbol{\mu}_{\text{new}}$
  $\boldsymbol{W} \leftarrow \boldsymbol{\Sigma}_{\text{new}}^{-1}\boldsymbol{\Sigma}\boldsymbol{W}, \boldsymbol{b} \leftarrow \boldsymbol{\Sigma}_{\text{new}}^{-1}(\boldsymbol{\Sigma}\boldsymbol{b} + \boldsymbol{\mu} - \boldsymbol{\mu}_{\text{new}})$ ▷ rescale $\boldsymbol{W}$ and $\boldsymbol{b}$
  $\boldsymbol{\Sigma} \leftarrow \boldsymbol{\Sigma}_{\text{new}}, \boldsymbol{\mu} \leftarrow \boldsymbol{\mu}_{\text{new}}$ ▷ update scale and shift
  $\boldsymbol{h} \leftarrow h_{\boldsymbol{\theta}}(X)$ ▷ store outputof $h_{\boldsymbol{\theta}}$
  $\boldsymbol{J} \leftarrow (\nabla_{\boldsymbol{\theta}} h_{\boldsymbol{\theta},1}(X), ..., \nabla_{\boldsymbol{\theta}} h_{\boldsymbol{\theta},m}(X))$ ▷ compute Jacobian of $h_{\boldsymbol{\theta}}$
  $\boldsymbol{\delta} \leftarrow \boldsymbol{W}\boldsymbol{h} + \boldsymbol{b} - \boldsymbol{\Sigma}^{-1}(Y - \boldsymbol{\mu})$ ▷ compute normalized error
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \alpha \boldsymbol{J}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{\delta}$ ▷ compute SGD update for $\boldsymbol{\theta}$
  $\boldsymbol{W} \leftarrow \boldsymbol{W} - \alpha \boldsymbol{\delta}\boldsymbol{h}^{\mathrm{T}}$ ▷ compute SGD update for $\boldsymbol{W}$
  $\boldsymbol{b} \leftarrow \boldsymbol{b} - \alpha \boldsymbol{\delta}$ ▷ SGD update for $\boldsymbol{b}$
end while

Algorithm 2 Normalized SGD

For a given differentiable function $h_{\boldsymbol{\theta}}$, initialize $\boldsymbol{\theta}$.
while learning **do**
  Observe input X and target Y
  Use Y to compute new scale $\boldsymbol{\Sigma}$
  $\boldsymbol{h} \leftarrow h_{\boldsymbol{\theta}}(X)$ ▷ store output of $h_{\boldsymbol{\theta}}$
  $\boldsymbol{J} \leftarrow (\nabla h_{\boldsymbol{\theta},1}, ..., \nabla h_{\boldsymbol{\theta},m}(X)))^{\mathrm{T}}$ ▷ compute Jacobian of $h_{\boldsymbol{\theta}}$
  $\boldsymbol{\delta} \leftarrow \boldsymbol{W}\boldsymbol{h} + \boldsymbol{b} - Y$ ▷ Compute unnormalized error
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \alpha \boldsymbol{J}(\boldsymbol{\Sigma}^{-1}\boldsymbol{W})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta}$ ▷ update $\boldsymbol{\theta}$ with scaled SGD
  $\boldsymbol{W} \leftarrow \boldsymbol{W} - \alpha \boldsymbol{\delta}\boldsymbol{g}^{\mathrm{T}}$ ▷ update $\boldsymbol{W}$ with SGD
  $\boldsymbol{b} \leftarrow \boldsymbol{b} - \alpha \boldsymbol{\delta}$ ▷ update $\boldsymbol{b}$ with SGD
end while

References

[1] Hado Van Hasselt, Arthur Guez, Matteo Hessel, Volodymyr Mnih, and David Silver. Learning values across many orders of magnitude. 2016.