Improving Stochastic Gradient Descent with Feedback

Jayanth Koushik, Hiroaki Hayashi
Language Technologies Institute, Carnegie Mellon University

Overview

Problem setting: First order optimization of smooth unconstrained non-convex problems.
Objective: Adapting to plateaus in the objective surface, and variance introduced by stochasticity.
Assumption: Minimum of objective function is 0.
Approach: For non-increasing f, tune learning rate of gradient descent with

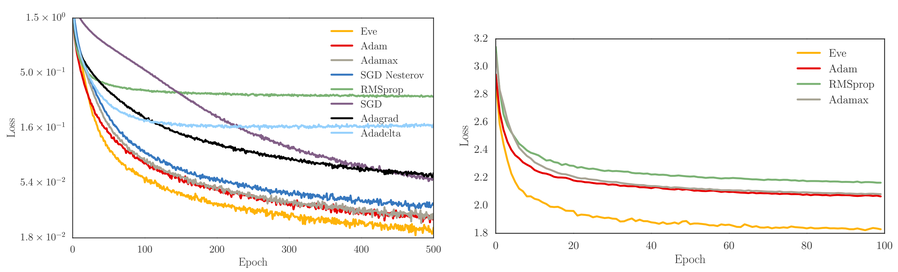
$$d_t = \beta d_{t-1} + (1-\beta)\frac{f_{t-2} - f_{t-1}}{f_{t-1}}$$

Algorithm

Algorithm 1 Eve: Adam with feedback.

$m_0 = v_0 = \tilde{f}_{-1} = t = 0$, $d_0 = 1$
while stopping condition is not reached **do**
  $t \leftarrow t + 1$
  $g_t \leftarrow \nabla_\theta f(\theta_{t-1})$
  $m_t \leftarrow \beta_1 m_{t-1} + (1-\beta_1) g_t$, $\bar{m}_t \leftarrow \frac{m_t}{(1-\beta_1^t)}$
  $v_t \leftarrow \beta_2 v_{t-1} + (1-\beta_2) g_t^2$, $\bar{v}_t \leftarrow \frac{v_t}{(1-\beta_2^t)}$
  if $t > 1$ **then**
    if $f(\theta_{t-1}) \geq \tilde{f}_{t-2}$ **then**
      $\delta_t \leftarrow k+1$, $\Delta_t \leftarrow K+1$
    else
      $\delta_t \leftarrow \frac{1}{K+1}$, $\Delta_t \leftarrow \frac{1}{k+1}$
    end if
    $\tilde{f}_{t-1} \leftarrow \tilde{f}_{t-2} \min\left\{\max\left\{\delta_t, \frac{f(\theta_{t-1})}{\tilde{f}_{t-2}}\right\}, \Delta_t\right\}$
    $r_t \leftarrow \frac{|\tilde{f}_{t-1} - \tilde{f}_{t-2}|}{\min\{\tilde{f}_{t-1}, \tilde{f}_{t-2}\}}$
    $d_t \leftarrow \beta_3 d_{t-1} + (1-\beta_3) r_t$
  else
    $\tilde{f}_{t-1} \leftarrow f(\theta_{t-1})$, $d_t \leftarrow 1$
  end if
  $\theta_t \leftarrow \theta_{t-1} - \alpha \frac{\bar{m}_t}{d_t \sqrt{\bar{v}_t} + \epsilon}$
end while

Results on Non-Convex Problems

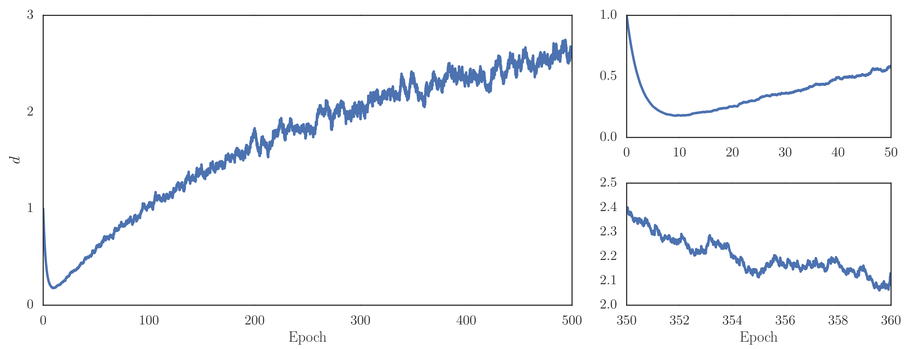


CNN on CIFAR10

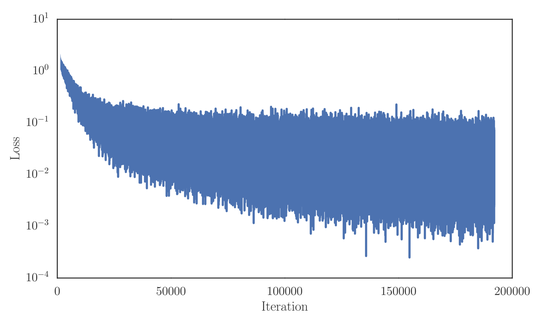
RNN on Penn TreeBank

Training loss for convolutional neural network on CIFAR10, and recurrent neural network on Penn TreeBank. In both cases, Eve achieves the best performance.

Tuning Coefficient Behavior

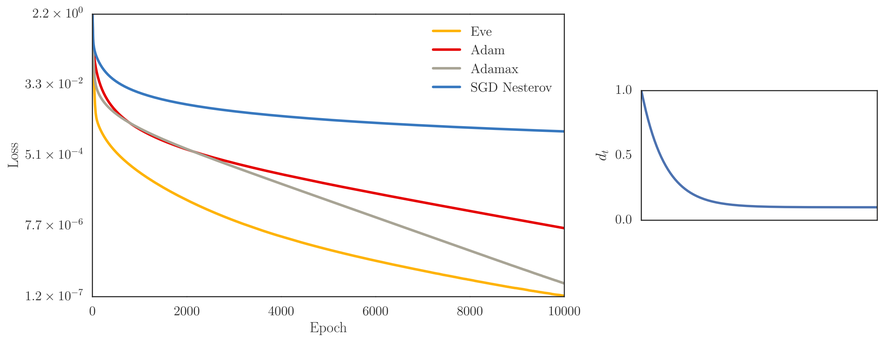


Behavior of the tuning coefficient d_t during the CIFAR10 experiment. There is an overall trend of acceleration followed by decay, but also more fine-grained behavior as indicated by the bottom-right plot.

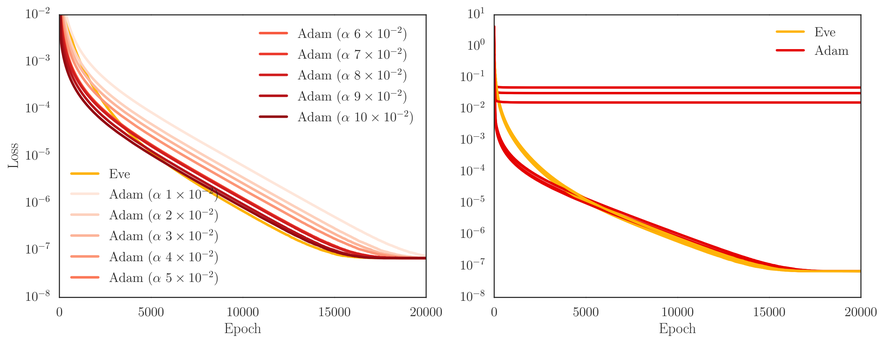


Minibatch losses for Eve during the CIFAR10 experiment. The variance in the losses increases throughout the training

Analysis on Convex Problem



Loss curves and tuning coefficient d_t for batch gradient descent training of a logistic regression model. For this convex case, d_t continuously decreases and converges to the lower threshold 0.1.



The left plot shows Adam with different learning rates, and Eve with learning rate 10^{-2}. The right plot shows 10 repetitions of Adam with learning rate 10^{-1}, and Eve with learning rate 10^{-2}. Although Adam with a larger learning rate can be almost identical to Eve, this is largely dependent on the initial values for Adam as shown in the plot on the right.

Conclusion

We investigated a simple and efficient method for incorporating feedback into stochastic gradient descent algorithms. We used this method to propose Eve, a modified version of the Adam algorithm. Experiments with a variety of models showed that the proposed method can help improve the optimization of deep neural networks.