Adaptive Gradient Methods with Dynamic Bound of Learning Rate

Liangchen Luo[1*] Yuanhao Xiong[2*] Yan Liu[3] Xu Sun[1]

[1]PKU [2]ZJU [3]USC *Equal Contribution

Dilemma

Adam conveges faster while generalizes worse;
SGD converges slower while generalizes better.

☆ Find out why Adam fails

Solution

AdaBound: Transforms from Adam to SGD by applying dynamic bound on learning rate.

1. Non-Convergence ⟸ (caused by) Extreme Learning Rate

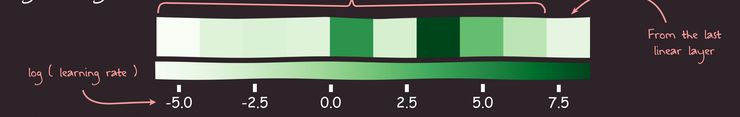
- Intuitive assuption (Wilson et al., 2017): The bad performance of adaptive methods may stem from UNSTABLE and EXTREME final learning rates.
- AMSGrad (Reddi et al., 2018), a new proposed optimizer, is claimed to have smaller learning rates compared with Adam, that may help abate the imparct of huge learning rates. However, it neglects possible effects of small ones.

} Previous attempts & oversevations

Still missing 1) convincing evidence of the assumption; 2) ways to address tiny learning rates

Evidence of the Assumption

- We sample learning rates of several weights and biases of ResNet-34 at the end of training on CIFAR-10 using Adam. Learning rates are composed of tiny ones less than 0.01 as well as huge ones greater than 1000.



Proof of the Assumption

- **Theorem 1.** *There is an online convex optimization problem where for any initial step size α, Adam has non-zero average regret i.e., $R_T/T \nrightarrow 0$ as $T \rightarrow \infty$.*
- **Theorem 2.** *For any constant $\beta_1, \beta_2 \in [0, 1)$ such that $\beta_1^2 < \beta_2$, there is an online convex optimization problem where for any initial step size α, Adam has non-zero average regret i.e., $R_T/T \nrightarrow 0$ as $T \rightarrow \infty$.*
- **Theorem 3.** *For any constant $\beta_1, \beta_2 \in [0, 1)$ such that $\beta_1^2 < \beta_2$, there is a stochastic convex optimization problem where for any initial step size α, Adam does not converge to the optimal solution.*

2. Dynamic Bound on Adam's Learning Rate

- Our aim is to devise a strategy that combines the benefits of adaptive methods, viz. fast initial progress, and the good final generalization properties of SGD.
- The key point is to restrict the actual learning rate of adaptive methods at the end of training. Make it neither too large nor too small.
- Inspired by Gradient Clipping, a popular technique for preventing from gradient explosion, we can also employ clipping on Adam's learning rates. An operation that clips the learning rate element-wisely such that the output is constrained to be in $[\eta_l, \eta_u]$. (lower/upper bound)

☆ Note: SGD(M) with a learning rate of α^* can be considered as the case where $\eta_l = \eta_u = \alpha^*$; as for Adam, $\eta_l = 0$ and $\eta_u = \infty$.

We propose AdaBound by employing dynamic bound rather than constant threshold:

Gradient: $g_t = \nabla f_t(x_t)$

First/Second-order momentum: $m_t = \beta_{1t} m_{t-1} + (1 - \beta_{1t}) g_t$

$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$

Lower bound function converges from 0 to α^*;
Upper bound function converges from ∞ to α^*;
which makes Adam gradually transforms to SGD.

$\hat{\eta}_t = \mathrm{Clip}(\alpha/\sqrt{v_t}, \eta_l(t), \eta_u(t))$ and $\eta_t = \hat{\eta}_t/\sqrt{t}$

Initial step size

$x_{t+1} = \Pi_{F, \mathrm{diag}(\eta_t^{-1})}(x_t - \eta_t \circ m_t)$

Clipped learning rate

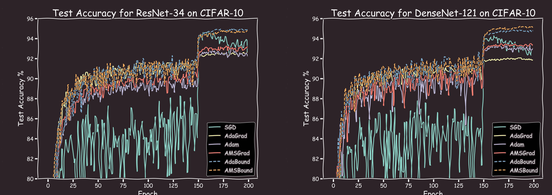
The algorithm is provable to have the following bound on regret:

$D_\infty = \max \| x - y \|_\infty$, $R_\infty = \eta_u(1)$, $G_2 = \max \| g_t \|$

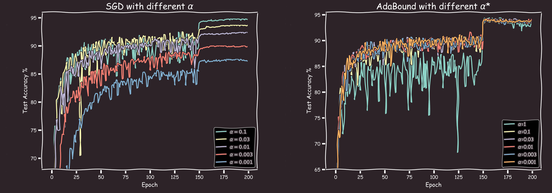
Regret: $R_T \le \frac{D_\infty^2 \sqrt{T}}{2(1-\beta_1)} \sum_{i=1}^{d} \hat{\eta}_{T,i}^{-1} + \frac{D_\infty^2}{2(1-\beta_1)} \sum_{t=1}^{T} \sum_{i=1}^{d} \beta_{1t} \eta_{t,i}^{-1} + (2\sqrt{T} - 1) \frac{R_\infty G_2^2}{1 - \beta_1}$

3. Experiments

Test accuracy for ResNet-34 and DenseNet-121 on CIFAR-10:



More rubust with different hyperparameters:



website | code | arXiv | slides

References

[1] S. Reddi, S. Kale, and S. Kumar. On the convergence of adam and beyond. In Proc. of ICLR, 2018.

[2] A. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht. The marginal value of adaptive gradient methods in machine learning. In Proc. of NeurIPS, 2017.