COSC 594/650
Clipping and adaptive gradient methods

Dongsheng Ding
dongshed@utk.edu

Department of Electrical Engineering and Computer Science
University of Tennessee, Knoxville

Recap: gradients can be badly scaled

Lecture 25: a weight's gradient is a product of one factor per layer above it

One factor: activation slope $\times$ weight

- FIFTY LAYERS ABOVE

factor per layer	0.9	1.0	1.1
product of 50 factors	0.005	1	117

- GRADIENT DESCENT: $\theta_{k+1} = \theta_k - \alpha \nabla f(\theta_k)$
  - $\theta \in \mathbb{R}^d$: all weights, updated together
  - each weight moves by α times its gradient: the step inherits its size

Gradients are cheap, but their size is uncontrolled

QUESTION: How should the step react to the size of the gradient?

to its norm along the run – clipping

to the scale of each coordinate (weight) – adaptive methods (Adam)

Clipped gradient descent

Issue: large gradients come with high curvature: a fixed step α overshoots

Parameters: step size $\alpha > 0$, step-length cap $\gamma > 0$

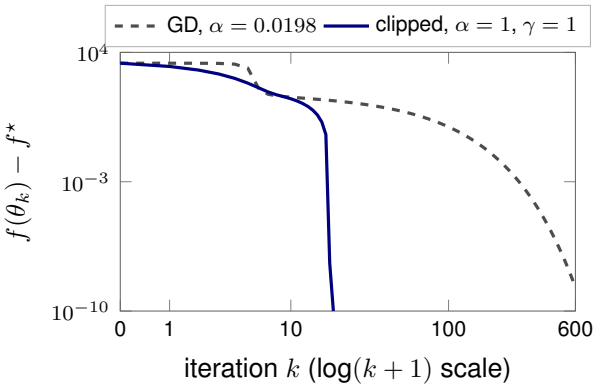
$$\theta_{k+1} \;=\; \theta_k - \alpha_k \nabla f(\theta_k), \qquad \alpha_k \;=\; \min\left\{\alpha,\; \frac{\gamma}{\|\nabla f(\theta_k)\|}\right\}$$

- TWO REGIMES
  - ⋆ large gradient, $\|\nabla f\| > \gamma/\alpha$: a move of length γ along $-\nabla f$
  - ⋆ small gradient: a plain gradient step with step size α

Clipping caps the length of the step

Experiment: a steep start

Example: $f(\theta) = \frac{1}{2}\|\theta\|^2 + \frac{1}{4}u^4$, u: first coordinate of θ, $\theta_0 = (10, 10)$



■ STEPS TO $f < 10^{-8}$

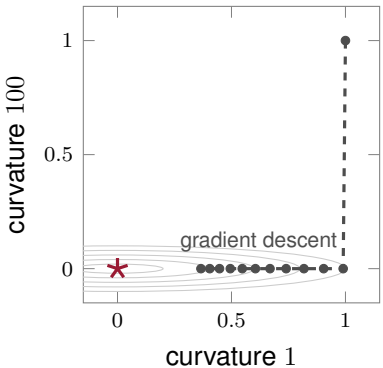
θ_0	GD	clipped
$(2, 2)$	19	4
$(10, 10)$	559	19
$(50, 50)$	15986	96

⋆ GD: the steep start forces $\alpha < 2/101$

Clipping: far fewer steps from a steep start

One step size cannot fit every coordinate

Example: $f(\theta) = \frac{1}{2}\theta^\top A\theta, \quad A = \mathrm{diag}(1, 100), \quad \theta_0 = (1, 1), \quad \alpha = 1/L = 0.01$



- GRADIENT DESCENT
  - steep coordinate: one step
  - flat coordinate: 1% per step
  - 882 steps to $f < 10^{-8}$
- IN A NETWORK (LECTURE 25)
  - layers: very different gradient sizes

One α is set by the steepest coordinate: scale each coordinate

One stepsize per coordinate

Stochastic gradient: $g_k = \nabla f_{i_k}(\theta_k) \in \mathbb{R}^d$

- FOR EACH COORDINATE $j = 1, \ldots, d$

$$v_{k,j} \;=\; \beta_2 v_{k-1,j} + (1-\beta_2)\, g_{k,j}^2, \qquad \theta_{k+1,j} \;=\; \theta_{k,j} - \underbrace{\frac{\alpha}{\sqrt{v_{k,j}} + \delta}}_{\text{stepsize of coordinate } j} g_{k,j}$$

  - ⋆ slide 5 at θ_0: gradient $(1, 100) \;\Rightarrow$ stepsizes $\propto (1,\, 1/100)$
  - ⋆ RMSprop: $\beta_2 = 0.9$
  - ⋆ AdaGrad: $v_{k,j} = g_{0,j}^2 + \cdots + g_{k,j}^2$; small $\delta > 0$

- ELEMENT-WISE : $\theta_{k+1} = \theta_k - \alpha\, g_k / (\sqrt{v_k} + \delta)$

Rescale each coordinate by its typical gradient size

Adam: adaptive moment estimation

Defaults: $\alpha = 10^{-3}, \quad \beta_1 = 0.9, \quad \beta_2 = 0.999, \quad \delta = 10^{-8}, \quad m_{-1} = v_{-1} = 0$

for $k = 0, 1, 2, \ldots$: $\quad g_k = \nabla f_{i_k}(\theta_k)$ vectors in $\mathbb{R}^d$, operations element-wise

$$m_k = \beta_1 m_{k-1} + (1 - \beta_1)\, g_k$$ estimates the mean of g

$$v_k = \beta_2 v_{k-1} + (1 - \beta_2)\, g_k^2$$ estimates the mean of g^2

$$\hat{m}_k = m_k / (1 - \beta_1^{k+1}), \quad \hat{v}_k = v_k / (1 - \beta_2^{k+1})$$ corrects the zero start

$$\theta_{k+1} = \theta_k - \alpha\, \hat{m}_k / \left(\sqrt{\hat{v}_k} + \delta\right)$$

WHY TWO MOMENTS

- $\hat{m}_k$: a less noisy direction – momentum, as in Lecture 9
- $\sqrt{\hat{v}_k}$: the scale of each coordinate – as in RMSprop

Adam: estimate the mean and mean square of each coordinate's gradient

Adam's steps do not see the gradient's scale

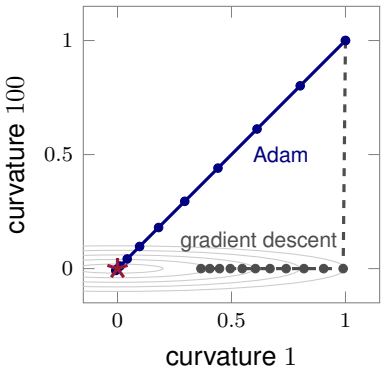
Step of coordinate j: $\theta_{k,j} - \theta_{k+1,j} = \alpha\, \hat{m}_{k,j} / \sqrt{\hat{v}_{k,j}} \quad (\delta = 0)$

- SCALE-FREE
  - ⋆ multiply every gradient by $c > 0$: the step does not change
- ABOUT α PER COORDINATE
  - ⋆ steady gradient: exactly α
  - ⋆ any gradients: at most $7.3\,\alpha$

Adam: a normalized step for every coordinate

Experiment: the ill-conditioned quadratic

Setup: $f(\theta) = \frac{1}{2}\theta^\top A\theta, \quad A = \mathrm{diag}(1, 100), \quad \theta_0 = (1, 1),$ Adam with $\alpha = 0.02$



- ADAM
  - ⋆ same pace in both coordinates
  - ⋆ 194 steps to $f < 10^{-8}$, GD: 882
- ROTATE THE PROBLEM
  - ⋆ GD: unchanged
  - ⋆ Adam: the gain disappears

Per-coordinate steps free the flat coordinate

Two cautions

- ADAM CAN FAIL
  - ⋆ even on a simple convex problem; a fix: AMSGrad

    Reddi, Kale, Kumar (2018), *On the Convergence of Adam and Beyond*.

- RARE, VERY LARGE STOCHASTIC GRADIENTS
  - ⋆ clip them: large models train with Adam plus clipping

    Zhang et al. (2020), *Why are Adaptive Methods Good for Attention Models?*

Adaptive methods: fast in practice, delicate in theory

Steps that react to the gradient

Method	Step $\theta_k - \theta_{k+1}$	Reacts to	Use when
Gradient descent	$\alpha \nabla f$	nothing	gradients are well scaled
Clipped GD	$\min\{\alpha, \gamma/\|\nabla f\|\}\, \nabla f$	gradient norm	steep regions, rare large gradients
Adam	$\alpha\, \hat{m}_k/(\sqrt{\hat{v}_k} + \delta)$	each coordinate	coordinates of different scales

all steps are vectors in $\mathbb{R}^d$; Adam's ratio is taken element-wise

React to the norm: clip; **to each coordinate**: Adam

Questions?

(**Email: dongshed@utk.edu**)