ECE 8101: Nonconvex Optimization for Machine Learning

Lecture Note 2-6: Adaptive First-Order Methods

Jia (Kevin) Liu

Assistant Professor
Department of Electrical and Computer Engineering
The Ohio State University, Columbus, OH, USA

Spring 2022

Outline

In this lecture:

- Key Idea of First-Order Methods with Adaptive Learning Rates
- AdaGrad, RMSProp, Adam, and AMSGrad
- Convergence Results

Motivation

- Recall that SGD has two hyber-parameter "control knobs" for convergence performance
  - Step-size
  - Batch-size
- A significant issue in SGD and variance-reduced versions: Tuning parameters
  - Time-consuming, particularly for training deep neural networks
  - Thus, adaptive first-order methods have received a lot of attention
- The most popular ones that spawn many variants:
  - AdaGrad: [Duchi et al. JMLR'11]
  - RMSProp: [Hinton, '12]
  - Adam: [Kingma & Ba, ICLR'15] (AMSGrad [Reddi et al. ICLR'18])
  - All of these methods still depend on some hyper-parameters, but they are more robust than other variants of SGD or variance-reduced methods
  - One can find PyTorch implementations of these popular adaptive first-order meth methods

AdaGrad

- AdaGrad stands for "adaptive gradient." It is the first algorithm aiming to remove the need for turning the step-size in SGD:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - s(\delta \mathbf{I} + \mathrm{Diag}\{\mathbf{G}_k\})^{-\frac{1}{2}} \mathbf{g}_k,$$

  where $\mathbf{G}_k = \sum_{t=1}^{k} \mathbf{g}_t \mathbf{g}_t^\top$, s is an initial learning rate, and $\delta > 0$ is a small value to prevent from the division by zero (typically on the order of 10^{-8})

- Entry-wise version: ($\mathbf{a}_{k,i}$ denotes the i-th entry of $\mathbf{a}_k$)

$$\mathbf{x}_{k+1,i} = \mathbf{x}_{k,i} - \frac{s_k}{\sqrt{\delta + G_{k,i}}} \mathbf{g}_{k,i},$$

  where $G_{k,i} = \sum_{t=1}^{k} (\mathbf{g}_{t,i})^2$. Typically, $s_k = s$, $\forall k$.

- AdaGrad can be viewed as a special case of SGD with an adaptively scaled step-size (learning rate) for each dimension (feature).

$G_{k,i} = \sum_{t=1}^{k} (g_{t,i})^2$ mono. $\uparrow$

$\left.\begin{array}{l} G_{k,i} \text{ big} \Rightarrow \text{step-size small.} \\ G_{k,i} \text{ small} \Rightarrow \text{step size large.} \end{array}\right\}$ balance the prog.

RMSProp

- A major limitation of AdaGrad: $g_{k,i} \neq 0$, for many iter.
  - Step-sizes could rapidly diminishing (particularly in dense settings), may get stuck in saddle points in nonconvex optimization

- RMSProp (root mean squared propagation)
  - First appeared in Hinton's Lecture 6 notes of the online course "Neural Networks for Machine Learning."
  - Motivated by RProp [Igel & Hüsken, NC'00] (resolving the issue that gradients may vary widely in magnitudes, only using the sign of the gradient)
  - Unpublished (and being famous because of this! ☺)
  - Idea: Keep an exponential moving average of squared gradient of each weight

$$\mathbb{E}[\mathbf{g}_{k+1,i}^2] = \beta \mathbb{E}[\mathbf{g}_{k,i}^2] + (1-\beta)(\nabla_i f(\mathbf{x}_k))^2, \quad \beta \in (0,1).$$
$$\mathbf{x}_{k+1,i} = \mathbf{x}_{k,i} - \frac{s_k}{(\delta + \mathbb{E}[\mathbf{g}_{k+1,i}^2])^{\frac{1}{2}}} \nabla_i f(\mathbf{x}_k).$$

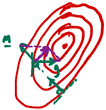
- RMSProp vs. AdaGrad
  - AdaGrad: Keep a running sum of squared gradients
  - RMSProp: Keep an exponential moving average of squared gradients

Adam

Regret: $R_T \triangleq \sum_{t=1}^{T} \left(f(\underline{x}_t) - f(\underline{x}^*)\right) = o(T)$ little-o. ← online opt.

$\frac{1}{T}\sum_{t=1}^{T} \left(f(\underline{x}_t) - f(\underline{x}^*)\right) \to 0.$

- Stands for adaptive momentum estimation
- Motivated by RMSProp, also aims to address the limitation of AdaGrad
- Algorithm: $(\mathbf{g}_k \triangleq \nabla f(\mathbf{x}_k))$



Heavy-Ball momentum.

$$\mathbf{m}_{k,i} = \beta_1 \mathbf{m}_{k-1,i} + (1-\beta_1)\mathbf{g}_{k,i}, \qquad \hat{\mathbf{m}}_{k,i} = \frac{\mathbf{m}_{k,i}}{1-(\beta_1)^k},$$

$$\mathbf{v}_{k,i} = \beta_2 \mathbf{v}_{k-1,i} + (1-\beta_2)(\mathbf{g}_{k,i})^2, \qquad \hat{\mathbf{v}}_{k,i} = \frac{\mathbf{v}_{k,i}}{1-(\beta_2)^2},$$

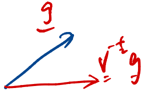
$$\mathbf{x}_{k+1,i} = \mathbf{x}_{k,i} - \frac{s_k}{\sqrt{\hat{\mathbf{v}}_{k,i}} + \delta}\hat{\mathbf{m}}_{k,i}, \qquad i = 1, \ldots, d.$$

- Parameters:
  - $\beta_1 \in [0,1)$: momentum parameter ($\beta_1 = 0.9$ by default, $\beta_1 = 0 \Rightarrow$ RMSProp)
  - $\beta_2 \in (0,1)$: exponential average parameter ($\beta_2 = 0.999$ in the original paper)
- A flaw in convergence proof spotted by [Reddi et al. ICLR'18], leading to...

AMSGrad

- To see the flaw of Adam (and RMSProp), consider a more generic view of adaptive methods: In each iteration k :

matrix-valued



vector-valued

PSD.

$$\mathbf{g}_k = \nabla f_k(\mathbf{x}_k)$$

$$\mathbf{m}_k = \phi_k(\mathbf{g}_1, \ldots, \mathbf{g}_k), \text{ and } \underline{\mathbf{V}_k} = \psi_k(\mathbf{g}_1, \ldots, \mathbf{g}_k)$$

$$\mathbf{x}_{k+1} = \mathbf{x}_k - s_k \mathbf{V}_k^{-\frac{1}{2}} \mathbf{m}_k$$

PSD: $\underline{A} = \underline{B}\,\underline{\Lambda}\,\underline{B}^T$

$\underline{A}^{-\frac{1}{2}} = \underline{B}\,\underline{\Lambda}^{-\frac{1}{2}}\,\underline{B}^T$, $\underline{\Lambda}^{-\frac{1}{2}} = \begin{bmatrix} \lambda_1^{-\frac{1}{2}} & 0 \\ 0 & \ddots \; \lambda_n^{-\frac{1}{2}} \end{bmatrix}$

  - SGD:

$$s_k = s, \quad \phi_k(\mathbf{g}_1, \ldots, \mathbf{g}_k) = \mathbf{g}_k, \quad \psi_k(\mathbf{g}_1, \ldots, \mathbf{g}_k) = \mathbf{I}$$

  - AdaGrad:

$$s_k = s, \quad \phi_k(\mathbf{g}_1, \ldots, \mathbf{g}_k) = \mathbf{g}_k, \text{ and } \psi_k(\mathbf{g}_1, \ldots, \mathbf{g}_k) = \mathrm{Diag}\Big(\sum_{t=1}^{k} \mathbf{g}_k \circ \mathbf{g}_k\Big)/k$$

Hadamard product.

  - Adam ($\beta_1 = 0$ reduces to RMSProp):

"entry-wise" product.

$$s_k = 1/\sqrt{k}, \quad \phi_k = (1 - \beta_1) \sum_{t=1}^{k} \beta_1^{k-t} \mathbf{g}_t,$$

$$\psi_k(\mathbf{g}_1, \ldots, \mathbf{g}_k) = (1 - \beta_2)\mathrm{Diag}\Big(\sum_{t=1}^{k} \beta_2^{k-t} \mathbf{g}_t \circ \mathbf{g}_t\Big).$$

AMSGrad

- A key quantity of interest in adaptive methods:

$$\mathbf{\Gamma}_{k+1} = \frac{\mathbf{V}_{k+1}^{\frac{1}{2}}}{s_{k+1}} - \frac{\mathbf{V}_{k}^{\frac{1}{2}}}{s_{k}}$$

  - Measure the change in the inverse of learning rate w.r.t. time
  - Require $\mathbf{\Gamma}_k \succeq 0$, $\forall k$, to ensure "non-increasing" learning rates
  - This is true for SGD and AdaGrad following their definitions
  - However, this is not necessarily true for Adam and RMSProp

- In [Reddi et al. ICLR'18], it was shown that for any $\beta_1, \beta_2 \in [0, 1)$ such that $\beta_1 < \sqrt{\beta_2}$, $\exists$ a stochastic convex optimization problem for which Adam does not converge to the optimal solution

- Implying that Adam needs dimension-dependent β_1 and β_2, which defeats the purpose of adaptive methods due to extensive parameter tuning!

AMSGrad

- Idea: Use a smaller learning rate and incorporate the intuition of slowly decaying the effect of past gradient as long as Γ_k is positive semidefinite
- The algorithm: In iteration k:

$$
\begin{aligned}
&\mathbf{g}_k = \nabla f_k(\mathbf{x}_k) \\
&\mathbf{m}_k = \beta_{1,k}\mathbf{m}_{k-1} + (1-\beta_{1,k})\mathbf{g}_k, \\
&\mathbf{v}_k = \beta_2\mathbf{v}_{k-1} + (1-\beta_2)\mathbf{g}_k \circ \mathbf{g}_k, \\
&\hat{\mathbf{v}}_k = \max(\hat{\mathbf{v}}_{k-1}, \mathbf{v}_k), \text{ and } \hat{\mathbf{V}}_k = \mathrm{Diag}(\hat{\mathbf{v}}_k) \\
&\mathbf{x}_{k+1} = \mathbf{x}_k - s_k \hat{\mathbf{V}}_k^{-\frac{1}{2}}\mathbf{m}_k
\end{aligned}
$$

- Maintain the maximum of all $\mathbf{v}_k$ until the present iteration and use the maximum to ensure non-increasing learning rate (i.e., $\Gamma_k \succeq 0$, $\forall k$)

Convergence of Adaptive First-Order Methods

- While faster convergence of adaptive methods over SGD has been widely observed, their best-known convergence rate bounds so far are the same (or even worse) than those of SGD $O\left(\frac{1}{(1-\beta_1)^{\alpha}}\right)$ $\alpha=1$ β_1
- We adopt the proof in [Défossez et al. '20] due to generality and simplicity
- A unified formulation used in [Défossez et al. '20] for AdaGrad and Adam $(0 < \beta_2 \leq 1$ and $0 \leq \beta_1 < \beta_2)$:

1st $\frac{1}{1-\beta_1}$ iters will be smaller than those in Adam. (e.g., if $\beta_1=0.9$, $\approx$ 50 iter), the rest are almost the same.

$$\begin{cases} \mathbf{m}_{k,i} = \beta_1 \mathbf{m}_{k-1,i} + \nabla_i f_k(\mathbf{x}_{k-1}), \\ \mathbf{v}_{k,i} = \beta_2 \mathbf{v}_{k-1,i} + (\nabla_i f_k(\mathbf{x}_{k-1}))^2, \\ \mathbf{x}_{k,i} = \mathbf{x}_{k-1,i} - s_k \dfrac{\mathbf{m}_{k,i}}{\sqrt{\delta + \mathbf{v}_{k,i}}}, \end{cases}$$

  - AdaGrad: $\beta_1 = 0$, $\beta_2 = 1$, and $s_k = s$
  - Adam: Take $s_k = s(1-\beta_1)\sqrt{\frac{1-\beta_2^k}{1-\beta_2}}$

Not exactly Adam:
1°. Drop $(1-\beta_2)$ factor in $\mathbf{v}_{k,i}$
2°. Drop $(1-\beta_1)$ factor in $\mathbf{m}_{k,i}$
3°. Add corrective term $\sqrt{1-\beta_2^k}$
4°. Drop corrective term $1-\beta_1^k$

Allows common treatment for AdaGrad and Adam.

Convergence of Adaptive First-Order Methods

- Consider a general expectation optimization problem

$$\min_{\mathbf{x}\in\mathbb{R}^d} F(\mathbf{x}) \triangleq \min_{\mathbf{x}\in\mathbb{R}^d} \mathbb{E}[f(\mathbf{x})]$$

- Notation: For a given time horizon $T \in \mathbb{N}$, let τ_T be a random index with value in $\{0,\ldots,T-1\}$ so that $\Pr[\tau_T = j] \propto 1-\beta_1^{T-j}$
  - $\beta_1 = 0$: Sampling τ_T uniformly in $\{0,\ldots,T-1\}$ (note: no momentum)
  - $\beta_1 > 0$: The fast few $\frac{1}{1-\beta_1}$ iterations are sampled relatively rarely and older iterations are sampled approximately uniformly

- Assumptions:
  - F is bounded from below: $F(\mathbf{x}) \geq F^*$, $\mathbf{x}\in\mathbb{R}^d$
  - ℓ_∞ norm of stochastic gradients is uniformly bounded almost surely: $\exists\epsilon > 0$ s.t. $\|\nabla f(\mathbf{x})\|_\infty \leq R - \sqrt{\epsilon}$ a.s.
  - L-smoothness: $\|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})\|_2 \leq L\|\mathbf{x}-\mathbf{y}\|_2$, $\forall \mathbf{x},\mathbf{y}\in\mathbb{R}^d$

Convergence of Adaptive First-Order Methods

Theorem 1 (~~AdaGrad~~ Adam w/o Momentum, i.e., AdaGrad)

Let the iterates $\{\mathbf{x}_k\}$ *be generated with* $\beta_2 = 1$, $s_k = s > 0$, *and* $\beta_1 = 0$. *Then for any* $T \in \mathbb{N}$, *we have:*

dimensional.

$$\mathbb{E}[\|\nabla F(\mathbf{x}_{\tau_T})\|^2] \leq 2R\frac{F(\mathbf{x}_0) - F^*}{s\sqrt{T}} + \frac{1}{\sqrt{T}}(4dR^2 + sdRL)\ln\left(1 + \frac{TR^2}{\epsilon}\right). = \tilde{O}\left(\frac{1}{\sqrt{T}}\right)$$

Theorem 2 (Adam w/o Momentum (RMSProp))

Let the iterates $\{\mathbf{x}_k\}$ *be generated with* $\beta_2 \in (0,1)$, $s_k = s\sqrt{\frac{1-\beta_2^k}{1-\beta_2}}$ *with* $s > 0$, *and* $\beta_1 = 0$. *Then for any* $T \in \mathbb{N}$, *we have:*

$$\mathbb{E}[\|\nabla F(\mathbf{x}_{\tau_T})\|^2] \leq 2R\frac{F(\mathbf{x}_0) - F^*}{sT} + C\left(\frac{1}{T}\ln\left(1 + \frac{R^2}{(1-\beta_2)\epsilon}\right) - \ln(\beta_2)\right),$$

$$= O\left(\frac{1}{T}\right).$$

where constant $C \triangleq \frac{4dR^2}{\sqrt{1-\beta_2}} + \frac{sdRL}{1-\beta_2}$. ← dep. d.

Convergence of Adaptive First-Order Methods

Theorem 3 (AdaGrad w/ Momentum)

Let the iterates $\{\mathbf{x}_k\}$ *be generated with* $\beta_2 = 1$, $s_k = s > 0$, *and* $\beta_1 \in (0,1)$. *Then for any* $T \in \mathbb{N}$ *such that* $T > \frac{\beta_1}{1-\beta_1}$, *we have:*

$$\mathbb{E}[\|\nabla F(\mathbf{x}_{\tau_T})\|^2] \leq 2R\sqrt{T}\frac{F(\mathbf{x}_0) - F^*}{s\tilde{T}} + \frac{\sqrt{T}}{\tilde{T}} C \ln\left(1 + \frac{TR^2}{\epsilon}\right). = \tilde{O}\left(\frac{1}{\sqrt{T}}\right)$$

where $\tilde{T} = T - \frac{\beta_1}{1-\beta_1}$ *and* $C = sdRL + \frac{12dR^2}{1-\beta_1} + \frac{2s^2dL^2\beta_1}{1-\beta_1}$.

Theorem 4 (Adam w/ Momentum)

Let $\{\mathbf{x}_k\}$ *be generated with* $\beta_2 \in (0,1)$, $\beta_1 \in [0, \beta_2)$, *and* $s_k = s(1-\beta_1)\sqrt{\frac{1-\beta_2^k}{1-\beta_2}}$ *with* $s > 0$. *Then for any* $T \in \mathbb{N}$ *such that* $T > \frac{\beta_1}{1-\beta_1}$, *we have:*

$$= O\left(\frac{1}{T}\right)$$

$$\mathbb{E}[\|\nabla F(\mathbf{x}_{\tau_T})\|^2] \leq 2R\frac{F(\mathbf{x}_0) - F^*}{sT} + C\left(\frac{1}{T}\ln\left(1 + \frac{R^2}{(1-\beta_2)\epsilon}\right) - \ln(\beta_2)\right),$$

where $\tilde{T} = T - \frac{\beta_1}{1-\beta_1}$ *and* $C = \frac{sdRL(1-\beta_1)}{(1-\frac{\beta_1}{\beta_2})(1-\beta_2)} + \frac{12dR^2\sqrt{1-\beta_1}}{(1-\frac{\beta_1}{\beta_2})^{3/2}\sqrt{1-\beta_2}} + \frac{2s^2dL^2\beta_1}{(1-\frac{\beta_1}{\beta_2})(1-\beta_2)^{3/2}}$.

Theorem 1 (~~AdaGrad~~ Adam w/o Momentum (i.e., AdaGrad))

Let the iterates $\{\mathbf{x}_k\}$ *be generated with* $\beta_2 = 1$, $s_k = s > 0$, *and* $\beta_1 = 0$. *Then for any* $T \in \mathbb{N}$, *we have:*

dimensional.

$$\mathbb{E}[\|\nabla F(\mathbf{x}_{\tau_T})\|^2] \leq 2R\frac{F(\mathbf{x}_0) - F^*}{s\sqrt{T}} + \frac{1}{\sqrt{T}}(4dR^2 + sdRL)\ln\left(1 + \frac{TR^2}{\epsilon}\right). = \tilde{O}\left(\frac{1}{\sqrt{T}}\right)$$

Theorem 2 (Adam w/o Momentum (RMSProp))

Let the iterates $\{\mathbf{x}_k\}$ *be generated with* $\beta_2 \in (0,1)$, $s_k = s\sqrt{\frac{1-\beta_2^k}{1-\beta_2}}$ *with* $s > 0$, *and* $\beta_1 = 0$. *Then for any* $T \in \mathbb{N}$, *we have:*

$$\mathbb{E}[\|\nabla F(\mathbf{x}_{\tau_T})\|^2] \leq 2R\frac{F(\mathbf{x}_0) - F^*}{sT} + C\left(\frac{1}{T}\ln\left(1 + \frac{R^2}{(1-\beta_2)\epsilon}\right) - \ln(\beta_2)\right),$$

$= O\left(\frac{1}{T}\right).$

where constant $C \triangleq \frac{4dR^2}{\sqrt{1-\beta_2}} + \frac{sdRL}{1-\beta_2}$. $\leftarrow$ dep. d.

Proof of Thm 1 & Thm 2:

step 1°: Establish the correlation bound btwn adaptive dir. and grad dir.

step 2°: Start from descent lemma $\Rightarrow$ bnd each iter descent.

$\Rightarrow$ telescoping $\Rightarrow$ bnd $\|\nabla F(x_{\tau_T})\|^2$ (interesting result to bnd "sum-of-ratios").

Notation: $\mathbb{E}_{k-1}[\cdot] \triangleq \mathbb{E}_{k-1}[\cdot \mid f_1(x_1), \cdots f_{k-1}(x_{k-1})]$

$$v_{k,i} = \beta_2 v_{k,i} + (\nabla_i f_k(x_{k-1}))^2$$

$$x_{k,i} = x_{k-1,i} - s_k \frac{\nabla_i f_k(x_{k-1})}{\sqrt{\delta + v_{k,i}}}$$

$$\tilde{v}_{k,i} = \beta_2 v_{k,i} + \mathbb{E}_{k-1}\left[(\nabla_i f_k(x_{k-1}))^2\right]$$

Lemma 1 (adaptive update approx. a descent dir.).

For all $k \in \mathbb{N}$ and $i \in [d]$ $\{1,\dots,d\}$, we have:

$$\mathbb{E}_{k-1}\left[\nabla_i F(x_{k-1}) \cdot \frac{\nabla_i f_k(x_{k-1})}{\sqrt{\delta + v_{k,i}}}\right] \geq \frac{(\nabla_i F(x_{k-1}))^2}{2\sqrt{\delta + \tilde{v}_{k,i}}} - 2R\, \mathbb{E}_{k-1}\left[\frac{(\nabla_i f_k(x_{k-1}))^2}{\delta + v_{k,i}}\right]$$

$\underline{v}^{-\frac{1}{2}} g$

Proof: For notational simplicity: Let $G = \nabla_i F(x_{k-1})$, $g = \nabla_i f_k(x_{k-1})$.

$v = \underline{v}_{k,i}$, $\tilde{v} = \underline{\tilde{v}}_{k,i}$, $\forall k, i$.

$$\mathbb{E}_{k-1}\left[\frac{Gg}{\sqrt{\delta+v}}\right] \overset{\substack{\text{add \& subtract} \\ \tilde{v}}}{=} \underbrace{\mathbb{E}_{k-1}\left[\frac{Gg}{\sqrt{\delta+\tilde{v}}}\right]}_{A} + \mathbb{E}_{k-1}\left[\underbrace{Gg\left(\frac{1}{\sqrt{\delta+v}} - \frac{1}{\sqrt{\delta+\tilde{v}}}\right)}_{B}\right] \quad (0).$$

Note that g and $\tilde{v}$ are cond. indep. given $f_1(x_1) \cdots f_{k-1}(x_{k-1})$, we have

$$\mathbb{E}_{k-1}\left[\frac{Gg}{\sqrt{\delta+\tilde{v}}}\right] = G\, \underbrace{\mathbb{E}_{k-1}[g]}_{G}\; \underbrace{\mathbb{E}_{k-1}\left[\frac{1}{\sqrt{\delta+\tilde{v}}}\right]}_{\text{itself}} = \frac{G^2}{\sqrt{\delta+\tilde{v}}}. \quad (1)$$

Next, to bnd B, we have:

$$B = Gg\left(\frac{1}{\sqrt{\delta+v}} - \frac{1}{\sqrt{\delta+\tilde{v}}}\right) = Gg\, \frac{\sqrt{\delta+\tilde{v}} - \sqrt{\delta+v}}{\sqrt{\delta+v}\,\sqrt{\delta+\tilde{v}}} = Gg\, \frac{\mathbb{E}_{k-1}[g^2] - g^2}{\sqrt{\delta+v}\,\sqrt{\delta+\tilde{v}}\,(\sqrt{\delta+v} + \sqrt{\delta+\tilde{v}})}$$

$|a-b| \leq |a| + |b|$

$$\leq |Gg|\, \frac{\mathbb{E}_{k-1}[g^2] + g^2}{\sqrt{\delta+v}\,\sqrt{\delta+\tilde{v}}\,(\sqrt{\delta+v} + \sqrt{\delta+\tilde{v}})}$$

$$\leq \underbrace{|Gg|\, \frac{\mathbb{E}_{k-1}[g^2]}{\sqrt{\delta+v}\,\sqrt{\delta+\tilde{v}}\,(\sqrt{\delta+v} + \sqrt{\delta+\tilde{v}})}}_{C} + \underbrace{|Gg|\, \frac{g^2}{\sqrt{\delta+v}\,\sqrt{\delta+\tilde{v}}\,(\sqrt{\delta+v} + \sqrt{\delta+\tilde{v}})}}_{D}.$$

1° For C: $C \le \frac{G^2}{4\sqrt{\delta+\tilde{v}}} + \frac{g^2 \mathbb{E}_{k-1}[g^2]^2}{(\delta+\tilde{v})^{3/2}(\delta+v)}$

$\left(\begin{array}{l} \text{Young's Ineq:} \\ ab \le \frac{\lambda a^2}{2} + \frac{b^2}{\lambda 2} \\ \lambda = \frac{\sqrt{\delta+\tilde{v}}}{2},\ a = \frac{|G|}{\sqrt{\delta+\tilde{v}}} \\ b = \frac{|g|\,\mathbb{E}_{k-1}[g^2]}{\sqrt{\delta+v}\,\sqrt{\delta+\tilde{v}}} \end{array} \right)$

Take cond. expectation, noting $\underline{\delta+\tilde{v} \ge \mathbb{E}_{k-1}[g^2]}$

$$\mathbb{E}_{k-1}[C] \le \frac{G^2}{4\sqrt{\delta+\tilde{v}}} + \underbrace{\frac{\mathbb{E}_{k-1}[g^2]}{\sqrt{\delta+\tilde{v}}}}_{\le R} \cdot \underbrace{\frac{\mathbb{E}_{k-1}[g^2]}{\delta+\tilde{v}}}_{\le 1} \cdot \mathbb{E}_{k-1}\left[\frac{[g^2]}{\delta+v}\right]$$

Also, since $\sqrt{\mathbb{E}_{k-1}[g^2]} \le \sqrt{\delta+\tilde{v}}$ and $\sqrt{\mathbb{E}_{k-1}[g^2]} \le R$.

So, we have: $\mathbb{E}_{k-1}[C] \le \frac{G^2}{4\sqrt{\delta+\tilde{v}}} + R\,\mathbb{E}_{k-1}\left[\frac{g^2}{\delta+v}\right]$ (2)

2° For D:

$D \overset{\text{Young's Ineq}}{\le} \frac{G^2}{4\sqrt{\delta+\tilde{v}}} \cdot \frac{g^2}{\mathbb{E}_{k-1}[g^2]} + \frac{\mathbb{E}_{k-1}[g^2]}{\sqrt{\delta+\tilde{v}}} \cdot \frac{g^4}{(\sqrt{\delta+v})^2}$

$\left(\begin{array}{l} \lambda = \frac{\sqrt{\delta+\tilde{v}}}{2\mathbb{E}_{k-1}[g^2]} \\ a = \frac{|Gg|}{\sqrt{\delta+\tilde{v}}} \\ b = \frac{g^2}{\delta+v}. \end{array} \right)$

Take cond. expection and note $\delta+v \ge g^2$. We have:

$$\mathbb{E}_{k-1}[D] \le \frac{G^2}{4\sqrt{\delta+\tilde{v}}} + \frac{\mathbb{E}_{k-1}[g^2]}{\sqrt{\delta+\tilde{v}}} \cdot \mathbb{E}_{k-1}\left[\frac{\cancel{g^4}\, g^2}{\delta+v}\right]$$

Using the same argument as in steps in "C", we have:

$\mathbb{E}_{k-1}[D] \le \frac{G^2}{4\sqrt{\delta+\tilde{v}}} + R\,\mathbb{E}_{k-1}\left[\frac{g^2}{\delta+v}\right]$ (3).

Adding (2) — (3) yields:

$\mathbb{E}_{k-1}[|B|] \le \frac{G^2}{2\sqrt{\delta+\tilde{v}}} + 2R\,\mathbb{E}_{k-1}\left[\frac{g^2}{\delta+v}\right]$ (4).

Plugging (4) and (1) into (0):

$$\mathbb{E}_{k-1}\left[\frac{Gg}{\sqrt{\delta+v}}\right] = \frac{G^2}{\sqrt{\delta+\tilde{v}}} + \mathbb{E}_{k-1}[|B|] \geq \frac{G^2}{\sqrt{\delta+\tilde{v}}} - \left[\frac{G^2}{2\sqrt{\delta+\tilde{v}}} + 2R\,\mathbb{E}_{k-1}\left[\frac{g^2}{\delta+v}\right]\right]$$

$$= \frac{G^2}{2\sqrt{\delta+\tilde{v}}} - 2R\,\mathbb{E}_{k-1}\left[\frac{g^2}{\delta+v}\right].$$

Proof of Lemma 1 is complete. $\blacksquare$

Proof of Thm 1 (AdaGrad):

Since $F(\cdot)$ is L-smooth, from the descent lemma:

$$F(x_{k+1}) \leq F(x_k) - s\nabla F(x_k)^T \underbrace{(x_{k+1}-x_k)}_{\triangleq\, u_k} + \frac{s^2 L}{2}\|x_{k+1}-x_k\|^2.$$

$$u_k = \frac{\nabla f(x_{k-1})}{\sqrt{\delta+\tilde{v}_k}}$$

Take cond. expectation and use Lemma 1:

$$\mathbb{E}_{k-1}[F(x_k)] \leq F(x_{k-1}) - s\nabla F(x_{k-1})^T\,\mathbb{E}_{k-1}\left[\frac{\nabla f_k(x_{k-1})}{2\sqrt{\delta+\tilde{v}_{k,i}}}\right] + \left(2sR + \frac{s^2 L}{2}\right)\mathbb{E}\left[\|u_k\|^2\right]. \quad (5)$$

Since the a.s. l_∞ bnd on grad, we have:

$$\sqrt{\delta+\tilde{v}_{k,i}} = \sqrt{\delta+\sum_{t=0}^{k-1} v_{t,i}^2} \leq \sqrt{\delta+R^2\cdot k} \leq R\sqrt{k}$$

Thus,
$$\frac{s(\nabla F_i(x_{k-1}))^2}{2\sqrt{\delta+\tilde{v}_{k,i}}} \geq \frac{s(\nabla_i F(x_{k-1}))^2}{2R\sqrt{k}}. \quad (6)$$

Plugging (6) into (5), we have:

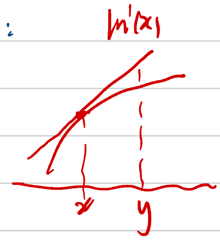
$$\mathbb{E}_{k-1}[F(x_k)] \le F(x_{k-1}) - \frac{s}{2R\sqrt{k}}\|\nabla F(x_{k-1})\|_2^2 + \left(2sR + \frac{s^2L}{2}\right)\mathbb{E}_{k-1}\left[\|u_k\|_2^2\right]$$

Summing this ineq. for all $k \in [T]$ $\{1 \cdots T\}$, taking full expectation, using $\sqrt{k} \le \sqrt{T}$, we have:

$$\mathbb{E}[F(x_T)] \le F(x_0) - \frac{s}{2R\sqrt{T}}\sum_{k=0}^{T-1}\mathbb{E}\left[\|\nabla F(x_k)\|_2^2\right] + \left(2sR + \frac{s^2L}{2}\right)\sum_{k=0}^{T-1}\mathbb{E}\left[\|u_k\|^2\right]$$

Lemma 2 (Sum of ratios w/ denominator being exp. avg of the history).
Suppose $0 < \beta_2 \le 1$. Consider a non-neg. seq. $\{a_k\}$. Let $b_k = \sum_{t=1}^{k}\beta_2^{k-t}a_t$. We have: $\sum_{t=1}^{T}\frac{a_t}{\delta + b_t} \le \ln\left(1 + \frac{b_T}{\delta}\right) - T\ln(\beta_2)$.

Proof of Lemma 2: Since $\ln(\cdot)$ is concave, it holds that:



$$\ln(y) \le \ln(x) + \ln'(x)(y-x) = \ln(x) + \frac{y-x}{x}$$

$$\Rightarrow \frac{x-y}{x} \le \ln(x) - \ln(y) \overset{\text{let } m = x-y}{\Longrightarrow} \frac{m}{m+y} \le \ln(m+y) - \ln(y).$$

Take $m = a_t$, $m+y = \delta + b_t \Rightarrow y = \delta + b_t - a_t$

$$\frac{a_t}{\delta + b_t} \le \ln(\delta + b_t) - \ln(\delta + b_t - a_t).$$

$$\overset{\text{def of } b_t}{=} \ln(\delta + b_t) - \ln(\delta + \beta_2 b_{t-1})$$

$$= \underbrace{\ln\left(\frac{\delta + b_t}{\delta + b_{t-1}}\right)}_{\text{telescoping series}} + \underbrace{\ln\left(\frac{\delta + b_{t-1}}{\delta + \beta_2 b_{t-1}}\right)}_{\approx -\ln\beta_2}$$

Summing over all $t \in [T]$ yields:

$$\sum_{t=1}^{T} \frac{a_t}{\delta + b_t} \le \ln\left(1 + \frac{b_T}{\delta}\right) - T\ln(\beta_2). \quad \blacksquare$$

(Continue on Thm 1):

Bounding the last term on the RHS and using Lemma 2 for each dimension, and rearranging terms, we arrive at the final result. $\blacksquare$

Proof of Thm 2: (Adam w/o Momentum, aka RMSProp):

Recall $s_k = s\sqrt{\frac{1-\beta_2^k}{1-\beta_2}}$, for some $s>0$. From L-smoothness and descent Lemma:

$$F(\underline{x}_k) \le F(\underline{x}_{k-1}) - s_k \nabla F(\underline{x}_{k-1})^T \underline{u}_k + \frac{s_k^2 L}{2}\|\underline{u}_k\|^2. \quad (7).$$

From a.s. l_∞ bnd on grad assump, we have:

$$\sqrt{\delta + \tilde{v}_{k,i}} \le R\sqrt{\sum_{t=0}^{k-1}\beta_2^t} \overset{\text{geometric series}}{=} R\sqrt{\frac{1-\beta_2^k}{1-\beta_2}}.$$

Thus, $$s_k \frac{(\nabla_i F(\underline{x}_{k-1}))^2}{2\sqrt{\delta + \hat{v}_{k,i}}} \ge s\cancel{\sqrt{\frac{1-\beta_2^k}{1-\beta_2}}} \cdot \frac{(\nabla_i F(\underline{x}_{k-1}))^2}{2R\cancel{\sqrt{\frac{1-\beta_2^k}{1-\beta_2}}}} = \frac{s(\nabla_i F(\underline{x}_{k-1}))^2}{2R}. \quad (8).$$

Taking cond. expectation w.r.t. $f_0(\underline{x}_0) \cdots f_{k-1}(\underline{x}_{k-1})$ on both sides of (7), applying Lemma 1, using (8), we have:

$$\mathbb{E}_{k-1}\left[F(x_k)\right] \le F(x_{k-1}) - \frac{s}{2R}\|\nabla F(x_{k-1})\|_2^2 + \left(2s_k R + \frac{s_k^2 L}{2}\right)\mathbb{E}_{k-1}\left[\|u_k\|_2^2\right]$$

Since $\beta_2 \le 1$, we have $s_k \le \frac{s}{\sqrt{1-\beta_2}}$. Summing the above ineq. and taking full expectation yields:

$$\mathbb{E}\left[F(x_T)\right] \le F(x_0) - \frac{s}{2R}\sum_{k=0}^{T-1}\mathbb{E}\left[\|\nabla F(x_k)\|_2^2\right] + \left(\frac{2sR}{\sqrt{1-\beta_2}} + \frac{s^2 L}{2(1-\beta_2)}\right)\sum_{k=0}^{T-1}\mathbb{E}\left[\|u_k\|_2^2\right]$$

Applying Lemma 2 and rearranging terms arrives at the stated result. $\blacksquare$

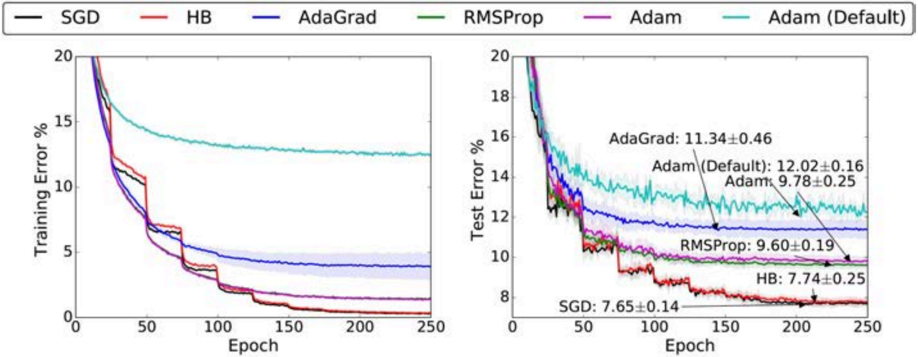
Theoretical Understanding of Adaptive Methods

- Pros:
  - [Zhang et al. NeurIPS'20]: Adam performs better than SGD when stochastic gradients are heavy-tailed since Adam does an "adaptive gradient clipping"
  - [Zhang et al. NeurIPS'20]: Also shows that SGD can fail to converge under heavy-tailed situations, while clipped-SGD can.
  - [Goodfellow & Bengio, '16]: Clipped-SGD works better than SGD in vicinity of extremely steep cliffs
  - [Zhang et al. ICML'20]: Clipped-GD converges without L-smoothness (with rate ϵ^{-2} while GD may converge arbitrarily slower

- Cons:
  - [Wilson et al. NeurIPS'17]: While converging faster in general, adaptive first-order methods does not have good test error and generalization performances in the over-parameterized regime. Adaptive methods often generalize significantly worse than SGD. So one may need to reconsider the use of adaptive methods to train deep neural networks

Limitations of Adaptive Methods

- [Wilson et al. NeurIPS'17]: VGG+BN+Dropout network for CIFAR-10



(a) CIFAR-10 (Train) **(b)** CIFAR-10 (Test)

Next Class

Federated and Decentralized Optimization