ECE 8101: Nonconvex Optimization for Machine Learning

Lecture Note 2-6: Adaptive First-Order Methods

Jia (Kevin) Liu

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

Autumn 2024

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 hyper-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

Bilevel Opt.

- 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 $\mathbf{a}_k$)

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

mono. ↑ as k ↑

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

$\sum g_{t,i}$: big ⇒ ss small
small ⇒ ss big

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

RMSProp

- A major limitation of AdaGrad:
  - Step-sizes could be 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

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

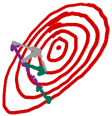
$$\mathbf{x}_{k+1,i} = \mathbf{x}_{k,i} - \frac{s_k}{(\delta + \mathbb{E}[\mathbf{g}_{k+1,i}^2])^{\frac{1}{2}}} \nabla_i f(\mathbf{x}_k).$$

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

Adam

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

HB-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 :

$$\mathbf{g}_k = \nabla f_k(\mathbf{x}_k)$$
$$\mathbf{m}_k = \phi_k(\mathbf{g}_1, \ldots, \mathbf{g}_k), \text{ and } \mathbf{V}_k = \psi_k(\mathbf{g}_1, \ldots, \mathbf{g}_k)$$
$$\boxed{\mathbf{x}_{k+1} = \mathbf{x}_k - s_k \mathbf{V}_k^{-\frac{1}{2}} \mathbf{m}_k}$$

  - 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:

  Hadamard.
  $$s_k = s, \quad \phi_k(\underline{\mathbf{g}_1, \ldots, \mathbf{g}_k}) = \mathbf{g}_k, \text{ and } \psi_k(\mathbf{g}_1, \ldots, \mathbf{g}_k) = \mathrm{Diag}(\sum_{t=1}^{k} \mathbf{g}_k \circ \mathbf{g}_k)/k$$
  - Adam ($\beta_1 = 0$ reduces to RMSProp):
  $$s_k = 1/\sqrt{k}, \quad \phi_k = (1-\beta_1)\sum\nolimits_{t=1}^{k} \beta_1^{k-t}\mathbf{g}_t,$$
  $$\psi_k(\mathbf{g}_1, \ldots, \mathbf{g}_k) = (1-\beta_2)\mathrm{Diag}(\sum\nolimits_{t=1}^{k} \beta_2^{k-t}\mathbf{g}_t \circ \mathbf{g}_t).$$

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
- 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 $0 \leq \beta_1 < \beta_2$):

$$\mathbf{m}_{k,i} = \beta_1 \mathbf{m}_{k-1,i} + \nabla_i f_k(\mathbf{x}_{k-1}),$$

$$\mathbf{v}_{k,i} = \beta_2 \mathbf{v}_{k-1,i} + (\nabla_i f_k(\mathbf{x}_{k-1}))^2,$$

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

1st: $\frac{1}{1-\beta_1}$ terms will small then in Adam [e.g., $\beta_1 = 0.9 \approx 50$ iter.) b/c $\beta_1^k \to 0$

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

1° Drop $1-\beta_2$ in $\underline{\mathbf{v}}_{k,i}$
2° Drop $1-\beta_1$ in $\underline{\mathbf{m}}_{k,i}$
3° Add corrective term $\sqrt{1-\beta_2^k}$ to SS.
4° Drop corrective term $1-\beta_1^k$

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: $\|\nabla F(\mathbf{x}) - F(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2$, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$

Convergence of Adaptive First-Order Methods

= AdaGrad.

Theorem 1 (~~AdaGrad~~ Adam w/o Momentum)

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:

$\tilde{O}(\frac{1}{\sqrt{T}})$

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

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:

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

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

= Ada...

Theorem 1 (~~AdaGrad~~ Adam w/o Momentum)

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: $\tilde{O}(\frac{1}{\sqrt{T}})$

$$\mathbb{E}[\|\nabla F(\mathbf{x}_{\tau_T})\|^2] \le 2R\frac{F(\mathbf{x}_0) - F^*}{s\sqrt{T}} + \overbrace{\frac{1}{\sqrt{T}}}^{=O(\frac{1}{\sqrt{T}})}(4\underline{d}R^2 + s\underline{d}RL)\,\underline{\ln}\left(1 + \frac{TR^2}{\epsilon}\right).$$

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: $=O(\frac{1}{T})$,

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

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

Proof. Step 1: Establish correlation bnd btwn adaptive dir and true grad dir, s.t. ensure enough descent.

Step 2: start some "descent lemma", $\Rightarrow$ bnd per-iter descent $\Rightarrow$ telescoping $\Rightarrow$ bnd $\|\nabla F(x_{\tau_T})\|$.

$\mathbb{E}_{k-1}[\cdot] \triangleq \mathbb{E}[\cdot \mid f_1, \dots, f_{k-1}]$

Lemma (Adaptive update is approx descent dir): For $k \in \mathbb{N}$ and $i \in [d] = \{1, \dots, d\}$, we have:

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

where: $\underline{v}_{k,i} = \beta_2 \underline{v}_{k-1,i} + (\nabla_i f_k(x_{k-1}))^2$.

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

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

For notation simplicity, let $G \triangleq \nabla_i F(x_{k-1})$, $g \triangleq \nabla_i f_k(x_{k-1})$.

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

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

Note that g and $\tilde{v}$ are indep give $f_1(x_1) \cdots f_{k-1}(x_{k-1})$.

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

Next, to bnd B, we have:

$$B = Gg\,\frac{\overbrace{\cancel{(\sqrt{\delta+\tilde{v}} - \sqrt{\delta+v})(\sqrt{\delta+\tilde{v}} + \sqrt{\delta+v})}}^{\tilde{v}-v}}{\sqrt{\delta+v}\,\sqrt{\delta+\tilde{v}}\,(\sqrt{\delta+\tilde{v}} + \sqrt{\delta+v})}$$

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

$|a-b| \le |a| + |b|$.

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

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

$$\left(ab \le \frac{\lambda}{2}a^2 + \frac{b^2}{2\lambda} \quad \lambda = \frac{\sqrt{\delta+\tilde{v}}}{2},\; a = \frac{|G|}{\sqrt{\delta+\tilde{v}}} \quad b = \frac{|g|\,\mathbb{E}_{k-1}[g^2]}{\sqrt{\delta+\tilde{v}}\,\sqrt{\delta+v}} \right)$$

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

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

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

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

2° For D:

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

Young's Ineq

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

Taking cond expectation, and noting $\delta+v \ge g^2$, we have

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

Using the same argument as in (2), we have:

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

Adding (2) and (3) yields:

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

$B \ge -[\,\cdot\,]$ (5).

Plugging (5) and (1) into (0):

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

Proof of Thm 1 (AdaGrad).

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

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

$$\underline{u}_{k-1} \leftarrow \frac{\nabla f(\underline{x}_k)}{\sqrt{\delta + \hat{v}_k}}$$

Take cond. exp w.r.t. $f_0(\underline{x}_0), \dots, f_{k-1}(\underline{x}_{k-1})$, and applying Lemma 1

$$\mathbb{E}_{k-1}[F(\underline{x}_k)] \overset{\text{Lemma 1}}{\le} F(\underline{x}_{k-1}) - s \nabla F(\underline{x}_{k-1})^T \begin{bmatrix} \vdots \\ \dfrac{\nabla_i F_i(\underline{x}_{k-1})}{2\sqrt{\delta + \hat{v}_{k-1,i}}} \\ \vdots \end{bmatrix} + \left(2sR + \frac{s^2 L}{2}\right) \mathbb{E}_{k-1}\left[\|\underline{u}_{k-1}\|^2\right] \qquad (5)$$

Since the a.s. l_∞ bound on grad (Assup), we have

$$\sqrt{\delta + \underbrace{\tilde{v}_{k-1,i}}_{= \beta \tilde{v} + g(\cdot)}} \le \sqrt{\delta + R^2 \cdot (k-1)} \le R\sqrt{k}$$

Thus $$\frac{1}{2} s \nabla_i F_i(\underline{x}_{k-1}) u_{k-1,i} = \frac{(\nabla_i F_i(\underline{x}_{k-1}))^2}{2\sqrt{\delta + \hat{v}_{k-1,i}}} \ge \frac{s(\nabla_i F(\underline{x}_{k-1}))^2}{2R\sqrt{k}} \qquad (6)$$

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

$$\mathbb{E}_{k-1}[F(\underline{x}_k)] \le F(\underline{x}_{k-1}) - \frac{s}{2R\sqrt{k}} \|\nabla F(\underline{x}_{k-1})\|^2 + \left(2sR + \frac{s^2 L}{2}\right) \mathbb{E}_{k-1}\left[\|\underline{u}_{k-1}\|^2\right]$$

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

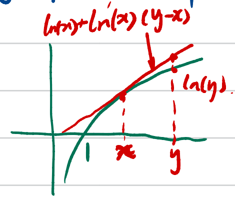
$$\mathbb{E}[F(\underline{x}_T)] \le F(\underline{x}_0) - \frac{s}{2R\sqrt{T}} \sum_{k=0}^{T-1} \mathbb{E}\left[\|\nabla F(\underline{x}_k)\|^2\right] + \left(2sR + \frac{s^2 L}{2}\right) \sum_{k=0}^{T-1} \mathbb{E}\left[\|\underline{u}_{k-1}\|^2\right]$$

To analyze (Δ), we first prove the following:

Lemma 2 (Sum of ratio w/ denominator take from history):

Suppose $0 < \beta_2 \le 1$. Consider a non-neg. seq. $\{a_k\}$. Let

$b_k \triangleq \sum_{t=1}^{k-t} \beta_2^{k-t} a_t$. 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. Since $\ln(\cdot)$ is concave, we have

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

$$\Rightarrow \frac{x-y}{x} \le \ln(x) - \ln(y).$$

Take $x = \delta + b_t$, $y = \delta + b_t - a_t$. Then, we have:

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

$$\overset{\text{def. of } b_t}{=} \underbrace{\ln(\delta + b_t) - \ln(\delta + \beta_2 b_{t-1})}_{\pm \ln(\delta + b_{t-1})} = \ln\left(\frac{\delta + b_t}{\delta + b_{t-1}}\right) - \overbrace{\ln\left(\frac{\delta + b_{t-1}}{\delta + \beta_2 b_{t-1}}\right)}^{\approx -\ln\beta_2}.$$

Bounding last term (Δ) in RHS using Lemma 2 for each dimension and rearranging terms arrives at the final result. $\blacksquare$

Proof of Thm 2 (Adam w/o Momentum, a.k.a RMSProp).

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

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

From a.s. l_∞ bound on grad assumption:

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

Thus, $$s_k \frac{(\nabla_i F(\underline{x}_{k-1}))^2}{2\sqrt{\delta+\tilde{v}_{k-1,i}}} \ge \frac{s(\nabla_i F(\underline{x}_{k-1}))^2}{2R}. \qquad (8).$$

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 and (8):

$$\mathbb{E}_{k-1}[F(\underline{x}_k)] \le F(\underline{x}_{k-1}) - \frac{s}{2R}\|\nabla F(\underline{x}_{k-1})\|^2 + \left(2s_k R + \frac{s_k^2 L}{2}\right)\mathbb{E}_{k-1}[\|\underline{u}_{k-1}\|^2].$$

Note that $s_k = s\sqrt{\frac{1-\cancel{\beta_2^k}}{1-\beta_2}} \le \frac{s}{\sqrt{1-\beta_2}}$. Summing the ineq above and taking full expectation:

$$\mathbb{E}[F(\underline{x}_T)] \le F(\underline{x}_0) - \frac{s}{2R}\sum_{k=0}^{T-1}\mathbb{E}[\|\nabla F(\underline{x}_k)\|_2^2] + \left(\frac{2sR}{\sqrt{1-\beta_2}} + \frac{s^2 L}{2(1-\beta_2)}\right)\sum_{k=0}^{T-1}\mathbb{E}[\|\underline{u}_{k-1}\|_2^2]$$

Applying Lemma 2 and rearranging arrive at the stated result. ∎

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:*

$= O(\frac{1}{\sqrt{T}}) \geq$

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

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:*

$= O(\frac{1}{T}) \geq$ "+"

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

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

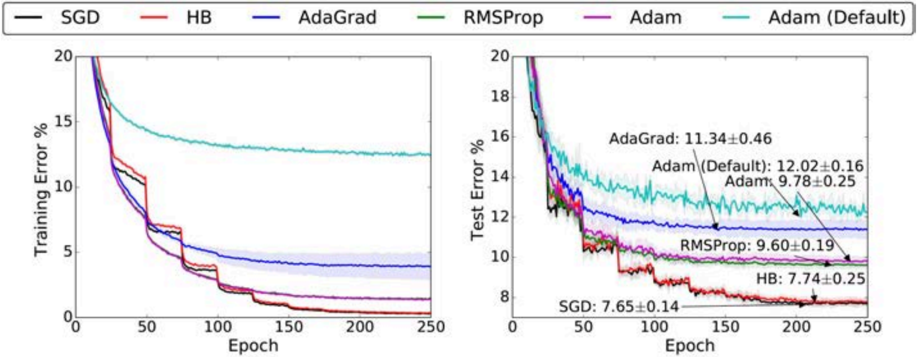
Theoretical Understanding of Adaptive Methods

- 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