Adaptive gradient methods for stochastic and online optimization

Based on the paper "On the convergence of Adam and beyond" by S. Reddi, S. Kale, S. Kumar (ICLR 2018)

Anton Rodomanov

16 February 2018

Seminar on Bayesian methods in machine learning

Stochastic gradient method (SGD)

Problem: Find $x^* \in Q$ such that $f(x^*) = \min_{x \in Q} f(x)$.

1. Q is a nonempty convex compact set in $\mathbb{R}^n$;
2. $f : Q \to \mathbb{R}$ is a subdifferentiable convex function.

Stochastic gradient method (SGD)

1. Choose a starting point $x_1 \in Q$;
2. Iterate for $t = 1, 2, \ldots$:
   (a) Choose a random g_t in $\mathbb{R}^n$ such that $\mathbb{E}(g_t \mid x_t) \in \partial f(x_t)$.
   (b) Set $x_{t+1} := \pi_Q(x_t - \alpha_t g_t)$ for some $\alpha_t \geq 0$.

Theorem (Convergence rate of SGD)

Suppose there exists $M \geq 0$ such that $\mathbb{E}\|g_t\|^2 \leq M^2$ for all $t \geq 1$. Let $D := \sup_{x,y \in Q} \|x - y\|$ be the diameter of Q, and let $\alpha_t := \frac{D}{M\sqrt{t}}$ for all $t \geq 1$. Also let $T \geq 1$, and let $\overline{x}_T := \frac{1}{T} \sum_{k=1}^{T} x_k$. Then

$$\mathbb{E} f(\overline{x}_T) - f(x^*) \leq \frac{3DM}{2\sqrt{T}}.$$

The AdaGrad method [Duchi et al., 2011]

Problem: Find $x^* \in Q$ such that $f(x^*) = \min_{x \in Q} f(x)$.

AdaGrad

1. Choose a starting point $x_1 \in Q$;
2. Iterate for $t = 1, 2, \ldots$:
   (a) Choose a random g_t in $\mathbb{R}^n$ such that $\mathbb{E}(g_t \mid x_t) \in \partial f(x_t)$.
   (b) Set $x_{t+1} := \pi_{Q,B_t}(x_t - \alpha B_t^{-1} g_t)$ for some $\alpha \geq 0$ and $B_t := \mathrm{Diag}(\sum_{k=1}^{t} g_k^2)^{1/2}$.

Theorem (Convergence rate of AdaGrad)

Suppose there exists $M_1, \ldots, M_n \geq 0$ *such that* $\mathbb{E} g_{t,j}^2 \leq M_j^2$ *for all* $1 \leq j \leq n$ *and all* $t \geq 1$. *Let* $D_\infty := \sup_{x,y \in Q} \|x - y\|_\infty$ *be the* l^∞ *diameter of* Q, *and let* $\alpha := D_\infty$. *Also let* $T \geq 1$, *and let* $\overline{x}_T := \frac{1}{T} \sum_{k=1}^{T} x_k$. *Then*

$$\mathbb{E} f(\overline{x}_T) - f(x^*) \leq \frac{3 D_\infty}{2\sqrt{T}} \sum_{j=1}^{n} M_j.$$

Compare SGD and AdaGrad

SGD: $\frac{3DM}{2\sqrt{T}}$. **AdaGrad:** $\frac{3D_\infty}{2\sqrt{T}} \sum_{j=1}^{n} M_j$,

where $\mathbb{E}\|g\|^2 \le M^2$, $\mathbb{E}g_j^2 \le M_j^2$.

- AdaGrad is efficient when $D_\infty \sum_{j=1}^n M_j \le DM$.
- Usually we have $M = (\sum_{j=1}^n M_j^2)^{1/2}$.
- By Cauchy-Schwarz, $\sum_{j=1}^n M_j \le \sqrt{n}M$.
- AdaGrad is efficient when $\sqrt{n}D_\infty \le D$ (e.g. $Q = [-1,1]^n$).

For the case $\sqrt{n}D_\infty = D$ (e.g. $Q = [-1,1]^n$),

AdaGrad is faster than SGD in $1 \le \frac{\sqrt{n}M}{\sum_{j=1}^n M_j} \le \sqrt{n}$ times.

Example: Robust regression

Let $Q := [-1,1]^n$, and let $f : Q \to \mathbb{R}$ be the function

$$f(x) := \frac{1}{m}\sum_{i=1}^{m} |\langle a_i, x\rangle - b_i|,$$

where $a_1, \ldots, a_m \in \mathbb{R}^n$, $b_1, \ldots, b_m \in \mathbb{R}$.

- At $x \in Q$, the vector
$$g := \operatorname{sign}(\langle a_{\hat{i}}, x\rangle - b_{\hat{i}})a_{\hat{i}}, \qquad \hat{i} \equiv \operatorname{Unif}\{1, \ldots, m\}$$
satisfies $\mathbb{E}g \in \partial f(x)$.
- For any $1 \le j \le n$, one has
$$\mathbb{E}g_j^2 = \frac{1}{m}\sum_{i=1}^{m} \operatorname{sign}(\langle a_i, x\rangle - b_i)^2 a_{i,j}^2 \le \frac{1}{m}\sum_{i=1}^{m} a_{i,j}^2 = M_j^2.$$
- Thus, M_j is the mean square in the jth column of A.

AdaGrad is efficient when A has many columns with small mean squares (e.g. sparse columns).

The Adam method [Kingma, Ba, 2014]

Adam

1. Choose a starting point $x_1 \in Q$; set $m_0 := 0$; $v_0 := 0$;
2. Iterate for $t = 1, 2, \ldots$:
   (a) Choose a random g_t in $\mathbb{R}^n$ such that $\mathbb{E}(g_t \mid x_t) \in \partial f(x_t)$.
   (b) Set $m_t := \beta_1 m_{t-1} + (1-\beta_1) g_t$ for some $0 \le \beta_1 < 1$;
   (c) Set $v_t := \beta_2 v_{t-1} + (1-\beta_2) g_t^2$ for some $0 \le \beta_2 < 1$;
   (d) Set $x_{t+1} := \pi_{Q,B_t}(x_t - \alpha_t B_t^{-1} m_t)$ for $B_t := \mathrm{Diag}(v_t)^{1/2}$ and some $\alpha_t \ge 0$.

- Note that

$$m_t = (1-\beta_1)\sum_{k=1}^{t} \beta_1^{t-k} g_k, \qquad v_t = (1-\beta_2)\sum_{k=1}^{t} \beta_2^{t-k} g_k^2.$$

- Compare with the **heavy ball method**

$$x_{t+1} := x_t - \alpha g_t + \beta(x_t - x_{t-1})$$

which can be written as

$$x_{t+1} = x_t - \alpha \sum_{k=1}^{t} \beta^{t-k} g_k.$$

Correcting Adam [Reddi et al., 2018]

The proof technique of SGD/AdaGrad/Adam exploits the monotonicity

$$\frac{B_t}{\alpha_t} \succeq \frac{B_{t-1}}{\alpha_{t-1}}.$$

AMSGrad

1. Choose a starting point $x_1 \in Q$.
2. Set $m_0 := 0$; $v_0 := 0$; $\hat{v}_0 := 0$.
3. Iterate for $t = 1, 2, \ldots$:
   (a) Choose a random g_t in $\mathbb{R}^n$ such that $\mathbb{E}(g_t \mid x_t) \in \partial f(x_t)$.
   (b) Set $m_t := \beta_{1t} m_{t-1} + (1 - \beta_{1t}) g_t$ for some $0 \leq \beta_{1t} < 1$;
   (c) Set $v_t := \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$ for some $0 \leq \beta_2 < 1$;
   (d) Set $\hat{v}_t := \max\{\hat{v}_{t-1}, v_t\}$;
   (e) Set $x_{t+1} := \pi_{Q,B_t}(x_t - \alpha_t B_t^{-1} m_t)$ for $B_t := \mathrm{Diag}(\hat{v}_t)^{1/2}$ and some $\alpha_t \geq 0$.

Convergence rate of AMSGrad [ICLR 2018 paper]

Theorem (Convergence rate of AMSGrad)

Suppose there exists $M_1, \ldots, M_n \geq 0$ *such that* $\mathbb{E} g_{t,j}^2 \leq M_j^2$ *for all* $1 \leq j \leq n$ *and all* $t \geq 1$. *Let* $D_\infty := \sup_{x,y \in Q} \|x - y\|_\infty$ *be the* l^∞ *diameter of* Q. *Let* β_1, β_2, $(\beta_{1t})_{t=1}^\infty$, $(\alpha_t)_{t=1}^\infty$ *be deterministic such that* $0 < \beta_1, \beta_2 < 1$, $\gamma := \frac{\beta_1}{\sqrt{\beta_2}} < 1$, $\beta_{11} = \beta_1$ *and* $\beta_{1t} \leq \beta_1$ *for all* $t \geq 1$. *Also let* $T \geq 1$, *and let* $\overline{x}_T := \frac{1}{T} \sum_{k=1}^T x_k$. *Then*

$$\mathbb{E} f(\overline{x}_T) - f(x^*) \leq \frac{D_\infty^2}{2(1-\beta_1) T \alpha_T} \sum_{j=1}^n \mathbb{E} \hat{v}_{T,j}^{1/2}$$

$$+ \frac{D_\infty^2}{2(1-\beta_1) T} \sum_{t=1}^T \frac{\beta_{1t}}{\alpha_t} \sum_{j=1}^n \mathbb{E} \hat{v}_{t,j}^{1/2}$$

$$+ \frac{1}{\sqrt{1-\beta_2}(1-\gamma) T} \sum_{t=1}^T \alpha_t \sum_{j=1}^n \mathbb{E} \left(\sum_{k=1}^t \beta_2^{t-k} g_{k,j}^2 \right)^{1/2}$$

Convergence rate of AMSGrad (continued)

Best scenario: $\hat{v} = v$. The bound becomes

$$\begin{aligned}
\mathbb{E}f(\overline{x}_T) - f(x^*) &\leq \frac{D_\infty^2 \sqrt{1-\beta_2}}{2(1-\beta_1)T\alpha_T} \sum_{j=1}^{n} \mathbb{E}\left(\sum_{t=1}^{T} \beta_2^{T-t} g_{t,j}^2\right)^{1/2} \\
&\quad + \frac{D_\infty^2 \sqrt{1-\beta_2}}{2(1-\beta_1)T} \sum_{t=1}^{T} \frac{\beta_{1t}}{\alpha_t} \sum_{j=1}^{n} \mathbb{E}\left(\sum_{k=1}^{t} \beta_2^{t-k} g_{k,j}^2\right)^{1/2} \\
&\quad + \frac{1}{\sqrt{1-\beta_2}(1-\gamma)T} \sum_{t=1}^{T} \alpha_t \sum_{j=1}^{n} \mathbb{E}\left(\sum_{k=1}^{t} \beta_2^{t-k} g_{k,j}^2\right)^{1/2} \\
&\leq \frac{D_\infty^2}{2(1-\beta_1)T\alpha_T} \sum_{j=1}^{n} M_j + \frac{D_\infty^2}{2(1-\beta_1)T} \sum_{t=1}^{T} \frac{\beta_{1t}}{\alpha_t} \sum_{j=1}^{n} M_j \\
&\quad + \frac{1}{(1-\beta_2)(1-\gamma)T} \sum_{t=1}^{T} \alpha_t \sum_{j=1}^{n} M_j.
\end{aligned}$$

Convergence rate of AMSGrad (continued 2)

- For the recommended choice $\alpha_t := \alpha/\sqrt{t}$ and $\beta_{1t} := \beta_1/t$, the bound becomes

$$\mathbb{E}f(\overline{x}_T) - f(x^*) \leq \left(\frac{3D_\infty^2}{2(1-\beta_1)\alpha} + \frac{2\alpha}{(1-\beta_2)(1-\gamma)} \right) \frac{1}{\sqrt{T}} \sum_{j=1}^{n} M_j$$

- With the best $\alpha := \frac{\sqrt{3}D_\infty\sqrt{(1-\beta_2)(1-\gamma)}}{2\sqrt{1-\beta_1}}$ this is

$$\frac{2\sqrt{3}}{\sqrt{(1-\beta_1)(1-\beta_2)(1-\gamma)}} \frac{D_\infty}{\sqrt{T}} \sum_{j=1}^{n} M_j.$$

- This is always worse than the corresponding AdaGrad bound

$$\frac{3D_\infty}{2\sqrt{T}} \sum_{j=1}^{n} M_j.$$

  The difference is an absolute multiplicative constant.

Conclusion

- For the sets with favourable geometry ($\sqrt{n}D_\infty \leq D$) adaptive stochastic gradient methods may be much more efficient than the basic non-adaptive SGD. The difference can reach $\sqrt{n}$ times.
- From the theoretical point of view, the more "complex" Adam method (or its corrections such as AMSGrad) are no better or even strictly worse than the AdaGrad method.
- There are **no theoretical results for the non-convex** smooth problems for adaptive gradient methods (however, in many cases, they seem to work well in practice).

Thank you!