Gradient Descent (proofs techniques)

Saturday, March 21, 2020 5:23 PM

There are many different types of results. Here, I try to unify some of them.

Algorithm / notation

Algo
$$
\begin{aligned}
&x_1 \text{ arbitrary} \\
&\text{for } t=1,2,\dots,T \\
&\qquad x_t = x_{t-1} - \eta v_t
\end{aligned}
$$

TBD
ex, $v_t = \nabla f(x_t)$
or $\mathbb{E}\, v_t \in \partial f(x_t)$

Lemma 14.1 (Shalev-Shwartz, Ben-David)

Let $\{v_t\}_{t=1}^T$ be arbitrary, f need not be convex nor smooth then Algo produces a sequence satisfying
$$\sum_{t=1}^T \langle x_t - x^*, v_t \rangle \le \frac{\|x^* - x_1\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^T \|v_t\|^2$$

Corollary If $\|v_t\| \le \rho\ \forall t$ (eg: f is ρ-Lipschitz) and $\|x^* - x_1\| \le B$ choosing $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$ gives $\frac{1}{T}\sum_{t=1}^T \langle x_t - x^*, v_t\rangle \le \rho \frac{B}{\sqrt{T}}$

(x^* denotes any minimizer of f)

<u>proof</u> sketch of lemma (just the good parts)

$$\sum_{t=1}^T \langle x_t - x^*, v_t\rangle = \frac{1}{2\eta}\sum \left(-\|x_{t+1} - x^*\|^2 + \|x_t - x^*\|^2 + \|v_t\|^2\right)$$
(via completing-the-square and algebra)

$$= \frac{1}{2\eta}\left(\|x_1 - x^*\|^2 - \|x_{T+1} - x^*\|^2\right) + \frac{\eta}{2}\sum \|v_t\|^2$$
since sum telescoped $\quad \|\cdot\| \ge 0$

$$\le \frac{1}{2\eta}\|x_1 - x^*\|^2 + \eta/2 \sum \|v_t\|^2 \quad \square$$

How to use this result?

Case: f is convex (and ρ-Lipschitz, so corollary applies)

Choose $v_t \in \partial f(x_t)$

then by convexity, $f(x_t) - \underbrace{f^*}_{= \min_x f(x) = f(x^*)} \leq \langle x_t - x^*, v_t \rangle$

So

Corollary 1: If f is convex and ρ-Lipschitz, running subgradient descent gives

$$\frac{1}{T}\sum_{t=1}^{T}\left(f(x_t) - f^*\right) \leq \frac{B\rho}{\sqrt{T}}$$

if $\|x_1 - x^*\| \leq B$ and stepsize $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$

How to apply this?

If we can easily evaluate $f(x_t)$, then

let $x_{best} = \operatorname{argmin}_{x \in \{x_1, \dots, x_T\}} f(x)$

and

Corollary 1a: $f(x_{best}) - f^* \leq \frac{B\rho}{\sqrt{T}}$

However, Sometimes it's not easy to evaluate $f(x)$

ex: $f(w) = L_D(w)$ (we can only sample from it... as may be the case in SGD)

In that case, define $\bar{x} = \frac{1}{T}\sum_{t=1}^{T} x_t$

then

$f(\bar{x}) \leq \frac{1}{T}\sum f(x_t)$ by Jensen's ineq., hence

Corollary 1b: $f(\bar{x}) - f^* \leq \frac{B\rho}{\sqrt{T}}$

Case: f is smooth (i.e. ∇f is β-Lipschitz)

Descent lemma: $f(x_{t+1}) \leq f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{\beta}{2}\|x_{t+1} - x_t\|^2$

(no convexity needed just smoothness)

and if we run gradient descent w/ stepsize $\eta = \frac{1}{\beta}$ then via some algebra,

$$(*) \quad f(x_{t+1}) \le f(x_t) - \frac{1}{2\beta} \| \underbrace{\nabla f(x_t)}_{\text{"}v_k\text{" in Algo}} \|^2$$

Case, part 1: f is smooth but not convex

Thm If f is β-smooth, then gradient descent w/ $\eta = \frac{1}{\beta}$ gives $\min_{t \in [T]} \|\nabla f(x_t)\|^2 \le \frac{2\beta}{T}\left(f(x_1) - f^*\right)$

(for nonconvex, we don't show convergence to a global or even local minimizer, just $\|\nabla f(x_t)\| \to 0$, ie., a stationary pt. where $\nabla f(x) = 0$)

proof Sum $(*)$ from $t = 1, \ldots, T$

$$\frac{1}{2\beta} \sum_{t=1}^{T} \|\nabla f(x_t)\|^2 \le \sum_{t=1}^{T} f(x_t) - f(x_{t+1})$$

(telescopes)

$$= f(x_1) - f(x_{T+1}) \le f(x_1) - f^*$$

and

$$\min_{t \in [T]} \|\nabla f(x_t)\|^2 \le \frac{1}{T} \sum_{t=1}^{T} \|\nabla f(x_t)\|^2 \quad (\min \le \text{avg}) \quad \square$$

Case, part 2: f is β-smoth and convex

As we already saw above, when f is convex, combined w/ Lemma 14.1 (but don't use the corollary yet)

$$\sum_{t=1}^{T} f(x_t) - f^* \le \frac{\|x_1 - x^*\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \| \underbrace{\nabla f(x_t)}_{v_t} \|^2$$

and by smoothness, descent lemma gives

$$f(x_{t+1}) + \frac{1}{2\beta} \|\nabla f(x_t)\|^2 \le f(x_t). \quad \text{Choose } \eta = \frac{1}{\beta}$$

then

$$\sum_{t=1}^{T} \left(f(x_{t+1}) + \frac{\eta}{2} \|\nabla f(x_t)\|^2 - f^* \right) \le \sum_{t=1}^{T} \left(f(x_t) - f^* \right)$$

$$\le \|x_1 - x^*\|^2 \ \eta \sum^{T} \ldots {}^2$$

$\frac{\quad}{2\eta}$ $+\frac{\eta}{2}\sum_{t=1} \|\nabla f(x_t)\|$

So (using $\eta = 1/\beta$)

Thm: $f(x_T) \le \frac{1}{T}\sum_{t=1}^{T}(f(x_t) - f^*) \le \frac{1}{T}\frac{\beta}{2}\|x_1 - x^*\|^2$

since $x_T = x_{best}$ in this case (descent lemma $\Rightarrow$ $f(x_{t+1}) \le f(x_t)$)

Case, part 3: f is β smooth and λ-strongly convex

First, define

"Polyak-Lojasiewicz Inequality" or just "PL"

$\forall x,\ \frac{1}{2}\|\nabla f(x)\|^2 \ge \lambda(f(x) - f^*)$

then f λ-strongly convex $\Rightarrow$ f is λ-PL

Then to analyze, start at descent lemma again

$$f(x_{t+1}) - f(x_t) \le \frac{-1}{2\beta}\|\nabla f(x_t)\|^2 \quad (\text{by } \beta\text{-smoothness})$$

$$\le \frac{-\lambda}{\beta}(f(x_t) - f^*) \quad (\text{by PL})$$

So re-arrange

$$f(x_{t+1}) - f^* \le \left(1 - \frac{\lambda}{\beta}\right)(f(x_t) - f^*)$$

$$\le \left(1 - \frac{\lambda}{\beta}\right)^t (f(x_0) - f^*)$$

So

Thm $f(x_{t+1}) - f^* \le c^t (f(x_0) - f^*)$

$c = (1 - \lambda/\beta) < 1$ "linear convergence"

Discussion of convergence rates

Error e_t. How many more iterations needed to go from $\varepsilon = 1$ accuracy to $\varepsilon = 0.01$ accuracy?

	Rate	Iter.	Examples
linear	1. $e_T \propto \frac{1}{\sqrt{T}}$ $(T = O(\varepsilon^{-2}))$	10,000 times more	Subgradient or gradient descent (not smooth); SGD
	2. $e_T \propto \frac{1}{T}$ $(T = O(\varepsilon^{-1}))$	100 times more	gradient descent (smooth)

sub. 3. $e_T \propto \frac{1}{T^2}$ $\left(T = O(\varepsilon^{-1/2})\right)$ 10 times more — accelerated gradient descent

linear 4. $e_T \propto c^T$, $c<1$ $\left(T = \log(\varepsilon^{-1})\right)$ 2·const more — gradient descent (smooth and strongly convex)

$e_{t+1} \le c \cdot e_t$

superlinear 5. $e_{t+1} \le c_t e_t$, $c_t < 1$, $c_t \to 0$

quadratic 6. $e_{t+1} \le c e_t^2$, $c>0$ $\left(T = \log(\log(\varepsilon^{-1}))\right)$ 1 more — Newton's method near a solution

Geometric Picture