Balls and Parabolas

Thursday, January 23, 2020 5:32 AM

Yudan

We have seen convergence rates of $(1-\frac{1}{n^2})$, $(1-\frac{1}{n})$ (in function value) for convex f

and $(1-\frac{\mu}{L})$ for strongly convex f with GD.

Is faster possible? $(1-\frac{1}{n})$ is best for general convex.

When $\frac{\mu}{L} > \frac{1}{n}$, the rate $(1-\frac{\mu}{L})$ is better.

Can we go faster?

Part 1. GD as a cutting plane method.

For convex,
$$f(x) \geq f(x^k) + \langle \nabla f(x^k), x - x^k \rangle$$

So
$$\{x : f(x) \leq f(x^k)\} \subseteq \underbrace{\{x : \langle \nabla f(x^k), x - x^k \rangle \leq 0\}}_{\text{Half space}}.$$

When f is strongly convex,
$$f(x^*) \geq f(x) + \langle \nabla f(x, x - x^* \rangle + \frac{\mu}{2} \| x^* - x \|^2$$

$$f(x^*) - f(x) \geq \frac{\mu}{2} \left\| x^* - x + \frac{\nabla f(x)}{\mu} \right\|^2 - \frac{2}{\mu} \| \nabla f(x) \|^2$$

i.e. with $x^+ = x - \frac{\nabla f(x)}{\mu}$
$$\| x^* - x^+ \|^2 \leq \frac{\| \nabla f(x) \|^2}{\mu^2} - \frac{2}{\mu} (f(x) - f(x^*))$$

Recall:
$$f(x^*) \leq f(x^+) \leq f(x - \frac{\nabla f(x^k)}{\mu}) \leq f(x) - \frac{1}{2\mu} \| \nabla f(x) \|^2$$

$$f(x^*) \le f(x^+) \le f\left(x - \frac{\nabla f(x^k)}{\mu}\right) \le f(x) - \frac{1}{2\mu}\|\nabla f(x)\|$$

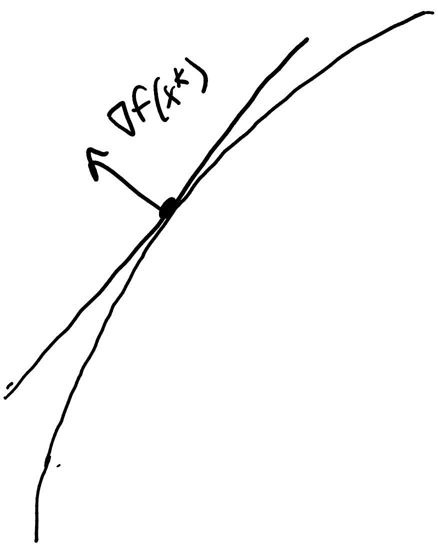
$$\therefore \quad \|x^* - x^+\|^2 \le \frac{\|\nabla f(x)\|^2}{\mu^2} - \frac{2}{\mu}\left(f(x) - f(x^+)\right) - \frac{2}{\mu}\left(f(x^+) - f(x^*)\right)$$

$$\le \frac{\|\nabla f(x)\|^2}{\mu^2}\left(1 - \frac{\mu}{L}\right) - \frac{2}{\mu}\left(f(x^+) - f(x^*)\right).$$

So the OPT x^* lies in the ball

$$B\left(x^+, \sqrt{1 - \frac{\mu}{L}} \cdot \frac{\|\nabla f(x)\|}{\mu}\right).$$

We now use balls instead of Ellipsoids.

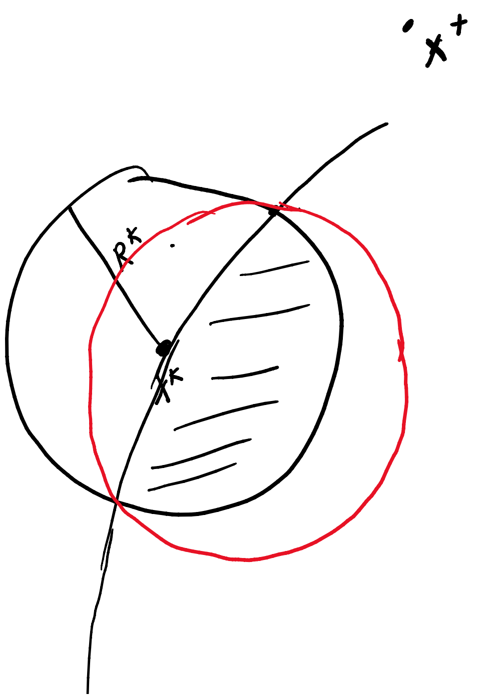


$$E_0 = B\left(0, \frac{\|\nabla f(x^0)\|}{\mu}\right)$$

$$E_k = B(x^k, R_k).$$

$$B^+ = B\left(x^+, \frac{\nabla f(x^k)}{\mu}\sqrt{1 - \frac{\mu}{L}}\right)$$

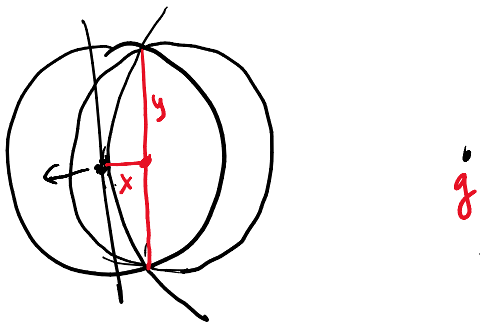
E_{k+1} = min ball containing $E_k \cap B^+$.



x^+

Lemma. $\exists x:\quad B(0,1) \cap B(g, \|g\|\sqrt{1-\varepsilon}) \subseteq B(x, \sqrt{1-\varepsilon})$.

Pf.



$$x^2+y^2=1$$
$$(x-g)^2+y^2 = g^2(1-\varepsilon)$$
$$\Rightarrow\ x^2+g^2-2gx+1-x^2 = g^2-\varepsilon g^2$$
$$x = \frac{1+\varepsilon g^2}{2g}.$$

and
$$y^2 = 1 - \frac{1}{4g^2} - \frac{\varepsilon^2 g^2}{4} - \frac{\varepsilon}{2}$$
$$= 1 - \frac{\varepsilon}{2} - \frac{1}{2}\left(\frac{1}{2g^2} + \frac{\varepsilon^2 g^2}{2}\right)$$
$$\geq 1 - \frac{\varepsilon}{2} - \frac{\varepsilon}{2} = 1-\varepsilon.$$

Applying this to $B(x^*, R_k)$, $B\left(x^k - \frac{\nabla f(x^k)}{\mu}, \frac{\|\nabla f(x^k)\|}{\mu}\left(1-\frac{\mu}{L}\right)^{\frac{1}{2}}\right)$

next ball is $B\left(x, R_k\sqrt{1-\frac{\mu}{L}}\right)$.

Lemma. $R_{k+1}^2 \leq \left(1-\frac{\mu}{L}\right) R_k^2$

$$\Rightarrow\quad R_k^2 \leq \left(1-\frac{\mu}{L}\right)^k \cdot R_0^2 \leq \left(1-\frac{\mu}{L}\right)^k.$$

Using $v(B_k) = R_k$

$$\min_i f(x^i) - f(x^*) \le \left(\sqrt{1-\frac{\mu}{L}}\right)^k \cdot (f(x^0) - f(x^*)).$$

Recovers GD rate for strongly concave f.

Can we do better?

Observe: Our cuts come from

$$\langle \nabla f(x), x^* - x \rangle \le 0$$

we start with 0
but this improves
as we get better estimates

$$\le f(x^{new}) - f(x)$$

i.e. we can "tighten" all previous planes!

accelerated ball method

$$x, \quad x^+ = x - \frac{\nabla f(x^0)}{L}$$

$$x^{++} = x - \frac{\nabla f(x^0)}{\mu} \qquad C_0 = x^{0++}$$

We maintain x^k, c^k, R_k. $\qquad R_0^2 = \frac{\|\nabla f(x^0)\|^2}{\mu^2}\left(1-\frac{\mu}{L}\right)$

$$x^{k+1} = \text{linesearch}(c^k, x^{k^+})$$

... terming

$x = \text{line search}(c, \dots)$

c^{k+1} is center of min ball containing

$$B\left(c^k, \sqrt{R_k^2 - \frac{\|\nabla f(x^{k+1})\|^2}{\mu^2} \cdot \frac{\mu}{L}}\right) \cap B\left(x^{k+1 ++}, \frac{\|\nabla f(x^{k+1})\|}{\mu} \sqrt{1 - \frac{\mu}{L}}\right)$$

Lemma. $\quad x^* \in B(c^k, R_k)$

$$R_{k+1}^2 \leq \left(1 - \sqrt{\frac{\mu}{L}}\right) R_k^2$$

Pf. We show by induction that

$$x^* \in B\left(c^k, \sqrt{R_k^2 - \frac{2}{\mu}\left(f(x^{k+}) - f(x^*)\right)}\right)$$

$k=0$. ✓

$$f(x^{(k+1)+}) \leq f(x^{k+1}) - \frac{\|\nabla f(x^{k+1})\|^2}{2L} \leq f(x^{k+}) - \frac{\|\nabla f(x^{k+1})\|^2}{2L}$$

$$R_{k+1}^2 \leq R_k^2 - \frac{2}{\mu}\left(f(x^{k+}) - f(x^*)\right)$$

$$= R_k^2 - \frac{2}{\mu}\left(f(x^{k+}) - f(x^{(k+1)+})\right) - \frac{2}{\mu}\left(f(x^{(k+1)+}) - f(x^*)\right)$$

$$\leq R_k^2 - \frac{\|\nabla f(x^{k+1})\|^2}{\mu^2} \cdot \frac{\mu}{L} - \frac{2}{\mu}\left(f(x^{(k+1)+}) - f(x^*)\right)$$

We also have

$$x^* \in B\left(x^{(k+1)++}, \sqrt{\frac{\|\nabla f(x^{k+1})\|^2}{\mu^2}\left(1 - \frac{\mu}{L}\right) - \frac{2}{\mu}\left(f(x^{(k+1)+}) - f(x^*)\right)}\right)$$

$$x^* \in B\left(x^{(k+1)++}, \sqrt{\frac{\|\nabla f(x^{k+1})\|^2}{\mu^2}\left(1-\frac{\mu}{L}\right) - \frac{2}{\mu}\left(f(x^{(k+1)+}) - f(x^*)\right)}\right)$$

Lemma.

$$B(0, \sqrt{1 - g^2\varepsilon - \delta}) \cap B(a, \sqrt{g^2(1-\varepsilon) - \delta})$$

$|a| > g$

$$\subseteq B(c, \sqrt{1 - \sqrt{\varepsilon} - \delta}).$$

$$\Rightarrow \quad R_{k+1}^2 \leq R_k^2\left(1 - \sqrt{\frac{\mu}{L}}\right).$$

$\therefore$ using $\mathcal{V}(\cdot) = R_k$

rate is $\left(1 - \sqrt{\frac{\mu}{L}}\right)$.

Thm. For any algorith "using only gradients"

i.e. $x^k \in \mathrm{Span}(x^0, \nabla f(x^0), \nabla f(x^1), \ldots \nabla f(x^{k-1}))$

best possible rate is $\min\left\{n, \sqrt{\frac{L}{\mu}}\right\}$.