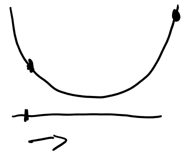
20 Gradient Descent Part II

Thursday, March 27, 2025 13:53



Last time:

$$x^{t+1} \leftarrow x^t - \eta \nabla f(x^t)$$

Lemma: f convex, L-smooth, and $\eta \le \frac{1}{L}$

then $f(x^{t+1}) \le f(x^t) - \frac{\eta}{2} \|\nabla f(x^t)\|^2$

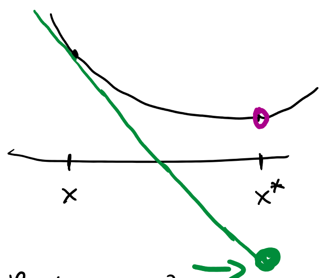
Theorem: $f(x^t) - f(x^*) \le \dfrac{\|x^0 - x^*\|^2}{2\eta \cdot t}$

$x^* := \arg\min_x f(x)$

$f(x^*) = \min_x f(x)$

Proof: $\underline{f(x^*)} \ge \underline{f(x) + \nabla f(x)^T (x^* - x)} \quad \forall x$

$$= f(x) - \nabla f(x)^T (x - x^*)$$



$$f(x) \le f(x^*) + \nabla f(x)^T (x - x^*)$$

$$f(x^{t+1}) \le f(x^t) - \frac{\eta}{2} \|\nabla f(x^t)\|^2$$

$$\le f(x^*) + \nabla f(x^t)^T (x^t - x^*) - \frac{\eta}{2} \|\nabla f(x^t)\|^2$$

(1) $f(x^{t+1}) - f(x^*) \le \underline{\nabla f(x^t)^T (x^t - x^*) - \frac{\eta}{2} \|\nabla f(x^t)\|^2}$

(2) $\|x^{t+1} - x^*\|^2 = \|\underline{x^t} - \underline{\eta \nabla f(x^t)} \underline{- x^*}\|^2$

$\|\underline{u} - \underline{v}\|^2 = \|u\|^2 - 2u^T v + \|v\|^2$

$$= \|\underline{x^t - x^*}\|^2 - 2 \underline{\eta \nabla f(x^t)}^T (\underline{x^t - x^*}) + \|\underline{\eta \nabla f(x^t)}\|^2$$

$$= 2\eta \left(\frac{\|x^t - x^*\|^2}{2\eta} - \underline{\nabla f(x^t)^T (x^t - x^*) + \frac{\eta}{2} \|\nabla f(x^t)\|^2} \right)$$

$$\Rightarrow -\frac{\|x^{t+1} - x^*\|^2}{2\eta} + \frac{\|x^t - x^*\|^2}{2\eta} = \underline{\nabla f(x^t)^T (x^t - x^*) - \frac{\eta}{2} \|\nabla f(x^t)\|^2}$$

$$\Rightarrow (1)+(2) \quad \underline{f(x^{t+1}) - f(x^*)} \le \left(\|x^t - x^*\|^2 - \|x^{t+1} - x^*\|^2 \right) \cdot \frac{1}{2\eta}$$

$$\sum_{k=1}^{t} \underline{f(x^k) - f(x^*)} \le \sum_{k=1}^{t} \left(\|x^{k-1} - x^*\|^2 - \|x^k - x^*\|^2 \right) \cdot \frac{1}{2\eta}$$

$$= \left(\|x^0 - x^*\|^2 - \|x^t - x^*\|^2 \right) \cdot \frac{1}{2\eta}$$

$$\le \frac{\|x^0 - x^*\|^2}{2\eta}$$

$$\frac{1}{t} \sum_{k=1}^{t} f(x^k) - f(x^*) \le \frac{\|x^0 - x^*\|^2}{2\eta t}$$

$$\min_{1 \le k \le t} f(x^k) - f(x^*) \le \frac{1}{t} \sum_{k=1}^{t} f(x^k) - f(x^*) \le \frac{\|x^0 - x^*\|^2}{2\eta t}$$

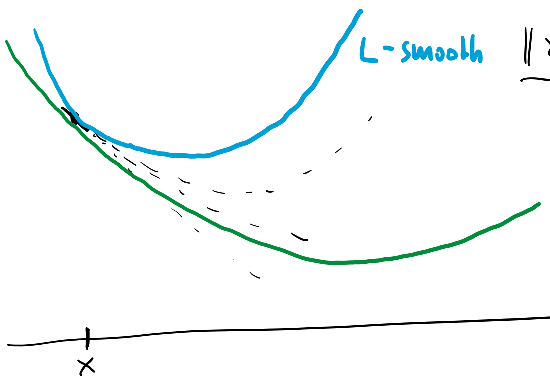
$$\parallel$$

$$f(x^t) - f(x^*)$$

$\parallel$

$f(x^t) - f(x^*)$

If f is convex and L-smooth then gradient descent converges in $\Theta(\frac{1}{\varepsilon})$ iterations



L-smooth

$$\frac{\|x^0 - x^*\|^2}{2\eta t} \le \varepsilon$$

$$t \ge \frac{\|x^0 - x^*\|^2}{2\eta \cdot \varepsilon} = \Theta(\tfrac{1}{\varepsilon})$$

μ- strongly convex

Def: f is μ-strongly conex if

$$f(x+\delta) \ge \underline{f(x) + \nabla f(x)^T \delta + \frac{\mu}{2}\|\delta\|^2} \qquad f''(x) \ge \mu$$

(recall L-smooth

$$f(x+\delta) \le \underline{f(x) + \nabla f(x)^T \delta + \frac{L}{2}\|\delta\|^2}, \qquad f''(x) \le L \Big)$$

Theorem: If f is L-smooth & μ-strongly convex, stepsize $\eta \le \frac{1}{L}$ then $t = O(\frac{1}{\eta\mu} \log \frac{1}{\varepsilon})$ iterations suffice to get error $\le \varepsilon$

Example: $f(x) = \|Ax - b\|^2 \qquad x^* = A^{-1}b$

$L = \lambda_{max}(A^TA)$ largest eigenvalue $\rightarrow \eta = \frac{1}{\lambda_{max}(A^TA)}$

$\mu = \lambda_{min}(A^TA)$ smallest eigenvalue

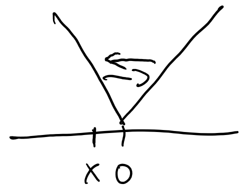
$$\Rightarrow O\left(\frac{1}{\eta\mu} \log \frac{1}{\varepsilon}\right) = O\left(\frac{\lambda_{max}(A^TA)}{\lambda_{min}(A^TA)} \log(\tfrac{1}{\varepsilon})\right)$$

$$= O\left(\kappa(A^TA) \log(\tfrac{1}{\varepsilon})\right)$$

condition number

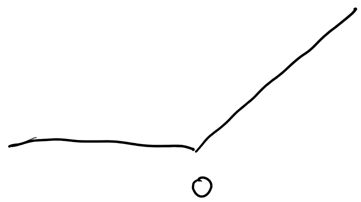
What if f is not smooth?

$f(x) = |x|$



gradient descent will just go back and forth if $x = \frac{\eta}{2}$

$f(x) = \max(0, x)$ ReLU



Def: L-Lipschitz if $|f(x) - f(y)| \le L \cdot \|x - y\|$ $\iff |f'(x)| \le L$

$\|\nabla f(x)\| \le L$

Theorem 1) $\left(\frac{\sum_{k=1}^{t} f(x^k)}{t}\right) - f(x^*) \le \frac{\eta L^2}{2} + \frac{\|x^0 - x^*\|^2}{2\eta t}$

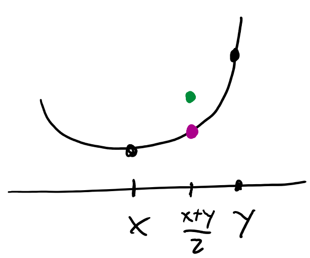
Theorem 2) $f\left(\frac{\sum_{k=1}^{t} x^k}{t}\right) - f(x^*) \le \frac{\eta L^2}{2} + \frac{\|x^0 - x^*\|^2}{2\eta t}$

$\left(\begin{array}{l} \text{error} \le \varepsilon \\ \eta = \frac{\|x^0 - x^*\|}{L\sqrt{t}} \\ t = O(1/\varepsilon^2) \end{array}\right)$

Thm 1 is better because Thm 1 implies Thm 2

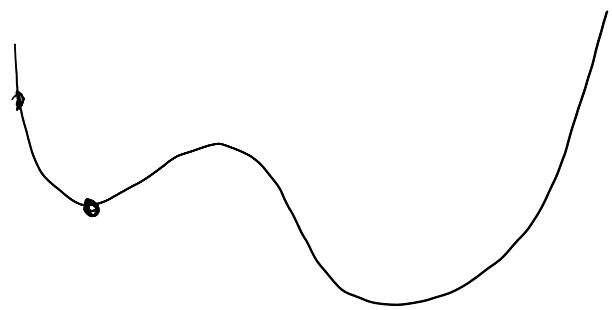
because f is convex

$f\left(\frac{\sum x^k}{t}\right) \le \frac{\sum f(x^k)}{t}$



$\frac{f(x) + f(y)}{2}$

$f\left(\frac{x+y}{2}\right)$



In non-convex setting gradient descent finds a <u>local</u> minimum.