Amath 515, Jan 20

* β-smoothness
* upper bounds
* Analysis of gradient descent for β-smooth fcns

For hw:

For problem 1, I recommend starting with simple facts:

(1) $(g(h(x)))' = g'(h(x))\,h'(x)$ for scalar fcns.

(2) Jacobian of $x \to Ax$ is A.

build gradient, hessian formulas for $g(a^T x)$ first, then $g(Ax)$

$$\min_x \underbrace{\sum \log(1+\exp(a_i^T x))}_{g(Ax)}$$

$$g(z) = \sum_i \log(1 + \exp(z_i))$$

Easy to differentiate

$$\nabla g(z) = \begin{bmatrix} \frac{\exp(z_1)}{1+\exp(z_1)} \\ \vdots \\ \text{derivatives} \quad z \quad z_n \end{bmatrix}$$

$$\nabla^2 g = \begin{bmatrix} \ddots & & \\ & \ddots & \\ & & \ddots \end{bmatrix}$$

Last time: ended with rates

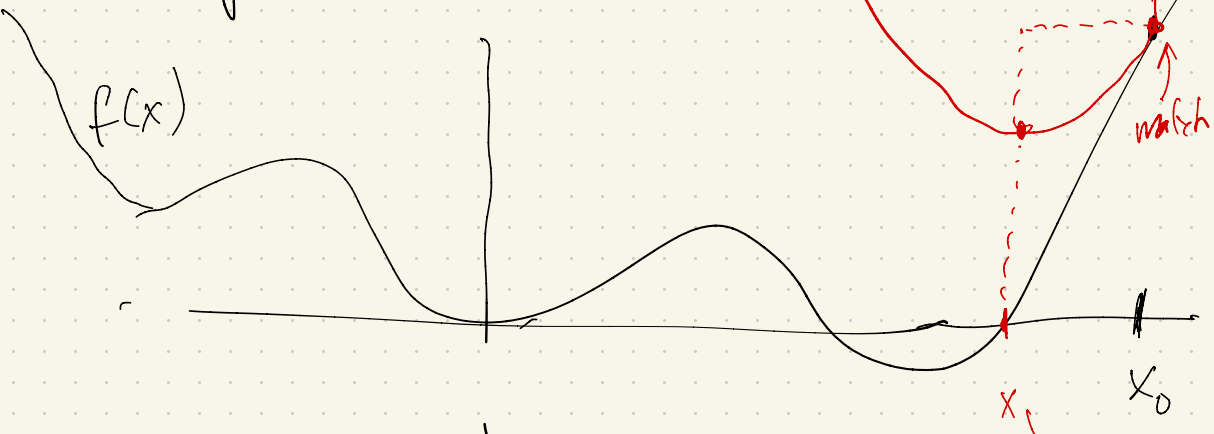
from sublinear to quadratic

Today: analyze gradient descent.

$$x_{k+1} = x_k - (\text{step})\,\nabla f(x_k)$$

for $\min_x f(x)$

* Then summarize other results
* go back to regression demo.



get $f(x_1) < f(x_0)$

for a global upper bound

Strategy:

(1) Build tight global upper bounds
   - match f and gradient at current point
   - above everywhere

(2) get iterates by minimizing these upper bounds.

Need a bit more notation:

Def: A function $F: \mathbb{R}^n \to \mathbb{R}^m$ is β-Lipschitz continuous if

$$\| F(x) - F(y) \| \leq \beta \| x - y \|$$

A function $f: \mathbb{R}^n \to \mathbb{R}$ is β-smooth if f is C^1 and ∇f is β-Lipschitz cont:

$$\| \nabla f(x) - \nabla f(y) \| \leq \beta \| x - y \|$$

Thm 1: If f is β-smooth, then

$$\rightarrow f(y) \leq \underbrace{f(x) + \langle \nabla f(x), y - x \rangle}_{\text{tangent line approx at } x} + \underbrace{\frac{\beta}{2}\|y-x\|^2}_{\text{quadratic}}$$

Thm 2: For f C^2, f is β-smooth if and only if $\nabla^2 f \preceq \beta I$

equiv, largest eigenvalue of $\nabla^2 f \leq \beta$

Ex: $\min_x \underbrace{\frac{1}{2}\|Ax-b\|^2}_{f(x)}$

$$\nabla f(x) = A^T(Ax - b)$$ Linear in x

So f is β-smooth, with β the largest eigenvalue of A^TA

Since $\nabla^2 f(x) = A^TA$.

Main point: β-smoothness gives us very simple global right upper bounds.

Call $m_x(y) = f(x) + \langle \nabla f(x), y - x \rangle + \frac{\beta}{2}\|y - x\|^2$

(1) Know $f(y) \le m_x(y)$ everywhere, by theorem 1.

(2) $m_x(x) = f(x)$ (yay)

(3) $\nabla m_x(y) = \nabla f(x) + \beta(y - x)$

at $y = x$, $\nabla m_x(x) = \nabla f(x)$

Strategy: minimize m_x at each iteration

$x^+ = \operatorname{argmin}_y m_x(y)$ "next point"

$x_{k+1} = \operatorname{argmin}_y m_{x_k}(y)$ "kth iterate"

Let's minimize $m_x(y)$ by completing the square

$$m_x(y) = f(x) + \langle \nabla f(x), y-x \rangle + \frac{\beta}{2}\|y-x\|^2$$

$$= f(x) + \beta\left(\langle \frac{1}{\beta}\nabla f(x), y-x \rangle + \frac{1}{2}\|y-x\|^2 \right)$$

$$= f(x) + \beta\left(\frac{1}{2}\left\| (y-x) + \frac{1}{\beta}\nabla f(x) \right\|^2 - \frac{1}{2\beta}\|\nabla f(x)\|^2 \right)$$

Check:

$$\beta\left(\frac{1}{2}\|y-x\|^2 + \langle y-x, \frac{1}{\beta}\nabla f(x) \rangle + \frac{1}{\beta^2}\|\nabla f(x)\|^2 \right)$$

$$= \frac{\beta}{2}\|y-x\|^2 + \langle y-x, \nabla f(x) \rangle + \frac{1}{2\beta}\|\nabla f(x)\|^2$$

$$m_x(y) = f(x) + \beta\left(\frac{1}{2}\left\| y - \underbrace{\left(x - \frac{1}{\beta}\nabla f(x)\right)}_{x^+}\right\|^2\right) - \frac{1}{2\beta}\|\nabla f(x)\|^2$$

$$m_x(y)$$

minimizer: $x - \frac{1}{\beta}\nabla f(x)$

Let $x^+ = x - \frac{1}{\beta}\nabla f(x)$

$$x_{k+1} = x_k - \frac{1}{\beta}\nabla f(x_k)$$

gradient descent.

Some nice observations:

$$f(x_+) \le m_x(x_+) = f(x) - \frac{1}{2\beta}\|\nabla f(x)\|^2$$

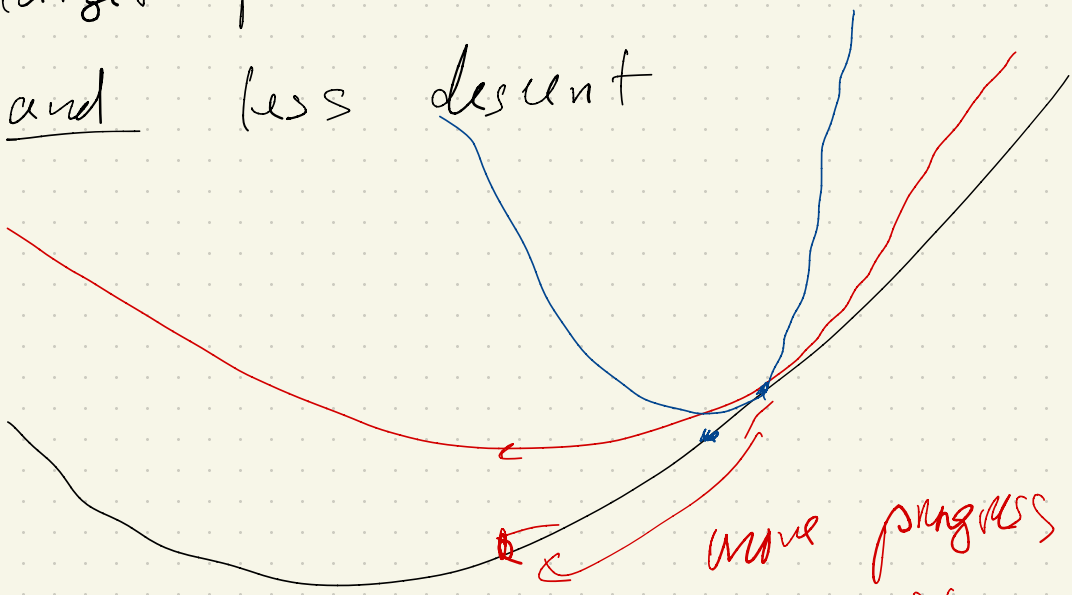
global upper bound

* we have $f(x_+)$ lower than $f(x)$ unless $\nabla f(x) = 0$, in which case we are at a stationary point.

step in our algorithm is $\frac{1}{\beta}$

larger β => smaller step.

and less descent



We have $f(x_{k+1}) \leq f(x_k) - \frac{1}{2\beta}\|\nabla f(x_k)\|^2$ (★)

for $x_{k+1} = x_k - \frac{1}{\beta}\nabla f(x_k)$

Rearrange ★, we end up with

$$\frac{1}{2\beta}\|\nabla f(x_k)\|^2 \leq f(x_k) - f(x_{k+1})$$

$$+$$

$$\frac{1}{2\beta}\|\nabla f(x_{k-1})\|^2 \leq f(x_{k-1}) - f(x_k)$$

telescoping

$$+ \vdots$$

$$\leq f(x_1) - f(x_2)$$

$$\frac{1}{2\beta}\|\nabla f(x_0)\|^2 \leq f(x_0) - f(x_1)$$

$$\frac{1}{2\beta}\sum_{j=0}^{k}\|\nabla f(x_j)\|^2 \leq f(x_0) - f(x_{k+1})$$

$$\leq f(x_0) - f^*$$

$$\frac{1}{K}\sum_{j=0}^{K} \|\nabla f(x_j)\|^2 \le \frac{2\beta\left(f(x_0)-f^*\right)}{K}$$

average value of observed quantities.

$$\min_{j=0,\dots,K} \|\nabla f(x_j)\|^2 \le \frac{2\beta}{K}\left(f(x_0)-f^*\right)$$

$$\min \le \text{avg.}$$

Result:

$$\min_{j=0,\dots,K} \|\nabla f(x_j)\| \le \frac{1}{\sqrt{K}} \sqrt{2\beta\left(f(x_0)-f^*\right)}$$

Sublinear rate, $O\left(\frac{1}{\sqrt{K}}\right)$

Our first analysis!

These results get stronger as we add more assumptions.

① We can talk about progress in terms of fcn values, distance to 'unique soln'

② Rates get better

Thm 2.25: If f is β-smooth and convex, x^* any minimizer, we have

$$\min_{j=0,\dots,k} \|\nabla f(x_j)\| \leq \frac{2\beta \|x_0 - x^*\|}{k}$$

not $\sqrt{k}$

$$\frac{f(x_k) - f^*}{\downarrow} \leq \frac{\beta \|x_0 - x^*\|^2}{2k}$$

fcn values, when f convex.

thm 2.26: f α-cvx + β-smooth

$$\|x_k - x^*\|^2 \le (*)^k \|x_0 - x^*\|^2$$

↓ Linear rate

$$* = \left(\frac{\frac{\beta}{\alpha} - 1}{\frac{\beta}{\alpha} + 1}\right)^2$$