ML and Opt Lecture 11

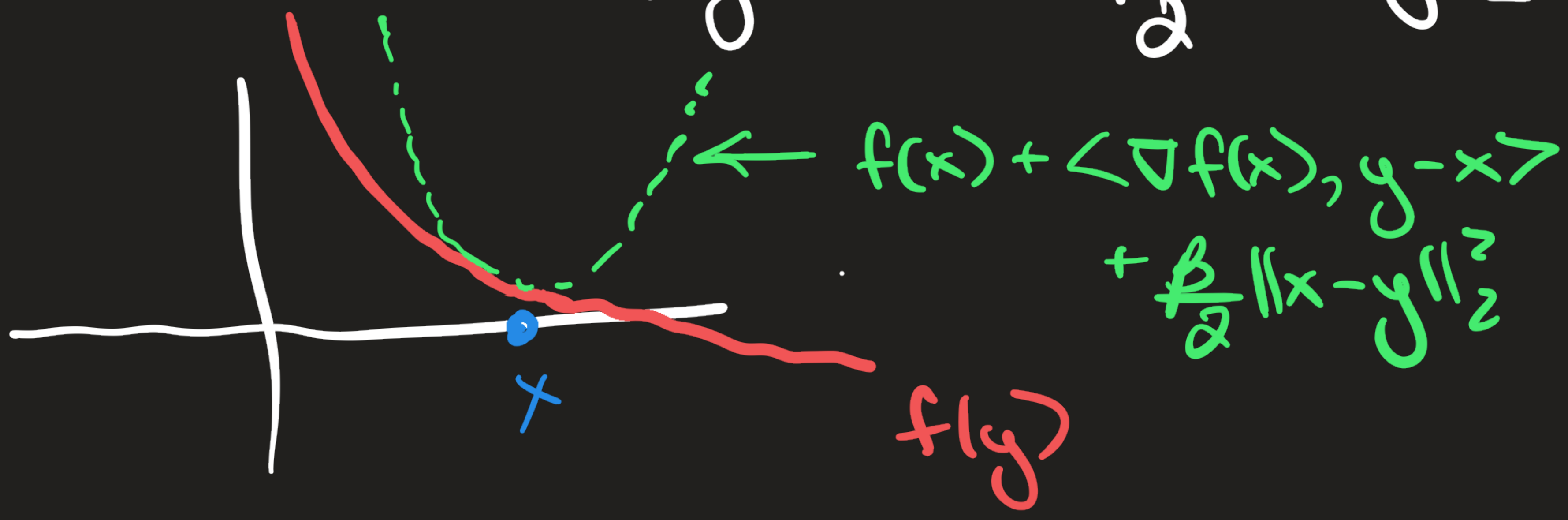
- Proj Sbxn due end of week of Feb 28
- Gradient Descent Convergence Rates for diff classes of cvx functions
- Strong convexity, convex condition numbers
- Polyak-Lojasiewicz (PL) Inequality
- Newton's Method

Recap: β-smoothness

A function f is β-smooth iff $\|\nabla f(x) - \nabla f(y)\|_2 \leq \beta\|x-y\|_2$

Consequence:

$$f(y) \leq f(x) + \langle \nabla f(x), y-x \rangle + \frac{\beta}{2}\|x-y\|_2^2$$



Consequence: **Descent Lemma**

when $x_{t+1} = x_t - \frac{1}{\beta}\nabla f(x_t)$

$$f(x_{t+1}) \leq f(x_t) - \frac{1}{2\beta}\|\nabla f(x_t)\|_2^2$$

α-Strong convexity

We say a function is α-strongly cvx if

$$f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle + \frac{\alpha}{2}\|x-y\|_2^2$$

$$D = \begin{bmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_2} \end{bmatrix}$$

$$\|Dx\|_2^2 = \lambda_1 x_1^2 + \lambda_2 x_2^2$$

where $\lambda_1 \ll \lambda_2$

$$\lambda_1 x_1^2 + \lambda_1 x_2^2 \leq \|Dx\|_2^2$$

note $$\|Dy\|_2^2 = \|Dx\|_2^2 + \langle 2D^2 x, y-x \rangle + \frac{1}{2}(y-x)^T(2D^2)(y-x\rangle$$

$$= \|Dx\|_2^2 + 2\langle D^2 x, y-x \rangle + \|D(y-x)\|_2^2$$

$$\|Dy\|_2^2 \geq \|Dx\|_2^2 + 2\langle D^2x, y-x\rangle$$
$$+ \lambda_1(y_1-x_1)^2 + \lambda_2(y_2-x_2)^2$$
$$\geq \|Dx\|_2^2 + 2\langle D^2x, y-x\rangle$$
$$+ \frac{2\lambda_1}{2}\|x-y\|_2^2$$

so $f(x) = \|Dx\|_2^2$ is $2\lambda_1$ strongly cvx

Ex function that is β-smooth and α-strongly cvx

$$f(x) = \frac{1}{2}\|Ax - b\|_2^2 \quad \text{where } \alpha \cdot I \preceq A^TA \preceq \beta \cdot I$$

then f is α-strongly cvx and β-smooth

Prf

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

Notice that since $A^TA \preceq \beta I$:

$$z^T A^T A z \leq \beta z^T z \quad \text{or} \quad \|Az\|_2^2 \leq \beta \|z\|_2^2$$

Take $z = x - y$

$$\frac{1}{2}(y-x)^T A^T A (y-x) \leq \frac{\beta}{2}\|x-y\|_2^2$$

Likewise $\alpha \cdot I \preceq A^T A$ implies

$$\forall z: \quad \alpha z^T z \leq z^T A^T A z$$

so take $z = y - x$

$$\frac{\alpha}{2}\|y-x\|_2^2 \leq \frac{1}{2}(y-x)^T A^T A (y-x)$$

Facts Let f be a twice-differentiable cvx function

1) If for all $x \in dom(f)$,

$$\nabla^2 f(x) \succeq \alpha I$$

then f is α-strongly cvx

2) If for all $x \in dom(f)$

$$\nabla^2 f(x) \preceq \beta I$$

then f is β-smooth

Gradient Descent converges <u>very</u> fast for β-smooth, α-strongly cvx functions

$$f(x_T) - f(x_*) \leq \left(1 - \frac{1}{K_f}\right)^T \left(f(x_0) - f(x_*)\right)$$

where $K_f = \frac{\beta}{\alpha}$ if we use GD w/ stepsize $\frac{1}{\beta}$

Fact: Strongly cvx functions satisfy the PL-inequality:

$$\frac{1}{2\alpha}\|\nabla f(x)\|_2^2 \geq (f(x) - f(x_*))$$

Prf

By strong cvxty we have that for all y>

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2}\|y - x\|_2^2$$

minimize the RHS as a function of y

$$\nabla f(x) + \alpha(y_* - x) = 0$$

$$\Leftrightarrow \quad y_* - x = -\frac{1}{\alpha}\nabla f(x)$$

$$\Leftrightarrow \quad y_* = x - \frac{1}{\alpha}\nabla f(x)$$

and at y_* the value of the RHS is

$$f(x) + \langle \nabla f(x), \left(-\frac{1}{\alpha}\nabla f(x)\right)\rangle$$

$$+ \frac{\alpha}{2}\left\|-\frac{1}{\alpha}\nabla f(x)\right\|_2^2$$

$$= f(x) - \frac{1}{\alpha}\|\nabla f(x)\|_2^2 + \frac{1}{2\alpha}\|\nabla f(x)\|_2^2$$

$$= f(x) - \frac{1}{2\alpha}\|\nabla f(x)\|_2^2$$

so for all y,

$$f(y) \geqslant f(x) - \frac{1}{2\alpha}\|\nabla f(x)\|_2^2$$

Take $y = x_*$

$$f(x_*) - f(x) \geqslant -\frac{1}{2\alpha}\|\nabla f(x)\|_2^2$$

Rearrange to see

$$\frac{1}{2\alpha}\|\nabla f(x)\|_2^2 \geqslant \left(f(x) - f(x_*)\right) >$$

i.e. PL-ineq.

Prf (Linear conv of GD for κ_f-conditioned cvx optimization)

Consider

$$f(x_{t+1}) \leq f(x_t) - \frac{1}{2\beta}\|\nabla f(x_t)\|_2^2$$

by the descent lemma, so

$$f(x_{t+1}) - f(x_t) \leq -\frac{1}{2\beta}\|\nabla f(x_t)\|_2^2$$

and by the PL-inequality,

$$\|\nabla f(x_t)\|_2^2 \geq 2\alpha\left(f(x_t) - f(x_*)\right)$$

so

$$f(x_{t+1}) - f(x_t) \leq -\frac{\alpha}{\beta}\left(f(x_t) - f(x_*)\right)$$

Add $f(x_t) - f(x_*)$ to both sides

$$f(x_{t+1}) - f(x_*) \leq -\frac{\alpha}{\beta}\left(f(x_t) - f(x_*)\right) + \left(f(x_t) - f(x_*)\right)$$

$$= \left(1 - \frac{\alpha}{\beta}\right)\left(f(x_t) - f(x_*)\right)$$

$$= \left(1 - \frac{1}{\kappa_f}\right)\left(f(x_t) - f(x_*)\right)$$

so

$$f(x_{t+1}) - f(x_*) \leq \left(1 - \frac{1}{\kappa_f}\right)\left(f(x_t) - f(x_*)\right)$$

and by induction,

$$f(x_T) - f(x_*) \leq \left(1 - \frac{1}{\kappa_f}\right)^T \left(f(x_0) - f(x_*)\right)$$

Convergence rates for different classes of cvx fcns using GD

properties	cvx	β-smooth	α-strongly cvx	$\kappa = \frac{\beta}{\alpha}$ conditioned
errors	$\frac{1}{\sqrt{T}}$	$\frac{\beta}{T}$	$\frac{1}{\alpha T}$	$e^{-\frac{T}{\kappa}}$
stepsize	$\frac{1}{\sqrt{T}}$	$\frac{1}{\beta}$	$\frac{1}{\alpha T}$	$\frac{1}{\beta}$