$\boxed{\text{Starting 9/21}}$ HW2 due date delayed to next Thursday.

Feel free to use computational aids (Wolframalpha) to help with computing derivatives going forward.

Examples: $f(x) = \frac{1}{2}\|x\|_2^2$ is 1-strongly convex
$$(\nabla^2 f(x) = I \succeq 1 \cdot I)$$

$f(x) = \frac{1}{2}\|Ax-b\|_2^2$ is $\lambda_{min}(A^TA)$-strongly convex (HW 2)

Lemma If $\{f_i$ is μ_i-strongly convex, then

$\sum f_i$ is $\sum \mu_i$-strongly convex.

Proof. ~~We know~~ (Lets assume each f_i is C^1).

Then for any x, y, we have
$$f_i(y) \geq f_i(x) + \nabla f_i(x)^T(y-x) + \frac{\mu_i}{2}\|y-x\|_2^2.$$

Summing over i
$$\sum f_i(y) \geq \sum f_i(x) + \underbrace{\left(\sum \nabla f_i(x)\right)}_{= \nabla(\sum f_i)(x)}{}^T(y-x) + \frac{\sum \mu_i}{2}\|y-x\|_2^2$$

This equivalent to $\sum f_i$ being $\sum \mu_i$-strongly convex. $\square$

Then we know the following are strongly convex.

1. Training Support Vector Machines:

Given observations (x_i, y_i), we want $w^T x \approx \text{sign}(y)$

y_i: ±1 as a label.

x_i: feature vector

$$\min_{w} \sum_i \left(\max\{0, 1 - y_i w^T x_i\}\right) + \underbrace{\frac{\lambda}{2}\|w\|_2^2}_{\lambda\text{-strongly convex}}$$

2. Sparse Regression (LASSO)

$$\min_{x} \underbrace{\|Ax-b\|_2^2}_{\lambda_{min}(A^TA)\text{-strongly convex}} + \lambda \|x\|_1 \qquad \|x\|_1 = \sum |x_i|$$

$\lambda_{min}(A^TA) = 0$

<u>Theorem</u> Let f be μ-strongly convex with L-Lipschitz gradient and has a minimizer x^*. Then GD with $\alpha_k = \frac{2}{\mu+L}$ has

$$\|x_k - x^*\|^2 \leq \left(1 - \frac{4\mu^2}{(\mu+L)^2}\right)^k \|x_0 - x^*\|^2.$$

Proof. We want to a contraction at each step.

$$\|x_{k+1} - x^*\|^2 = \|x_k - \frac{2}{\mu+L}\nabla f(x_k) - x^*\|^2$$

$$= \|x_k - x^*\|^2 + \frac{4}{(\mu+L)^2}\|\nabla f(x_k)\|^2 - \frac{4}{\mu+L}\nabla f(x_k)^T(x_k - x^*)$$

Recall $(\nabla f(x_k) - \nabla f(x^*))^T(x_k - x^*) \geq \frac{1}{L}\|\nabla f(x_k) - \nabla f(x^*)\|^2$

(by smoothness equiv condition)

$\frac{L}{\mu+L} \nearrow$

$\frac{\mu}{\mu+L} \nearrow$ $(\nabla f(x_k) - \nabla f(x^*))^T(x_k - x^*) \geq \mu\|x_k - x^*\|^2$

$$\Rightarrow (\nabla f(x_k) - \nabla f(x^*))^T(x_k - x^*)$$

$$\geq \frac{L}{\mu+L}\cdot\frac{1}{L}\|\nabla f(x_k) - \nabla f(x^*)\|^2 + \frac{\mu}{\mu+L}\mu\|x_k - x^*\|^2$$

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

$$\leq \|x_k - x^*\|^2 + \frac{4}{(\mu+L)^2}\|\nabla f(x_k) - \underbrace{\nabla f(x^*)}_{=0}\|^2 - \frac{4}{\mu+L}\left(\frac{1}{\mu+L}\|\nabla f(x_k) - \nabla f(x^*)\|^2 + \frac{\mu^2}{(\mu+L)}\|x_k - x^*\|^2\right)$$

$$= \|x_k - x^*\|^2 - \frac{4\mu^2}{(\mu+L)^2}\|x_k - x^*\|^2. \qquad \square$$

6+7 Complexity Lowerbounds and Acceleration

We have shown GD has

$$f(x_K) - f(x^*) \le \frac{2L\|x_0 - x^*\|^2}{K}.$$

Lets imagine a wider class of algorithms following gradient directions

Assumption 1 The given method produce points x_K satisfying

$$x_K \in x_0 + \mathrm{Lin}\{\nabla f(x_0), \dots, \nabla f(x_{K-1})\}.$$

(For example, $x_K = x_0 - \sum_{i=0}^{K-1} \alpha_i \nabla f(x_i)$)
is the gradient descent sequence.

Theorem For any $1 \le k \le \frac{1}{2}(d-1)$ and $L \ge 0$,
there exists a convex function $f: \mathbb{R}^d \to \mathbb{R}$ that L-smooth such any algorithm satisfying Assumption 1 has

$$f(x_K) - f(x^*) \ge \frac{3L\|x_0 - x^*\|^2}{32(K+1)^2}$$

$$\|x_K - x^*\|^2 \ge \frac{1}{32}\|x_0 - x^*\|^2.$$

where x^* minimizes f.

It turns out we can give a faster method
(for smooth, convex optimization).

Nesterov's Accelerated Gradient Method (1983)

Let $y_0 = x_0$. Then iterate

$$y_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$$

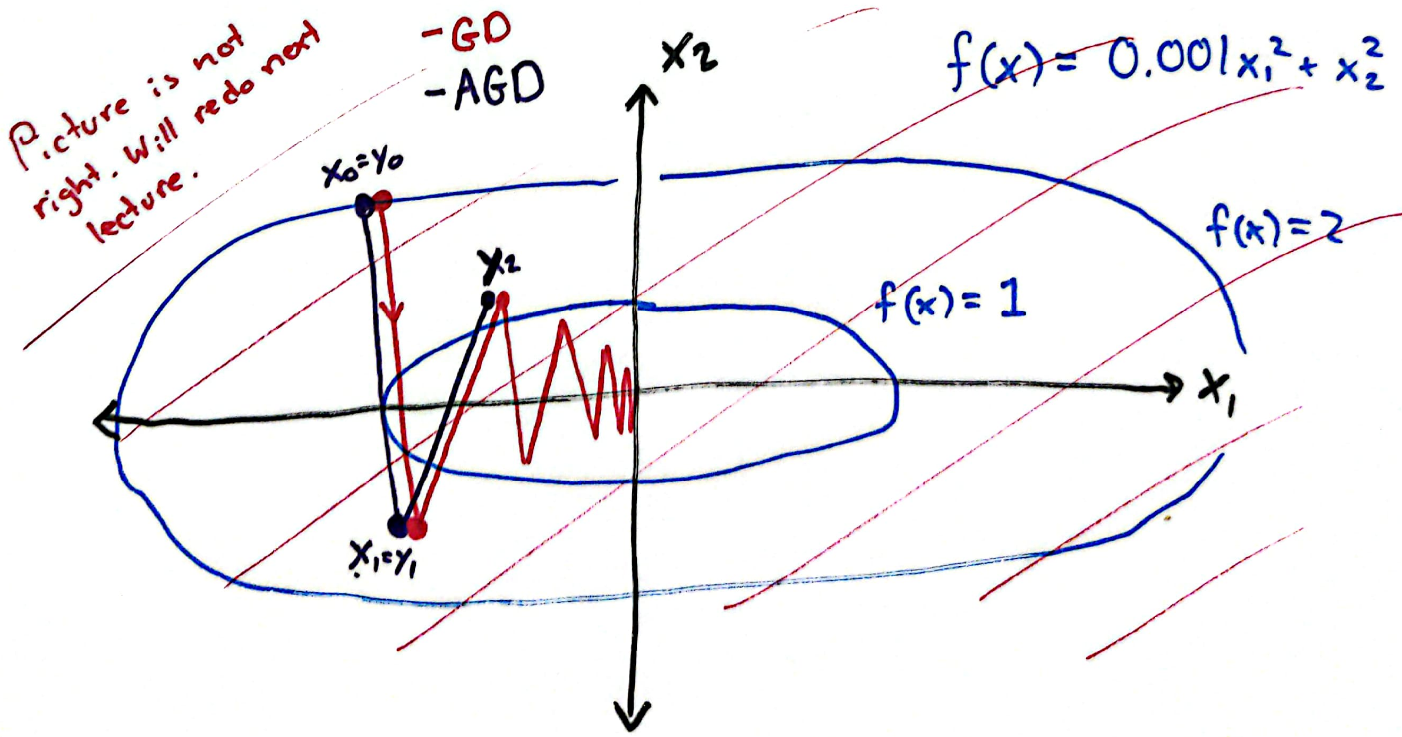
$$x_{k+1} = y_{k+1} + \left(\frac{\lambda_k - 1}{\lambda_{k+1}}\right)(y_{k+1} - y_k)$$

$$\frac{\lambda_k - 1}{\lambda_{k+1}} \approx \frac{k}{k+3}$$

where $\lambda_0 = 0$

$$\lambda_{k+1} = \frac{1+\sqrt{1+4\lambda_k^2}}{2}.$$

Note $x_k \in x_0 + \mathrm{Lin}\{\nabla f(x_0), \dots \nabla f(x_{k-1})\}$



Theorem Let f be convex with L-Lipschitz grad. Then for any minimizer x^*,

$$f(y_k) - f(x^*) \le \frac{2L\|x_0 - x^*\|^2}{k^2}.$$