From Previous Lec:

1 Basic descent methods

Take the form

$$x_{k+1} = x_k + \alpha_k p_k, \qquad k = 0, 1, \ldots$$

Definition 1. $p \in \mathbb{R}^d$ is a *descent direction* for f at x if

$$f(x + tp) < f(x)$$

for all sufficiently small $t > 0$.

Proposition 1. *If f is continuously differentiable (in a neighborhood of x), then any p such that $\langle \nabla f(x), p \rangle < 0$ is a descent direction.*

Proof. By Taylor's theorem:

$$f(x + tp) = f(x) + t \langle \nabla f(x + \gamma t p), p \rangle$$

for some $\gamma \in (0, 1)$. We know that $\langle \nabla f(x), p \rangle < 0$. As ∇f is continuous, for all sufficiently small $t > 0$,

$$\langle \nabla f(x + \gamma t p), p \rangle < 0,$$

hence $f(x + tp) < f(x)$. □

2 Gradient descent

What would be a good descent direction? Could try to move in the direction of $-\nabla f(x)$, since

$$-\frac{\nabla f(x)}{\|\nabla f(x)\|_2} = \arg \max_{\|p\|_2 = 1} \langle -\nabla f(x), p \rangle .$$

"Simplest" descent algorithm:

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

where α_k is the step size. Ideally, choose α_k small enough so that

$$f(x_{k+1}) < f(x_k)$$

when $\nabla f(x_k) \neq 0$.

Known as "gradient method", "gradient descent", "steepest descent" (w.r.t. the ℓ_2 norm).

3 Analysis of Gradient descent

Consider the gradient descent (GD) iteration with *constant* stepsize:

$$x_{k+1} = x_k - \underline{\alpha} \nabla f(x_k), \qquad \forall k = 0, 1, \ldots$$

Assumptions for this part:

(A1) $\boxed{f \text{ is } L\text{-smooth}}$ for $L < \infty$ (thus also continuously differentiable.)

(A2) $\mathcal{X} = \mathbb{R}^d$, i.e., the problem is $\boxed{\text{unconstrained}}$.

Note: we do *not* assume f is convex, until explicitly stated otherwise.

By property of L-smooth,

$$\forall y,\ f(y) \le \underbrace{f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + \frac{L}{2}\|y - x_k\|^2}_{\text{RHS}}. \quad \circledast$$

Set $x_{k+1} = \underset{y \in \mathbb{R}^d}{\arg\min}\ \text{RHS} = x_k - \frac{1}{L}\nabla f(x_k)$

$g(y) = \langle \nabla f(x_k), y - x_k \rangle + \frac{L}{2}\|y - x_k\|^2$ is cvx.

$$\nabla_y g(y) = 0 \Rightarrow \nabla f(x_k) + L\|y - x_k\| \cdot \frac{y - x_k}{\|y - x_k\|} = 0. \quad y^* = x_k - \frac{1}{L}\nabla f(x_k) = x_{k+1}.$$

$$\text{RHS}_{\min} = f(x_k) + g\left(x_k - \frac{1}{L}\nabla f(x_k)\right) = f(x_k) + \left\langle \nabla f(x_k), -\frac{1}{L}\nabla f(x_k)\right\rangle + \frac{L}{2}\left\|-\frac{1}{L}\nabla f(x_k)\right\|_2^2$$

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

Plug $\text{RHS}_{\min}$ back to $\circledast$, we have

$$f(x_{k+1}) \le f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|_2^2, \quad \text{when } x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$$

More generally, we have

Lemma 1 (Descent Lemma). *If* $x_{k+1} = x_k - \alpha \nabla f(x_k), \alpha \in (0, \frac{1}{L}]$, *then*

$$f(x_{k+1}) \le f(x_k) - \frac{\alpha}{2}\|\nabla f(x_k)\|_2^2.$$

Pf: By L-smoothness,

$$f(x_{k+1}) \le f(x_k) + \langle \nabla f(x_k), -\alpha \nabla f(x_k) \rangle + \frac{L}{2}\|-\alpha \nabla f(x_k)\|_2^2$$

$$= f(x_k) + \left(\frac{L}{2}\alpha^2 - \alpha\right)\|\nabla f(x_k)\|_2^2 \qquad 0 < \alpha \le \frac{1}{L} \Rightarrow L \le \frac{1}{\alpha}$$

$$\le f(x_k) - \frac{\alpha}{2}\|\nabla f(x_k)\|_2^2.$$

Remark 1. Eq. (1) gives an alternative way of deriving GD: we $\boxed{\text{minimize}}$ a upper bound of f, where the upper bound is constructed using the local information $\nabla f(x_k)$.

3.1. Case of General Smooth Function.

Repeatedly use Lemma I,

$$f(x_{k+1}) \le f(x_k) - \frac{\alpha}{2}\|\nabla f(x_k)\|_2^2$$

$$\le f(x_{k-1}) - \frac{\alpha}{2}\left(\|\nabla f(x_{k-1})\|_2^2 + \|\nabla f(x_k)\|_2^2\right)$$

$$\le \cdots \le f(x_0) - \frac{\alpha}{2}\sum_{i=0}^{k}\|\nabla f(x_i)\|_2^2$$

At $x_0, \cdots, x_k$, $f(x)$ is always smooth.

$$\Rightarrow f(x_0) - f(x_{k+1}) \ge \frac{\alpha}{2}\sum_{i=0}^{k}\|\nabla f(x_i)\|_2^2 .$$

Let $f^* = \inf_x f(x) > -\infty \quad \Rightarrow \quad f(x_0) - f^* \ge f(x_0) - f(x_{k+1})$

Meanwhile, $\frac{\alpha}{2}\sum_{i=0}^{k}\|\nabla f(x_i)\|_2^2 \ge \frac{\alpha(k+1)}{2}\min_{i\in[k]}\|\nabla f(x_i)\|_2^2$

$$\Rightarrow f(x_0) - f^* \ge \frac{\alpha(k+1)}{2}\min_{i\in[k]}\|\nabla f(x_i)\|_2^2 .$$

Got a bound for min grad norm:

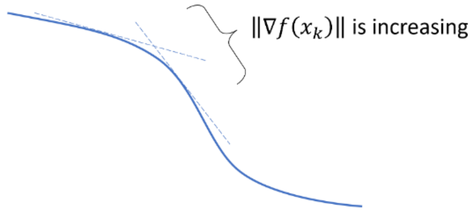
$$\min_{i\in[k]}\|\nabla f(x_i)\|_2^2 \le \frac{2(f(x_0)-f^*)}{\alpha(k+1)}$$

$$\Leftrightarrow \min_{i\in[k]}\|\nabla f(x_i)\| \le \sqrt{\frac{2(f(x_0)-f^*)}{\alpha(k+1)}}$$

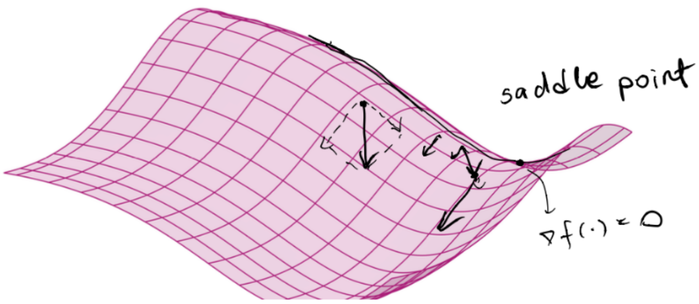
Set #Iter: $k+1 \ge \frac{2(f(x_0)-f^*)}{\alpha\varepsilon^2} \Rightarrow \min_{i\in[k]}\|\nabla f(x_i)\| \le \varepsilon. \Rightarrow$ We got ε-near stationary pt.

$$\frac{2(f(x_0)-f^*)}{\alpha\varepsilon^2} = O\left(\frac{C}{\varepsilon^2}\right) \text{ iterations.}$$

Remark 2. While function value $f(x_k)$ is decreasing in k, the gradient $\nabla f(x_k)$ need not.



Remark 3. When $\nabla f(x) = 0$, x may be a local min or a saddle point. Without further assumption, finding a $\boxed{\text{stationary point}}$ is the best we can hope for (recall the hard case mentioned at the end of Lecture 4). Under certain assumptions (which exclude the hard case), we can show that randomly initialized GD usually converges to a local min.[1] [2]



[1] "Gradient Descent Converges to Minimizers", Jason Lee, Max Simchowitz, Michael Jordan, Benjamin Recht, 2016.
[2] Plot by Jelena Diakonikolas

3.2 Convex Case.

Convexity gives us a bound:

for $x^* \in \arg\min_{x \in \mathbb{R}^d} f(x)$, $\quad \forall x, \; f(x^*) \ge f(x) + \langle \nabla f(x), x^* - x \rangle$

Goal: Bound the optimality gap $:= f(x_{k+1}) - f(x^*)$.

Recall GD's update rule:

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

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

$$(a-b)(c-a) = \tfrac{1}{2}(c-b)^2 - \tfrac{1}{2}(a-b)^2 - \tfrac{1}{2}(c-a)^2$$

$$f(x^*) \ge f(x_k) + \langle \nabla f(x_k), x^* - x_k \rangle$$

$$= f(x_k) + \frac{1}{\alpha}\langle x_k - x_{k+1}, x^* - x_k \rangle$$

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

$$= \underline{f(x_k) - \frac{\alpha}{2}\|\nabla f(x_k)\|_2^2} + \frac{1}{2\alpha}\|x_{k+1} - x^*\|^2 - \frac{1}{2\alpha}\|x_k - x^*\|^2$$

Descent lemma $\quad \ge f(x_{k+1}) + \frac{1}{2\alpha}\|x_{k+1} - x^*\|^2 - \frac{1}{2\alpha}\|x_k - x^*\|^2$

$$\Rightarrow \quad \|x_{k+1} - x^*\|_2^2 - \|x_k - x^*\|^2 \le 2\alpha\left(f(x^*) - f(x_{k+1})\right) \le 0.$$

Conclusion: GD never moves further away from the set of minimizers.

Further,

$$\sum_{k=0}^{K} f(x^*) - f(x_{k+1}) \ge \frac{1}{2\alpha}\sum_{k=1}^{K}\left(\|x_{k+1} - x^*\|^2 - \|x_k - x^*\|^2\right) = \frac{1}{2\alpha}\left(\|x_{K+1} - x^*\|^2 - \|x_0 - x^*\|^2\right)$$

$$\sum_{k=0}^{K} f(x_{k+1}) - f(x^*) \le \frac{1}{2\alpha}\left(\|x_0 - x^*\|^2 - \|x_{k+1} - x^*\|^2\right) \le \frac{1}{2\alpha}\|x_0 - x^*\|^2$$

$$\because f(x_1) > f(x_2) > \cdots > f(x_{K+1}) \qquad \therefore \text{LHS} \ge (K+1)\left(f(x_{K+1}) - f(x^*)\right)$$

$$\Rightarrow \quad f(x_{K+1}) - f(x^*) \le \frac{\|x_0 - x^*\|^2}{2\alpha(K+1)} \overset{\text{set}}{\le} \varepsilon.$$

$$\Rightarrow \quad k = \left\lceil \frac{\|x_0 - x^*\|^2}{2\alpha\varepsilon} \right\rceil. \quad \text{\# iterations.}$$

	# Iter
GD, smooth :	$O(\frac{1}{\varepsilon^2})$
GD, smooth + CVX :	$O(\frac{1}{\varepsilon})$

Remark 4 (Telescoping Sum). We just saw a pattern that will appear many times in the proofs this semester. We summarize this argument below:

Lemma 2. *Let* $\{a_k\}_{k\ge 0}$ *and* $\{D_k\}_{k\ge 0}$ *be sequences of real numbers, with* D_k *non-negative. If*

$$a_k \le D_k - D_{k+1} \quad \textit{for all } k,$$

then

$$\min_{0\le i\le k} a_i \le \frac{D_0}{k+1} \quad \textit{for all } k.$$

If in addition a_k *is non-increasing in* k*, then*

$$a_k \le \frac{D_0}{k+1} \quad \textit{for all } k.$$

Proof. Observe that

$$(k+1)\cdot \min_{0\le i\le k} a_i \le \sum_{i=0}^{k} a_i \le \sum_{i=0}^{k}(D_i - D_{i+1}) = D_0 - D_{k+1} \le D_0.$$

Moreover, when a_i is non-increasing in i, we have $\min_{0\le i\le k} a_i \ge a_k$. □

3.3. Strongly Convex Case. (Implicitly, there must be $m \leq L$)

m: strongly-cvx. L: smoothness

A similar bound is provided:

$$f(x^*) \geq f(x_k) + \langle \underbrace{\nabla f(x_k)}_{= \frac{1}{\alpha}(x_k - x_{k+1})}, x^* - x_k \rangle + \frac{m}{2}\|x^* - x_k\|^2$$

$$\geq f(x_{k+1}) + \frac{1}{2\alpha}\|x_{k+1} - x^*\|^2 - \frac{1}{2\alpha}\|x_k - x^*\|^2 + \frac{m}{2}\|x_k - x^*\|^2 \quad \cdots \text{same as cvx, } + \frac{m}{2}\|x^* - x_k\|^2$$

$$= f(x_{k+1}) + \frac{1}{2\alpha}\|x_{k+1} - x^*\|^2 + \left(\frac{m}{2} - \frac{1}{2\alpha}\right)\|x_k - x^*\|^2$$

$$\Rightarrow \quad \frac{1}{2\alpha}\|x_{k+1} - x^*\|^2 \leq \left(\frac{1}{2\alpha} - \frac{m}{2}\right)\|x_k - x^*\|^2 + \underbrace{f(x^*) - f(x_{k+1})}_{\leq 0}$$

$$\Rightarrow \quad \|x_{k+1} - x^*\|^2 \leq (1 - m\alpha)\|x_k - x^*\|^2.$$

$\begin{cases} \alpha \in (0, \frac{1}{L}] \\ m \leq L \end{cases} \Rightarrow m\alpha \leq L\alpha < 1.$

$\Rightarrow 0 < 1 - m\alpha < 1.$

$$\Rightarrow \quad \|x_{k+1} - x^*\|^2 \leq (1 - m\alpha)^{k+1}\|x_0 - x^*\|^2 \leq \varepsilon^2$$

set

$$k = O\left(\frac{1}{m\alpha}\log\left(\frac{\|x_0 - x^*\|}{\varepsilon}\right)\right) \quad \text{\# iterations.}$$

Above, we find 3 useful judge for "convergence".

Smooth: 1. $\|\nabla^2 f(x_k)\| \leq \varepsilon$. $\Rightarrow$ The best is to find local min. (ε-near stationary)

Smooth+cvx 2. $\|f(x_k) - f(x^*)\| \leq \varepsilon$.

Smooth + strongly cvx 3. $\|x_k - x^*\| \leq \varepsilon$.

Exercise 1. Show that we also have

$$f(x_{k+1}) - f(x^*) \le (1 - m\alpha)^{k+1} (f(x_0) - f(x^*)).$$

How about $\|\nabla f(x_{k+1})\|_2$?

Before this, we prove a lemma:

For smooth, strongly-cvx f, if x^* is its optimal sol, then,

$$\|f(x) - f(x^*)\| \le \frac{1}{2m} \|\nabla f(x)\|^2$$

Lemma pf:

By strong-cvx, $f(x^*) \ge f(x) + \langle \nabla f(x), x^* - x \rangle + \frac{m}{2} \|x^* - x\|^2$

$$f(x) - f(x^*) \le \langle \nabla f(x), x - x^* \rangle - \frac{m}{2} \|x - x^*\|^2.$$

Let $u = x - x^*$, $v = \nabla f(x)$.

$RHS = g(u) = v^T u - \frac{m}{2}\|u\|^2$. To get the tightest bound,

$\nabla g(u) = v - m\|u\| \cdot \frac{u}{\|u\|} = v - mu = 0$. $u^* = \frac{1}{m} v$

$x - x^* = \frac{1}{m} \nabla f(x)$.

$RHS_{min} = \frac{1}{m} \|\nabla f(x)\|^2 - \frac{1}{2m} \|\nabla f(x)\|^2 = \frac{1}{2m} \|\nabla f(x)\|^2$

$\Rightarrow \forall x$, $f(x) - f(x^*) \le \frac{1}{2m} \|\nabla f(x)\|^2$

Apply this lemma,

$$f(x_{k+1}) - f(x^*) \underset{\substack{\uparrow \\ \text{descent} \\ \text{lemma}}}{\le} f(x_k) - \frac{\alpha}{2} \underbrace{\|\nabla f(x_k)\|^2}_{\ge 2m(f(x_k) - f(x^*))} - f(x^*)$$

$$\leq f(x_k) - f(x^*) - m\alpha(f(x_k) - f(x^*)) = (1-m\alpha)(f(x_k) - f(x^*))$$

$$\Rightarrow f(x_{k+1}) - f(x^*) \leq \underline{(1-m\alpha)^{k+1}}(f(x_0) - f(x^*))$$

$\searrow$ This converges faster than cvx case $\left(\frac{1}{\alpha(k+1)}\right)$

About $\|\nabla f(x_{k+1})\|_2$:

Another Lemma :

For L-smooth f, if x^* is a local min $(\nabla f(x^*) = 0)$

Then $$\boxed{\forall x, \ \|\nabla f(x)\|^2 \leq 2L(f(x) - f(x^*))}$$

pf: $x - \frac{1}{L}\nabla f(x)$ is not necessarily optimal $\Rightarrow f(x^*) \leq f(x - \frac{1}{L}\nabla f(x))$

Apply descent lemma to $f(x - \frac{1}{L}\nabla f(x))$. $\alpha \in (0, \frac{1}{L}]$, $m \leq L$

$$f(x^*) \leq f(x - \tfrac{1}{L}\nabla f(x)) \leq f(x) + \langle \nabla f(x), -\tfrac{1}{L}\nabla f(x)\rangle + \underbrace{\tfrac{m}{2}}_{\leq \frac{L}{2}}\left\|-\tfrac{1}{L}\nabla f(x)\right\|_2^2$$

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

$$\Rightarrow \|\nabla f(x)\|^2 \leq 2L(f(x) - f(x^*))$$

Then, apply this lemma, follow above,

$$\|\nabla f(x)\|^2 \leq 2L(f(x_{k+1}) - f(x^*)) \leq 2L(1-m\alpha)^{k+1}(f(x_0) - f(x^*))$$