From last lec:

Exercise 1. For a simple quadratic function $f(x) = \|x - x^*\|_2^2$, all measures of optimality (optimality gap $f(x) - f(x^*)$, gradient norm $\|\nabla f(x)\|_2^2$ and distance to optimum $\|x - x^*\|_2^2$) are equivalent up to constants. The same is true for a function that is both strongly convex and smooth, as such a function is sandwiched between two quadratics. With this in mind, you can try to prove geometric convergence in function value:

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

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

Remark 1. The bounds in (1) and (2) depend on $\boxed{m\alpha}$, which equals $\boxed{\frac{m}{L}}$ if we take $\alpha = \frac{1}{L}$. Note that $\frac{L}{m}$ is (an upper bound of) the condition number of the Hessian $\nabla^2 f$. Fast convergence if $\nabla^2 f$ is well-conditioned.

$\alpha \leq \frac{1}{L}$

1.2 Unknown L

All previous analysis is valid when we use a stepsize $\alpha \leq \frac{1}{L}$, which requires knowing L, or at least an upper bound of L. How to choose α if we don't know L?

1.2.1 Trial and error

For example:

- Choose the largest α for which GD does not diverge.
- Use your lucky number as the initial value of α. Adjust and see if it works better.

Popular among machine learning practitioners. For example, PyTorch, a popular package for training neural networks, implements several variants of GD with default stepsizes like 0.01 or 0.001, which is the starting point for most users.

Methods to choose l.r. :

1.2.2 Exact line search

Choose α as the solution to the *one-dimensional* optimization problem

$$\min_{\alpha > 0} f(x_k - \alpha \nabla f(x_k)) .$$

That is, we find the exact minimum of f along the half line $\{x_k - \alpha \nabla f(x_k) : \alpha > 0\}$.

This method is most useful when f has some special structure so that the above 1-D problem can be solved efficiently at low cost.

1.2.3 Backtracking line search

Start with some initial α_0. Sequentially try stepsize $\alpha_0, \frac{1}{2}\alpha_0, \frac{1}{4}\alpha_0, \frac{1}{8}\alpha_0, \ldots$ until the descent condition

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

is satisfied. Backtracking terminates before or when $2^{-t}\alpha_0 \leq \frac{1}{L}$ is satisfied for the first time, so it requires no more than $\boxed{O(\log(\alpha_0 L))}$ function evaluations of f (and one gradient computation at x_k).

This method is useful when function evaluation is easy but the exact linear search problem is costly to solve.

1. **Preconditioned methods:**

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

$S_k \succeq 0$, $\forall \lambda \in \sigma(S_k)$ $\gamma_1 \le \lambda \le \gamma_2$

where S_k is a symmetric positive definite matrix with all eigenvalues in $[\gamma_1, \gamma_2]$, $0 < \gamma_1 < \gamma_2 < \infty$.
From properties of L-smooth functions (Lemma 1 in Lecture 4):

$$\begin{aligned}
f(x_{k+1}) &\le f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|_2^2 \\
&= f(x_k) - \alpha \underbrace{\langle S_k \nabla f(x_k), \nabla f(x_k) \rangle}_{\ge \gamma_1 \|\nabla f(x_k)\|_2^2} + \frac{L}{2}\alpha^2 \underbrace{\|S_k \nabla f(x_k)\|_2^2}_{\le \gamma_2^2 \|\nabla f(x_k)\|_2^2} \\
&\le f(x_k) - \underbrace{\left(\alpha\gamma_1 - \frac{L}{2}\gamma_2^2\alpha^2\right)}_{>0 \text{ for sufficiently small } \alpha} \|\nabla f(x_k)\|_2^2.
\end{aligned}$$

$\lambda_i' = \frac{1}{\lambda_i}$

Newton's method uses $S_k = (\nabla^2 f(x_k))^{-1}$; need $\nabla^2 f(x_k)$ to have positive eigenvalues for this to work.
With appropriately chosen S_k, preconditioned methods can converge substantially faster near x^* than GD.

2. **Gauss-Southwell (aka greedy coordinate descent):**

$\nabla f(x_k)$

Only 1 coordinate update each iteration.

$$x_{k+1} = x_k - \alpha \underbrace{\nabla_{i_k} f(x_k) e_{i_k}}_{-p_k}$$

$p_k = -\nabla_{i_k} f(x_k) e_{i_k} = \|\nabla f(x_k)\|_\infty e_{i_k}$

where $i_k = \arg\max_{1 \le i \le d} \{-\nabla_i f(x_k)\}$, and $e_{i_k} = [0, 0, \ldots, \underbrace{1}_{i_k \text{ position}}, \ldots, 0]$ is the i_k-th standard basis vector in $\mathbb{R}^d$. Note that

$$\|p_k\|_2^2 \ge \frac{1}{d}\|\nabla f(x_k)\|_2^2,$$

$\|\nabla f(x_k)\|_2^2 = \sum_{i=1}^{d} |\nabla_i f(x_k)|^2 \le d\,\|\nabla f(x_k)\|_\infty^2 = d\,\|p_k\|_2^2$

hence one can show that (exercise) for $\alpha = \frac{1}{L}$,

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

This algorithm is most useful when i_k and $\nabla_{i_k} f(x_k)$ are much easier to compute than the full gradient $\nabla f(x_k)$.
Can be viewed as greedy descent w.r.t. ℓ_1 norm.

$\ell_1 \to \ell_\infty$. dual
$\Downarrow$
measure for grad!

Pros: cost effective when single coor grad is easy to compute.

cons: As dim $\uparrow$, converges slower. ($\frac{1}{2Ld}$ dominates) $\alpha = \frac{1}{L}$

By L-smoothness,

$$\begin{aligned}
f(x_{k+1}) &\le f(x_k) + \langle \nabla f(x_k), \alpha p_k \rangle + \frac{L}{2}\|\alpha p_k\|^2 \\
&= f(x_k) - \alpha \nabla_{i_k} f(x_k) \langle \nabla f(x_k), e_{i_k} \rangle + \frac{L\alpha^2}{2}|\nabla_{i_k} f(x_k)|^2 \\
&= f(x_k) + \left(\frac{L}{2}\alpha^2 - \alpha\right)(\nabla_{i_k} f(x_k))^2 \\
&= f(x_k) - \frac{1}{2L}(\nabla_{i_k} f(x_k))^2 \le f(x_k) - \frac{1}{2Ld}\|\nabla f(x_k)\|^2
\end{aligned}$$

3. **Randomized coordinate descent.** Similar to above, with i_k chosen uniformly at random from $\{1, 2, \ldots, d\}$. See HW2.

Randomly, uniformly pick $i_k \in \{1, 2, \cdots, d\}$.

$$\begin{cases} x_{k+1}^{(i_k)} = x_k^{(i_k)} - \alpha_k [\nabla f(x_k)]_{i_k} \\ x_{k+1}^{(j)} = x_k^{(j)} & j \neq x_k. \end{cases}$$

$$\mathbb{E}[f(x_{k+1}) | x_k] = \frac{1}{d} \sum_{i=1}^{d} f(x_k - \alpha_k [\nabla f(x_k)]_i e_i)$$

$$\leq \frac{1}{d} \sum_{i=1}^{d} \left(f(x_k) + \langle -\alpha_k [\nabla f(x_k)]_i e_i, \nabla f(x_k) \rangle + \frac{L}{2} \| -\alpha_k [\nabla f(x_k)]_i e_i \|_2^2 \right)$$

Let $P_i = -[\nabla f(x_k)]_i e_i$. Similarly we have $\|P_i\|^2 \geq \frac{1}{d} \|\nabla f(x_k)\|^2$

$$= \frac{1}{d} \sum_{i=1}^{d} \left(f(x_k) + \langle \alpha_k P_i, \nabla f(x_k) \rangle + \frac{L\alpha_k^2}{2} \|P_i\|^2 \right)$$

$$= \frac{1}{d} \sum_{i=1}^{d} \left(f(x_k) + \left(\frac{L}{2}\alpha_k^2 - \alpha_k\right) (\nabla_i f(x_k))^2 \right)$$

$\alpha_k = \frac{1}{L}$ $\forall k$ $$= f(x_k) - \frac{1}{2Ld} \sum_{i=1}^{d} (\nabla_i f(x_k))^2 = f(x_k) - \frac{1}{2Ld} \|\nabla f(x_k)\|^2.$$

2.3 收敛速度区别 + 联系:

1. 都是 $f(x_k) - \frac{1}{2Ld} \|\nabla f(x_k)\|^2$ ... rate.

2. Gauss–Southwell 要求计算各 coor grad. $\Rightarrow$ Converge in value (更强)
   Uniform rand coor grad ... sample 1 coor (cost effective) $\Rightarrow$ converge in expectation

4. **Stochastic gradient descent,** where

$$x_{k+1} = x_k - \alpha g(x_k, \xi_k),$$

where $\boxed{\xi_k\text{'s}}$ are i.i.d. random variable (sample) satisfying $\mathbb{E}_{\xi_k}[g(x_k, \xi_k)] = \nabla f(x_k)$. That is, $g(x_k, \xi_k)$ is an unbiased (but potentially very noisy) estimate of the true gradient at x_k. Under certain assumptions it satisfies the descent condition *in expectation*.

"General GD".

5. **Gradient descent w.r.t. ℓ_p norm**, where

$$x_{k+1} = \arg\min_u \left\{ f(x_k) + \langle \nabla f(x_k), u - x_k \rangle + \frac{1}{\alpha} \|u - x_k\|_p^2 \right\}.$$

See HW2.

(proximal point of view)

Note: standard GD: $X_{k+1} = \arg\min_u \{ f(x_k) + \langle \nabla f(x_k), u - x_k \rangle + \frac{1}{2\alpha} \|u - x_k\|_2^2 \}$.

$\iff X_{k+1} = X_k - \alpha \nabla f(x_k)$

pf: Let $v = u - x_k$ and $g(v) = f(x_k) + \langle \nabla f(x_k), v \rangle + \frac{1}{2\alpha} \|v\|_2^2$

$g(v)$ is cvx w.r.t v. $\nabla g(v) = 0$. $\Rightarrow \nabla f(x_k) + \frac{1}{\alpha} v^* = 0$

$\Rightarrow v^* = -\alpha \nabla f(x_k) = u^* - x_k = x_{k+1} - x_k$.

$x_{k+1} = x_k - \alpha \nabla f(x_k)$. $\longrightarrow$ l_2-norm primal space is the same as its dual space. $(\frac{1}{2} + \frac{1}{2} = 1)$

Hence this simplified version is okay.

Understand

$$X_{k+1} = \arg\min_u \{ \underbrace{f(x_k) + \langle \nabla f(x_k), u - x_k \rangle}_{} + \underbrace{\frac{1}{2\alpha} \|u - x_k\|_p^2}_{} \}.$$

① linear approximate of f at x_k.

$\Downarrow$

We want smaller $f(x_k)$

② proximity term. (regularizer)

$\Downarrow$

We want x_{k+1}, x_k are close.

6. **Mirror descent,** where

$$x_{k+1} = \arg\min_{u \in \mathcal{X}} \left\{ f(x_k) + \langle \nabla f(x_k), u - x_k \rangle + \frac{1}{\alpha_k} D_\psi(u, x_k) \right\},$$

and $D_\psi(\cdot, \cdot)$ is the Bregman divergence generated by ψ. See HW2.

$D_\varphi(y, x) := \varphi(y) - \varphi(x) - \langle \nabla \varphi(x), y - x \rangle$. ... A new metric for distance.

defined for $\boxed{\text{cvx}}$ φ.

Properties:

1. non-negative. φ is cvx $\Rightarrow$ $D_\varphi(y,x) = \varphi(y) - \varphi(x) - \langle \nabla\varphi(x), y-x \rangle \geq 0$.

$D_\varphi(y,x) = 0$ iff $y = x$. $\Rightarrow$ Can act as a "distance"

2. for fixed x, $D_\varphi(y,x)$ is cvx w.r.t. y. (obvious).

$\Rightarrow$ $x_{k+1} = \arg\min_u \{ f(x_k) + \langle u, x_k \rangle + \frac{1}{\alpha_k} D_\varphi(x_k, u) \}$ is a cvx problem when x_k fixed, which is ensured as Mirror Descent (MD) is iterative.

① Proximal Point of View (simplified):

$$x_{k+1} = \arg\min_u \{ \alpha_k \langle u, x_k \rangle + D_\varphi(x_k, u) \}.$$

② Mirror Space point of view:

$x_t \in$ primal space, $\nabla f(x_t) \in$ dual space.

Mirror map:

$\nabla\varphi$: primal $\mapsto$ dual.

Algo:

1. Map to dual space. $\theta_k = \nabla\varphi(x_k)$
2. GD on dual space. $\theta_{k+1} = \theta_k - \alpha_k \nabla f(x_k)$
3. Map back to primal. $\overline{x_{k+1}} = (\nabla\varphi)^{-1}(\theta_{k+1})$
4. Project to constraint. $x_{k+1} = \arg\min_{x\in K} D_\varphi(x, \overline{x_{k+1}})$

Proximal point of view $\Leftrightarrow$ Mirror space point of view.

pf: ($\Leftarrow$ revertible)

$$x_{k+1} = \arg\min_{x\in K} D_\varphi(x, \overline{x_{k+1}})$$

$$= \underset{x\in K}{\arg\min}\ \varphi(x) - \langle \nabla\varphi(x_{k+1}), x\rangle$$

$$= \underset{x\in K}{\arg\min}\ \varphi(x) - \langle \nabla\varphi(x_k) - \alpha_k \nabla f(x_k), x\rangle$$

const.

$$= \underset{x\in K}{\arg\min}\ \alpha_k\langle \nabla f(x_k), x\rangle + \varphi(x) - \underline{\varphi(x_k)} - \langle \nabla\varphi(x_k), x\rangle$$

$$= \underset{x\in K}{\arg\min}\ \alpha_k\langle \nabla f(x_k), x\rangle + D_\varphi(x, x_k)$$

... proximal pt. of view

3 Convergence of descent methods

Consider any iterative method that generates a sequence $x_0, x_1, \ldots$ satisfying the descent condition

$$f(x_{k+1}) \leq f(x_k) - \frac{\beta}{2}\|\nabla f(x_k)\|_2^2, \quad \forall k \geq 0 \tag{3}$$

for some $\beta > 0$.

For $x_{k+1} = x_k - \beta\nabla f(x_k)$. f is L-smooth & $\beta \in (0, \frac{1}{L}]$. This cond holds.

Q: Only descent methods enjoy this?

3.1 General case

Assume f is bounded below: $f(x) \geq f_* > -\infty, \forall x$. The same analysis from previous lecture applies and gives

$$\min_{0\leq i\leq k} \|\nabla f(x_i)\|_2 \leq \sqrt{\frac{2(f(x_0) - f_*)}{\beta(k+1)}}.$$

Apply descent lemma inductively.

Next: Bounds without update rule (不要求 GD 及其 gen version).

Only assume $f(x_{k+1}) \leq f(x_k) - \frac{\beta}{2}\|\nabla f(x_k)\|_2^2$ holds.

3.2. Cvx + Smooth Loss. + Bounded data pts.

Let $R_0 := \max\{\|x-x^*\|_2 : f(x) \le f(x_0)\} < \infty$.

$$\Delta_k = f(x_k) - f(x^*) \underset{\substack{\uparrow \\ \text{cvx}}}{\le} \langle \nabla f(x_k), x_k - x^* \rangle \le R_0 \|\nabla f(x_k)\|_2. \qquad \|\nabla f(x_k)\|_2 \ge \frac{\Delta_k}{R_0}$$

Plug to descent Lemma,

$$f(x_{k+1}) \le f(x_k) - \frac{\beta}{2}\|\nabla f(x_k)\|_2^2 \le f(x_k) - \frac{\beta}{2R_0^2}\Delta_k^2$$

$$\Rightarrow f(x_{k+1}) - f(x^*) \le f(x_k) - f(x^*) - \frac{\beta}{2R_0^2}\Delta_k^2$$

$$\Rightarrow \Delta_{k+1} \le \Delta_k\left(1 - \frac{\beta}{2R_0^2}\Delta_k\right) \underset{\substack{\uparrow \\ 1-x \le \frac{1}{1+x}}}{\le} \Delta_k \frac{1}{1+\frac{\beta}{2R_0^2}\Delta_k} = \frac{1}{\frac{1}{\Delta_k} + \frac{\beta}{2R_0^2}}$$

$$\Rightarrow \frac{1}{\Delta_{k+1}} \ge \frac{1}{\Delta_k} + \frac{\beta}{2R_0^2} \ge \cdots \ge \underbrace{\frac{1}{\Delta_0}}_{\text{probably small.}} + \frac{\beta(k+1)}{2R_0^2} \ge \frac{\beta(k+1)}{2R_0^2}$$

$$\therefore \quad \Delta_{k+1} = f(x_{k+1}) - f(x^*) \le \frac{2R_0^2}{\beta(k+1)}.$$

Compare with the bound for GD (cvx, smooth): $f(x_{k+1}) - f(x^*) \le \dfrac{\|x_0 - x^*\|^2}{2\alpha(k+1)}$

3.3. Strongly cvx + Smooth Loss

$\Rightarrow$ unique minimizer x^*.

Recall: $\boxed{\text{For } m\text{-strongly cvx } f,\ \|\nabla f(x)\|_2^2 \ge 2m(f(x) - f(x^*))}$

Plug this in descent lemma,

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

$$\le f(x_k) - m\beta\left(f(x_k) - f(x^*)\right)$$

$$f(x_{k+1}) - f(x^*) \le (1-m\beta)\left(f(x_k) - f(x^*)\right) \le \cdots \le (1-m\beta)^{k+1}\left(f(x_0) - f(x^*)\right)$$

Remark:

$\underbrace{\|\nabla f(x)\|_2^2 \ge 2m(f(x) - f(x^*))}$ holds for some f that is not strongly convex.

$\Downarrow$

PL condition / Grad domination cond.

Ex 2. Prove $f(x) = \frac{1}{2}x^TAx$, $A \succeq 0$, singular. (f is not strongly-cvx) satisfies PL cond with $m = ?$

pf: $\nabla f(x) = Ax$. $f(x^*) = 0$.

$$\frac{\|\nabla f(x)\|^2}{2f(x)} = \frac{\|Ax\|^2}{x^TAx} = \frac{x^TA^TAx}{x^TAx} = \frac{x^TA^2x}{x^TAx} = \frac{x^TA^{\frac{1}{2}}AA^{\frac{1}{2}}x}{x^TA^{\frac{1}{2}}A^{\frac{1}{2}}x} = \frac{y^TAy}{y^Ty} \ge \lambda_{min}(A). \checkmark$$

Let $y = A^{\frac{1}{2}}x$

Rayleigh Quotient.

Ex3. Find a non cvx function that satisfies PL-cond.

4 Other generalizations of strong convexity

A strongly convex function cannot be flat near the minimum: the function value must grow when moving away from the minimizer. There are several other conditions that also control the growth of a function and hence can be viewed as generalizations of strong convexity.

Recall the definition of strong convexity:

$$f((1-\alpha)x+\alpha y) \le (1-\alpha)f(x)+\alpha f(y) - \frac{m}{2}(1-\alpha)\alpha \|y-x\|_2^2, \qquad \forall x,y,\forall \alpha \in (0,1). \tag{6}$$

$$\Longleftrightarrow f(y) \ge f(x) + \langle \nabla f(x), y-x\rangle + \frac{m}{2}\|y-x\|_2^2, \qquad \forall x,y. \tag{7}$$

One may ① replace the ℓ_2 norm on the right hand side by another norm $\|\cdot\|$, or ② by another polynomial of norm $\|y-x\|^r$ (*uniform convexity*).

There are further generalization to nonconvex functions. We have talked about the PL condition (5). $\boxed{\text{PL}}$ can be generalized further to the *Kurdyka-Łojasiewicz (KL)* condition, which is (5) with $\|\nabla f(x)\|^r$ on the LHS. Another generalization is known as the *sharpness* condition or *Holderian error bounds*: a function is called (r,m)-sharp if

$$f(x) - \min_y f(y) \ge \frac{m}{r} \min_{x^* \in \mathcal{X}^*} \|x-x^*\|^r, \quad \forall x$$

where $\mathcal{X}^* := \arg\min_{x\in\mathbb{R}^d} f(x)$ denotes the set of minimizers.

Ex 4. prove a m-strongly cvx function is (2,m)-sharp.

pf: $f(x) - f(x^*) \ge \frac{m}{2}\|x-x^*\|^2$ unique x^*.

Directly derived by stronly convexity & $\nabla f(x^*) = 0$.

All those conds. enables faster convergence (than merely assuming smoothness).

5 Generalization of smoothness (optional)

Complementary to the above "growth" conditions, the smoothness condition stipulates that a function cannot grow/fluctuate too quickly. One may generalize smoothness by replacing Lipschitz-continuity of gradient by $\boxed{\text{Holder-continuity.}}$

Definition 1. A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is called (κ, L)-weakly smooth for $\kappa \in [1,2]$ w.r.t. a norm $\|\cdot\|$ if there exists a constant $L < \infty$ such that

$$\|\nabla f(x) - \nabla f(y)\|_* \le L\|x-y\|^{\kappa-1}, \qquad \forall x, y.$$

(2,L) weak smooth ⟺ L-smooth.

(1,L) weak smooth ⟺ L-lipschitz

$(2,L)$-weak smoothness is the same as the usual L-smoothness. $(1,L)$-weak smoothness means $\|\nabla f(x) - \nabla f(y)\|_* \le L$, which implies Lipschitz continuity of f.

Example 1. Examples of (weak) smoothness:

1. The log-sum-exp (soft-max) function $f(x) = \log \sum_{i=1}^d e^{x_i}$ is 1-smooth w.r.t. $\|\cdot\|_\infty$.
2. $\frac{1}{2}\|x\|_p^2$ with $p \ge 2$ is $(p-1)$-smooth w.r.t. $\|\cdot\|_p$.
3. $\frac{1}{2}\|x\|_p^p$ with $p \in [1,2]$ is $(p,1)$-weakly smooth w.r.t. $\|\cdot\|_p$.