- Subgradient methods

$$x^{(k+1)} = x^{(k)} - \alpha_k \cdot g^{(k)}$$

$g^{(k)}$ is any subgradient of f at x

not a descent method. need to keep track of the best point so far: $f_{best}^{(k)} = \min_i f(x^{(i)})$

- Step-size rules

Step sizes are fixed ahead of time

1. Costant step size: $\alpha_k = \alpha$ (costant)
2. Costant step length: $\alpha_k = \gamma / \|g^{(k)}\|_2$ (so $\|x^{(k+1)} - x^{(k)}\|_2 = \gamma$)
3. square summable but not summable:
   step sizes satisfy: $\sum_{k=1}^{\infty} a_k^2 < \infty$ $\sum_{k=1}^{\infty} a_k = \infty$
4. nonsummable diminishing:
   step sizes satisfy: $\lim_{k\to\infty} a_k = 0$ $\sum_{k=1}^{\infty} a_k = \infty$

- Assumptions

$f^* = \inf_x f(x) > -\infty$ with $f(x^*) = f^*$

$\|g\|_2 \le G$ for all $g \in \partial f$ (equivalent to Lipschitz condition on f)

$R \geq \|x^{(1)} - x^*\|_2$ (can take "=" here)

- Convergence results

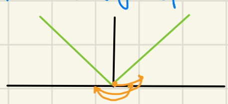
define $\bar{f} = \lim_{k\to\infty} f_{best}^{(k)}$

constant step size: $\bar{f} - f^* \le G^2 \cdot \alpha/2$

converges to $G^2 \cdot \alpha/2$ suboptimal

converges to f^* if f differential and α small enough

eg. $\min |x|$, returning $\partial f(x) = 0.75$ at $x=0$, which is a valid subgradient



end up oscillating

constant step length: $\bar{f} - f^* \le G\gamma/2$ converges to $G\gamma/2$ suboptimal

diminishing step size: $\bar{f} = f^*$ converges

eg. $\alpha^{(k)} = 1/k$

· Convergence proof

Euclidean distance to the optimal set decreases (not the function value)

let x^* be minimizer of f

$$\|x^{(k+1)} - x^{*}\|_2^2 = \|x^{(k)} - a_k g^{(k)} - x^*\|_2^2$$

$$= \|x^{(k)} - x^*\|_2^2 - 2a_k \cdot g^{(k)T}(x^{(k)} - x^*) + a_k^2 \|g^{(k)}\|_2^2 \qquad f(x^*) \geq f(x^{(k)}) + g^{(k)T}(x^* - x^{(k)})$$

$$\leq \|x^{(k)} - x^*\|_2^2 \underbrace{- 2a_k (f(x^{(k)}) - f(x^*))}_{\text{good term}} + \underbrace{a_k^2 \|g^{(k)}\|_2^2}_{\text{bad term}} \qquad f(x^*) - f(x^{(k)}) \geq g^{(k)T}(x^* - x^{(k)})$$

$\Downarrow$ apply recursively

$$\|x^{(k+1)} - x^*\|_2^2 \leq \|x^{(1)} - x^*\|_2^2 - \sum_{i}^{k} 2a_i (f(x^{(i)}) - f^*) + \sum_i a_i^2 \|g^{(i)}\|_2^2$$

$$\leq R^2 - \sum_i^k 2a_i (f(x^{(i)}) - f^*) + G^2 \cdot \sum_i a_i^2$$

(subgradient does not necessarily decrease, e.g. $|x|$)

$$\sum_{i=1}^{k} a_k (f(x^{(i)}) - f^*) \geq (f_{best}^{(k)} - f^*) \cdot (\sum_i a_i)$$

$$\leq R^2 - 2 \cdot (\sum_i a_i) \cdot (f_{best}^{(k)} - f^*) + G^2 \cdot (\sum_i a_i^2)$$

$\Downarrow$

$$f_{best}^{(k)} - f^* \leq \frac{R^2 - \|x^{(k+1)} - x^*\|_2^2 + G^2 \sum_i a_i^2}{2\sum_i a_i}$$

$$f_{best}^{(k)} - f^* \leq \frac{R^2 + G^2 \sum_i^k a_i^2}{2\sum_i^k a_i} \qquad \left(\sum_i^{\infty} a_i = \infty \qquad \sum_i^{\infty} a_i^2 < \infty \right)$$

1. constant step size: $a_k = a$

$$f_{best}^{(k)} - f^* \leq \frac{R^2 + G^2 \cdot k a^2}{2ka} \rightsquigarrow \frac{aG^2}{2} \text{ as } k \to +\infty$$

2. constant step length: $a_k = \gamma / \|g^{(k)}\|_2$

$$f_{best}^{(k)} - f^* \leq \frac{R^2 + \sum_i^k a_i^2 \|g^{(i)}\|_2^2}{2\sum_i^k a_i} \leq \frac{R^2 + \gamma^2 k}{2\gamma k / G} \rightsquigarrow \frac{\gamma G}{2} \text{ as } k \to +\infty$$

3. square-summable but not summable $(1/k)$

$$\sum_i^{\infty} a_i = \infty \qquad \sum_i^{\infty} a_i^2 < \infty$$

$$f_{best}^{(k)} - f^* \leq \frac{R^2 + G^2 \cdot \sum_i a_i^2}{2\sum_i a_i} \rightsquigarrow 0 \text{ as } k \to +\infty$$

• Stopping criterion

terminate when $\frac{R^2 + G^2 \cdot \sum_i^k a_i^2}{2\sum_i^k a_i} \le \epsilon$ is very slow

optimal choice of a_i to achieve $\frac{R^2 + G^2 \cdot \sum_i^k a_i^2}{2\sum_i^k a_i} \le \epsilon$ for smallest k

choice of a_i's for fixed k that minimizes $\frac{R^2 + G^2 \cdot \sum_i^k a_i^2}{2\sum_i^k a_i} \ge f_{best}^{(k)} - f^*$

$\underbrace{\min_a (R^2 + G^2 \cdot \sum_i^k a_i^2) / 2 (\sum_i^k a_i)}$ quadratic-over-linear, $\sum_i^k a_i > 0$ } hence convex

$\Downarrow$

symmetric over a

if elements of a^* is not all equal, and a^* is optimal

a permutation of a^* is $P(a^*) \ne a^*$ is still optimal (symmetric over a)

thus there're 2 different optimal

$\min_a (R^2 + (ka^2) G^2) / 2ka$

$\min_a \frac{R^2}{2ka} + G^2 a/2$ gradient $-\frac{R^2}{2k} 1/a^2 + \frac{G^2}{2} = 0$

$a_i = (R/G)/\sqrt{k}$

number of steps required: $((RG)/\epsilon)^2$

there is not a good stopping criterion for subgradient method

eg. Piecewise linear minimization

$\min.\ f(x) = \max_i (a_i^T x + b_i)$

a subgradient of f: find index j for which

$a_j^T x + b_j = \max_i a_i^T x + b_i$

take $g = a_j$

$x^{(k+1)} = x^{(k)} - a_k \cdot g$

· Optimal step size if f^* is known

Choice due to Polyak:

$$a_k = \frac{f(x^{(k)}) - f^*}{\|g^{(k)}\|_2^2} \qquad \text{(can also use when } f^* \text{ is estimated)}$$

Motivation:

$$\|x^{(k+1)} - x^*\|_2^2 = \|x^{(k)} - x^*\|_2^2 - 2a_k g^{(k)T}(x^{(k)} - x^*) + \|a_k \cdot g^k\|_2^2$$

$$f(x^*) \geq f(x^{(k)}) + g^T(x^* - x^{(k)})$$

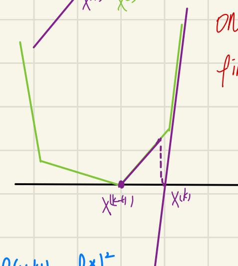
$$\leq \|x^{(k)} - x^*\|_2^2 - 2a_k(f(x^{(k)}) - f^*) + a_k^2 \cdot \|g^k\|_2^2 \qquad g^T(x^* - x^{(k)}) \leq f(x^*) - f(x^{(k)})$$

take the grad of rhs: $2(f^* - f(x^{(k)})) + 2a_k \cdot \|g^{(k)}\|_2^2 := 0$

$f(x^{(k)}) + g^T(x - x^{(k)}) := 0$

$$a_k = \frac{f(x^{(k)}) - f^*}{\|g^{(k)}\|_2^2}$$

this is a step size appropriate for non-differentiable funcs, not that appropriate for differentiable functions



one step for absolute value

finite steps for piecewise linear

$$\|x^{(k+1)} - x^{(k)}\|_2^2 \leq \|x^{(k)} - x^*\|_2^2 - \frac{(f(x^{(k)}) - f^*)^2}{\|g^{(k)}\|_2^2}$$

$\|x^{(k)} - x^*\|_2^2$ decreases each steps

goes down by a lot if x^k is quite suboptimal

applying recursively:

$$\sum_{i=1}^{k} \frac{(f(x^{(i)}) - f^*)^2}{\|g^{(i)}\|_2^2} \leq R^2$$

$$\sum_{i=1}^{k} (f(x^{(i)}) - f^*)^2 \leq R^2 G^2$$

$$f(x^{(i)}) \longrightarrow f^*$$

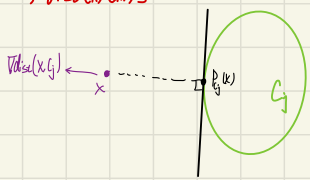
eg. find a point in the intersection of convex sets

$C = C_1 \cap C_2 \cdots \cap C_m$ is nonempty. $C_1, \cdots C_m \in \mathbb{R}^n$ close and convex

find a point in C by min. $\max\{\mathrm{dist}(x, C_1), \cdots, \mathrm{dist}(x, C_m)\}$

with $\mathrm{dist}(x, C_j) = f(x)$, a subgradient of f is

$$g = \nabla \mathrm{dist}(x, C_j) = \frac{x - P_{C_j}(x)}{\|x - P_{C_j}(x)\|_2}$$



Subgradient update with optimal step size

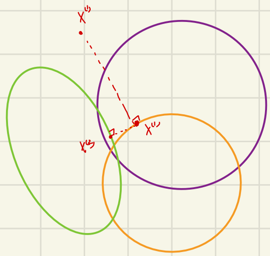
$$x^{(k+1)} = x^{(k)} - \alpha_k \cdot g^{(k)}$$

$$= x^{(k)} - \frac{f(x^{(k)}) - 0}{\|g^{(k)}\|_2^2} \cdot \frac{x^{(k)} - P_{C_j}(x^{(k)})}{\|x^{(k)} - P_{C_j}(x^{(k)})\|_2}$$

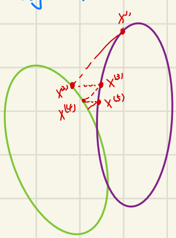
$$= x^{(k)} - f(x^{(k)}) \cdot \frac{x^{(k)} - P_{C_j}(x^{(k)})}{\|x^{(k)} - P_{C_j}(x^{(k)})\|_2}$$

$$= x^{(k)} - \|x^{(k)} - P_{C_j}(x^{(k)})\|_2 \cdot \frac{x^{(k)} - P_{C_j}(x^{(k)})}{\|x^{(k)} - P_{C_j}(x^{(k)})\|_2}$$

$$= P_{C_j}(x^{(k)})$$



alternating projection



the theorem does not say that you get convergence in a finite number of steps, but f^* goes down to 0

Sets that can be easily projected on:

- affine sets: least square
- euclidean ball: go to the center until hit the ball
- non-negative orthant: truncate all negative numbers
- rectangle: truncate or saturate
- polyhedron: solve a QP $\quad \min.\ \|x - x^{(0)}\|_2^2 \quad \text{s.t.}\ Ax \le b$
- simplex: not obvious
- cone of psd matrices

eg. positive semidefinite matrix completion

- some entries of matrix in S^n fixed ;
  find values for others so that the completed matrix is psd.

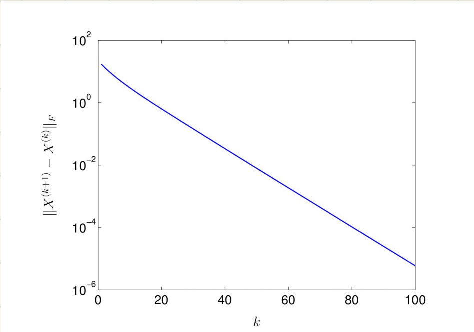
- $C_1 = S^n_+$, C_2 is an affine set in S^n with specified fixed entries
- projection onto C_1 by eigenvalue decomposition

$$P_{C_1}(X) = \sum_{i=1}^{n} \max\{0, \lambda_i\} q_i q_i^T$$ choose the basis with positive eigenvalue

(eg. project a non-symmetric matrix onto the unit ball in spectral norm (max singular value)
form the SVD, $U \Lambda V^T \Rightarrow U \tilde{\Lambda} V^T$ by truncating the singular values)

- projection onto C_2 by resetting specified entries to the fixed values)

things like EM algorithm !!!



in even steps, we get tiny negative eigenvalue
in odd steps, the number in fixed positions are slightly off

can project on a cone where the eigenvalues are bigger than ϵ

- speed up subgradient methods

$$x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)} + \beta_k (x^{(k)} - x^{(k-1)})$$

heavy ball methods (momentum)

other ideas: localization methods, conjugate directions

- Project subgradient method

solves constrained optimization problem

$$\min.\ f(x) \qquad f: \mathbb{R}^n \to \mathbb{R} \text{ is convex}$$
$$\text{s.t. } x \in C \qquad C \subseteq \mathbb{R}^n \text{ is convex}$$

$$x^{(k+1)} = P(x^{(k)} - \alpha_k g^{(k)})$$

projection does not increase the distance to x^*

(once you have a merit function, you can modify your algorithm way as long as whatever you do does not increase that function

e.g. linear equality constraints

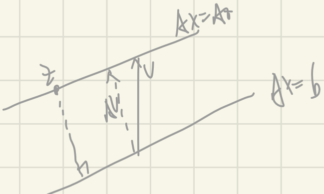
$$\min.\ f(x)$$
$$\text{s.t. } Ax = b$$

projection z onto $\{x \mid Ax = b\}$

$$p(z) = z - A^T(AA^T)^{-1}(Az - b)$$
$$= [I - A^T(AA^T)^{-1}A]z + A(AA^T)^{-1}b$$

projected subgradient update is

$$x^{(k+1)} = P(x^{(k)} - a_k g^{(k)})$$
$$= x^{(k)} - \alpha_k P(g^{(k)})$$
$$= x^{(k)} - \alpha_k [I - A^T(AA^T)^{-1}A]\, g^{(k)}$$
$$= x^{(k)} - \alpha_k P_{N(A)}(g^k)$$



e.g. $\min \|x\|_1$

s.t. $Ax = b$

$$g(x) = \text{sign}(x)$$

$$x^{(k+1)} = x^{(k)} - a_k(I - A^T(AA^T)^{-1}A)\ \text{sign}(x)$$

$$AV = Az - b$$
$$A^T(AA^T)^{-1}AV = A^T(AA^T)^{-1}(Az - b)$$
$$V = A^T(AA^T)^{-1}(Az - b)$$