Back to Optimization with constraints

$$\min f(x)$$
$$\text{s.t. } g_i(x) \le 0 \quad i = 1..m$$
$$\& \quad h_j(x) = 0 \quad j = 1..l$$

Temporarily absorb h_j's as 2 neg constraints

We need $-h_j$ & h_j both convex $\therefore$ affine h_j

(1) From midsem Question

Define: $I_{g_i}(x) = 0$ if $g_i(x) \le 0$ &
$= \infty$ o/w

$$\min_x f(x) + \sum_i I_{g_i}(x)$$:- Convex function But not diff.

Solve by either analysing optimality conditions in terms of subgradients or employ subgradient descent...

(2) Write it equivalently as a cone program (yet to be analysed)... MIDSEM

③ Replace $I_{g_i}(x)$ with a more "graceful" penalty function

$$\min_{x} f(x) - \sum_{i=1}^{m} \lambda_i \log(-g_i(x))$$

iteratively decrease $\lambda_i \geq 0$

④ Instead consider the Lagrangian fn

$$L(x,\lambda) = f(x) + \sum \lambda_i g_i(x)$$

We will briefly visit ① & then ④ & later ② & ③

① Recall gradient descent & Newton:

$$x^{k+1} = \min_{x} f(x^k) + \nabla^T f(x^k)(x - x^k) + \frac{t}{2}(x-x^k)^T M (x-x^k)$$

$M = I$ for gradient desc & $\nabla^2 f(x^k)$ for Newton

"Proximal" / "Mirror descent" / Projection algos treat problem of finding x^{k+1} as that of locating next iterate as close as possible to x^k

In the sense of an approximation or in the sense of minimizing constraint violation etc

Recap from Midsem for (1)

(10 Marks)

5. Consider a constrained convex optimization problem:

$$\begin{aligned} \text{minimize} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & g_i(\mathbf{x}) \le 0 \quad \text{for } i = 1 \dots m \end{aligned} \tag{2}$$

where f and g_i's are closed convex functions.

We will discuss two ways of reformulating this problem:

Let $S_{g_i} = \{y \mid g_i(y) \le 0\}$

S_{g_i} is a convex fn (sublevel set of convex fn)

(a) ~~Consider indicator function $I_{g_i}(\mathbf{x})$ associated with $g_i(\mathbf{x})$ for each $i =$~~ $1 \dots m$ such that $I_{g_i}(\mathbf{x}) = 0$ iff $g_i(\mathbf{x}) \le 0$ and $I_{g_i}(\mathbf{x}) = 1$ otherwise.

i. Prove that $\partial I_{g_i}(\mathbf{x})$ is a convex cone. Is it closed?

ii. Pose (2) as an equivalent unconstrained convex optimization problem making use of $I_{g_i}(\mathbf{x})$.

iii. ~~Now derive a necessary and sufficient condition for global con-~~strained optimality of (2) at a point $\mathbf{x}^*$.

(5 Marks)

(a) (i) We can prove that $\partial I_{g_i}(x) = \{g \mid g^T y \le g^T x \ \forall y\}$

$$I_{g_i}(y) \ge I_{g_i}(x) + \text{subgrad}^T(x)(y-x) \quad \forall y$$

If: $g_i(x) \le 0$ then $0 \ge \text{subgrad}^T(x)(y-x)$ is necessary & sufficient

ie $\partial I_{g_i}(x) = \{g \mid g^T y \le g^T x \ \forall y \in S_{g_i}\} = \bigcap_{y \in S_{g_i}} \{g \mid g^T(y-x) \le 0\}$ = intersection of half spaces of infinite hyperplane through origin = closed convex cone

$\therefore I_{g_i}(x)$ is a closed convex cone $\forall x : g_i(x) \le 0$

$I_{g_i}(x) = \phi$ if $x : g_i(x) > 0$

(ii) There are multiple ways to achieve it:-

$$\lim_{\lambda \to \infty} \min_x f(x) + \lambda \sum_i I_{g_i}(x) \ \dots$$

Equivalent to redefining $I_{g_i}(x) = 0$ if $g_i(x) \le 0$, $= \infty$ o/w

& using $\min_x f(x) + \sum_i I_{g_i}(x)$

OR

$$\min_x \max_{\lambda_i} f(x) + \sum_i \lambda_i I_{g_i}(x)$$

(iii) Necessary & sufficient conditions for optimality

(from Q4) : Let $I_{g_i}(x) = 0$ if $g_i(x) \leq 0$

$= \infty$ o/w

$0 \in \partial\left(f(x) + \sum I_{g_i}(x)\right)$

$\partial I_{g_i}(x) = \{g \mid g^T y \leq g^T x \;\; \forall y : g_i(y) \leq 0\}$ if $g_i(x) \leq 0$

& $\partial I_{g_i}(x) = \phi$ o/w

Subgradient Methods

- subgradient method and stepsize rules
- convergence results and proof
- optimal step size and alternating projections
- speeding up subgradient methods

g_x is a subgradient at x if

$f(y) \geq f(x) + g_x^T(y-x) \quad \forall y \in \mathrm{dmn}(f)$

Subgradient method

subgradient method is simple algorithm to minimize nondifferentiable convex function f

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

- $x^{(k)}$ is the kth iterate
- $g^{(k)}$ is **any** subgradient of f at $x^{(k)}$
- $\alpha_k > 0$ is the kth step size

not a descent method, so we keep track of best point so far

$$\boxed{f_{\rm best}^{(k)} = \min_{i=1,\ldots,k} f(x^{(i)})}$$

$f(x^k) \neq f^{(k)}_{best}$

not necessarily

Step size rules

step sizes are fixed ahead of time

- *constant step size:* $\alpha_k = \alpha$ (constant)
- *constant step length:* $\alpha_k = \gamma / \|g^{(k)}\|_2$ (so $\|x^{(k+1)} - x^{(k)}\|_2 = \gamma$)
- *square summable but not summable:* step sizes satisfy

$$\sum_{k=1}^{\infty} \alpha_k^2 < \infty, \qquad \sum_{k=1}^{\infty} \alpha_k = \infty$$

- *nonsummable diminishing:* step sizes satisfy

$$\lim_{k \to \infty} \alpha_k = 0, \qquad \sum_{k=1}^{\infty} \alpha_k = \infty$$

Assumptions

- $f^\star = \inf_x f(x) > -\infty$, with $f(x^\star) = f^\star$
- $\|g\|_2 \leq G$ for all $g \in \partial f$ (equivalent to Lipschitz condition on f)
- $\|x^{(1)} - x^\star\|_2 \leq R$

H/W

$|f(y) - f(x)| \leq G\|x-y\|$

these assumptions are stronger than needed, just to simplify proofs

Convergence results

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

- *constant step size:* $\bar{f} - f^\star \leq G^2\alpha/2$, $i.e.$,
  converges to $G^2\alpha/2$**-suboptimal**
  (converges to $f^\star$ if f differentiable, α small enough)

- *constant step length:* $\bar{f} - f^\star \leq G\gamma/2$, $i.e.$,
  converges to $G\gamma/2$**-suboptimal**

- *diminishing step size rule:* $\bar{f} = f^\star$, $i.e.$, **converges**

(Recall: If f is Lipschitz, gradient descent gives $O(1/\epsilon^2)$ convergence: That proof comes from here)

Convergence proof

key quantity: *Euclidean distance to the optimal set*, not the function value

let $x^\star$ be any minimizer of f

$$\begin{aligned}\|x^{(k+1)} - x^\star\|_2^2 &= \|x^{(k)} - \alpha_k g^{(k)} - x^\star\|_2^2 \\ &= \|x^{(k)} - x^\star\|_2^2 - 2\alpha_k g^{(k)T}(x^{(k)} - x^\star) + \alpha_k^2 \|g^{(k)}\|_2^2 \\ &\le \|x^{(k)} - x^\star\|_2^2 - 2\alpha_k (f(x^{(k)}) - f^\star) + \alpha_k^2 \|g^{(k)}\|_2^2\end{aligned}$$

using $f^\star = f(x^\star) \ge f(x^{(k)}) + g^{(k)T}(x^\star - x^{(k)})$

apply recursively to get

$$\begin{aligned}\|x^{(k+1)} - x^\star\|_2^2 &\leq \|x^{(1)} - x^\star\|_2^2 - 2\sum_{i=1}^{k} \alpha_i (f(x^{(i)}) - f^\star) + \sum_{i=1}^{k} \alpha_i^2 \|g^{(i)}\|_2^2 \\ &\leq R^2 - 2\sum_{i=1}^{k} \alpha_i (f(x^{(i)}) - f^\star) + G^2 \sum_{i=1}^{k} \alpha_i^2\end{aligned}$$

now we use

$$\sum_{i=1}^{k} \alpha_i (f(x^{(i)}) - f^\star) \geq (f_{\rm best}^{(k)} - f^\star) \underbrace{\left(\sum_{i=1}^{k} \alpha_i\right)}$$

Note why summability of $\sum \alpha_i$ becomes important

to get

$$f_{\rm best}^{(k)} - f^\star \leq \frac{R^2 + G^2 \sum_{i=1}^{k} \alpha_i^2}{2\sum_{i=1}^{k} \alpha_i}.$$

$f_{best}^{[k]} = \min_{i=0..k} f(x^{(i)})$

constant step size: for $\alpha_k = \alpha$ we get

we always had $\|x^{(1)} - x^\star\|_2 = R$

$$f_{\rm best}^{(k)} - f^\star \le \frac{R^2 + G^2 k\alpha^2}{2k\alpha}$$

righthand side converges to $G^2\alpha/2$ as $k \to \infty$

constant step length: for $\alpha_k = \gamma/\|g^{(k)}\|_2$ we get

$$f_{\rm best}^{(k)} - f^\star \le \frac{R^2 + \sum_{i=1}^k \alpha_i^2 \|g^{(i)}\|_2^2}{2\sum_{i=1}^k \alpha_i} \le \frac{R^2 + \gamma^2 k}{2\gamma k/G},$$

righthand side converges to $G\gamma/2$ as $k \to \infty$

square summable but not summable step sizes:
suppose step sizes satisfy

$$\sum_{k=1}^{\infty} \alpha_k^2 < \infty, \qquad \sum_{k=1}^{\infty} \alpha_k = \infty$$

then

$$f_{\rm best}^{(k)} - f^\star \leq \frac{R^2 + G^2 \sum_{i=1}^k \alpha_i^2}{2 \sum_{i=1}^k \alpha_i}$$

as $k \to \infty$, numerator converges to a finite number, denominator converges to ∞, so $f_{\rm best}^{(k)} \to f^\star$

Stopping criterion

- terminating when $\dfrac{R^2 + G^2 \sum_{i=1}^k \alpha_i^2}{2 \sum_{i=1}^k \alpha_i} \leq \epsilon$ is really, really, slow

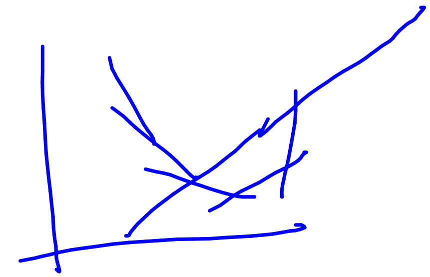
- optimal choice of α_i to achieve $\dfrac{R^2 + G^2 \sum_{i=1}^k \alpha_i^2}{2 \sum_{i=1}^k \alpha_i} \leq \epsilon$ for smallest k:

$$\alpha_i = (R/G)/\sqrt{k}, \quad i = 1, \ldots, k$$

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

- the truth: there really isn't a good stopping criterion for the subgradient method . . .

Example: Piecewise linear minimization



$$\text{minimize} \quad f(x) = \max_{i=1,\ldots,m}(a_i^T x + b_i)$$

to find a subgradient of f: find index j for which

$$a_j^T x + b_j = \max_{i=1,\ldots,m}(a_i^T x + b_i)$$

and take $g = a_j$

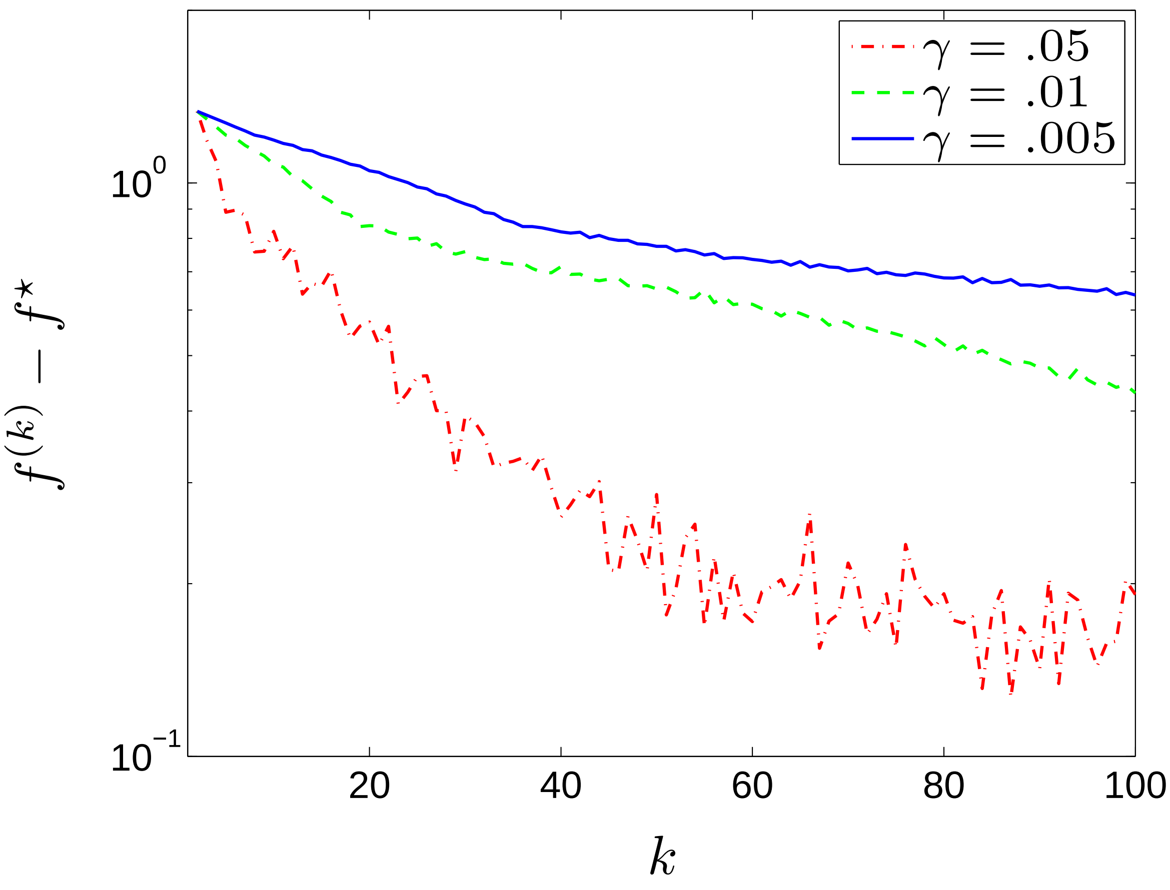
subgradient method: $x^{(k+1)} = x^{(k)} - \alpha_k a_j$

Note: This problem is equivalent to the following constrained opt problem:

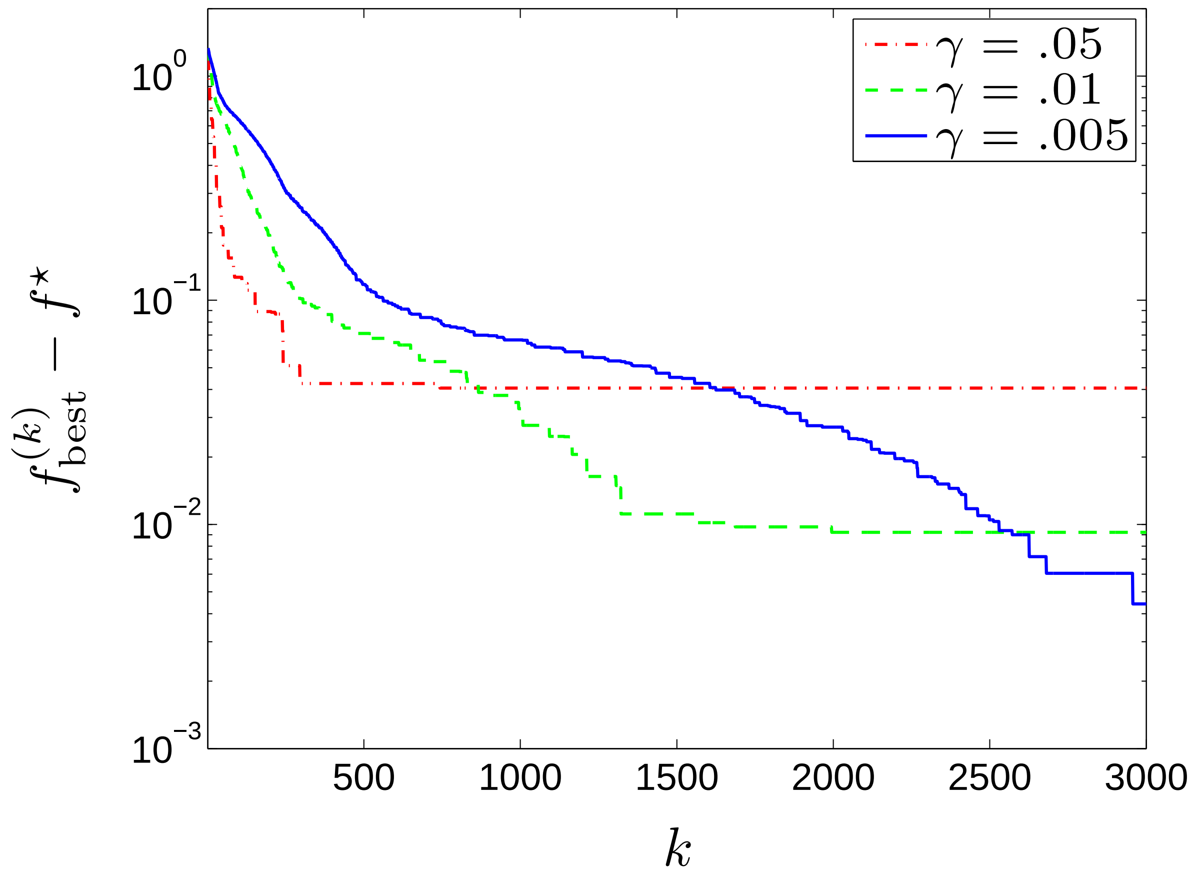
$$\min \ t$$
$$a_i^T x + b_i - t \le 0 \quad \forall i$$

problem instance with $n = 20$ variables, $m = 100$ terms, $f^\star \approx 1.1$

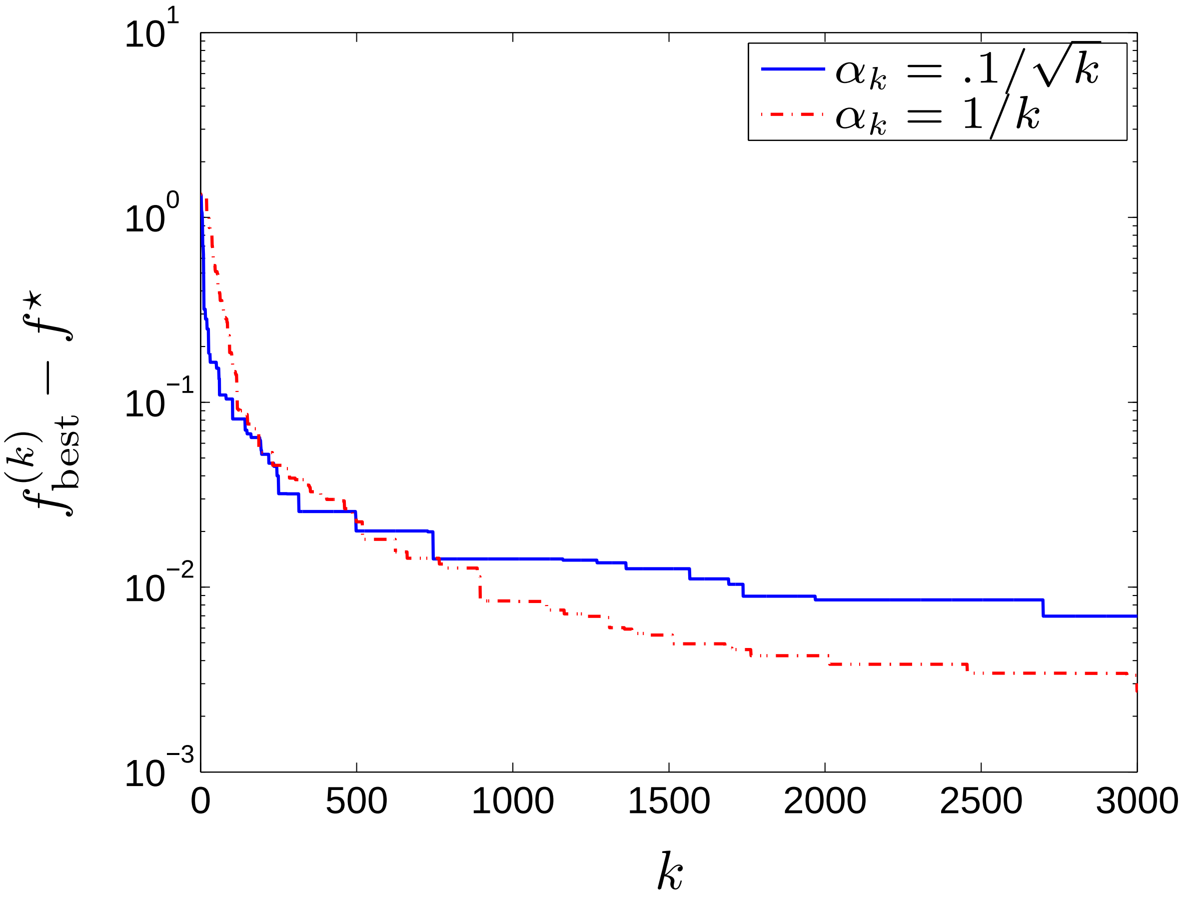
constant step length, $\gamma = 0.05, 0.01, 0.005$, first 100 iterations



$f_{\rm best}^{(k)} - f^\star$, constant step length $\gamma = 0.05, 0.01, 0.005$



diminishing step rule $\alpha_k = 0.1/\sqrt{k}$ and square summable step size rule $\alpha_k = 1/k$



Suppose: $\min f(x)$

s.t $g_i(x) \le 0$

$$\min f(x) + \eta \max_i g_i(x)$$

(we let η iteratively tend to ∞)

You need to find the formulation of the constrained opt problem for which the subgradient can be discovered easily.

H/W Eg Lasso: $\min_x \|Ax - b\|_2^2$

$\|x\|_1 \le \theta$

$Ax - b \rightarrow$ Regression loss / error

Finding a point in the intersection of convex sets

$C_i = \{x \mid g_i(x) \le 0\}$

$C = C_1 \cap \cdots \cap C_m$ is nonempty, $C_1, \ldots, C_m \subseteq \mathbf{R}^n$ closed and convex

find a point in C by minimizing

x could be an iterate obtained using gradient descent

$$f(x) = \max\{\mathbf{dist}(x, C_1), \ldots, \mathbf{dist}(x, C_m)\}$$

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

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

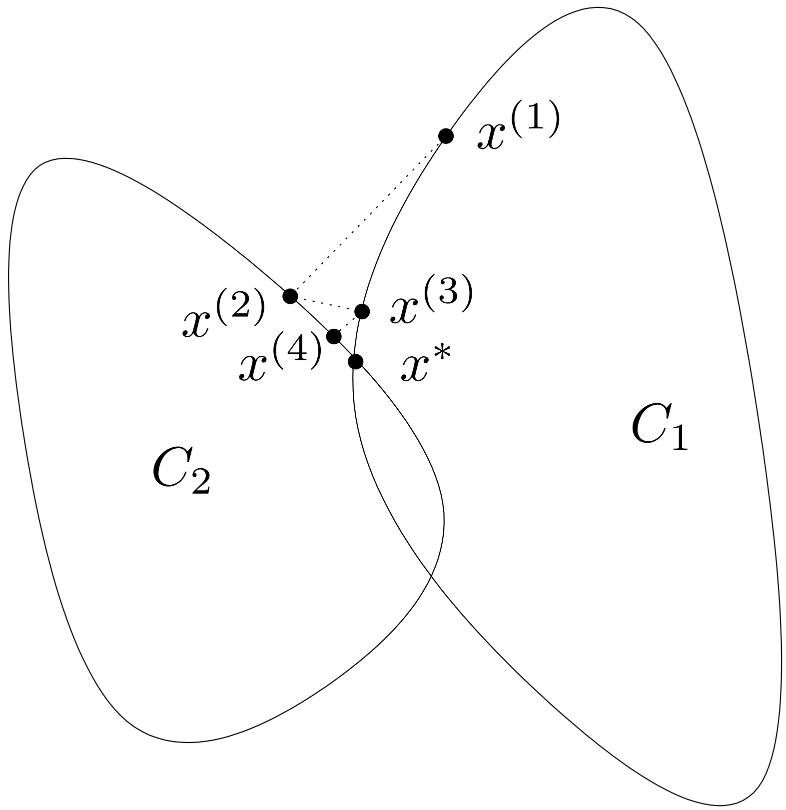
subgradient update with optimal step size:

$$\begin{aligned} x^{(k+1)} &= x^{(k)} - \alpha_k g^{(k)} \\ &= x^{(k)} - f(x^{(k)}) \frac{x^{(k)} - P_{C_j}(x^{(k)})}{\|x^{(k)} - P_{C_j}(x^{(k)})\|_2} \\ &= P_{C_j}(x^{(k)}) \end{aligned}$$

- a version of the famous *alternating projections* algorithm
- at each step, project the current point onto the farthest set
- for $m = 2$ sets, projections alternate onto one set, then the other
- convergence: $\mathbf{dist}(x^{(k)}, C) \to 0$ as $k \to \infty$

Alternating projections

first few iterations:



. . . $x^{(k)}$ eventually converges to a point $x^* \in C_1 \cap C_2$

Speeding up subgradient methods

- subgradient methods are very slow

- often convergence can be improved by keeping memory of past steps

$$x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)} + \beta_k \underbrace{(x^{(k)} - x^{(k-1)})}_{\text{keeping track of several previous iterates}}$$

(heavy ball method)

other ideas: localization methods, $\underbrace{\text{conjugate directions}}_{\text{only approximately conjugate for non quadratic problems.}}$, . . .

Now visiting ...

4 $L(x,\lambda,\mu) = f(x) + \sum_i \lambda_i g_i(x) + \sum_j \mu_j h_j(x)$

for
$$\min f(x)$$
$$\text{s.t } g_i(x) \le 0 \quad i = 1..m$$
$$h_j(x) = 0 \quad j = 1..l$$

$\min L(x,\lambda,\mu)$ --- you should ideally have $\lambda_i \ge 0$ to penalize $g_i(x) > 0$

$$\min_{\substack{x \\ g_i(x)\le 0 \\ h_j(x) = 0}} f(x) \ge \min_{\substack{x \\ g_i(x) \le 0,\ \lambda_i \ge 0 \\ h_j(x) = 0}} f(x) + \sum_i \lambda_i g_i(x) + \sum_j \mu_j h_j(x) \ge \min_{x,\ \lambda_i \ge 0} L(x,\lambda,\mu)$$

$$\min_{x} f(x) \geq \max_{\lambda_i \geq 0} \min_{x} L(x, \lambda, u)$$

s.t. $g_i(x) \leq 0$

$h_j(x) = 0$

Pushes up the lower bound from previous inequality.

$$\begin{cases} \min f(x) \\ \text{s.t } g_i(x) \le 0 \quad i = 1 \dots m \\ h_j(x) = 0 \quad j = 1 \dots k \end{cases}$$

We will generalize the inequalities & equalities

$$\min_x f(x) \quad \text{s.t } g_i(x) \le 0, \; h_j(x) = 0 \;\geq\; \min_x \max_{\lambda,\mu} \underbrace{f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{k} \mu_j h_j(x)}_{L(x,\lambda,\mu)}$$

$$\text{s.t } g_i(x) \le 0, \quad h_j(x) = 0, \quad \lambda_i \ge 0, \; \mu_j \in \mathbb{R}$$

$$\geq \min_x \max_{\lambda,\mu} L(x,\lambda,\mu) \quad \lambda_i \ge 0, \; \mu_j \in \mathbb{R}$$

under strong duality

$\lambda_i^* g_i(x^*) = 0 \quad \forall i$

$\lambda_i^* \ge 0$

$\mu_j^* h_j(x^*) = 0 \quad \forall j$

$$\geq \max_{\lambda,\mu} \underbrace{\min_x L(x,\lambda,\mu)}_{L^*(\lambda,\mu) \text{ or lagrange dual fn.}} \quad \lambda_i \ge 0, \; \mu_j \in \mathbb{R}$$

$$= \underbrace{\max_{\lambda,\mu,\lambda\ge 0} L^*(\lambda,\mu)}_{\text{Dual opt problem}}$$

General weak duality result

$$\min_{x} f(x) \le \max_{\substack{\lambda_i \ge 0 \\ \mu \in \mathbb{R}}} L^*(\lambda, \mu)$$

$$\text{s.t } g_i(x) \le 0$$

$$h_j(x) = 0$$

$\underline{Q1}$: Did we require f, g_i's & h_j's to be convex or affine? Ans: No

$\underline{Q2}$: Is L^* concave irrespective of f, g_i's & h_j's? Note: $L(x,\lambda,\mu)$ is affine in λ, μ

$$L^* = \min_{x} \underbrace{L(x,\lambda,\mu)}_{L_x(\lambda,\mu)}$$

min of affine fns is concave

$\min_{x \in D} f(x)$

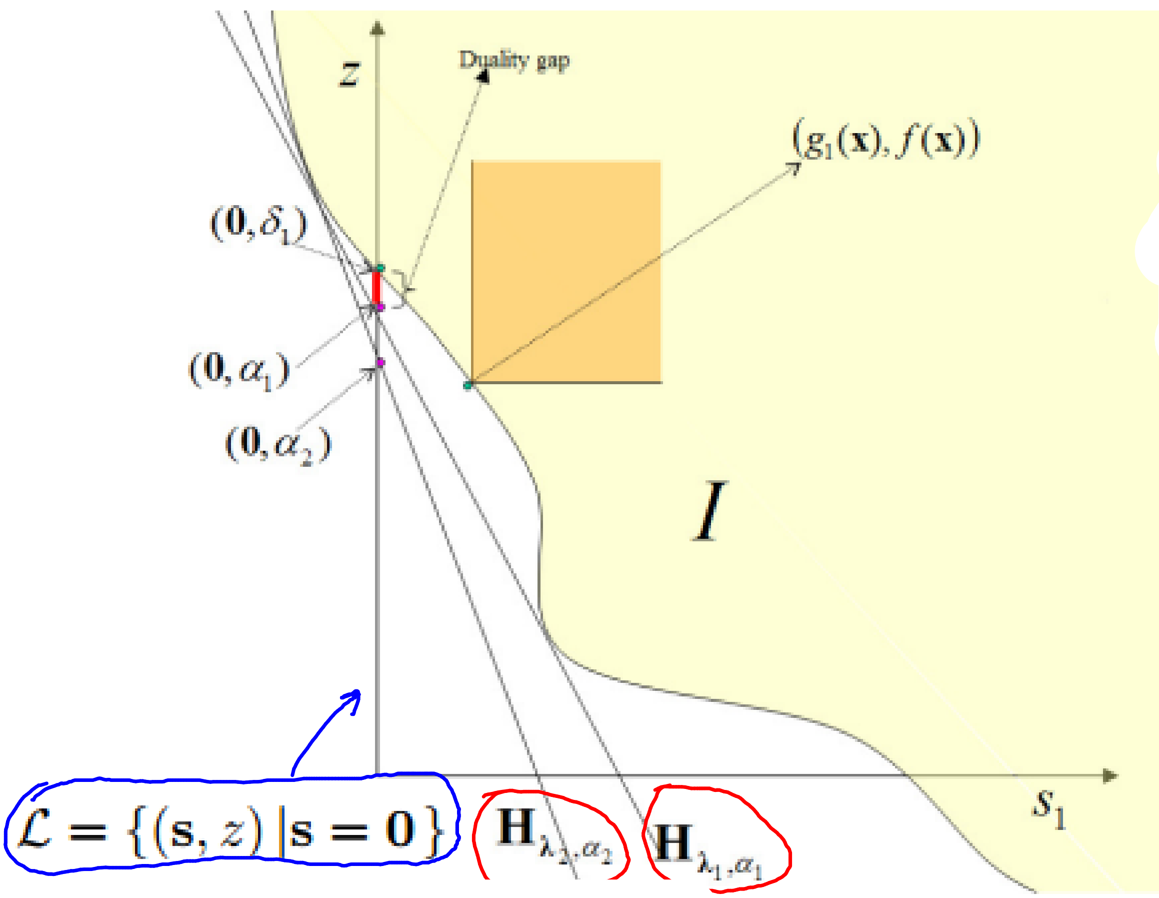
s.t $g_i(x) \le 0 \quad i = 1 \ldots m$

(The general dual problem & its geometric interpretation)

pg 292, sec 4.4.3 of http://www.cse.iitb.ac.in/~cs709/notes/BasicsOfConvexOptimization.pdf

Consider the set:

$$\mathcal{I} = \{(\mathbf{s}, z) \mid \mathbf{s} \in \Re^m,\ z \in \Re,\ \exists \mathbf{x} \in \mathcal{D}\ with\ g_i(\mathbf{x}) \le s_i\ \forall 1 \le i \le m,\ f(\mathbf{x}) \le z\}$$



$$H_{\lambda,\alpha} = \{(\mathbf{s}, z) \mid \lambda^T.\mathbf{s} + z = \alpha\}$$