Constraint Convex Optimization

Summarized and presented by Yuan Zhong

zhong.yu@husky.neu.edu

Outline

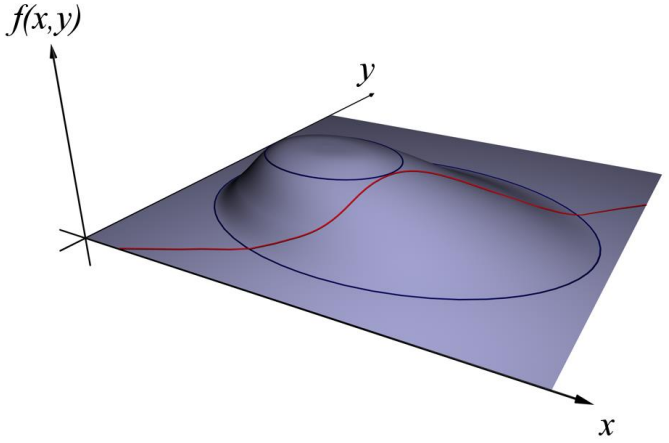
- Unconstraint convex optimization
- Equality constraint convex optimization
- Inequality constraint convex optimization
- Algorithm
  - Interior point methods
    - Barrier method
    - Primal-dual interior point method
    - Primal-dual infeasible interior point method
  - Projected gradient descent

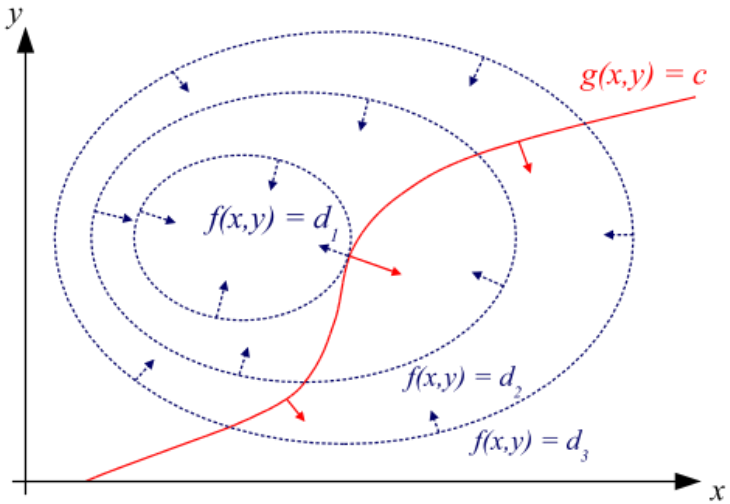
Unconstraint convex optimization

- Unconstrained convex optimization problem
  - minimize $f(x)$, where $f: \mathcal{R}^n \rightarrow \mathcal{R}$
- Fermat theorem
  - If $f: (a, b) \rightarrow \mathcal{R}$ be a function and suppose that $x_0 \in (a, b)$ is a point where f has a local extremum. If f is differentiable at x_0, then $f'(x_0) = 0$
- Solution
  - Vanilla gradient descent
  - Newton method

Equality constraint convex optimization

- Equality constraint convex optimization
  - minimize $f(x)$, where $f: \mathcal{R}^n \rightarrow \mathcal{R}$
  - $l_i(x) = 0, i = 1, \cdots, r$
- Lagrange multiplier
  - Find the local maxima and minima of a function s.t. equality constraint
  - Lagrange function $\mathcal{L}(\boldsymbol{\lambda}, x) = f(x) + \sum_{i=1}^{r} \lambda_i \, l_i(x)$
  - Stationary point $(x_0, \boldsymbol{\lambda}_0)$





Inequality constraint convex optimization

- General minimization problem
  - minimize $f(x)$, where $f: \mathcal{R}^n \rightarrow \mathcal{R}$
  - $h_i(x) \leq 0, i = 1, \cdots, m$
  - $l_j(x) = 0, j = 1, \cdots, r$
- Lagrange dual
- Weak and strong duality
- KKT

Lagrangian

Consider general minimization problem

$$\begin{array}{ll} \min\limits_{x} & f(x) \\ \text{subject to} & h_i(x) \leq 0, \;\; i = 1, \ldots m \\ & \ell_j(x) = 0, \;\; j = 1, \ldots r \end{array}$$

Need not be convex, but of course we will pay special attention to convex case

We define the Lagrangian as

$$L(x, u, v) = f(x) + \sum_{i=1}^{m} u_i h_i(x) + \sum_{j=1}^{r} v_j \ell_j(x)$$

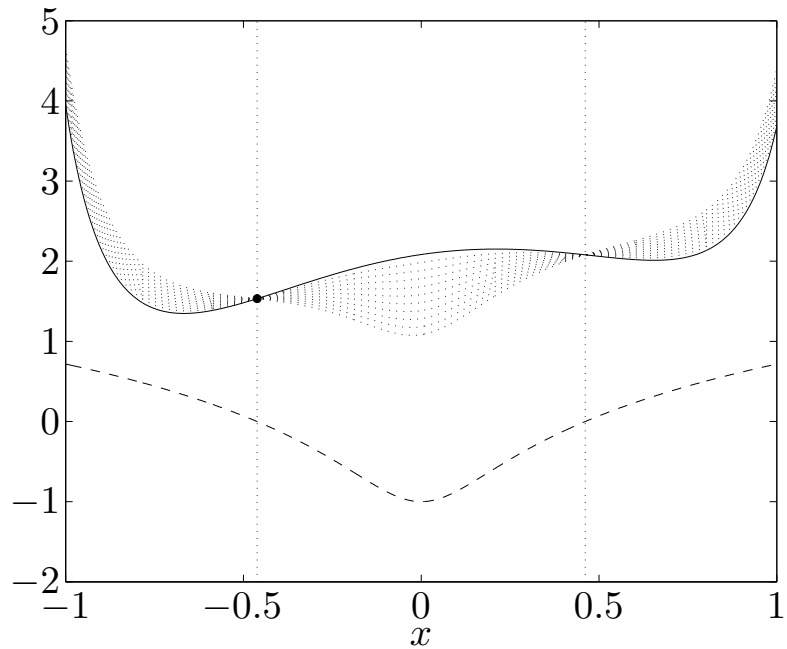
New variables $u \in \mathbb{R}^m, v \in \mathbb{R}^r$, with $u \geq 0$ (implicitly, we define $L(x, u, v) = -\infty$ for $u < 0$)

Important property: for any $u \geq 0$ and v,

$$f(x) \geq L(x, u, v) \quad \text{at each feasible } x$$

Why? For feasible x,

$$L(x, u, v) = f(x) + \sum_{i=1}^{m} u_i \underbrace{h_i(x)}_{\leq 0} + \sum_{j=1}^{r} v_j \underbrace{\ell_j(x)}_{=0} \leq f(x)$$



- Solid line is f
- Dashed line is h, hence feasible set $\approx [-0.46, 0.46]$
- Each dotted line shows $L(x, u, v)$ for different choices of $u \geq 0$ and v

(From B & V page 217)

Lagrange dual function

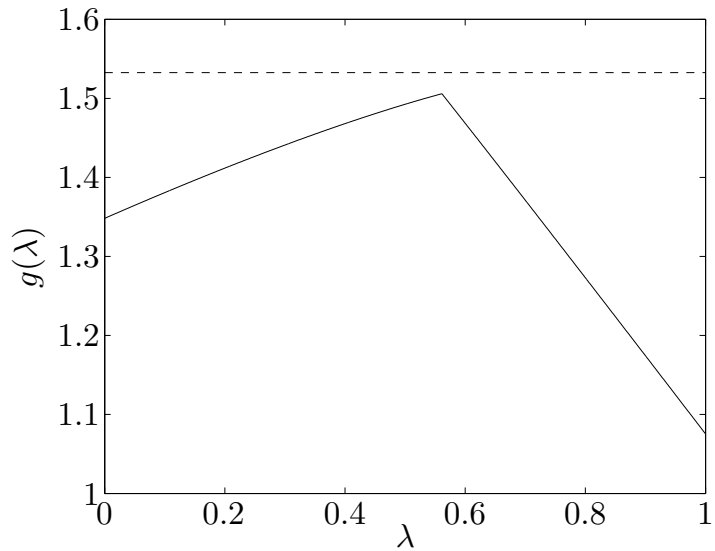
Let C denote primal feasible set, $f^\star$ denote primal optimal value. Minimizing $L(x, u, v)$ over all x gives a lower bound:

$$f^\star \;\geq\; \min_{x \in C} \; L(x, u, v) \;\geq\; \min_{x} \; L(x, u, v) \;:=\; g(u, v)$$

We call $g(u, v)$ the Lagrange dual function, and it gives a lower bound on $f^\star$ for any $u \geq 0$ and v, called dual feasible u, v

- Dashed horizontal line is $f^\star$
- Dual variable λ is (our u)
- Solid line shows $g(\lambda)$

(From B & V page 217)



Example: quadratic program

Consider quadratic program:

$$\begin{array}{ll} \min\limits_{x \in \mathbb{R}^n} & \frac{1}{2} x^T Q x + c^T x \\ \text{subject to} & Ax = b,\ x \geq 0 \end{array}$$

where $Q \succ 0$. Lagrangian:

$$L(x, u, v) = \frac{1}{2} x^T Q x + c^T x - u^T x + v^T (Ax - b)$$

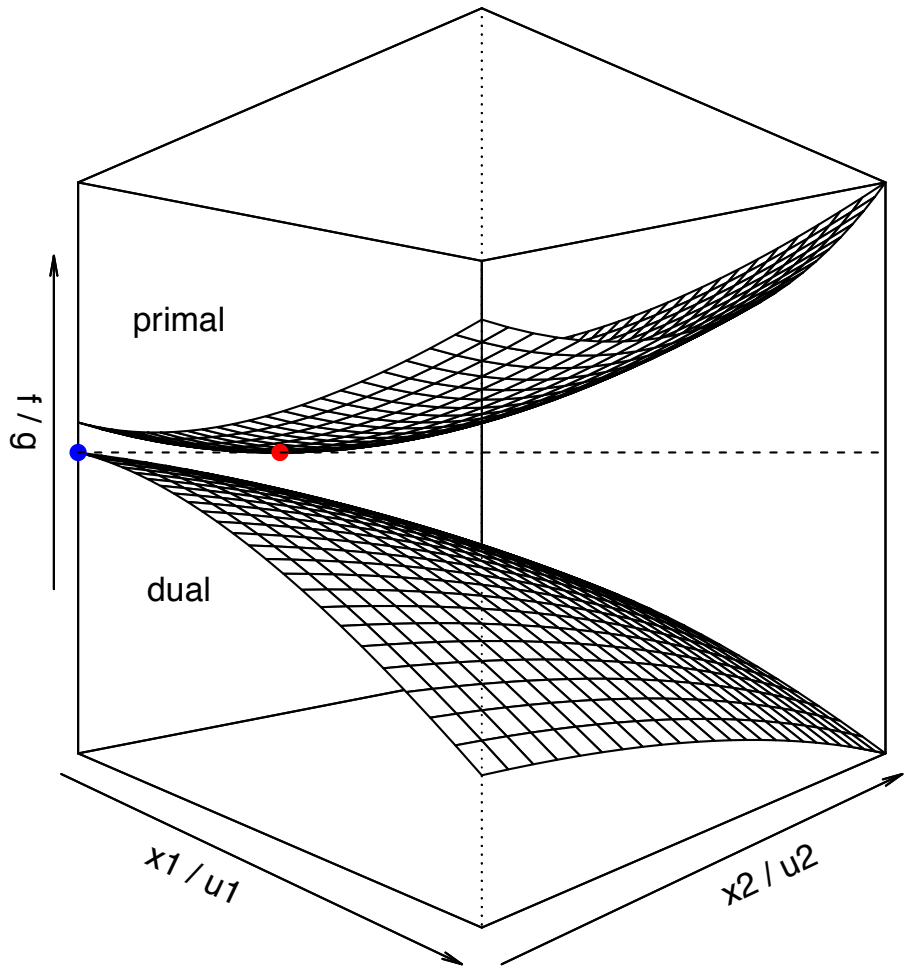
Lagrange dual function:

$$g(u, v) = \min_{x \in \mathbb{R}^n} L(x, u, v) = -\frac{1}{2}(c - u + A^T v)^T Q^{-1} (c - u + A^T v) - b^T v$$

For any $u \geq 0$ and any v, this is lower a bound on primal optimal value $f^\star$

Example: quadratic program in 2D

We choose $f(x)$ to be quadratic in 2 variables, subject to $x \geq 0$. Dual function $g(u)$ is also quadratic in 2 variables, also subject to $u \geq 0$



Dual function $g(u)$ provides a bound on $f^\star$ for every $u \geq 0$

Largest bound this gives us: turns out to be exactly $f^\star$ ... coincidence?

More on this later, via KKT conditions

Lagrange dual problem

Given primal problem

$$\begin{array}{ll} \min\limits_{x} & f(x) \\ \text{subject to} & h_i(x) \leq 0, \quad i = 1, \ldots m \\ & \ell_j(x) = 0, \quad j = 1, \ldots r \end{array}$$

Our constructed dual function $g(u, v)$ satisfies $f^\star \geq g(u, v)$ for all $u \geq 0$ and v. Hence best lower bound is given by maximizing $g(u, v)$ over all dual feasible u, v, yielding Lagrange dual problem:

$$\begin{array}{ll} \max\limits_{u,v} & g(u, v) \\ \text{subject to} & u \geq 0 \end{array}$$

Key property, called weak duality: if dual optimal value is $g^\star$, then

$$f^\star \geq g^\star$$

Note that this always holds (even if primal problem is nonconvex)

Another key property: the dual problem is a convex optimization problem (as written, it is a concave maximization problem)

Again, this is always true (even when primal problem is not convex)

By definition:

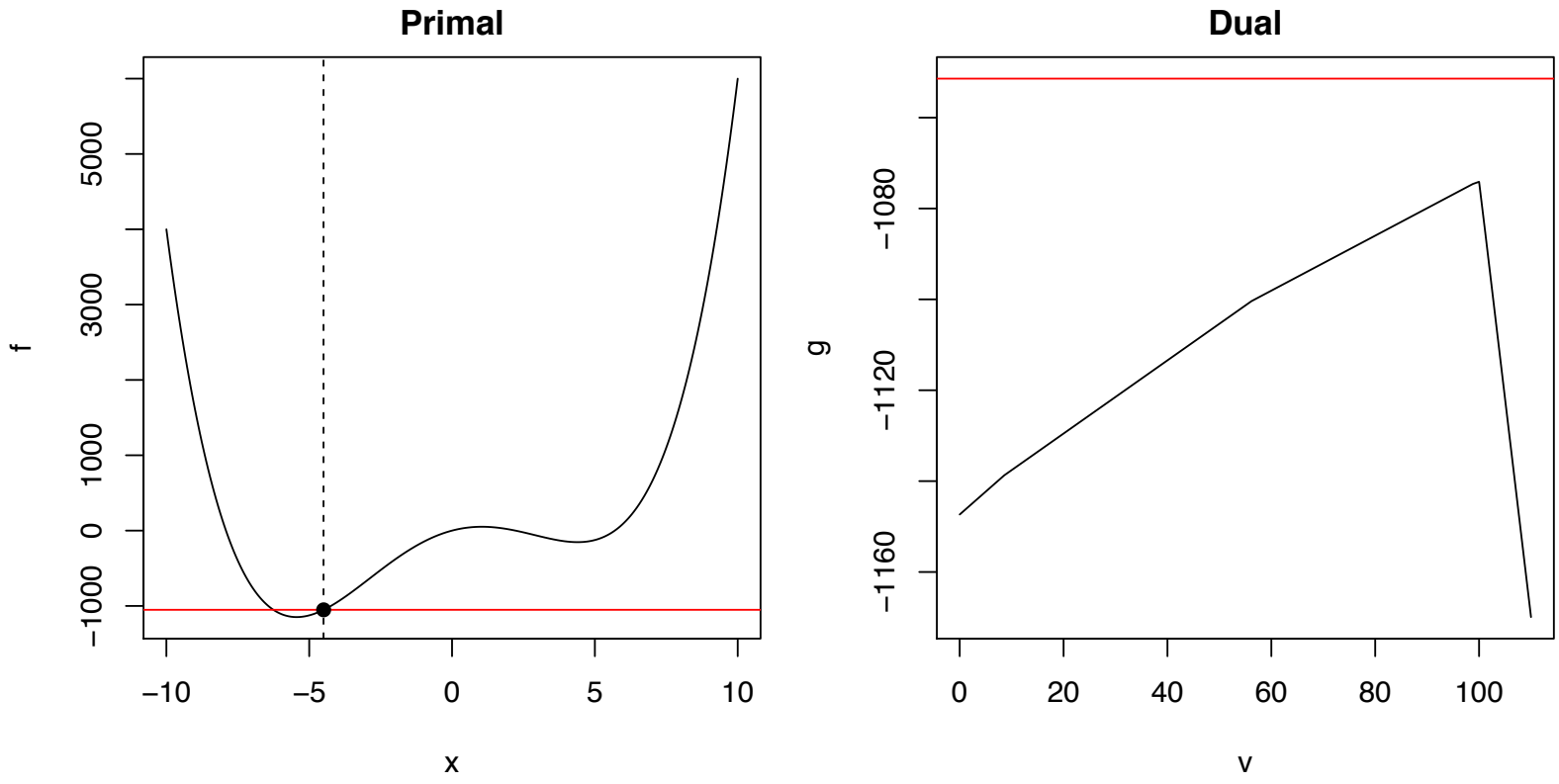
$$g(u,v) = \min_x \left\{ f(x) + \sum_{i=1}^m u_i h_i(x) + \sum_{j=1}^r v_j \ell_j(x) \right\}$$

$$= \underbrace{- \max_x \left\{ -f(x) - \sum_{i=1}^m u_i h_i(x) - \sum_{j=1}^r v_j \ell_j(x) \right\}}_{\text{pointwise maximum of convex functions in } (u,v)}$$

I.e., g is concave in (u, v), and $u \geq 0$ is a convex constraint, hence dual problem is a concave maximization problem

Example: nonconvex quartic minimization

Define $f(x) = x^4 - 50x^2 + 100x$ (nonconvex), minimize subject to constraint $x \geq -4.5$



Dual function g can be derived explicitly, via closed-form equation for roots of a cubic equation

Form of g is quite complicated:

$$g(u) = \min_{i=1,2,3} F_i^4(u) - 50F_i^2(u) + 100F_i(u),$$

where for $i = 1, 2, 3$,

$$F_i(u) = \frac{-a_i}{12 \cdot 2^{1/3}} \left(432(100-u) - \left(432^2(100-u)^2 - 4 \cdot 1200^3 \right)^{1/2} \right)^{1/3}$$
$$-100 \cdot 2^{1/3} \frac{1}{\left(432(100-u) - \left(432^2(100-u)^2 - 4 \cdot 1200^3 \right)^{1/2} \right)^{1/3}},$$

and $a_1 = 1$, $a_2 = (-1 + i\sqrt{3})/2$, $a_3 = (-1 - i\sqrt{3})/2$

Without the context of duality it would be difficult to tell whether or not g is concave ... but we know it must be!

Strong duality

Recall that we always have $f^\star \geq g^\star$ (weak duality). On the other hand, in some problems we have observed that actually

$$f^\star = g^\star$$

which is called strong duality

Slater's condition: if the primal is a convex problem (i.e., f and $h_1, \ldots h_m$ are convex, $\ell_1, \ldots \ell_r$ are affine), and there exists at least one strictly feasible $x \in \mathbb{R}^n$, meaning

$$h_1(x) < 0, \ldots h_m(x) < 0 \quad \text{and} \quad \ell_1(x) = 0, \ldots \ell_r(x) = 0$$

then strong duality holds

This is a pretty weak condition. (Further refinement: only require strict inequalities over functions h_i that are not affine)

Example: support vector machine dual

Given $y \in \{-1, 1\}^n$, $X \in \mathbb{R}^{n \times p}$, rows $x_1, \ldots x_n$, recall the support vector machine problem:

$$\begin{aligned} \min_{\beta, \beta_0, \xi} \quad & \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^n \xi_i \\ \text{subject to} \quad & \xi_i \geq 0, \; i = 1, \ldots n \\ & y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i, \; i = 1, \ldots n \end{aligned}$$

Introducing dual variables $v, w \geq 0$, we form the Lagrangian:

$$\begin{aligned} L(\beta, \beta_0, \xi, v, w) = \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^n \xi_i - \sum_{i=1}^n v_i\xi_i + \\ \sum_{i=1}^n w_i\big(1 - \xi_i - y_i(x_i^T\beta + \beta_0)\big) \end{aligned}$$

Minimizing over β, β_0, ξ gives Lagrange dual function:

$$g(v,w) = \begin{cases} -\frac{1}{2} w^T \tilde{X} \tilde{X}^T w + 1^T w & \text{if } w = C1 - v,\ w^T y = 0 \\ -\infty & \text{otherwise} \end{cases}$$

where $\tilde{X} = \mathrm{diag}(y)X$. Thus SVM dual problem, eliminating slack variable v, becomes

$$\begin{aligned} \max_w \quad & -\frac{1}{2} w^T \tilde{X} \tilde{X}^T w + 1^T w \\ \text{subject to} \quad & 0 \le w \le C1,\ w^T y = 0 \end{aligned}$$

Check: Slater's condition is satisfied, and we have strong duality. Further, from study of SVMs, might recall that at optimality

$$\beta = \tilde{X}^T w$$

This is not a coincidence, as we'll later via the KKT conditions

Duality gap

Given primal feasible x and dual feasible u, v, the quantity

$$f(x) - g(u, v)$$

is called the duality gap between x and u, v. Note that

$$f(x) - f^\star \leq f(x) - g(u, v)$$

so if the duality gap is zero, then x is primal optimal (and similarly, u, v are dual optimal)

From an algorithmic viewpoint, provides a stopping criterion: if $f(x) - g(u, v) \leq \epsilon$, then we are guaranteed that $f(x) - f^\star \leq \epsilon$

Very useful, especially in conjunction with iterative methods ... more dual uses in coming lectures

Karush-Kuhn-Tucker conditions

Given general problem

$$\begin{array}{ll} \min\limits_{x} & f(x) \\ \text{subject to} & h_i(x) \leq 0, \quad i = 1, \ldots m \\ & \ell_j(x) = 0, \quad j = 1, \ldots r \end{array}$$

The Karush-Kuhn-Tucker conditions or KKT conditions are:

- $0 \in \partial f(x) + \sum_{i=1}^{m} u_i \partial h_i(x) + \sum_{j=1}^{r} v_j \partial \ell_j(x)$ (stationarity)
- $u_i \cdot h_i(x) = 0$ for all i (complementary slackness)
- $h_i(x) \leq 0$, $\ell_j(x) = 0$ for all i, j (primal feasibility)
- $u_i \geq 0$ for all i (dual feasibility)

Necessity

from Slater's condition to KKT

Let $x^\star$ and $u^\star, v^\star$ be primal and dual solutions with zero duality gap (strong duality holds, e.g., under Slater's condition). Then

$$\begin{aligned} f(x^\star) &= g(u^\star, v^\star) \\ &= \min_x \; f(x) + \sum_{i=1}^m u_i^\star h_i(x) + \sum_{j=1}^r v_j^\star \ell_j(x) \\ &\leq f(x^\star) + \sum_{i=1}^m u_i^\star h_i(x^\star) + \sum_{j=1}^r v_j^\star \ell_j(x^\star) \\ &\leq f(x^\star) \end{aligned}$$

In other words, all these inequalities are actually equalities

Two things to learn from this:

- The point $x^\star$ minimizes $L(x, u^\star, v^\star)$ over $x \in \mathbb{R}^n$. Hence the subdifferential of $L(x, u^\star, v^\star)$ must contain 0 at $x = x^\star$—this is exactly the stationarity condition
- We must have $\sum_{i=1}^m u_i^\star h_i(x^\star) = 0$, and since each term here is ≤ 0, this implies $u_i^\star h_i(x^\star) = 0$ for every i—this is exactly complementary slackness

Primal and dual feasibility hold by virtue of optimality. Therefore:

> If $x^\star$ and $u^\star, v^\star$ are primal and dual solutions, with zero duality gap, then $x^\star, u^\star, v^\star$ satisfy the KKT conditions

(Note that this statement assumes nothing a priori about convexity of our problem, i.e., of f, h_i, ℓ_j)

Sufficiency

from KKT to Slater's condition

If there exists $x^\star, u^\star, v^\star$ that satisfy the KKT conditions, then

$$\begin{aligned} g(u^\star, v^\star) &= f(x^\star) + \sum_{i=1}^{m} u_i^\star h_i(x^\star) + \sum_{j=1}^{r} v_j^\star \ell_j(x^\star) \\ &= f(x^\star) \end{aligned}$$

where the first equality holds from stationarity, and the second holds from complementary slackness

Therefore the duality gap is zero (and $x^\star$ and $u^\star, v^\star$ are primal and dual feasible) so $x^\star$ and $u^\star, v^\star$ are primal and dual optimal. Hence, we've shown:

> If $x^\star$ and $u^\star, v^\star$ satisfy the KKT conditions, then $x^\star$ and $u^\star, v^\star$ are primal and dual solutions

Putting it together

In summary, KKT conditions:

- always sufficient
- necessary under strong duality

Putting it together:

For a problem with strong duality (e.g., assume Slater's condition: convex problem and there exists x strictly satisfying non-affine inequality contraints),

$$x^\star \text{ and } u^\star, v^\star \text{ are primal and dual solutions}$$
$$\iff x^\star \text{ and } u^\star, v^\star \text{ satisfy the KKT conditions}$$

(Warning, concerning the stationarity condition: for a differentiable function f, we cannot use $\partial f(x) = \{\nabla f(x)\}$ unless f is convex)

Example: quadratic with equality constraints

Consider for $Q \succeq 0$,

$$\begin{array}{ll} \min\limits_{x \in \mathbb{R}^n} & \frac{1}{2} x^T Q x + c^T x \\ \text{subject to} & Ax = 0 \end{array}$$

E.g., as we will see, this corresponds to Newton step for equality-constrained problem $\min_x \ f(x) \ \text{ subject to } \ Ax = b$

Convex problem, no inequality constraints, so by KKT conditions: x is a solution if and only if

$$\begin{bmatrix} Q & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} = \begin{bmatrix} -c \\ 0 \end{bmatrix}$$

for some u. Linear system combines stationarity, primal feasibility (complementary slackness and dual feasibility are vacuous)

Example: water-filling

Example from B & V page 245: consider problem

$$\begin{array}{ll} \min\limits_{x\in\mathbb{R}^n} & -\sum\limits_{i=1}^{n}\log(\alpha_i+x_i) \\ \text{subject to} & x\geq 0,\ 1^Tx=1 \end{array}$$

Information theory: think of $\log(\alpha_i+x_i)$ as communication rate of ith channel. KKT conditions:

$$-1/(\alpha_i+x_i)-u_i+v=0,\quad i=1,\ldots n$$
$$u_i\cdot x_i=0,\quad i=1,\ldots n,\quad x\geq 0,\quad 1^Tx=1,\quad u\geq 0$$

Eliminate u:

$$1/(\alpha_i+x_i)\leq v,\quad i=1,\ldots n$$
$$x_i(v-1/(\alpha_i+x_i))=0,\quad i=1,\ldots n,\quad x\geq 0,\quad 1^Tx=1$$

Can argue directly stationarity and complementary slackness imply

$$x_i = \begin{cases} 1/v - \alpha_i & \text{if } v < 1/\alpha_i \\ 0 & \text{if } v \geq 1/\alpha_i \end{cases} = \max\{0, 1/v - \alpha_i\}, \quad i = 1, \ldots n$$

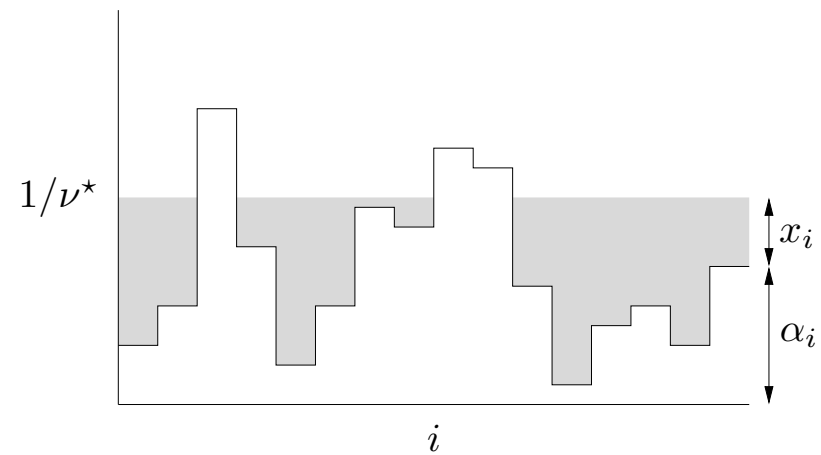
Still need x to be feasible, i.e., $1^T x = 1$, and this gives

$$\sum_{i=1}^{n} \max\{0, 1/v - \alpha_i\} = 1$$

Univariate equation, piecewise linear in $1/v$ and not hard to solve

This reduced problem is called water-filling

(From B & V page 246)



Example: support vector machines

Given $y \in \{-1, 1\}^n$, and $X \in \mathbb{R}^{n \times p}$, the support vector machine problem is:

$$\begin{aligned} \min_{\beta, \beta_0, \xi} \quad & \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^n \xi_i \\ \text{subject to} \quad & \xi_i \geq 0, \ i = 1, \ldots n \\ & y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i, \ i = 1, \ldots n \end{aligned}$$

Introduce dual variables $v, w \geq 0$. KKT stationarity condition:

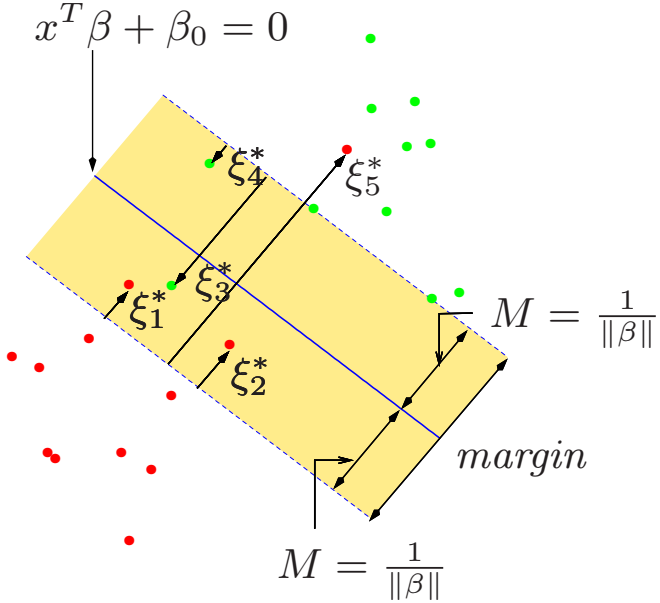
$$0 = \sum_{i=1}^n w_i y_i, \quad \beta = \sum_{i=1}^n w_i y_i x_i, \quad w = C1 - v$$

Complementary slackness:

$$v_i \xi_i = 0, \ w_i\big(1 - \xi_i - y_i(x_i^T\beta + \beta_0)\big) = 0, \quad i = 1, \ldots n$$

Hence at optimality we have $\beta = \sum_{i=1}^{n} w_i y_i x_i$, and w_i is nonzero only if $y_i(x_i^T \beta + \beta_0) = 1 - \xi_i$. Such points i are called the support points

- For support point i, if $\xi_i = 0$, then x_i lies on edge of margin, and $w_i \in (0, C]$;
- For support point i, if $\xi_i \neq 0$, then x_i lies on wrong side of margin, and $w_i = C$



KKT conditions do not really give us a way to find solution, but gives a better understanding

In fact, we can use this to screen away non-support points before performing optimization

Barrier method

Equality and Inequality Constrained Minimization

- **Problem to be solved:**

$$\begin{aligned} \min_x \quad & f_0(x) \\ \text{s.t.} \quad & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & Ax = b \end{aligned}$$

- **Reformulation via indicator function,**

$$\min_x \quad f_0(x) + \sum_{I=1}^{m} I_-(f_i(x)) \qquad I_-(u) = \begin{cases} 0 & \text{if } u \leq 0 \\ \infty & \text{otherwise} \end{cases}$$
$$Ax = b$$

→ No inequality constraints anymore, but very poorly conditioned objective function

Log barrier function

Consider the convex optimization problem

$$\begin{array}{ll} \min\limits_{x} & f(x) \\ \text{subject to} & h_i(x) \leq 0, \;\; i = 1, \ldots m \\ & Ax = b \end{array}$$

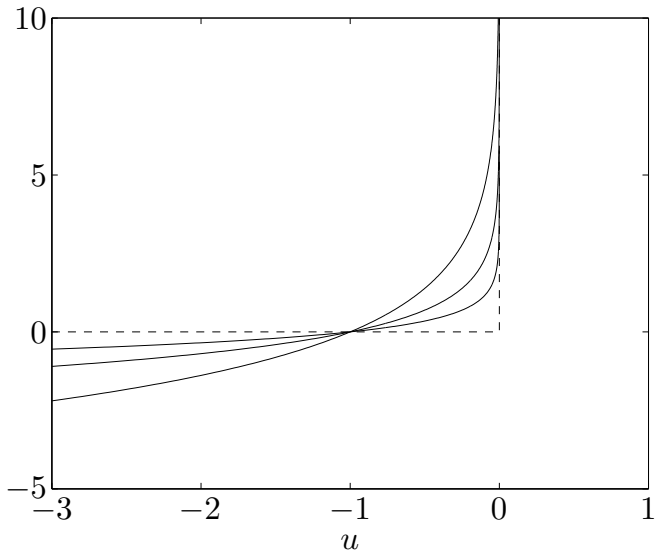
We will assume that f, $h_1, \ldots h_m$ are convex, twice differentiable, each with domain $\mathbb{R}^n$. The function

$$\phi(x) = -\sum_{i=1}^{m} \log(-h_i(x))$$

is called the log barrier for the above problem. Its domain is the set of strictly feasible points, $\{x : h_i(x) < 0, \; i = 1, \ldots m\}$, which we assume is nonempty

Ignoring equality constraints for now, our problem can be written as

$$\min_x \quad f(x) + \sum_{i=1}^m I_{\{h_i(x) \leq 0\}}(x)$$



We approximate this representation by adding the log barrier function:

$$\min_x \quad f(x) - (1/t) \cdot \sum_{i=1}^m \log(-h_i(x))$$

where $t > 0$ is a large number

This approximation is more accurate for larger t. But for any value of t, the log barrier approaches ∞ if any $h_i(x) \to 0$

Log barrier calculus

For the log barrier function

$$\phi(x) = -\sum_{i=1}^{m} \log(-h_i(x))$$

let us write down its gradient and Hessian, for future reference:

$$\nabla\phi(x) = -\sum_{i=1}^{m} \frac{1}{h_i(x)} \nabla h_i(x)$$

and

$$\nabla^2\phi(x) = \sum_{i=1}^{m} \frac{1}{h_i(x)^2} \nabla h_i(x) \nabla h_i(x)^T - \sum_{i=1}^{m} \frac{1}{h_i(x)} \nabla^2 h_i(x)$$

computed using the chain rule

Central path

Consider minimizing our problem, after replacing hard inequalities with barrier term:

$$\begin{array}{ll} \min\limits_{x} & tf(x) + \phi(x) \\ \text{subject to} & Ax = b \end{array}$$

(Here we switched placement of t, but its role is the same.) The central path is defined as the solution $x^\star(t)$ as a function of $t > 0$. These solutions are characterized by the KKT conditions:

$$Ax^\star(t) = b, \quad h_i(x^\star(t)) < 0, \quad i = 1, \ldots m$$

$$t\nabla f(x^\star(t)) - \sum_{i=1}^{m} \frac{1}{h_i(x^\star(t))} \nabla h_i(x^\star(t)) + A^T w = 0$$

for some $w \in \mathbb{R}^m$. As $t \to \infty$, hope is that $x^\star(t) \to x^\star$, solution of our original problem

Dual points from central path

From central points, we can derive feasible dual points for original problem. Given $x^\star(t)$ and corresponding w, we define

$$u_i^\star(t) = -\frac{1}{th_i(x^\star(t))}, \quad i = 1, \ldots m, \quad v^\star(t) = w/t$$

We claim $u^\star(t), v^\star(t)$ are dual feasible for original problem. Why?

- Note that $u_i^\star(t) > 0$ since $h_i(x^\star(t)) < 0$ for all i
- Further, the point $(u^\star(t), v^\star(t))$ lies in domain of Lagrange dual function $g(u, v)$, since by definition

$$\nabla f(x^\star(t)) + \sum_{i=1}^{m} u_i(x^\star(t)) \nabla h_i(x^\star(t)) + A^T v^\star(t) = 0$$

  I.e., $x^\star(t)$ minimizes Lagrangian $L(x, u^\star(t), v^\star(t))$ over x, so $g(u^\star(t), v^\star(t)) > -\infty$

This allows us to bound suboptimality of $f(x^\star(t))$, with respect to original problem, via the duality gap. We compute

$$\begin{aligned} g(u^\star(t), v^\star(t)) &= f(x^\star(t)) + \sum_{i=1}^m u_i^\star(t) h_i(x^\star(t)) + \\ &\qquad\qquad v^\star(t)^T (Ax^\star(t) - b) \\ &= f(x^\star(t)) - m/t \end{aligned}$$

That is, we know that $f(x^\star(t)) - f^\star \leq m/t$

This will be very useful as a stopping criterion; it also confirms the fact that $x^\star(t) \to x^\star$ as $t \to \infty$

Barrier method

The barrier method solves a sequence of problems

$$\begin{array}{ll} \min\limits_{x} & tf(x) + \phi(x) \\ \text{subject to} & Ax = b \end{array}$$

for increasing values of $t > 0$, until $m/t \leq \epsilon$. We start at a value $t = t^{(0)} > 0$, and solve the above problem using Newton's method to produce $x^{(0)} = x^\star(t)$. Then for a barrier parameter $\mu > 1$, we repeat, for $k = 1, 2, 3, \ldots$

- Solve the barrier problem at $t = t^{(k)}$, using Newton's method initialized at $x^{(k-1)}$, to produce $x^{(k)} = x^\star(t)$
- Stop if $m/t \leq \epsilon$
- Else update $t^{(k+1)} = \mu t$

The first step above is called a centering step (since it brings $x^{(k)}$ onto the central path)

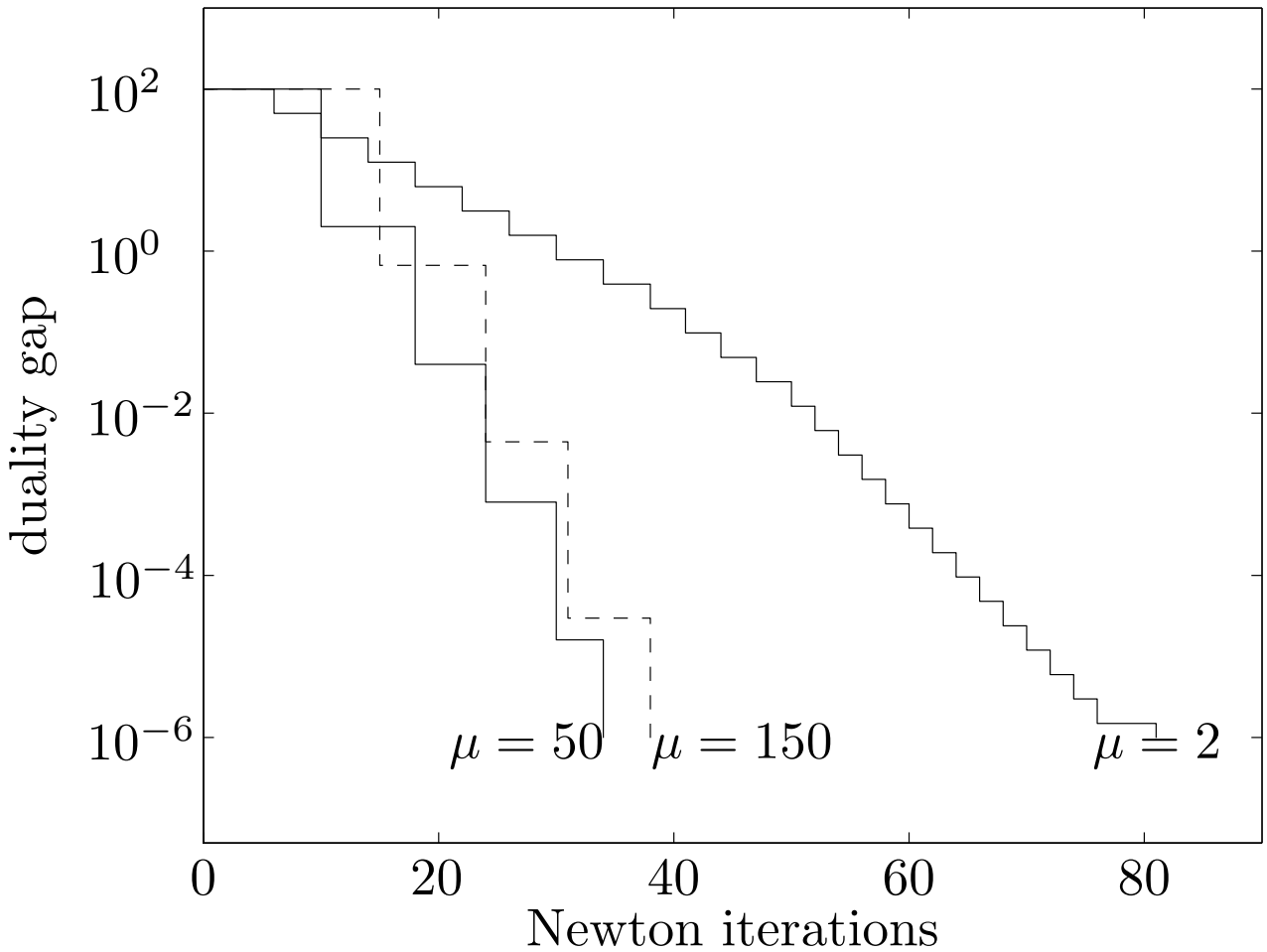
Considerations:

- Choice of μ: if μ is too small, then many outer iterations might be needed; if μ is too big, then Newton's method (each centering step) might take many iterations to converge
- Choice of $t^{(0)}$: if $t^{(0)}$ is too small, then many outer iterations might be needed; if $t^{(0)}$ is too big, then the first Newton's solve (first centering step) might require many iterations to compute $x^{(0)}$

Fortunately, the performance of the barrier method is often quite robust to the choice of μ and $t^{(0)}$ in practice

(However, note that the appropriate range for these parameters is scale dependent)

Example of a small LP in $n = 50$ dimensions, $m = 100$ inequality constraints (from B & V page 571):



Projected gradient descent

- minimize $f(x), f(x) \in C^1$
- $x \in \Omega$, a nonempty closed convex set

Theorem 1.1. *[The Projection Theorem for Convex Sets]*
Let $x_0 \in \mathbb{R}^n$ and let $\Omega \subset \mathbb{R}^n$ be a nonempty closed convex set. Then $\overline{x} \in \Omega$ solves the problem

$$\min\{\frac{1}{2}\|x - x_0\|_2^2 : x \in \Omega\}$$

if and only if

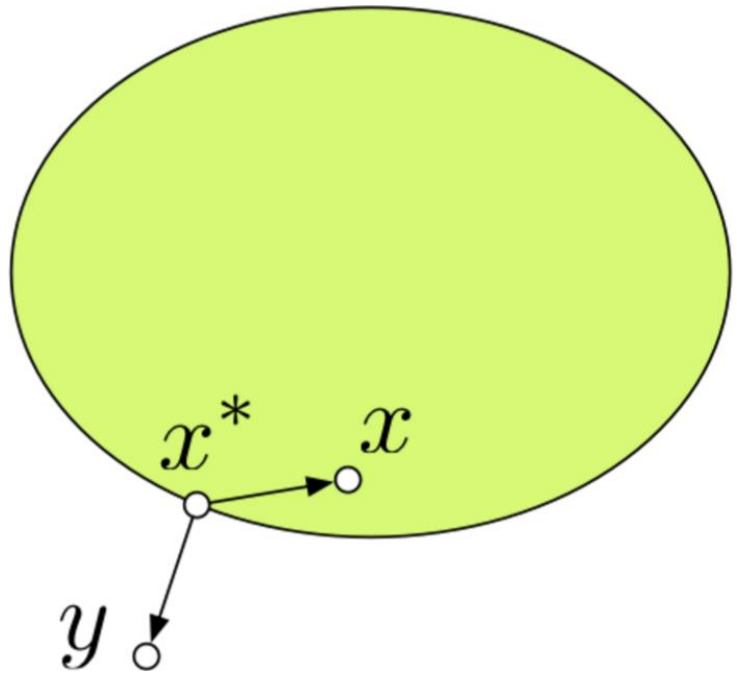
$$(\overline{x} - x_0)^T(y - \overline{x}) \geq 0 \tag{2}$$

for all $y \in \Omega$. Moreover, the solution $\overline{x}$ always exists and is unique.

Definition 1.1 (The Projection Mapping). *Let $\Omega \subset \mathbb{R}^n$ be nonempty closed convex. We define the projection into Ω to be the mapping $P_\Omega : \mathbb{R}^n \to \Omega$ given by*

$$\frac{1}{2}\|P_\Omega(x) - x\|_2^2 = \min\{\frac{1}{2}\|y - x\|_2^2 : y \in \Omega\}.$$

Observe that P_Ω is well-defined by Theorem 1.1.



Projection instances

- Box constraints

Let us suppose that Ω is given by $\Omega := \{x \in \mathbb{R}^n : \ell \leq x \leq u\}$, where $\ell, u \in \overline{\mathbb{R}}^n$ with $\overline{\mathbb{R}} = \Omega \cup \{+\infty, -\infty\}$ and $\ell :\leq u$, $i = 1, \ldots, n, \ell_i \neq +\infty$ $i = 1, \ldots, n$ and $u_i \neq -\infty$ $i = 1, \ldots, n$. Then P_Ω can be expressed componentwise as

$$[P_\Omega(x)]_i := \begin{cases} \ell_i & \text{if } x_i \leq \ell_i \\ x_i & \text{if } \ell_i < x_i < u_i \\ u_i & \text{if } u_i \leq x_i \end{cases}$$

Thus, for example, if $\Omega = \mathbb{R}^n_+$, then

$$P_\Omega(x) = x_+.$$

- Polyhedron constraint

Let Ω be the polyhedron given by

$$\Omega := \{x \in \mathbb{R}^n : a_i^T x \leq \alpha_i, i = 1, \ldots, 3, a_i^T x = \alpha_i, i = s+1, \ldots, m\}.$$

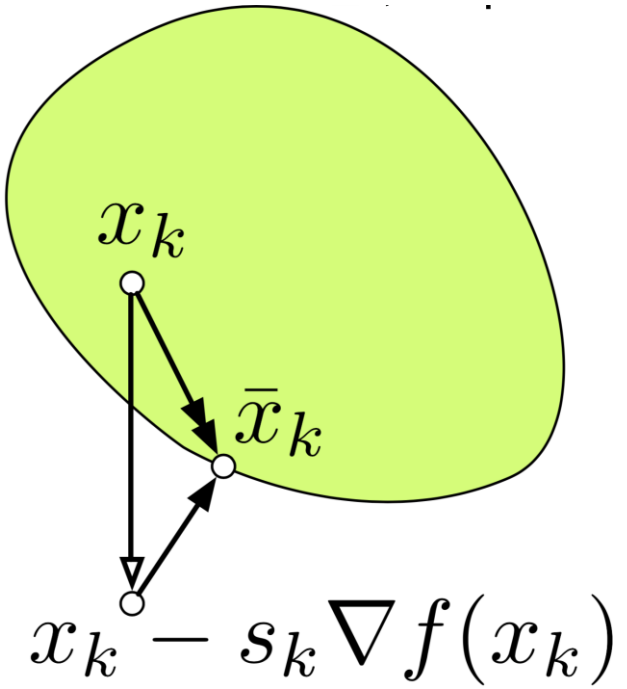
Then P_Ω is determined by solving the quadratic program

$$\begin{array}{ll} \min \frac{1}{2}\|x - y\|_2^2 & \\ \text{subject to} & a_i^T x \leq \alpha_i \quad i = 1, \ldots, s \\ & a_i^T x = \alpha_i \quad i = s+1, \ldots, m. \end{array}$$

Projected gradient descent

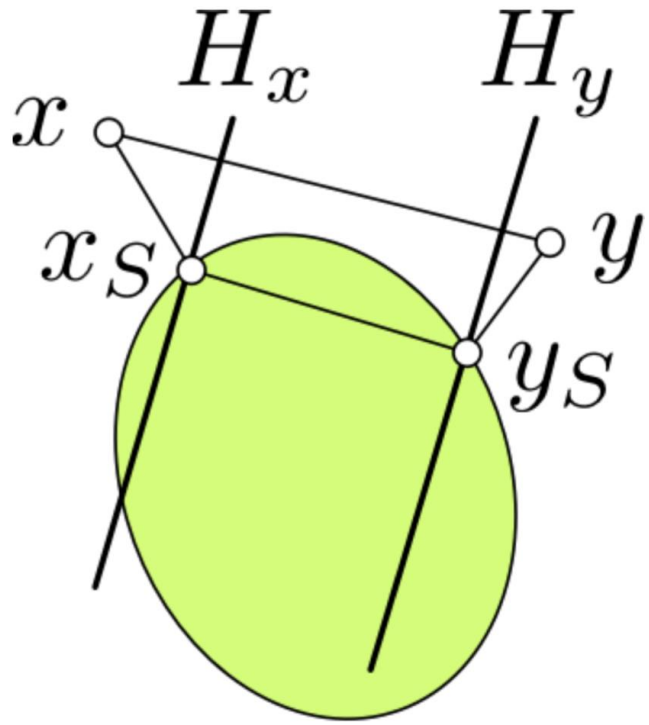
- $\bar{x}_k = P_S\left(x_k - s_k \nabla f(x_k)\right), s_k > 0$
- $x_{k+1} = x_k + \alpha_k(\bar{x}_k - x_k), \alpha_k \in (0, 1]$

- s_k is learning rate, when $\alpha_k = 1$
  - $x_{k+1} = P_S\left(x_k - s_k \nabla f(x_k)\right)$
- α_k provides another learning rate or line search process



Convergence of projected gradient descent

- Convergence of vanilla gradient descent
  - $\|x_{k+1} - x^*\| \leq M\|x_k - x^*\|$
  - $0 < M < 1$
- Convergence of projected gradient descent
  - $\|P_S(x_{k+1}) - x^*\| = \|P_S(x_{k+1}) - P_S(x^*)\| \quad (x^* \in S)$
    $\leq \|x_{k+1} - x^*\|$
    $\leq M\|x_k - x^*\|$



- $S \subset \mathcal{R}^n$ is nonempty closed convex set, The projection mapping $P_S : x \mapsto \underset{y \in S}{\operatorname{argmin}} \|x - y\|$ is a **contraction**. I.e. $\|P_S(x) - P_S(y)\| \leq \|x - y\|, \quad \forall x, y \in \mathbb{R}^n$

Projected gradient descent application

- LASSO

$$\min_{w} \sum_{i=1}^{N} (w^T x_i - y_i)^2 + \lambda \sum_{j=1}^{n} |w^j|$$

- From unconstrained problem to constrained problem

$$w_+^j = \max\{w^j, 0\} \qquad w_-^j = \max\{-w^j, 0\}$$

$$w^j = w_+^j - w_-^j \qquad |w^j| = w_+^j + w_-^j$$

$$\min_{w_+, w_-} \sum_{i=1}^{N} \left((w_+ - w_-)^T x_i - y_i \right)^2 + \lambda \sum_{j=1}^{n} (w_+^j + w_-^j)$$

$$s.t. \quad w_+ \succeq 0, w_- \succeq 0$$

Projected gradient descent application

- LASSO $$\min_{w_+, w_-} \quad \|Xw_+ - Xw_- - y\|^2 + \lambda w_+^T \mathbf{1} + \lambda w_-^T \mathbf{1}$$

$$s.t. \quad w_+ \succeq 0, w_- \succeq 0$$

$$\min_{\tilde{w}} \quad \|\tilde{X}\tilde{w} - y\|^2 + \lambda \tilde{w}^T \mathbf{1}$$

$$s.t. \quad \tilde{w} \succeq 0$$

Q & A!

Thanks!