ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

M.Sc. Program in Data Science
Department of Informatics

Optimization Techniques

Convex Optimization

Problems with Constraints – Optimality Conditions

Instructor: G. ZOIS
georzois@aueb.com

Optimization Problems with Constraints

Nonlinear Optimization under Constraints

Example:

Consider the problem

$\min x_1 + 2x_2$

s.t.

$x_1^2 + x_2^2 = 1$

How do we handle nonlinear constraints?

- Once we have such constraints, we cannot a priori use the iterative methods we have seen so far
- We would need to ensure that the sequence of points produced satisfy the set of constraints

Nonlinear Optimization under Constraints

Example:

Consider the problem

min $x_1 + 2x_2$

s.t.

$x_1^2 + x_2^2 = 1$

A first attempt: Lagrange method

- A method that can be applied when we have few variables and/or constraints
- Define the Lagrange function:

$$L(x_1, x_2, \lambda) = x_1 + 2x_2 + \lambda(x_1^2 + x_2^2 - 1)$$

- λ is called the Lagrange multiplier
- Corresponds to a "dual" variable as we will see later

Nonlinear Optimization under Constraints

Lagrange method:

- We now try to optimize the Lagrange function instead of the original one
- This is an unconstrained optimization problem
- Hence, at its minimum, it should hold that $\nabla L(x_1, x_2; \lambda) = 0$
- This will give us x_1 and x_2 as functions of λ
- The constraint will then tell us how to set λ

$$\nabla L(x_1, x_2; \lambda) = 0 \Rightarrow (1 + 2\lambda x_1, 2 + 2\lambda x_2) = (0,0) \Rightarrow (x_1, x_2) = (-\frac{1}{2\lambda}, \frac{1}{\lambda})$$

- Using now the constraint, we get a value for λ
- Substituting, we eventually have $x_1 = -2/\text{sqrt}(5)$, $x_2 = 2/\text{sqrt}(5)$

Nonlinear Optimization under Constraints

What if we have multiple equality constraints?

- We can now use one Lagrange multiplier per constraint
- Again, we will solve for $\nabla L(x; \lambda_1, \lambda_2, \ldots, \lambda_p) = 0$
- We will express x as a function of the multipliers
- The constraints will tell us the final solution

min $f(x)$

s.t.

$h_i(x) = 0, i=1,\ldots,p$

$\Rightarrow$

- Let $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_p)$
- The Lagrange function becomes:

$$L(x; \lambda) = f(x) + \Sigma_i \lambda_i h_i(x)$$

Nonlinear Optimization under Constraints

- The method does not always succeed
- Need to be careful with the values of the multipliers

Theorem:

Given an optimization problem with equality constraints, let $L(x; \lambda)$ be the Lagrange function

•If there exists a vector of Lagrange multipliers λ such that

$$\min_x L(x; \lambda) > -\infty \text{ and attained at some } x^*$$

•And if x^* satisfies the equality constraints

Then x^* is a solution to our original minimization problem

- Hence, need to determine first the acceptable range for λ
- In our example, we could not have accepted a negative value for λ

Nonlinear Optimization under Constraints

- How was this method derived?
- What is the meaning of Lagrange multipliers?

Answer:

- It is a consequence of duality theory for nonlinear programs
- The multipliers correspond to “dual” variables

Lagrange Duality and the KKT Optimality Conditions

Lagrange Duality

Consider again the more general form of optimization problems (not restricting ourselves to convex problems):

$$\min \; f(x)$$

s. t.:

$$g_i(x) \leq 0, \quad i = 1, 2, \ldots, m$$
$$h_i(x) = 0, \quad i = 1, 2, \ldots, p$$

Lagrange multipliers:

- $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_p)$ for the equality constraints
- $\mu = (\mu_1, \mu_2, \ldots, \mu_m)$ for the inequality constraints, with $\mu_i \geq 0$

The Lagrange function:

$$L(x; \lambda, \mu) = f(x) + \sum_{i=1}^{p} \lambda_i h_i(x) + \sum_{i=1}^{m} \mu_i g_i(x)$$

Lagrange Duality

The dual function is now defined as:

$$d(\lambda, \mu) = \inf_x L(x; \lambda, \mu)$$

Note that it can be the case that $d(\lambda, \mu) = -\infty$ for some values of λ and μ

Observation (essentially weak duality):

If p^* is the value of the optimal solution to the primal problem, then

$$d(\lambda, \mu) \leq p^* \text{ for any } \lambda \text{ and any } \mu \geq 0$$

Proof: If x is any feasible solution to our problem, then

$$L(x; \lambda, \mu) \leq f(x)$$

Then the infimum should also be $\leq f(x)$

But this should also hold for the optimal x

Lagrange Duality

Let us compare with linear programming duality

- In linear programming, we started with a maximization primal problem
- We searched for upper bounds on the optimal solution
- Now we have a minimization primal program
- Hence, we are interested in finding lower bounds on the optimal solution

Q: What is the best lower bound that can be derived from the dual function?

Lagrange Duality

The dual (maximization) problem corresponding to the primal

$$\max \; d(\lambda, \mu)$$

s. t.:

$$\mu_i \geq 0, \quad i = 1, 2, \ldots, m$$

Observations:

- $d(\lambda, \mu)$ is concave
- Maximizing a concave function is equivalent to minimizing a convex function
- Hence, the Lagrange dual problem is a convex optimization problem even if the primal problem is not a convex one!
- Very useful property if the primal problem is not easy to handle

Lagrange Duality

A pair (λ, μ) is called dual feasible if $\mu \geq 0$ and $d(\lambda, \mu) > -\infty$

Weak Duality:

- The same as in linear programming
- The optimal solution to the dual is the best lower bound we can hope to get
- Hence, if p^* and d^* are the optimal solutions to the primal and dual respectively, then

$$d^* \leq p^*$$

Notes:

- Weak duality holds even if the primal problem is not convex
- Inequality holds also when p^* or d^* are infinite
- E.g., if $p^* = -\infty$, then we must have that $d^* = -\infty$, hence there is no dual feasible solution, the dual problem is infeasible
- If $d^* = +\infty$, then the primal problem is infeasible

Lagrange Duality

What about strong duality?

- Does it hold that $p^* = d^*$?
- Unlike linear programming, strong duality does not hold in general
- $p^* - d^*$ = duality gap

HOWEVER:

- When we have a convex optimization problem, strong duality holds in most cases
- There are various results specifying conditions under which strong duality holds

Slater's condition:

- If we have a **convex optimization problem**,
- and **there exists a feasible point such that $g_i(x) < 0$ for i=1,...,m, and $h_i(x) = 0$ for i=1,...,p**, then the dual optimal value is attained when $d^\star > -\infty$, i.e., there exists a dual feasible (λ^*, μ^*) such that:

$$d(\lambda^*, \mu^*) = p^* = d^*$$

Lagrange Duality

Example with LP

Consider the following form of a linear program

min $c^T x$		min $c^T x$
s.t.	Convert to the more convenient form $\Rightarrow$	s.t.
$Ax = b$		$A_i x - b_i = 0$, $i=1,\dots,m$
$x \geq 0$		$-x_i \leq 0$, $i=1,\dots,n$

- The Lagrange function:

$$\begin{aligned} L(x;\lambda,\mu) &= c^T \cdot x + \sum_{i=1}^{m} \lambda_i (A_i \cdot x - b_i) - \sum_{i=1}^{n} \mu_i x_i \\ &= -b^T \cdot \lambda + c^T \cdot x + \lambda^T \cdot A \cdot x - \mu^T \cdot x \\ &= -b^T \cdot \lambda + (c + A^T \cdot \lambda - \mu)^T \cdot x \end{aligned}$$

Lagrange Duality

Example with LP

- The dual function:

$$d(\lambda, \mu) = \inf_x L(x; \lambda, \mu)$$

- If $c + A^T\lambda - \mu$ is not identically 0, then the infimum is $-\infty$
- Otherwise, it is equal to $-b^T \lambda$

$$d(\lambda, \mu) = \begin{cases} -b^T \cdot \lambda, & \text{if } A^T \cdot \lambda + c - \mu = 0 \\ -\infty & \text{otherwise} \end{cases}$$

Lagrange Duality

Example with LP

- For the dual to be feasible, we need $A^T\lambda - \mu + c = 0$
- Since, we have the constraint $\mu \geq 0$, this means $A^T \lambda + c \geq 0$
- The dual then becomes:

$$\begin{aligned} &\max \; -b^T \cdot \lambda \\ &\text{s. t.:} \\ &\qquad A^T \cdot \lambda + c \geq 0 \end{aligned}$$

- Set now $y = -\lambda$
- We then get the same LP as we would get with the more standard way of producing the dual LP

$$\begin{aligned} &\max \; b^T \cdot y \\ &\text{s. t.:} \\ &\qquad A^T \cdot y \leq c \end{aligned}$$

Optimality Conditions

- duality framework ⇒ optimality conditions for 2 candidate primal and dual solutions (in analogy to the complementary slackness conditions in linear programming)
- Suppose all functions are differentiable and strong duality holds (duality gap is zero and the dual optimum is attained)
- Let x and (λ, μ) be primal and dual feasible solutions
- **If they are optimal solutions**, they must satisfy the **KKT optimality conditions**

$$g_i(x) \leq 0, \quad i = 1, 2, \ldots, m \tag{1}$$

$$h_i(x) = 0, \quad i = 1, 2, \ldots, p \tag{2}$$

$$\mu_i \geq 0, \quad i = 1, 2, \ldots, m \tag{3}$$

$$\mu_i \cdot g_i(x) = 0, \quad i = 1, 2, \ldots, m \tag{4}$$

$$\nabla f(x) + \sum_{i=1}^{p} \lambda_i \nabla h_i(x) + \sum_{i=1}^{m} \mu_i \nabla g_i(x) = 0, \tag{5}$$

Optimality Conditions

The KKT conditions

- Independently derived by Karush (1939, M.Sc. thesis) and by Kuhn and Tucker (1951).
- For convex optimization problems where strong duality holds, these are necessary and sufficient conditions for optimiality

Condition (4) - Complementarity condition

- $\mu_i \cdot g_i(x) = 0$
- Either the dual variable $\mu_i = 0$, or the i-th inequality constraint must be tight
- In analogy to linear programming

Optimality Conditions - Examples

1. Write the KKT conditions and find the optimal solution for the problem:

$$\min x - 2y$$

s.t.

$$x^2 + 2y^2 \leq 1$$

2. Write the KKT conditions for the problem:

$$\min -\Sigma_i \ln(\alpha_i + x_i)$$

s.t.

$$\Sigma_i x_i = 1$$

$$x_i \geq 0, i=1,\ldots, n$$

Convex Optimization Problems with Equality Constraints

Optimization with equality constraints

- If the problem has a relatively simple form, we can use the KKT conditions to derive the optimal solution
- The complementarity condition is now absent, since we do not have inequality constraints
- All KKT conditions are equalities, hence we might hope to solve this system in simple cases
- We exhibit such a solution for convex quadratic programs

Convex Quadratic Minimization with equality constraints

A convex quadratic program:

$$\min \; f(x) = \frac{1}{2}x^T \cdot P \cdot x + q^T \cdot x + r$$

s. t.:

$$A \cdot x = b$$

where

- $x = (x_1, x_2, \ldots, x_n)$
- P is a symmetric PSD n x n matrix
- q is a n-dimensional vector
- r is a constant

Convex Quadratic Minimization with equality constraints

- The KKT conditions yield:

$Px + q + A^T\lambda = 0$

$Ax = b$

In a more concise form:

$$\begin{bmatrix} P & A^T \\ A & 0 \end{bmatrix} \cdot \begin{bmatrix} x \\ \lambda \end{bmatrix} = \begin{bmatrix} -q \\ b \end{bmatrix}$$

- Called the KKT matrix
- If non-singular, we have a unique solution

Solving the dual

- Another approach would be to construct the dual
- Useful in certain cases but not always successful
- Advantage: since we have no inequalities, the dual is an unconstrained optimization problem!
  - Recall the dual requires that $\mu \geq 0$ only when we have inequality constraints
- Disadvantages: It may not be easy to describe the dual
- It may also not be easy to recover the primal solution from the dual

Newton's Method

- Suppose now we have a non-quadratic problem
- We will see a generalization of Newton's method in the presence of equality constraints
- Almost the same approach except that:
  - We now need to start with a feasible solution
  - We need to ensure the update will continue to be feasible
- We will use the fact that quadratic problems can be solved via the KKT conditions
- We typically write the problem in the form:

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

Newton's Method

- We start with an initial feasible solution
  - Hence we first pick a point $x^{(0)}$ such that $Ax^{(0)} = b$
- How can we perform the updates and maintain feasibility?
  - Idea: it suffices to ensure in every iteration k that $A \cdot \Delta x^{(k)} = 0$
  - Then $A \cdot x^{(k+1)} = A \cdot (x^{(k)} + \alpha_k \Delta x^{(k)}) = A \cdot x^{(k)} = b$

Newton's Method

- Recall the 2nd order Taylor approximation for a function of n variables, at a given point $x^{(k)}$

$$f(x^{(k)} + \delta) = f(x^{(k)}) + \nabla f(x^{(k)})^T \cdot \delta + ½\, \delta^T \cdot H(f, x^{(k)}) \cdot \delta$$

- Hence each step of the procedure reduces to:

$\min f(x^{(k)} + \delta)$		$\min f(\delta)$
s.t.	$\Rightarrow$	s.t.
$A(x^{(k)} + \delta) = b$		$A \cdot \delta = 0$

- But this is precisely a quadratic program with linear constraints
- When the KKT matrix is non-singular, we can solve this

Newton's Method

- Finding the search direction (or Newton step) for the next iteration
- We need to solve

$$\begin{bmatrix} H(f, x^{(k)}) & A^T \\ A & 0 \end{bmatrix} \cdot \begin{bmatrix} \Delta x^{(k)} \\ \lambda \end{bmatrix} = \begin{bmatrix} -\nabla f(x^{(k)}) \\ 0 \end{bmatrix}$$

- If the KKT matrix is not invertible, we can make some small perturbation on the values
- We can then use backtracking line search to determine the step size
- If the problem is quadratic we will be done in 1 iteration
- If the function is nearly quadratic, we have made good progress towards the optimal solution

Newton's Method

- Convergence analysis similar to the unconstrained case
- For strongly convex functions, the analysis yields upper bounds on the number of iterations till we are ε-close to the optimal solution
- Usually very high accuracy with only few iterations
- Infeasible start Newton method:
  - A variant where the initial point is not a feasible solution

Convex Optimization Problems with Inequality Constraints

Optimization under inequality constraints

Consider again the general form of convex optimization problems

$$\min \; f(x)$$

$$\text{s. t.:}$$

$$g_i(x) \leq 0, \quad i = 1, 2, \ldots, m$$

$$h_i(x) = 0, \quad i = 1, 2, \ldots, p$$

- Where each h_i is a linear function
- Each g_i is a convex function
- The main problem in solving this comes from the inequality constraints

Optimization under inequality constraints

Interior Point Methods

Main ideas:

- We want to prevent each g_i from becoming positive
- We will work with feasible solutions that are "away from the boundary" of the feasible region
- How can we enforce this?
  - We will incorporate into the objective a function of the inequality constraints
  - The function will be appropriately chosen so that it "penalizes" solutions close to the boundary
  - The new objective is usually referred to as the barrier function

Interior point methods for linear programming

- It is convenient to illustrate the main ideas for solving LPs
- Suppose we have a LP in the form

$$\min c^T x$$
s.t.
$$A \cdot x \geq b$$
$$x \geq 0$$

- Let's add slack variables and bring it into the form:

$$\min c^T x$$
s.t.
$$A \cdot x = b$$
$$x \geq 0$$

Interior point methods for linear programming

- How could we enforce that each $x_j > 0$?
- Idea: For each constraint $x_j > 0$, we add to the objective function the term $-\log(x_j)$
- This penalizes each x_j from going close to 0
  - For a minimization problem, x_j should better be away from 0 if we have the negative of a logarithm into the objective function
- The barrier function with parameter $\mu > 0$:

$$B_\mu(x) = c^T x - \mu \, \Sigma_j \log(x_j)$$

Logarithmic barrier functions

- The barrier function with $\mu > 0$:

$$B_\mu(x) = c^T x - \mu \sum_j \log(x_j)$$

- When μ becomes large, the logarithmic terms are dominating
- As μ approaches 0, the logarithmic terms become negligible and we are back to the original problem

Barrier problems

- Family of non-linear optimization problems, parameterized by $\mu>0$ (called barrier problems):

$$\min B_\mu(x) = c^T x - \mu \sum_j \log(x_j)$$

s.t.

$$A \cdot x = b$$

BP(μ)

Facts:

- BP(μ) always has a unique optimal solution, because $B_\mu(x)$ is strongly convex
- Given μ, let $x(\mu)$ be the optimal solution of BP(μ)
- $\lim_{\mu \to 0} x(\mu)$ = optimal solution to the initial LP problem

Barrier problems

- As μ varies from $+\infty$ down to 0, the optimal solution $x(\mu)$ moves along a trajectory called the central path
- When $\mu = +\infty$, the problem becomes equivalent to:

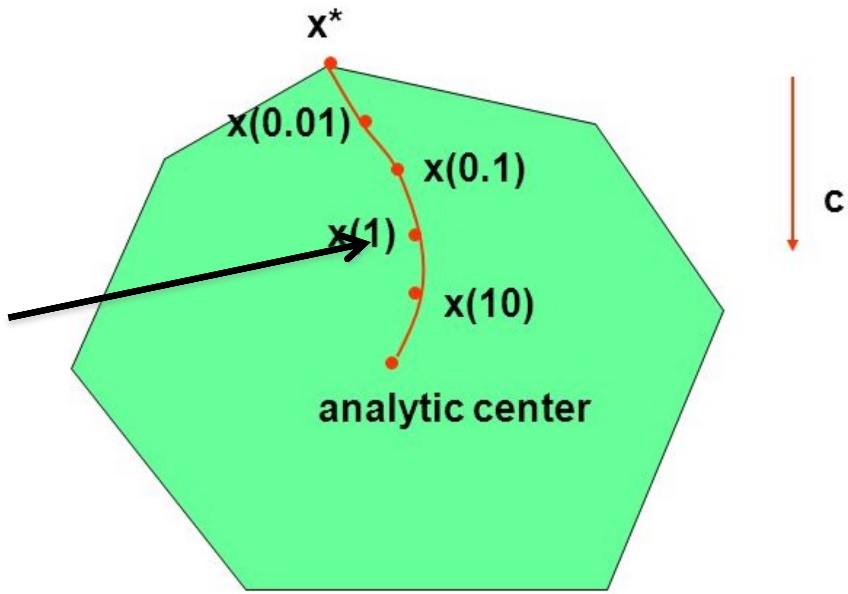
$$\min \; -\Sigma_j \log(x_j)$$

s.t.

$$A \cdot x = b$$

- The optimal solution to this problem is called the analytic center of the feasible region

Barrier problems



x^* = optimal solution of the LP

The central path

- The central path starts at the analytic center and ends at the optimal solution that we want to compute for the initial LP problem

[Reading material: Sections 11.2,11.3 from the book of Boyd & Vandenberghe, and Lecture notes from the course on Machine Learning by Ryan Tibshirani: https://www.stat.cmu.edu/~ryantibs/convexopt/lectures/barr-method.pdf]

Path following interior point algorithms

- The barrier problem is non-linear
- BUT: it is a convex optimization problem with linear equality constraints
- We could solve it with Newton's method
- It suffices to come close to $x(\mu)$ for any fixed $\mu>0$
- This means we stay "near" the central path
- By decreasing μ and repeating the process, we gradually approach the optimal solution

Path following interior point algorithms

Description of path following algorithms

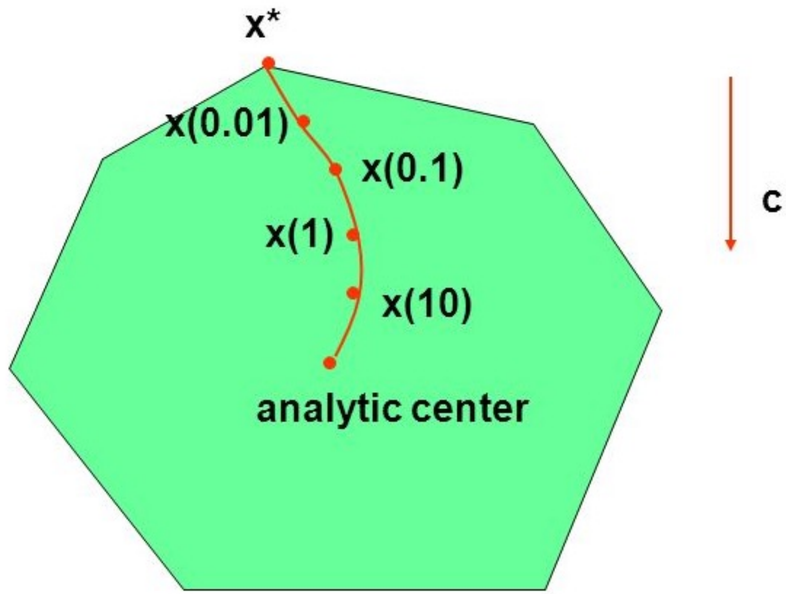
- Initialization: Start with a feasible solution in the interior of the feasible region, and fix a value for μ
- Repeat:
  - Check if the stopping criterion is met
  - If not, find x(μ) (perhaps approximately) using Newton's method
    - Called the centering step
  - Update μ: Set μ= αμ, for α with $0 < \alpha < 1$
    - μ is decreasing geometrically
    - Typically $\alpha \in [1/20, 1/10]$

Path following interior point algorithms

Main properties

- Can be adjusted to solve together the primal and the dual LP
- Several variants for the initial choice of primal and dual feasible solutions
- Choice of α: it involves a trade-off
  - If α is small, μ decreases fast but we may do more Newton iterations in each step
  - If α is large, μ decreases slowly, we need a higher number of updates on μ, but fewer Newton iterations within each step
  - In practice, it works well if $\alpha = 1/t$, with $t \in \{10, 11, ..., 20\}$

Geometric interpretation



x^* = optimal solution of the LP

- We may not be able to compute the exact value of x(μ), for μ>0
- We may also run just a few Newton iterations in each step
- So, we may not be moving on the central path
- But, we always move very close to the central path
- We call such algorithms path following algorithms

Path following algorithms for convex problems

- The idea for solving LPs can be used for convex optimization problems as well

$$
\begin{aligned}
&\min \ f(x) \\
&\text{s. t.:} \\
&\qquad g_i(x) \leq 0, \quad i = 1, 2, \ldots, m \\
&\qquad h_i(x) = 0, \quad i = 1, 2, \ldots, p
\end{aligned}
$$

- We need now to produce a path in the interior of the feasible region
  - Maintain < in the inequality constraints

Path following algorithms for convex problems

$$\begin{aligned} &\min \; f(x) \\ &\text{s. t.:} \\ &\qquad g_i(x) \leq 0, \quad i = 1, 2, \ldots, m \\ &\qquad h_i(x) = 0, \quad i = 1, 2, \ldots, p \end{aligned}$$

- New barrier function:

$$B_\mu(x) = c^T x - \mu \, \Sigma_i \log(-g_i(x))$$

- The central path and the analytic center are defined in the same way
- We can again use Newton's method within each iteration

Conclusions on interior point algorithms

- Initial variants not as fast
- Currently, for LPs they have comparable performance with simplex
- For convex optimization, one of the best methods to solve non-quadratic problems
- Provably polynomial running time
  - For the exact solution in LP
  - For an ε-approximate solution in general convex problems