Tutorial Session on
Clustering Large and High-Dimensional Data

Organizers

Jacob Kogan, UMBC

Charles Nicholas, UMBC

Marc Teboulle, Tel-Aviv University

International Conference on Information and Knowledge Management – New Orleans, November 3-8, 2003

A Tutorial on Modern Optimization: Theory, Algorithms and Applications

Marc Teboulle

School of Mathematical Sciences

Tel-Aviv University

`www.math.tau.ac.il/~teboulle`

presented at

Tutorial Session on Clustering Large and High-Dimensional Data
CIKM 2003

International Conference on Information and Knowledge Management – New Orleans, November 3-8, 2003

Planning
Data Networks
Finance-Economics
VLSI Design
....
Pattern Recognition
Data Mining
Ressource Allocation
Machine Learning
Signal Processing
Tomography
Human Behavior....

OPTIMIZATION APPEARS TO BE PRESENT ”ALMOST” EVERYWHERE....

Outline of the Talk

- **Ideas and Principles**
- **Constrained Problems: Difficulties**
- **Convexity and Duality: A Working Horse in Optimization**
- **Some Fundamental/Useful Optimization Models**

♣♣♣♣♣

- **Devising Optimization Algorithms**
- **Convergence and Complexity issues**
- **Basic Iterative Schemes for Unconstrained Problems**
- **Some Classical and Modern Algorithms for Constrained Problems**

History of Optimization....

- Fermat (1629): Unconstrained Minimization Principle
- ...+160...Lagrange (1789) Equality Constrained Problems (Mechanics)
- Calculus of Variations, 18-19th Century [Euler, Lagrange, Legendre, Hamilton...]
- ...+150...Karush (1939), Fritz-John (47), Kuhn-Tucker (1951)
- KKT Theorem for Inequality Constraints: Modern Optimization Theory
- Engineering Applications (1960)
- Optimal Control Bellman, Pontryagin...
- Major Algorithmic Developments (50's with LP) and 60-80's for NLP
- Polynomial Interior Points Methods for Convex Optimization Nesterov-Nemirovsky (1988)
- Combinatorial Problems via continuous approximations 90's
- ....More Theory, Algorithmic **and much more applications** .... A young, and vibrant area of research.

General Formulation: Nonlinear Programming

$$(O) \quad \text{minimize}\{f(x) : \; x \in X \cap C\}$$

$X \subset \mathbb{R}^n \equiv n$-dimensional Euclidean space, (implicit or simple constraints)
C is a set of explicit constraints described by constraints

$$\begin{aligned} C = \{x \in \mathbb{R}^n : g_i(x) &\leq 0, i = 1, \dots m, \\ h_i(x) &= 0, i = 1, \dots, p\}. \end{aligned}$$

All the functions in problem (O) are real valued functions on $\mathbb{R}^n$, and the set X can describe more abstract constraints of the problem.

Very Important Special Case: Unconstrained Problem $X \cap C \equiv \mathbb{R}^n$

$$(U) \quad \text{minimize}\{f(x) : \; x \in \mathbb{R}^n\}$$

Many (if not most) methods for constrained problems based on solving some type of problem (U).

Definitions and Terminology

$$(O) \quad \text{minimize}\{f(x) : \ x \in X \cap C\}$$

- A point $x \in X \cap C$ is called a feasible solution of (O).
- An optimal solution is any feasible point where the local or global minimum of f relative to $X \cap C$ is actually attained.

Definition

$$\begin{aligned} x^* \text{ local mininum } f(x^*) &\leq f(x), \ \forall x \in N_\epsilon(x^*) \\ x^* \text{ global minimum } f(x^*) &\leq f(x), \ \forall x \in \mathbb{R}^n \end{aligned}$$

Note: There are also "max" problems...But $\max F \equiv -\min[-F]$

How to Solve an Optimization Problem?

- Analytically/Explicitly: Very rarely....or Never....
- We try to generate an **Iterative-Descent Algorithm** to **approximately** solve the problem to a prescribed accuracy.

Algorithm: a map $\mathcal{A} : x \rightarrow y$ (start with x to get new point y)
Iterative: generate a sequence of pts calculated on prior point or points
Descent: Each new point y is such that $f(y) < f(x)$

A Powerful Algorithm!

Set $k = 0$

While $x^k \in \mathcal{D} \equiv \{$set of desisable Points$\}$ Do $\{$

$$\begin{aligned} x^{k+1} &= \mathcal{A}(x^k) \\ k &\leftarrow k + 1\} \end{aligned}$$

Stop

Expected Output(s): $\{x^k\}$ *is a minimizing sequence*: as $k \to \infty$

- $f(x^k) \to f_*$, (optimal value)
- or/and even more, $x^k \to x^*$ (optimal solution)

Some Basic Questions

- How do we pick the initial starting point?
- How to construct $\mathcal{A}$ so that x^k converges to optimal x^*?
- How do we stop the algorithm?
- How close is the approximate solution to the optimal one? (that we do not know!)
- How sensitive is the whole process to data perturbations?
- How fast the algorithm converges to optimality?
- What is the computational cost? The complexity ?

Emerging Topics and Tools

To answer these questions, we need an appropriate mathematical foundation. For example:

- Existence of optimal solutions
- Optimality conditions
- Convexity and Duality
- Convergence and Numerical Analysis
- Error and Complexity Analysis

While each algorithm for each type of problem will often require a specific analysis (exploiting special structures of the problem), the above tools will remain essential and fundamental.

Optimality for Unconstrained Minimization

$(U) \quad \inf\{f(x) : x \in \mathbb{R}^n\}$ $f : \mathbb{R}^n \to \mathbb{R}$ is a *smooth* function.

Fermat Principle Let $x^* \in \mathbb{R}^n$ be a local minimum. Then,

$$\spadesuit \qquad \nabla f(x^*) = 0, \quad \text{Zero Slope}$$

This is a **First Order Necessary condition**

Second Order Necessary Condition: Nonnegative curvature at x^*

The Hessian Matrix $\nabla^2 f(x^*) \succeq 0$ positive semidefinite

Sufficient conditions for x^* to be a local min.
Replace $\nabla^2 f(x^*) \succeq 0$ by $\nabla^2 f(x^*) \succ 0$

Whenever f is assumed ***convex***, then $\spadesuit$ becomes a **sufficient condition** for x^* to be a **global** minimum for f.

Convexity

$S \subset \mathbb{R}^n$ is convex if the line segment joining any two different points of S is contained in it:

$$\forall x, y \in S,\ \forall \lambda \in [0, 1] \implies \lambda x + (1 - \lambda)y \in S$$

$f : S \to \mathbb{R}$ is convex if for any $x, y \in S$ and any $\lambda \in [0, 1]$,

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)$$

A Key Fact: Local Minima are also Global under convexity

Convexity plays a fundamental role in optimization
Even in Non convex problems!

Equality constraints:Lagrange Theorem

$$(E) \quad \min\{f(x) : \ h(x) = 0, \ x \in \mathbb{R}^n\}$$

with $f : \mathbb{R}^n \to \mathbb{R}, \ h : \mathbb{R}^n \to \mathbb{R}^p$.

Lagrange Theorem (necessary conditions) Let x^* be a local minimum for problem (E). Assume:

$$(A) \quad \{\nabla h_1(x^*), \ldots, \nabla h_p(x^*)\} \text{ are linearly independent.}$$

Then there exists a unique $y^* \in \mathbb{R}^p$ satisfying:

$$\nabla f(x^*) + \sum_{k=1}^{p} y_k^* \nabla h_k(x^*) = 0.$$

A system of $(n+p)$ nonlinear equations in $(n+p)$ variables (x^*, y^*)

Inequality constraints lead to more complications....

Inequality Constraints: The Lagrangian

$$(P) \quad f_* := \inf\{f(x) : \; g(x) \le 0, \; x \in \mathbb{R}^n\}$$

with $f : \mathbb{R}^n \to \mathbb{R}, \; g : \mathbb{R}^n \to \mathbb{R}^m$ are given data.

We assume that there exists a *feasible* solution for (P) and $f_* \in \mathbb{R}$.

Observation : Problem (P) is equivalent to

$$\inf_{x \in \mathbb{R}^n} \sup_{y \ge 0} \{f(x) + \langle y, g(x) \rangle\}$$

which leads to the Lagrangian associated with (P) $L : \mathbb{R}^n \times \mathbb{R}^m_+ \to \mathbb{R}$:

$$L(x, y) = f(x) + \langle y, g(x) \rangle \equiv f(x) + \sum_{i=1}^{m} y_i g_i(x).$$

Hidden in this equivalent min-max formulation of (P) is another problem called the **DUAL**. This in turn is also at the origin of optimality conditions.

Definition A vector $y^* \in \mathbb{R}^m$ is called a Lagrangian multiplier for (P) if

$$y^* \ge 0, \text{ and } f_* = \inf\{L(x, y^*) : x \in \mathbb{R}^n\}$$

.

Lagrangian Duality

$$L(x,\lambda) = f(x) + \sum_{i=1}^{m} y_i g_i(x).$$

and

$$(P) \iff \inf_{x \in S} \sup_{y \in \mathbb{R}^m_+} L(x,y)$$

Suppose we can reverse the $\inf \sup$ **operations**, that is consider

$$\sup_{y \in \mathbb{R}^m_+} \inf_{x \in C} L(x,y)$$

Define the **Dual Function:**

$$h(y) := \inf_{x \in S} L(x,y), \quad \operatorname{dom} h = \{y \in \mathbb{R}^m : \ h(y) > -\infty\}.$$

and the **Dual Problem:**

$$(D) \quad h_* := \sup\{h(y) : y \in \mathbb{R}^m_+ \cap \operatorname{dom} h\}$$

Note: In general the dual problem consists of simple nonnegativity constraints. **But**, to avoid $h(\cdot)$ to be $-\infty$, *additional constraints* might also emerge through $y \in \operatorname{dom} h$.

Dual problem Properties

The dual Problem Uses the same data

$$(D) \qquad h_* = \sup_y \{h(y) : y \in \mathbb{R}^m_+ \cap \operatorname{dom} h\}, \quad h(y) = \inf_x L(x, y)$$

Properties of (P)-(D)

- Dual is **always convex** (ax max of concave func.)
- **Weak duality holds:** $f_* \geq h_*$ for any feasible pair (P)-(D)

Valid for **any** optimization problem. **No convexity assumed or/and, any other assumptions!**

Duality: Key Questions for the pair (P)-(D)

$$\boxed{f_* = \inf\{f(x) : \; g(x) \leq 0, \; x \in \mathbb{R}^n\}; \; h_* = \sup_y \{h(y) : y \in \mathbb{R}^m_+\}}$$

- **Zero Duality Gap:** when $f_* = h_*$?
- **Strong Duality:** when $\inf / \sup$ attained?
- **Structure/Relations of Primal-Dual Optimal Sets/Solutions**

Convex data + a *Constraint Qualification*,on constraints e.g.,

$\exists \widehat{x} \in \mathbb{R}^n : \quad g(\widehat{x}) < 0$

deliver the answers.

Linear equality constraints can also be treated easily.

Proof based on a simple and powerful geometric argument: Any point outside a closed convex set can be separated by a hyperplane.

An Example: Least Squares Optimization

$$(P) \qquad \min_x \|Ax-b\|^2 \iff \min_{x,z}\{\|z\|^2 : \ Ax-b=z\}$$

$$(D) \qquad \max\{\|b\|^2 - \|y-b\|^2 : \ A^T y = 0\}$$

Strong Duality holds: $\min(P) = \max(D)$

(distance to subspace $R(A))^2 +$ (distance to $N(A^T))^2 = \|b\|^2$

An Example: Least Squares Optimization

$$(P) \qquad \min_x \|Ax - b\|^2 \iff \min_{x,z}\{\|z\|^2 : \ Ax - b = z\}$$

$$(D) \qquad \max\{\|b\|^2 - \|y - b\|^2 : \ A^T y = 0\}$$

Strong Duality holds: $\min(P) = \max(D)$

(distance to subspace $R(A))^2 +$ (distance to $N(A^T))^2 = \|b\|^2$

$\boxed{\textbf{...THIS PROVES PYTHAGORAS THEOREM !}}$

Primal-Dual Optimal Solutions

Definition The pair $(x^*, y*) \in S \times \mathbb{R}^m_+$ is called a saddle point for L if

$$L(x^*, y) \leq L(x^*, y^*) \leq L(x, y^*), \ \ \forall x \in S, \ \forall y \in \mathbb{R}^m_+.$$

Proposition (Saddle point characterization)

$$(x^*, y^*) \in S \times \mathbb{R}^m_+$$

is a saddle point for L iff
(a) $x^* = \text{argmin}_{x \in S} L(x, y^*)$ (L-optimality)
(b) $x^* \in S, \ g(x^*) \leq 0$ (Primal feasibility)
(c) $y^* \in \mathbb{R}^m_+$ (Dual feasibility)
(d) $y_i^* g_i(x^*) = 0, \ i = 1, \ldots, m$ (Complementarity).

Note that the above is valid with **0-assumptions on the problem's data!**

Proposition (Sufficient condition for optimality) If $(x^*, y^*) \in S \times \mathbb{R}^m_+$ is a saddle point for L, then x^* is a global optimal solution for NLP.

Once again this result is very general and holds for **any** optimization problem. However for nonconvex problem it is in general difficult to find a saddle point.

The KKT Theorem

$$(P) \quad \inf\{f(x) : \ g(x) \le 0, \ x \in \mathbb{R}^n\}$$

Let x^* be a local minimum for problem (P) and assume that a (CQ) holds. Then there exists a Lagrange multipier $y^* \in \mathbb{R}^m_+$ s.t.

$$\begin{array}{rcl} \nabla f(x^*) + \sum_{i=1}^m \lambda_i^* \nabla g_i(x^*) & = & 0, \ \text{[Saddle pt. in } x^*\text{]} \\ g_i(x^*) & \le & 0, \ \forall i \in [1,m], \ \text{[Feasibility} \equiv \text{Saddle pt. in } y^*\text{]} \\ y_i^* g_i(x^*) & = & 0, \ i = 1, \dots, m. \end{array}$$

The system of equations and inequalities is called the KKT system.

With *convex data* + (CQ), the KKT conditions become **necessary and sufficient for *global* optimality**...Closing the loop....Equiv. to strong duality....

Useful Convex Models: Conic Problems

$$\min\{\langle c, x\rangle : \mathcal{A}(x) = b, x \in \mathcal{K}\}$$

- $\mathcal{K}$ is a closed convex cone in some finite dimensional space X
- $\langle \cdot, \cdot \rangle$ appropriate inner product on X
- $\mathcal{A}$ is a linear map

Example: Linear Programming

$X \equiv \mathbb{R}^n,\ x \in \mathbb{R}^n$ decision variables
$K \equiv \mathbb{R}^n_+$, the nonnegative orthant
$A \in \mathbb{R}^{m\times n},\ b \in \mathbb{R}^m, c \in \mathbb{R}^n$ and $\langle \cdot, \cdot \rangle$ the usual scalar product in $\mathbb{R}^n$

....Other Examples...?

Semidefinite Programming-Primal Dual Forms

$$\min_{x \in \mathbb{R}^m} \{c^T x : A(x) \succeq 0\};$$

$$\max_{Z \in S_n} \{-\operatorname{tr} A_0 Z : \ \operatorname{tr} A_i Z = c_i, \ i \in [1, m], \ Z \succeq 0\}$$

Here

$$A(x) := A_0 + \sum_{i=1}^{m} x_i A_i, \text{ each } A_i \in S_n \equiv \text{ symmetric}$$

- **Primal :** $x \in \mathbb{R}^n$ decision variables. $A(x) \succeq 0$ is a linear matrix inequality.
- **Dual in Conic Form:** $Z \in S_n$ decision variables, $\mathcal{K} \equiv S_n^+$ is the closed convex cone of p.s.d. matrices, tr trace of a matrix

SDP Features and Applications

♢ Features

- SDP are special classes of convex (nondifferentiable) problems
- Computationally tractable: Can be approximately solved to a desired accuracy in polynomial time
- Include linear and quadratic programs
- A very active research area since mid 90's

♢ Applications–A Short list!

- Combinatorial optimization
- Control theory
- Statistics
- Computational Geometry
- Classification and Clustering problems

Related conic convex problems

Other models arising in many applications include

- Second order cone programming
- max-determinant optimization problems
- Eigenvalue problems

Convex Optimization–Summary

- Local minima are global

Convex Optimization–Summary

- Local minima are global
- Computationally Tractable: Can be **approximately solved** to a desired accuracy in **polynomial time** [Self-Concordance Theory–Nemirovski-Nesterov]

Convex Optimization–Summary

• Local minima are global

• Computationally Tractable: Can be **approximately solved** to a desired accuracy in **polynomial time** [Self-Concordance Theory–Nemirovski-Nesterov]

• Model many more problems than one might think!

Convex Optimization–Summary

- Local minima are global
- Computationally Tractable: Can be **approximately solved** to a desired accuracy in **polynomial time** [Self-Concordance Theory–Nemirovski-Nesterov]
- Model many more problems than one might think!
- Enjoy a powerful Duality Theory that can be used to find bounds for **hard** problems

Tractability is a key Issue

- Drawing a line between **Easy [Convex]and Hard [Nonconvex] Problems**

- Convexity plays a key role in this distinction.

Easy/Hard: Example

$$(P1) \qquad \max\{\sum_{j=1}^{n} x_j : \; x_j^2 - x_j = 0, j = 1, \ldots, n; \; x_i x_j = 0 \;\forall i \neq j \in \Gamma\}$$

$(P2)$ $\quad$ $\inf \;\; x_0$ subject to

$$\begin{aligned}
\sum_{j=1}^{m} x_j = 1, \; \sum_{j=1}^{m} a_j x_j^l \;&=\; b^l, l = 1, \ldots, k \\
\lambda_{\min} \begin{pmatrix} x_1 & & & & x_1^l \\ & \cdot & & & \cdot \\ & & \cdot & & \cdot \\ & & & \cdot & \cdot \\ & & & x_m & x_m^l \\ x_1^l & \cdot & \cdot & \cdot \; x_m^l & x_0 \end{pmatrix} \;&\geq\; 0, l = 1, \ldots, k \\
x \in \mathbb{R}^{m+1}, \; x^l \in \mathbb{R}^m, \; l \;&=\; 1, \ldots, k
\end{aligned}$$

(P1) ”looks” much easier than (P2)...

Easy/Hard: Example

$$(P1) \qquad \max\{\sum_{j=1}^{n} x_j : \; x_j^2 - x_j = 0, j = 1, \ldots, n; \; x_i x_j = 0 \; \forall i \neq j \in \Gamma\}$$

$$(P2) \; \min\{x_0 : \; \lambda_{\min}(A(x, x^l)) \geq 0, \sum_{j=1}^{m} a_j x_j^l = b^l, l = 1, \ldots, k, \; \sum_{j=1}^{m} x_j = 1\}$$

where $A(x, x^l)$ is affine in $x_0, x_1, \ldots, x_m, x_1^l, \ldots, x_m^l$.

♠ **(P1) easy formulation** but: is as **difficult** as an optimization problem can be! Worst case computational effort within absolute inaccuracy 0.5, for $n = 256$ is $2^{256} \approx 10^{77} \approx +\infty$!

♠ **(P2) complicated formulation** but: **easy to solve!** For $m = 100, k = 6 \implies$ 701 variables ($\approx$ 3 times larger) solved in less than 2 minutes for 6 digits accuracy!

convex (P2)[slow $\nearrow (n, \varepsilon)$] vs. nonconvex (P1) [very fast $\nearrow (n, \varepsilon)$]

A Bird's-Eye View of Classical and Modern Algorithms

A Generic Unconstrained Minimization Algorithm

$$(U) \quad \min\{f(x) : x \in \mathbb{R}^n\}$$

Start with $x \in \mathbb{R}^n$ such that $\nabla f(x) \neq 0$.

Compute new point $x^+ = x + td$ where

- $d \in \mathbb{R}^n$ is a *descent direction*: $\langle d, \nabla f(x) \rangle < 0$
- $t \in (0, +\infty)$ is a *stepsize*. How far to go in direction d such that for t small one guarantees

$$f(x^+) = f(x + td) < f(x)$$

Basic Gradient Iterative Schemes

$$\boxed{x^0 \in \mathbb{R}^n, \; x^{k+1} = x^k + t_k W^k d^k}$$

where

$$W^k \succ 0, \quad t_k \simeq \underset{t}{\text{argmin}}\, f(x^k + tW^k d^k)$$

- $W^k \equiv I, d^k \equiv -\nabla f(x^k)$, *Steepest Descent Method*;
  Slow but **Globally convergent**
- $W^k \equiv \nabla^2 f(x^k)^{-1}$, Newton's Method; **Fast** but **Locally convergent**
- Global Rate of convergence depends on information and topological properties of $\nabla f, \; \nabla^2 f$.

Three fundamental algorithms in applications which are gradient based

- Clustering: **The k-means algorithm**
- Neuro-computation: **The backpropagation (perceptron) algorithm**
- **The EM (Expectation-Maximization)** algorithm in statistical estimation

Constrained Optimization Algorithms

Richer but much more Difficult....

In most algorithms

- either we will solve a nonlinear system of equations and inequalities
- or we will have to solve a sequence of unconstrained minimization problems.
- Thus, the importance of having efficient linear algebra packages and a fast and reliable unconstrained routine.

Some Classes of Constrained Optimization Algorithms...

- Penalty and Barrier Methods
- Sequential Quadratic Programming
- Multiplier Methods
- Active set methods
- Dual Methods
- Interior point/primal-dual Methods
- ....and more...

Penalty Methods: Courant 1943, Ablow-Brigham 1955.

$$\text{(C)} \qquad \min\{f(x) : x \in S \subset \mathbb{R}^n\}$$

Idea: Replace (C) by **a family of unconstrained problems**

$$(C_t) \qquad \min_{x\in\mathbb{R}^n}\{f(x) + tP(x)\} \quad (t > 0)$$

Let

$$x(t) = \operatorname{argmin}\{f(x) + tP(x)\}$$

- $P(\cdot) \geq 0$ and $= 0$ if and only if x feasible.
  P is a **Penalty** we pay for constraints violation.
- For large t the minimum of (C_t) will be in a region where P is small.
  We thus expect that as $t \to \infty$:

$$\begin{aligned} tP(x(t)) &\to 0 \\ x(t) &\to x^* \quad \text{optimal solution of (C)} \end{aligned}$$

Examples of Penalty Functions

For Inequality Constraints $S = \{x : g_i(x) \leq 0, i = 1, \ldots, m\}$

$$P(x) = \sum_{i=1}^{m} \max(0, g_i(x)); \quad P(x) = \sum_{i=1}^{m} \max^2(0, g_i(x)) \leftarrow \textit{smooth}$$

For Equality Constraints $S = \{x : h_i(x) = 0, i = 1, \ldots, m\}$

$$P(x) = ||h(x)||^2, \quad h : \mathbb{R}^n \to \mathbb{R}^m$$

The Penalty Algorithm

Let $0 < t_k < t_{k+1}, \;\; \forall k$ with $c_k \to \infty$.

For each k solve $x_k = \text{argmin}_x\{f(x) + t_k P(x)\}$.

Convergence

Every limit point of $\{x_k\}$ is a solution of (C).

Barrier Methods: Frish 58, Fiacco-McCormick 68

Similar idea, but acting from the **interior** (for inequality constraints only!)

Let $S := \{x : g_i(x) \leq 0, i = 1, \dots, m\}$

Assume S has nonempty interior.

A **Barrier** function for S is a continuous function s.t.

$$B(x) \to \infty \ \text{ as } \ x \to \text{boundary} S$$

B is a barrier on bdy$S$ preventing leaving the feasible region. The constrained problem is replaced by the unconstrained

$$x(\varepsilon) = \text{argmin}\{f(x) + \varepsilon B(x)\} \in \text{int} S$$

Examples:

$$B(x) = -\sum_{i=1}^{m} \frac{1}{g_i(x)}, \ B(x) = -\sum_{i=1}^{m} \log(-g_i(x))$$

Barrier Algorithm

Let $0 < \varepsilon_{k+1} < \varepsilon_k \ \ \forall k$ with $\varepsilon_k \to 0$.

For each k solve

$$x_k = \text{argmin}_x \{f(x) + \varepsilon_k B(x)\}.$$

Convergence Every limit point of $\{x_k\}$ is a solution of (C).

In both Penalty/Barrier Methods:Compromise

- $t(\varepsilon)$ must be chosen sufficiently large (small) so that $x(t)(x(\varepsilon))$ will approach S from the exterior (interior).
- **BUT**, if $t(\varepsilon)$ is chosen too large (small), then *Ill-Conditionning* may occurs.

Avoid IC, do not send $t \to \infty, \varepsilon \to 0$.

.....use augmented Lagrangian/Multiplier methods.....

A Basic Multiplier Method for Equality Constraints

$$\min\{f(x):\ h(x)=0\}\quad h:\mathbb{R}^n\rightarrow\mathbb{R}^m$$

Lagrangian: $L(x,u)=f(x)+u^T h(x)$

Augmented L: $A(x,u,c))=L(x,u)+2^{-1}c||h(x)||^2$

AL = Penalized Lagrangian

Multiplier Method Given $\{u^k, c^k\}$

1. Find $x^{k+1} = \text{argmin}\{A(x,u^k,c^k):\ x\in\mathbb{R}^n\}$

2. Update Rule: $u^{k+1}=u^k+c^k h(x^{k+1})$

3. Increase $c^k > 0$ if necessary.

Features of Multipliers Method

- A key Advantage: **it is not** necessary to increase c^k to ∞, for convergence (as opposed to ”Penalty/Barrier method”)
- As a result, A is ”less subject to ill-conditionning”, and more ”robust”.
- The AL depends on c **but also** on the *dual* multiplier u : faster convergence can be expected (rather than keeping u constant)
- Extendible to inequality constrained problems

Multiplier Methods for Inequality Constrained Problems

$$(C) \quad \min\{f(x):\ g_i(x) \leq 0, i = 1, \ldots, m\}, \quad g := (g_1, \ldots, g_m)^T$$

Quadratic Method of Multipliers

$$\begin{aligned} x^{k+1} &\in \operatorname{argmin}\{L(x, u^k, c^k):\ x \in \mathbb{R}^n\} \\ u^{k+1} &= (u^k + c^k g(x^{k+1}))_+, \quad (c^k > 0) \end{aligned}$$

with $z_+ := \max\{0, z\}$, (componentwise)

$$L(x, u, c) := f(x) + (2c)^{-1}\{||(u + cg(x))_+||^2 - ||u||^2\}$$

More recent and modern approaches allow for constructing **smooth Lagrangians** so that Newton's method can be applied for the unconstrained minimization.

Interior Point Methods

Idea goes back to Barrier Methods, but within a different methodology, eliminating the ill-conditioning drawback.

Basically the idea is to approximately follow the *central path* generated within the interior of the corresponding feasible set.

Computation of Central Path

$$x^*(\mu) = \underset{x}{\text{argmin}}\{\mu\langle c, x\rangle + S(x)\}$$

Where S is a ***Self-Concordant Barrier*** for the feasible set of the given optimization problem .

- $x^*(\mu)$ remains strictly feasible for every $\mu > 0$
- $x^*(\mu) \to x^*$ optimal for $\mu \to \infty$
- Can be computed in polynomial time with Newton method

This relies on the fundamental theory of Selconcordance developed by Nesterov-Nemirovsky (1990)s. [Idea: to make the convergence analysis *coordinate invariant*]

Interior Point Methods for SC-Convex Problems

For self-concordant convex problems

- IPM can be proven to be **polynomially solvable** for a prescribed accuracy ϵ.
- Worst case complexity: # Newton steps $\leq$ square root of problem size
- **Each iteration requires forming gradient, Hessian and solving a linear system**

Mathematical and Computational Challenges

- Convex problems appears in applications more than we (use to) think
- Convex optimization can be used to *approximate* (finding bounds) hard problems
- Convex problems can be solved efficiently, namely with polynomial time algorithms

.......BUT.........

- Polynomial algorithms are highly sophisticated and require informations on the Hessians of objective and constraints, often not available.
- Require heavy computational cost at **each iteration**
- For large scale problems with no particular structures, ... even **ONE ITERATION** cannot be completed...!

Challenge: to solve very large scale optimization problems emerging from applied world,
keeping in mind the trade off between
Efficiency versus Practicality

Thus the needs to

- further study potential **direct/simple methods** (e.g., first order methods, using function or/and gradient infos only).
- Produce **faster algorithms** within these methods

Conclusion

Optimizers are not (yet!) out of job......

Thank you for listening!