Advanced Methods for Sequence Analysis

G. Rätsch[1], C.S. Ong[1,2] and P. Philips[1]

[1] Friedrich Miescher Laboratory, Tübingen
[2] Max Planck Institute for Biological Cybernetics, Tübingen

Vorlesung WS 2006/2007
Eberhard Karls Universität Tübingen

10 January 2007

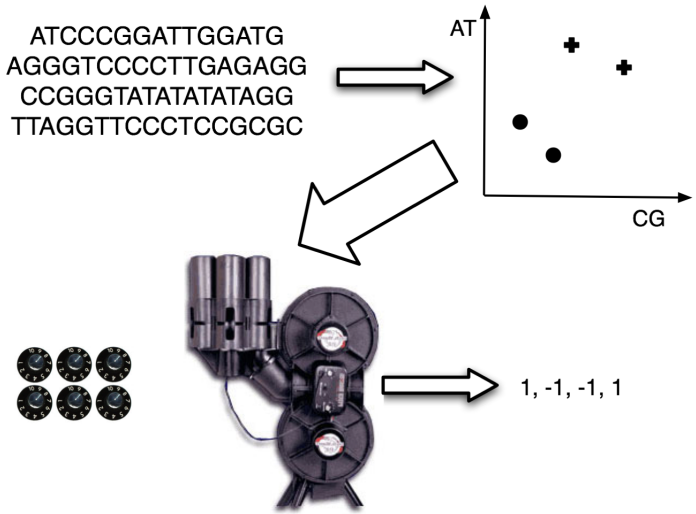
http://www.fml.mpg.de/raetsch/lectures/amsa

Convex Optimization

- Machine Learning as Numerical Optimization
- Unconstrained Optimization
  - Gradient Descent
  - Newton Method
- Constrained Optimization
  - Some History
  - Convex Functions and Convex Sets
  - Common Problem Formulations
  - KKT conditions
- http://www.stanford.edu/~boyd/cvxbook/

Recall the SVM

1. How are examples represented?
2. How are labels represented?
3. What are the inputs to the SVM?
4. What does SVM training output?



Recall the SVM

1. How are examples represented?
   - By the kernel matrix K.
2. How are labels represented?
   - By the label vector y.
3. What are the inputs to the SVM?
   - K, y
4. What does SVM training output?
   - α

Recall: Representer Theorem

$$K\alpha = y$$

Linear Equations

Motivation

$$K\alpha = y$$

Linear Algebra Golub and van Loan [1996]

- Gaussian Elimination
- LU Factorization
- QR Factorization
- Eigenvalue methods
- Lanczos methods
- Conjugate gradient methods

Variational Formulation

Motivation

Solve a scalar function instead of a matrix equation.

Objective

$$\min_{\alpha} \frac{1}{2}\alpha^\top K\alpha - y^\top \alpha$$

The gradient at optimality

$$K\alpha - y = 0$$

gives the original problem.

Observations

- Second derivavtive of objective is K.
- K is positive semidefinite hence the problem is convex.

Gradient Descent

Gradient

For a multivariate function $f : \mathbb{R}^n \to \mathbb{R}$, define the gradient of f to be

$$\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}, \ldots, \frac{\partial f(x)}{\partial x_n}, \right).$$

Algorithm Schölkopf and Smola [2002]

For an intial value x_0 and precision ε,

$k = 0$

while $(\|\nabla f(x_k)\| \geqslant \varepsilon)$

Compute $g = \nabla f(x_k)$

Perform line search on $f(x_k - \gamma g)$ for optimal γ

$x_{k+1} = x_k - \gamma g$

$k = k + 1$

Newton Method (1)

Hessian

For a multivariate function $f : \mathbb{R}^n \to \mathbb{R}$,

$$\nabla^2 f(x)_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}, \quad i, j = 1, \ldots, n$$

Motivation

The second order Taylor approximation $\hat{f}$ of f at x is

$$\hat{f}(x+v) = f(x) + \nabla f(x)^\top v + \frac{1}{2} v^\top \nabla^2 f(x) v,$$

which is a convex quadratic function of v, with minimizer

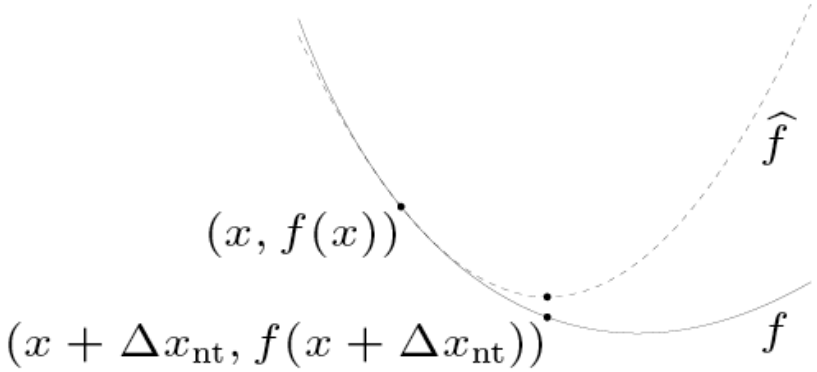
$$v^* = -\nabla^2 f(x)^{-1} \nabla f(x).$$

Newton step

The vector v^* above is called the Newton step for f at x.

Newton Method (2)

Motivation



Algorithm

For an intial value x_0 and precision ε,

$k = 0$

while $(\|\nabla f(x_k)\| \geqslant \varepsilon)$

$x_{k+1} = x_k - \nabla^2 f(x)^{-1} \nabla f(x)$

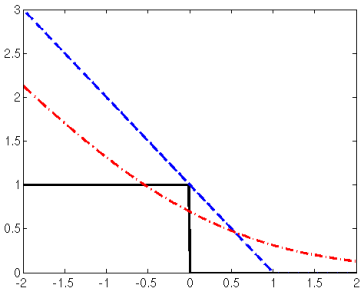
$k = k + 1$

Non-smooth functions

Motivation

Non differentiable functions f have to be treated carefully. For example the hinge loss

$$\ell(f(\mathbf{x}_i), y_i) := \max\{0, 1 - y_i f(\mathbf{x}_i)\}.$$



Piecewise differentiable

For piecewise differentiable functions, we can treat each piece individually, and then we can apply the above methods.

Constrained Optimization

$$\begin{aligned} \min_x \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leqslant 0 \text{ for all } i \\ & g_j(x) = 0 \text{ for all } j \end{aligned}$$

- $x \in \mathbb{R}^n$ is the optimization variable
- $f_0 : \mathbb{R}^n \to \mathbb{R}$ is the objective or cost function
- $f_i : \mathbb{R}^n \to \mathbb{R}$ are inequality constraint functions
- $g_j : \mathbb{R}^n \to \mathbb{R}$ are equality constraint functions

Some History

Theory (Convex Analysis) ca. 1900–1970

Algorithms

- 1947: simplex algorithm for linear programming (Dantzig)
- 1960s: early interior-point methods (Fiacco & McCormick, Dikin, . . .)
- 1970s: ellipsoid method and other subgradient methods
- 1980s: polynomial-time interior-point methods for linear programming (Karmarkar 1984)
- late 1980s-now: polynomial-time interior-point methods for nonlinear convex optimization (Nesterov & Nemirovski 1994)

Convex Set

line segment between x_1 and x_2: all points

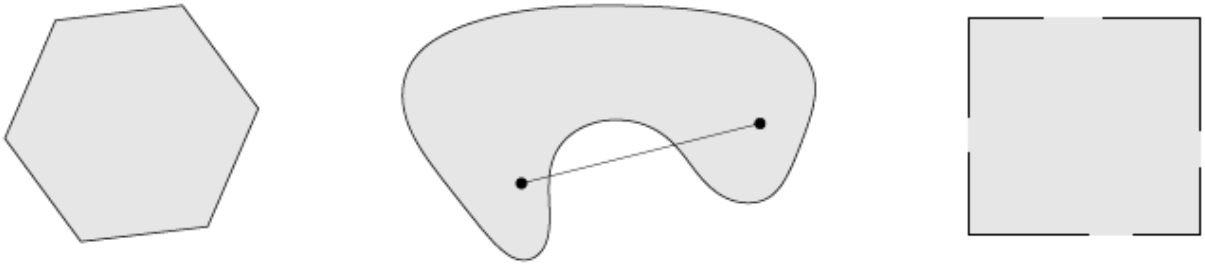
$$x = \theta x_1 + (1 - \theta) x_2$$

with $0 \leqslant \theta \leqslant 1$.

convex set : contains line seqment between any two points in the set

$$x_1, x_2 \in \mathcal{C}, 0 \leqslant \theta \leqslant 1 \Longrightarrow x = \theta x_1 + (1 - \theta) x_2 \in \mathcal{C}.$$

Examples (one convex, two non-convex sets)



How to check

Use definition

To show that a set $\mathcal{C}$ is convex, show that $\mathcal{C}$ is obtained from simple convex sets (where we establish convexity by the definition).

Applying operations that preserve convexity

- intersection
- affine functions
- perspective functions
- linear-fractional functions

Convex Function

$f : \mathbb{R}^n \to \mathbb{R}$ is convex if the domain of f is a convex set and

$$f(\theta x_1 + (1-\theta)x_2) \leqslant \theta f(x_1) + (1-\theta) f(x_2)$$

for all x_1, x_2 in the domain of f, $0 \leqslant \theta \leqslant 1$.

- affine: $ax + b$ on $\mathbb{R}$, for any $a, b \in \mathbb{R}$.
- affine: $a^\top x + b$ on $\mathbb{R}^n$, for any $a \in \mathbb{R}^n, b \in \mathbb{R}$.
- exponential: $\exp^{ax}$ for any $a \in \mathbb{R}$.
- powers: x^a on $\mathbb{R}_{++}$, for $a \geqslant 1$ or $a \leqslant 0$.
- powers of absolute value: $|x|^a$ on $\mathbb{R}$, for $a \geqslant 1$.
- negative entropy: $x \log x$ on $\mathbb{R}_{++}$.
- norms: $\|x\|_p = (\sum_{i=1}^n |x_i|^p)^{\frac{1}{p}}$ for $p \geqslant 1$.

How to check (1)

Restrict to line

$f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if the function $g : \mathbb{R} \to \mathbb{R}$,

$$g(t) = f(x + tv), \quad \mathbf{dom}\ g = \{t | x + tv \in \mathbf{dom}\ f\}$$

is convex (in t) for any $x \in \mathbf{dom}\ f, v \in \mathbb{R}^n$.

First-order condition

The first order approximation of f is a global underestimator.

$$\nabla f(x) = \left(\frac{\partial f(x)}{\partial x_1}, \frac{\partial f(x)}{\partial x_2}, \ldots, \frac{\partial f(x)}{\partial x_n}, \right)$$

f with convex domain is convex if and only if

$$f(x_2) \geqslant f(x_1) + \nabla f(x_1)^\top (x_2 - x_1)$$

for all $x_1, x_2 \in \mathbf{dom}\ f$.

How to check (2)

Second-order condition

The hessian is positive semi-definite.

$$\nabla^2 f(x)_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}, \quad i, j = 1, \ldots, n$$

f is convex if and only if

$$\nabla^2 f(x) \succeq 0 \text{ for all } x \in \mathbf{dom}\ f.$$

How to check (3)

Show that f is obtained from simple convex functions by operations that preserve convexity

- nonnegative weighted sum
- composition with affine function
- pointwise maximum and supremum
- composition
- minimization
- perspective

Convex Optimization

Constrained Optimization (generally hard)

$$\begin{aligned} \min_x \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leqslant 0 \text{ for all } i \\ & g_j(x) = 0 \text{ for all } j \end{aligned}$$

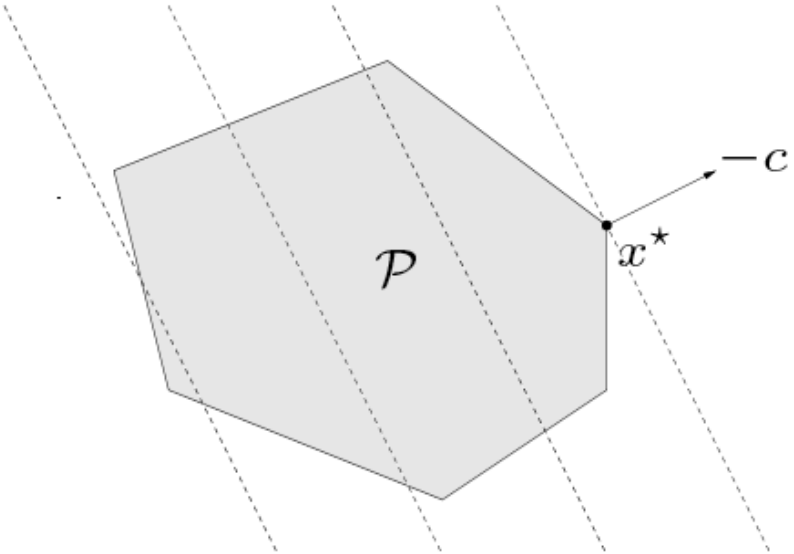
Convex Optimization (generally easy)

$$\begin{aligned} \min_x \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leqslant 0 \text{ for all } i \\ & a_j^\top x = b_j \text{ for all } j \end{aligned}$$

$f_0, f_1, \ldots, f_m$ are convex, and the equality constraints are affine Boyd and Vandenberghe [2004].

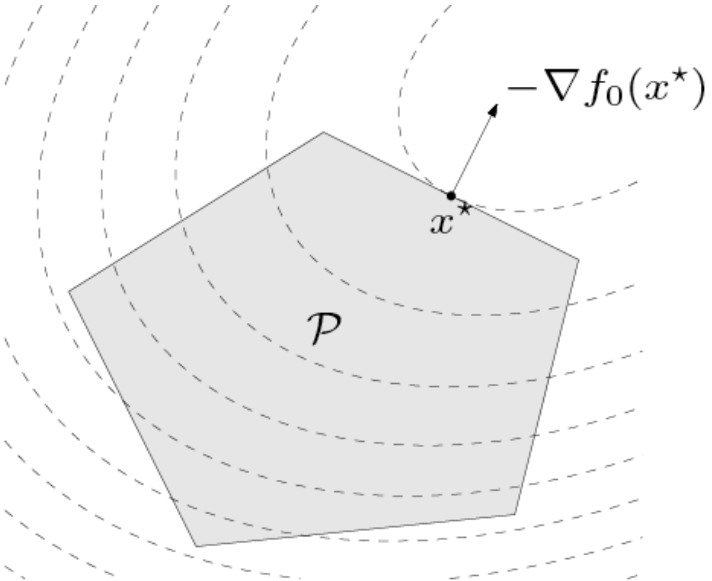
Linear Program

$$\begin{aligned} \min_x \quad & c^\top x + d \\ \text{subject to} \quad & Gx \leqslant h \\ & Ax = b \end{aligned}$$



Quadratic Program

$$\begin{aligned} \min_x \quad & \tfrac{1}{2} x^\top P x + q^\top x + r \\ \text{subject to} \quad & Gx \leqslant h \\ & Ax = b \end{aligned}$$



Some other programs

Quadratically constrained quadratic program

$$\begin{aligned} \min_x \quad & \tfrac{1}{2}x^\top P_0 x + q_0^\top x + r_0 \\ \text{subject to} \quad & \tfrac{1}{2}x^\top P_0 x + q_0^\top x + r_0 \leqslant 0 \text{ for all } i \\ & Ax = b \end{aligned}$$

Second order cone program

$$\begin{aligned} \min_x \quad & f^\top x \\ \text{subject to} \quad & \|A_i x + b_i\|_2 \leqslant c_i^\top x + d_i \text{ for all } i \\ & Fx = g \end{aligned}$$

Semidefinite program

$$\begin{aligned} \min_x \quad & c^\top x \\ \text{subject to} \quad & x_1 F_1 + x_2 F_2 + \ldots + x_n F_n + G \preceq 0 \\ & Ax = b \end{aligned}$$

Problem reformulations

- equivalent formulations of a problem can lead to very different duals
- reformulating the primal problem can be useful when the dual is difficult to derive, or uninteresting.

Common reformulations

- introduce new variables and equality constraints
- make explicit constraints implicit and vice versa
- transform objective or constraint functions

Equivalent convex problems (1)

Two problems are (informally) equivalent if the solution of one is readily obtained from the solution of the other, and vice-versa.

Eliminating equality constraints

$$\begin{aligned} \min_x \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leqslant 0 \text{ for all } i \\ & Ax = b \end{aligned}$$

is equivalent to

$$\begin{aligned} \min_z \quad & f_0(Fz + x_0) \\ \text{subject to} \quad & f_i(Fz + x_0) \leqslant 0 \text{ for all } i \\ & Ax = b \end{aligned}$$

where F and x_0 are such that $Ax = b \Leftrightarrow x = Fz + x_0$ for some z.

Equivalent convex problems (2)

Introducing Equality Constraints

$$\begin{aligned} \min_x \quad & f_0(A_0 x + b) \\ \text{subject to} \quad & f_i(A_i x + b) \text{ for all } i \end{aligned}$$

is equivalent to

$$\begin{aligned} \min_{x,y_i} \quad & f_0(y_0) \\ \text{subject to} \quad & f_i(y_i) \leqslant 0, \text{ for all } i \\ & y_i = A_i x + b_i \end{aligned}$$

Equivalent convex problems (3)

Introducing slack variables

$$\begin{aligned} \min_x \quad & f_0(x) \\ \text{subject to} \quad & a_i^\top x \leqslant b_i \text{ for all } i \end{aligned}$$

is equivalent to

$$\begin{aligned} \min_{x,s} \quad & f_0(x) \\ \text{subject to} \quad & a_i^\top x + s_i = b_i \text{ for all } i \\ & s_i \geqslant 0 \end{aligned}$$

Equivalent convex problems (4)

Epigraph form

$$\begin{aligned} \min_x \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leqslant 0 \text{ for all } i \\ & Ax = b \end{aligned}$$

is equivalent to

$$\begin{aligned} \min_{x,t} \quad & t \\ \text{subject to} \quad & f_0(x) - t \leqslant 0 \\ & f_i(x) \leqslant 0 \text{ for all } i \\ Ax = b & \end{aligned}$$

Equivalent convex problems (5)

Minimizing over some variables

$$\begin{array}{rl} \min_{x_1,x_2} & f_0(x_1,x_2) \\ \text{subject to} & f_i(x_1) \leqslant 0 \text{ for all } i \end{array}$$

is equivalent to

$$\begin{array}{rl} \min_{x_1} & \tilde{f}_0(x_1) \\ \text{subject to} & f_i(x_1) \leqslant 0 \text{ for all } i \end{array}$$

where $\tilde{f}_0(x_1) = inf_{x_2} f_0(x_1,x_2)$.

Lagrange Duality

Lagrangian

$\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ with

$$\mathcal{L}(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{p} \nu_j g_j(x).$$

- weighted sum of objective and constraint functions
- λ_i is the Lagrange multiplier associated with $f_i(x) \leqslant 0$.
- ν_j is the Lagrange multiplier associated with $g_j(x) = 0$.

Dual function

Lagrange dual function

$h : \mathbb{R}^m \times \mathbb{R}^p \longrightarrow \mathbb{R}$,

$$h(\lambda, \nu) = \inf_x \mathcal{L}(x, \lambda, \nu).$$

h is concave.

Lagrange dual problem

$$\begin{aligned} &\max_{\lambda,\nu} \; h(\lambda, \nu) \\ &\text{subject to} \; \lambda \geqslant 0 \end{aligned}$$

Checking for Optimality

If x, λ, ν satisfy the KKT conditions for a convex problem, then they are optimal.
The following four conditions are called the Karush-Kuhn-Tucker (KKT) conditions:

- primal constraints: $f_i(x) \leqslant 0, g_i(x) = 0$
- dual constraints: $\lambda \geqslant 0$
- complementary slackness: $\lambda_i f_i(x) = 0$
- gradient of Lagrangian with respect to x vanishes:

$$\nabla f_0(x) + \sum_i \lambda_i \nabla f_i(x) + \sum_j \nu \nabla g_j(x) = 0$$

Summary

- Machine Learning as Numerical Optimization
- Unconstrained Optimization
  - Gradient Descent
  - Newton Method
- Constrained Optimization
  - Convex Functions and Convex Sets
  - Common Problem Formulations
  - KKT conditions
- http://www.stanford.edu/~boyd/cvxbook/

References

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Gene H. Golub and Charles F. van Loan. *Matrix Computations*. Johns Hopkins, 3rd edition, 1996.

B. Schölkopf and A.J. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.