Data Science - Convex optimization and application

Summary

We begin by some illustrations in challenging topics in modern data science. Then, this session introduces (or reminds) some basics on optimization, and illustrate some key applications in supervised classification.

1 Data Science

1.1 What is data science :

Extract from data some knowledge for industrial or academic exploitation. It generally involves :

1. Signal Processing (how to record the data and represent it ?)
2. Modelisation (What is the problem, what kind of mathematical model and answer ?)
3. Statistics (reliability of estimation procedures ?)
4. Machine Learning (what kind of efficient optimization algorithm ?)
5. Implementation (software needs)
6. Visualization (how can I represent the resulting knowledge ?)

In its whole, this sequence of questions are at the core of Artificial Intelligence and may also be referred to as Computer Science problems. In this lecture, we will address some issues raised in red items. Each time, practical examples will be provided

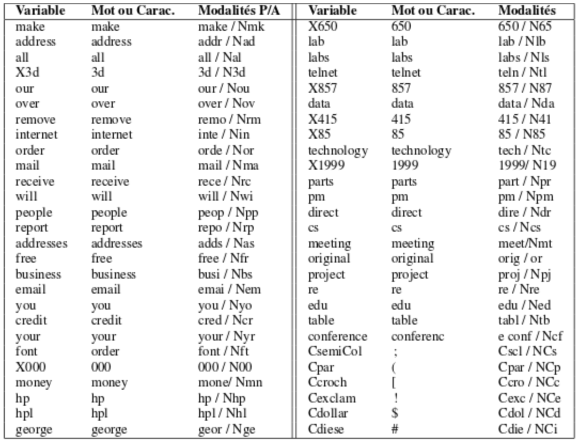
Most of our motivation comes from the *Big Data* world, encountered in image processing, finance, genetics and many other fields where knowledge extraction is needed, when facing many observations described by many variables.

n : number of observations - p : number of variables per observations

$$p >> n >> O(1).$$

1.2 Several examples

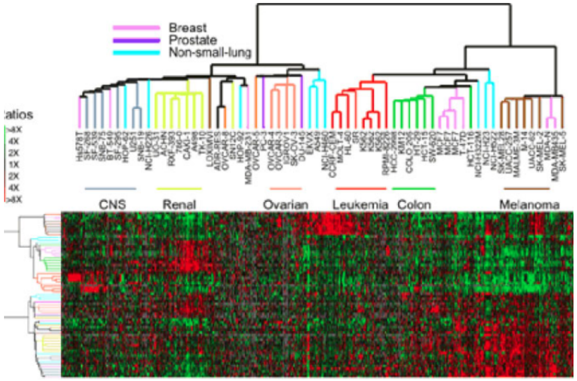
Spam detection From a set of labelled messages (spam or not), build a classification for automatic spam rejection.



Variable	Mot ou Carac.	Modalités P/A	Variable	Mot ou Carac.	Modalités
make	make	make / Nmk	X650	650	650 / N65
address	address	addr / Nad	lab	lab	lab / Nlb
all	all	all / Nal	labs	labs	labs / Nls
X3d	3d	3d / N3d	telnet	telnet	teln / Ntl
our	our	our / Nou	X857	857	857 / N87
over	over	over / Nov	data	data	data / Nda
remove	remove	remo / Nrm	X415	415	415 / N41
internet	internet	inte / Nin	X85	85	85 / N85
order	order	orde / Nor	technology	technology	tech / Ntc
mail	mail	mail / Nma	X1999	1999	1999/ N19
receive	receive	rece / Nrc	parts	parts	part / Npr
will	will	will / Nwi	pm	pm	pm / Npm
people	people	peop / Npp	direct	direct	dire / Ndr
report	report	repo / Nrp	cs	cs	cs / Ncs
addresses	addresses	adds / Nas	meeting	meeting	meet/Nmt
free	free	free / Nfr	original	original	orig / or
business	business	busi / Nbs	project	project	proj / Npj
email	email	emai / Nem	re	re	re / Nre
you	you	you / Nyo	edu	edu	edu / Ned
credit	credit	cred / Ncr	table	table	tabl / Ntb
your	your	your / Nyr	conference	conferenc	e conf / Ncf
font	order	font / Nft	CsemiCol	;	Cscl / NCs
X000	000	000 / N00	Cpar	(	Cpar / NCp
money	money	mone/ Nmn	Ccroch	[	Ccro / NCc
hp	hp	hp / Nhp	Cexclam	!	Cexc / NCe
hpl	hpl	hpl / Nhl	Cdollar	$	Cdol / NCd
george	george	geor / Nge	Cdiese	#	Cdie / NCi

- Select among the words meaningful elements ?
- Automatic classification ?

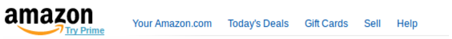
Gene expression profiles analysis One measures micro-array datasets built from a huge amount of profile genes expression. Number of genes p (of order thousands). Number of samples n (of order hundred).



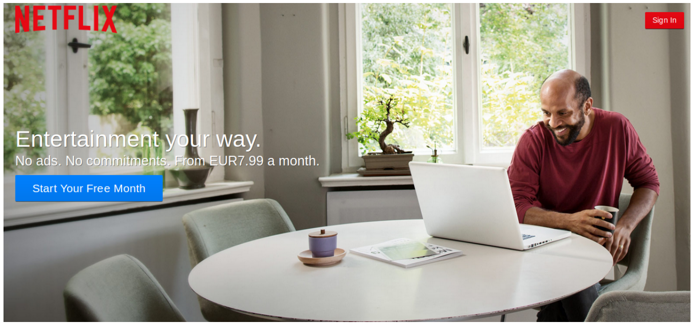
Diagnostic help : healthy or ill ?

- Select among the genes meaningful elements ?
- Automatic classification ?

Recommandation problems



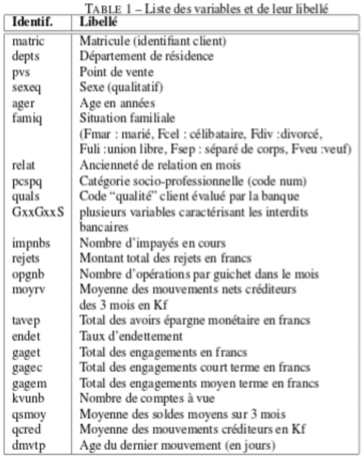
And more recently :



- What kind of database ?
- Reliable recommandation for clients ?
- Online strategy ?

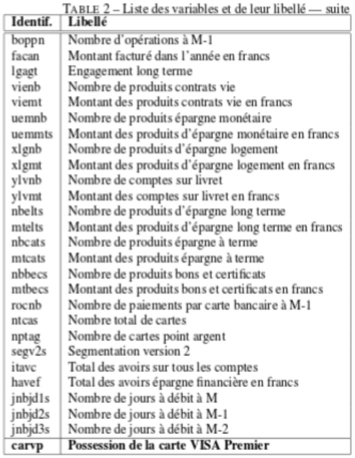
Credit scoring Build an indicator (Q score) from a dataset for the probability of interest in a financial product (Visa premier credit card).

TABLE 1 – Liste des variables et de leur libellé



Identif.	**Libellé**
matric	Matricule (identifiant client)
depts	Département de résidence
pvs	Point de vente
sexeq	Sexe (qualitatif)
ager	Age en années
famiq	Situation familiale (Fmar : marié, Fcel : célibataire, Fdiv :divorcé, Fuli :union libre, Fsep : séparé de corps, Fveu :veuf)
relat	Ancienneté de relation en mois
pcspq	Catégorie socio-professionnelle (code num)
quals	Code “qualité” client évalué par la banque
GxxGxxS	plusieurs variables caractérisant les interdits bancaires
impnbs	Nombre d'impayés en cours
rejets	Montant total des rejets en francs
opgnb	Nombre d'opérations par guichet dans le mois
moyrv	Moyenne des mouvements nets créditeurs des 3 mois en Kf
tavep	Total des avoirs épargne monétaire en francs
endet	Taux d'endettement
gaget	Total des engagements en francs
gagec	Total des engagements court terme en francs
gagem	Total des engagements moyen terme en francs
kvunb	Nombre de comptes à vue
qsmoy	Moyenne des soldes moyens sur 3 mois
qcred	Moyenne des mouvements créditeurs en Kf
dmvtp	Age du dernier mouvement (en jours)

TABLE 2 – Liste des variables et de leur libellé — suite



Identif.	**Libellé**
boppn	Nombre d'opérations à M-1
facan	Montant facturé dans l'année en francs
lgagt	Engagement long terme
vienb	Nombre de produits contrats vie
viemt	Montant des produits contrats vie en francs
uemnb	Nombre de produits épargne monétaire
uemmts	Montant des produits d'épargne monétaire en francs
xlgnb	Nombre de produits d'épargne logement
xlgmt	Montant des produits d'épargne logement en francs
ylvnb	Nombre de comptes sur livret
ylvmt	Montant des comptes sur livret en francs
nbelts	Nombre de produits d'épargne long terme
mtelts	Montant des produits d'épargne long terme en francs
nbcats	Nombre de produits épargne à terme
mtcats	Montant des produits épargne à terme
nbbecs	Nombre de produits bons et certificats
mtbecs	Montant des produits bons et certificats en francs
rocnb	Nombre de paiements par carte bancaire à M-1
ntcas	Nombre total de cartes
nptag	Nombre de cartes point argent
segv2s	Segmentation version 2
itavc	Total des avoirs sur tous les comptes
havef	Total des avoirs épargne financière en francs
jnbjd1s	Nombre de jours à débit à M
jnbjd2s	Nombre de jours à débit à M-1
jnbjd3s	Nombre de jours à débit à M-2
carvp	**Possession de la carte VISA Premier**

1. Define a model, a question ?
2. Use a supervised classification algorithm to rank the best clients.
3. Use logistic regression to provide a score.

1.3 What about maths ?

Various mathematical fields we will talk about :

- Analysis : Convex optimization, Approximation theory
- Statistics : Penalized procedures and their reliability
- Probabilistic methods : concentration inequalities, stochastic processes, stochastic approximations

Famous keywords :

- Lasso
- Boosting
- Convex relaxation
- Supervised classification
- Support Vector Machine

- Aggregation rules
- Gradient descent
- Stochastic Gradient descent
- Sequential prediction
- Bandit games, minimax policies
- Matrix completion

In this session : We will slightly talk about optimization, that are mainly convex in our statistical worl. Non-convex problems are also very interesting even though much more difficult to deal with from a numerical point of view.

2 Standard Convex optimisation procedures

2.1 Convex functions

We recall some background material that is necessary for a clear understanding of how some machine learning procedures work. We will cover some basic relationships between convexity, positive semidefiniteness, local and global minimizers.

DÉFINITION 1. — *[Convex sets, convex functions] A set D is convex if and only if for any $(x_1, x_2) \in D^2$ and all $\alpha \in [0,1]$,*

$$x = \alpha x_1 + (1-\alpha)x_2 \in D.$$

A function f is convex if

- *its domain D is convex*
- $f(x) = f(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f(x_1) + (1-\alpha) f(x_2)$.

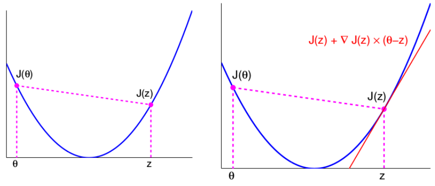
DÉFINITION 2. — *[Positive Semi Definite matrix (PSD)] A $p \times p$ matrix H is (PSD) if for all $p \times 1$ vectors z, we have $z^t H z \geq 0$.*

There exists a strong link between SDP matrix and convex functions, given by the following proposition.

PROPOSITION 3. — *A smooth $\mathcal{C}^2(D)$ function f is convex if and only if $D^2 f$ is SDP at any point of D.*

The proof follows easily from a second order Taylor expansion.

2.2 Example of convex functions



- Exponential function : $\theta \in \mathbb{R} \longmapsto \exp(a\theta)$ on $\mathbb{R}$ whatever a is.
- Affine function : $\theta \in \mathbb{R}^d \longmapsto a^t\theta + b$
- Entropy function : $\theta \in \mathbb{R}_+ \longmapsto -\theta \log(\theta)$
- p-norm : $\theta \in \mathbb{R}^d \longmapsto |\theta\|_p := \sqrt[p]{\sum_{i=1}^{d} \|\theta_i|^p}$ with $p \geq 1$.
- Quadratic form : $\theta \in \mathbb{R}^d \longmapsto \theta^t P\theta + 2q^t\theta + r$ where P is symetric and positive.

2.3 Why such an interest in convexity ?

From external motivations :

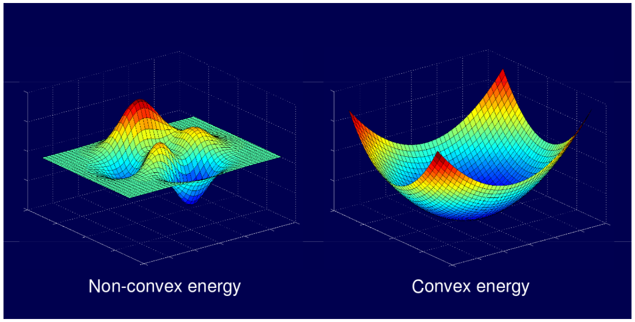
- Many problems in machine learning come from the minimization of a convex criterion and provide meaningful results for the statistical initial task.
- Many optimization problems admit a convex reformulation (SVM classification or regression, LASSO regression, ridge regression, permutation recovery, . . .).

From a numerical point of view :

- Local minimizer = global minimizer. It is a powerful point since in general, descent methods involve $\nabla f(x)$ (or something related to), which is a local information on f.
- x is a local (global) minimizer of f if and only if $0 \in \partial f(x)$.
- Many fast algorithms for the optimization of convex function exist, and sometimes are independent on the dimension d of the original space.

2.4 Why convexity is powerful ?

Two kinds of optimization problems :



- On the left : non convex optimization problem, use of Travelling Salesman type method. Greedy exploration step (simulated annealing, genetic algortihms).
- On the right : convex optimization problem, use local descent methods with gradients or subgradients.

DÉFINITION 4. — *[Subgradient (nonsmooth functions ?)] For any function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$, and any x in $\mathbb{R}^d$, the subgradient $\partial f(x)$ is the set of all vectors g such that*

$$f(x) - f(y) \leq \langle g, x - y \rangle.$$

This set of subgradients may be empty. Fortunately, it is not the case for convex functions.

PROPOSITION 5. — *$f : \mathbb{R}^d \longrightarrow \mathbb{R}$ is convex if and only if $\partial f(x) \neq \emptyset$ for any x of $\mathbb{R}^d$.*

3 Gradient descent method

3.1 Projected descent

In either constrained or unconstrained problems, descent methods are powerful with convex functions. In particular, consider constrained problems in $\mathcal{X} \subset \mathbb{R}^d$. The most famous local descent method relies on

$$y_{t+1} = x_t - \eta g_t \qquad \text{where} \qquad g_t \in \partial f(x_t),$$

and

$$x_{t+1} = \Pi_{\mathcal{X}}(y_{t+1}),$$

where $\eta > 0$ is a fixed step-size parameters.

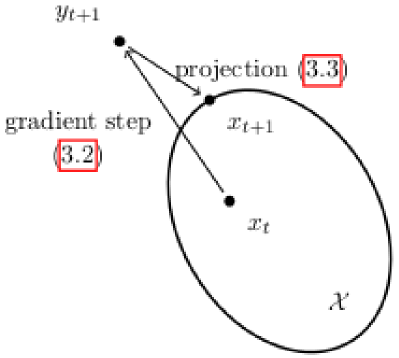


Fig. 3.2 Illustration of the Projected Subgradient Descent method.

THÉORÈME 6. — *[Convergence of the projected gradient descent method, fixed step-size] If f is convex over $\mathcal{X}$ with $\mathcal{X} \subset B(0, R)$ and $\|\partial f\|_\infty \leq L$, the choice $\eta = \frac{R}{L\sqrt{t}}$ leads to*

$$f\left(\frac{1}{t}\sum_{s=1}^{t} x_s\right) - \min f \leq \frac{RL}{\sqrt{t}}$$

3.2 Smooth unconstrained case

Results can be seriously improved with smooth functions with bounded second derivatives.

DÉFINITION 7. — *f is β smooth if ∇f is β Lipschitz :*

$$\|\nabla f(x) - \nabla f(y)\| \le \beta \|x - y\|.$$

Standard gradient descent over $\mathbb{R}^d$ becomes

$$x_{t+1} = x_t - \eta \nabla f(x_t),$$

THÉORÈME 8. — *[Convergence of the gradient descent method, β smooth function] If f is a convex and β-smooth function, then $\eta = \frac{1}{\beta}$ leads to*

$$f\left(\frac{1}{t}\sum_{s=1}^{t} x_s\right) - \min f \le \frac{2\beta\|x_1 - x_0\|^2}{t-1}$$

Remarque. —

- Note that the two past results do not depend on the dimension of the state space d.
- The last result can be extended to the constrained situation.

3.3 Constrained case

Elements of the problem :

- θ unknown vector of $\mathbb{R}^d$ to be recovered
- $J : \mathbb{R}^d \mapsto \mathbb{R}$ function to be minimized
- f_i and g_i differentiable functions defining a set of constraints.

Definition of the problem :

- $\min_{\theta \in \mathbb{R}^d} J(\theta)$ such that :
- $f_i(\theta) = 0, \forall i = 1, \ldots, n$ and $g_i(\theta) \le 0, \forall i = 1, \ldots, m$

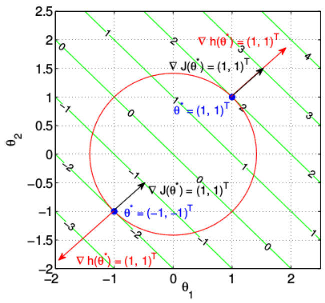
Set of admissible vectors :

$$\Omega := \left\{\theta \in \mathbb{R}^d \,|\, f_i(\theta) = 0, \forall i \text{ and } g_j(\theta) \le 0, \forall j\right\}$$

Typical situation :

Exemple

$$\min_{\theta \in \mathbb{R}^2} \quad \theta_1 + \theta_2$$
$$\text{s.c.} \quad \theta_1^2 + \theta_2^2 - 2 = 0$$



Ω : circle of radius $\sqrt{2}$

Optimal solution : $\theta^\star = (-1, -1)^t$ and $J(\theta^\star) = -2$.

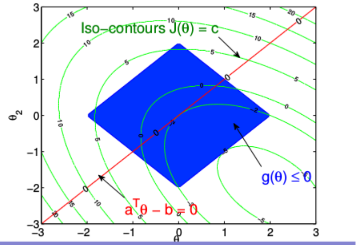
Important restriction : we will restrict our study to convex functions J.

DÉFINITION 9. — *A constrained problem is convex iff*

- *J is a convex function*
- *f_i are linear or affine functions and g_i are convex functions*

Exemple

$$\min_{\theta \in \mathbb{R}^2} \quad \theta_1^2 + \theta_2^2 + \theta_1\theta_2$$
$$-2\theta_1 + 2\theta_2 - 2$$
$$\text{s.c.} \quad \theta_1 - \theta_2 = 0$$
$$\|\theta\|_1 - 2 \le 0$$



Example

$$\min_{\theta} J(\theta) \qquad \text{such that} \qquad a^t\theta - b = 0$$

- Descent direction h : $\nabla J(\theta)^t h < 0$.
- Admissible direction h : $a^t(\theta + h) - b = 0 \iff a^t h = 0$.

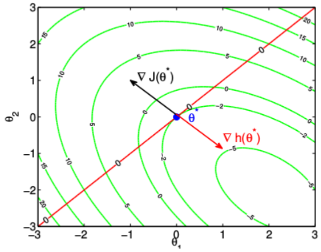
Optimality θ^* is *optimal if there is no admissible descent direction starting from θ^**. The only possible case is when $\nabla J(\theta^*)$ and a are linearly dependent :

$$\exists \lambda \in \mathbb{R} \qquad \nabla J(\theta^*) + \lambda a = 0.$$

Exemple

$$\min_{\theta\in\mathbb{R}^2} \quad \theta_1^2+\theta_2^2+\theta_1\theta_2 -2\theta_1+2\theta_2-2$$

$$\text{s.c.} \quad \theta_1-\theta_2=0$$



In this situation :

$$\nabla J(\theta) = \begin{pmatrix} 2\theta_1+\theta_2-2 \\ \theta_1+2\theta_2+2 \end{pmatrix} \qquad \text{and} \qquad a = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

Hence, we are looking for θ such that $\nabla J(\theta) \propto a$. Computations lead to $\theta_1 = -\theta_2$. Optimal value reached for $\theta_1 = 1/2$ (and $J(\theta^*) = -15/4$).

3.4 Lagrangian function

$$\min_\theta J(\theta) \qquad \text{such that} \qquad f(\theta) := a^t\theta - b = 0$$

We have seen the important role of the scalar value λ above.

DÉFINITION 10. — *[Lagrangian function]*

$$L(\lambda,\theta) = J(\theta) + \lambda f(\theta)$$

λ is the Lagrange multiplier. The optimal choice of (θ^*, λ^*) corresponds to

$$\nabla_\theta L(\lambda^*,\theta^*) = 0 \qquad \text{and} \qquad \nabla_\lambda L(\lambda^*,\theta^*) = 0.$$

Argument : θ^* is optimal if there is no admissible descent directions h. Hence, ∇J and ∇f are linearly dependent. As a consequence, there exists λ such that

$$\nabla_\theta L(\lambda^*,\theta^*) = \nabla J(\theta) + \lambda \nabla f(\theta) = 0 \qquad \text{(Dual equation)}$$

Since θ must be admissible, we have

$$\nabla_\theta L(\lambda^*,\theta^*) = f(\theta^*) = 0 \qquad \text{(Primal equation)}$$

3.5 Inequality constraint

Case of a unique inequality constraint :

$$\min_\theta J(\theta) \qquad \text{such that} \qquad g(\theta) \le 0$$

- Descent direction h : $\nabla J(\theta)^t h < 0$.
- Admissible direction h : $\nabla g(\theta)^t h \le 0$ guarantees that $g(\theta+\alpha h)$ is decreasing with α.

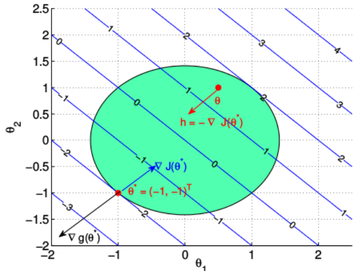
Optimality θ^* is *optimal if there is no admissible descent direction starting from* θ^*. The only possible case is when $\nabla J(\theta^*)$ and $\nabla g(\theta^*)$ are linearly dependent and opposite :

$$\exists\lambda\in\mathbb{R} \qquad \nabla J(\theta^*) = -\mu\nabla g(\theta^*) \qquad \text{with} \qquad \mu \ge 0.$$

Exemple

$$\min_{\theta\in\mathbb{R}^2} \quad \theta_1+\theta_2$$

$$\text{s.c.} \quad g(\theta) = \theta_1^2+\theta_2^2-2 \le 0$$



We can check that $\theta^* = (-1,-1)$.

3.5.1 *Lagrangian in general settings*

We consider the minimization problem :

- $\min_\theta J(\theta)$ such that
- $g_j(\theta) \le 0, \forall j = 1,\ldots,m$ and $f_i(\theta) = 0, \forall i = 1,\ldots,n$

DÉFINITION 11. — *[Lagrangian function] We associate to this problem the Lagrange multipliers* $(\lambda,\mu) = (\lambda_1,\ldots,\lambda_n,\mu_1,\ldots,\mu_m)$.

$$L(\theta,\lambda,\mu) = J(\theta) + \sum_{i=1}^n \lambda_i f_i(\theta) + \sum_{j=1}^m \mu_j g_j(\theta)$$

- θ primal variables
- (λ,μ) dual variables

3.5.2 KKT Conditions

DÉFINITION 12. — *[KKT Conditions] If J and f, g are smooth, we define the Karush-Kuhn-Tucker (KKT) conditions as*

- *Stationarity : $\nabla_\theta L(\lambda, \mu, \theta) = 0$.*
- *Primal Admissibility : $f(\theta) = 0$ and $g(\theta) \leq 0$.*
- *Dual admissibility $\mu_j \geq 0, \forall j = 1, \ldots, m$.*

THÉORÈME 13. — *A convex minimization problem of J under convex constraints f and g has a solution θ^* if and only if there exists λ^* and μ^* such that KKT conditions hold.*

Example :

$$J(\theta) = \frac{1}{2}\|\theta\|_2^2 \qquad \text{s.t.} \qquad \theta_1 - 2\theta_2 + 2 \leq 0$$

We get $L(\theta, \mu) = \frac{\|\theta\|_2^2}{2} + \mu(\theta_1 + 2\theta_2 + 2)$ with $\mu \geq 0$.
Stationarity : $(\theta_1 + \mu, \theta_2 - 2\mu) = 0$.

$$\theta_2 = -2\theta_1 \qquad \text{with} \qquad \theta_2 \leq 0.$$

We deduce that $\theta^* = (-2/5, 4/5)$.

3.5.3 Dual function

We introduce the *dual* function :

$$\mathcal{L}(\lambda, \mu) = \min_\theta L(\theta, \lambda, \mu).$$

We have the following important result

THÉORÈME 14. — *Denote the optimal value of the constrained problem $p^* = \min\{J(\theta)|f(\theta) = 0, g(\theta) \leq 0\}$, then*

$$\mathcal{L}(\lambda, \mu) \leq p^*.$$

Remark :

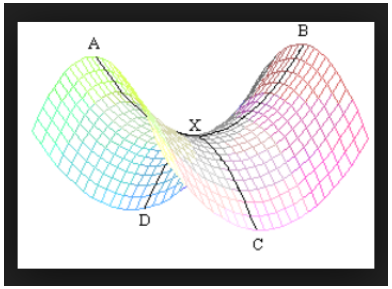
- The dual function $\mathcal{L}$ is lower than p^*, for any $(\lambda, \mu) \in \mathbb{R}^n \times \mathbb{R}_+^m$
- We aim to make this lower bound as close as possible to p^* : idea to maximize w.r.t. λ, μ the function $\mathcal{L}$.

DÉFINITION 15. — *[Dual problem]*

$$\max_{\lambda \in \mathbb{R}^n, \mu \in \mathbb{R}_+^m} \mathcal{L}(\lambda, \mu).$$

$L(\theta, \lambda, \mu)$ affine function on λ, μ and thus convex. Hence, $\mathcal{L}$ is convex and almost unconstrained.

- Dual problems are easier than primal ones (because of almost constraints omissions).
- Dual problems are equivalent to primal ones : maximization of the dual $\Leftrightarrow$ minimization of the primal (not shown in this lecture).
- Dual solutions permit to recover primal ones with KKT conditions (Lagrange multipliers).



Example :

- Lagrangian : $L(\theta, \mu) = \frac{\theta_1^2 + \theta_2^2}{2} + \mu(\theta_1 - 2\theta_2 + 2)$.
- Dual function $\mathcal{L}(\mu) = \min_\theta L(\theta, \mu) = -\frac{5}{2}\mu^2 + 2\mu$.
- Dual solution : $\max \mathcal{L}(\mu)$ such that $\mu \geq 0$: $\mu = 2/5$.
- Primal solution : KKT $\Longrightarrow \theta = (-\mu, 2\mu) = (-2/5, 4/5)$.

To obtain further details, see the Minimax von Neuman's Theorem . . .

3.6 Take home message from convex optimization

- Big Data problems arise in a large variety of fields. They are complicated for a computational reason (and also for a statistical one, see later).

- Many Big Data problems will be traduced in an optimization of a convex problem.
- Efficient algorithms are available to optimize them :

  independently on the dimension of the underlying space.
- Primal - Dual formulations are important to overcome some constraints on the optimization.
- Numerical convex solvers are widely and freely distributed.

4 Applications & Homework

Length limitation : 5 pages !
Deadline : 8th of February.
Group of 2 students allowed.

- This report should be short : strictly less than 5 pages, including the references.
- The work relies either on an academic widespread subject or on a group of selected papers. In any case, you have to highlight the relationship between the concerned chapter and the theme you selected.

For the chosen subject, the report should be organized as follows

1. First motivate the problem with a concrete application and propose a reasonnable modelisation.
2. Second, the report should explain the mathematical difficulties to solve the model and some recent developments to bypass these difficulties. You can also describe the behaviour of some algorithms.
3. Third, the report should propose either :
   - numerical simulations using packages found on the www or your own experiments.
   - some sketch of proofs of baseline theoretical results
   - a discussion part that present alternative methods (with references), exposing pros and cons of each methods.

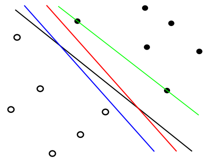
You can choose to only exploit a subsample of the proposed references, as soon as the content of your work is interesing enough. You can also complement your report with a reproducible set of simulations (use R or Matlab please) that can be inspired from existing packages. (If packages are not public, send the whole source files). These simulations are not accounted in the 4/5 pages of the report.

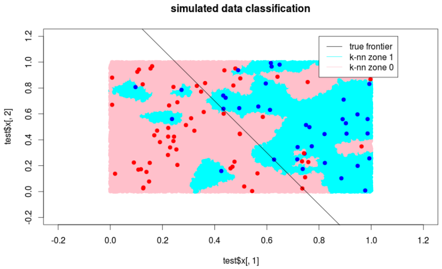
The report files should be named lastname.doc or lastname.pdf and expected in my mailbox before 8th of February.

And to do this, anything is fair game (you can do what you want and find sources everywhere, but take care to avoid a plagiat !)

4.1 Classification with NN & SVM

The supervised classification problem is a long-standing issue in statistics and machine learning and many algorithms can be found to deal with this standard framework. After a brief introduction and a concrete example, a modelisation of this statistical problem, explain the important role of the Bayes classifier and of the NN rule. Then, present the geometric interpretation of the SVM classifier, the role of convexity and the maths behind. After, discuss on the influence of the several parameters : number of observations, dimension of the ambiant space, etc.

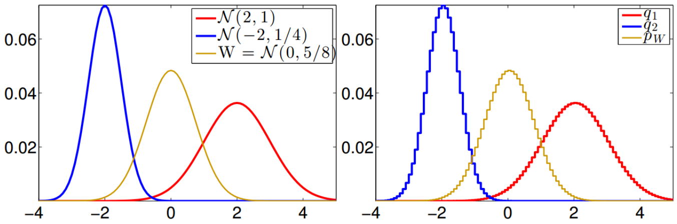




References :

- CRAN repository
- Journal of Statistical Software webpage
- Hastie Tibshirani and Friedman, *The elements of statistical learning data mining inference and prediction*
- Gyorfi, Lugosi, *A Probabilistic Theory of Pattern Recognition*
- My website `perso.math.univ-toulouse.fr/gadat/`
- Wikistat `wikistat.fr/`

4.2 Transport problems



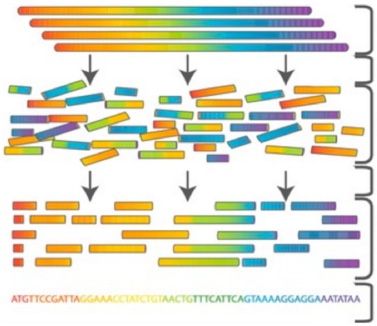
Optimal transportation problem is a growing field of interest in machine learning, big data and statistics. You will find below some interesting readings. Try to mainly understand the motivations, the mathematical tools, and the nature of the several applications. You can also find some softwares.

References

- Cran repository : lpsolve and Transport packages
- Gabriel Peyré's webpage (Matlab softwares)
- Mario Cuturi's webpage
- Nicolas Papadakis's webpage
- Benamou, Carlier, Cuturi, Nenna and Peyré *Iterative Bregman Projections for Regularized Transportation Problems*
- Cuturi and Peyré *A smoothed dual approach for variationnal Wasserstein problems*
- Cuturi *Sinkhorn Distances : Lightspeed Computation of Optimal Transport*
- Cuturi and Doucet *Fast Computation of Wasserstein Barycenters*

4.3 Permutation recovery

The statistical recovery of a permutation is a perfect example of NP-hard problem, for which a non trivial convex relaxation should be studied. You can either propose to focus on global optimization with simulated annealing or genetic algorithm, or convex methods for solving relaxed convex problems. The problem of permutation recovery is useful in seriation, graphs, ... Instead of focusing on the statistical part, focus on the optimization problem, the principle of the relaxation and the potential applications.



References :

- Cran repository : lpsolve and Transport packages
- Francis Bach's webpage (Matlab softwares)
- Alexandre d'Aspremont's webpage
- Fogel,Jenatton, Bach, D'Aspremont *Convex relaxations for permutation problems*
- Lim and Wright *Beyond the Birkhoff Polytope : Convex Relaxations for Vector Permutation Problems*
- Collier and Dalalyan *Minimax rates in permutation estimation for feature matching*