Convex Optimization in Machine Learning and Computational Finance Lecture 8: More Applications

Dr. Michel Baes, Pr. Patrick Cheridito

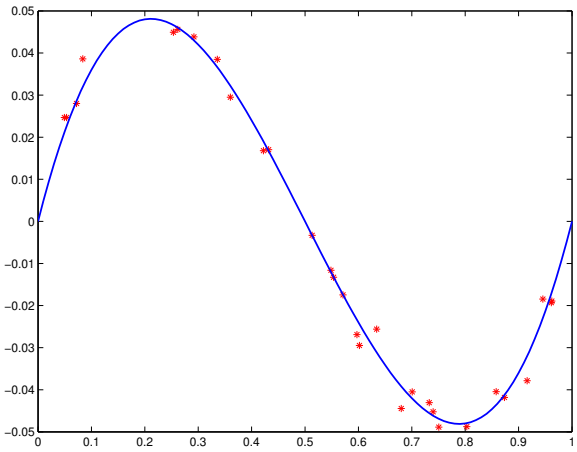
RiskLab / ETH Zürich

Quick recall of last week's lecture

- Important classes of convex problems:
  - Second-order cone problems
    - CQr sets and functions
  - Semidefinite problems
    - SDr sets and functions
- Applications of SOCP:
  - Robust LP, Minimal surface problem
- Applications of SPD:
  - Dealing with univariate polynomial constraints
  - Splitting a graph in two (MaxCut)
  - Ensuring stability of some system (S-Lemma)

I. Approximation

Polynomial approximation (among others)



Let $(t_1, v_1), \ldots (t_N, v_N)$ in $\mathbb{R}^2$. Find a polynomial of degree d approximating these points as well as possible.

Find $p_0, \ldots, p_d$ such that $v_i \approx \sum_{j=0}^{d} p_j t_i^j$.

With $[A]_{ij} := t_i^j$, we will try to get $v \approx Ap$.

Denote $r_i := (Ap)_i - v_i$ (error at the ith point)

Different strategies:

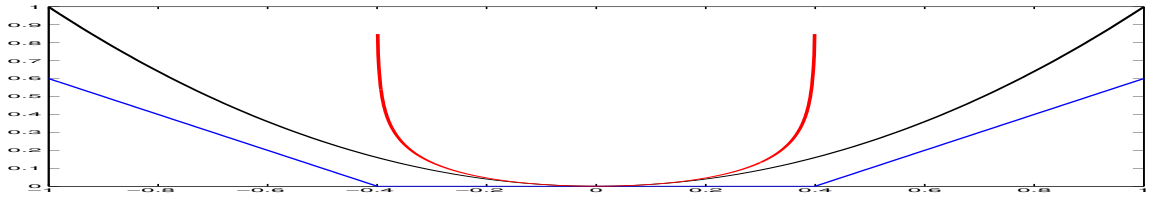
- Chebysev approximation: $\min \max_i |r_i|$. It is an LP:
  $\min\{t : -t \leq r_i \leq t, \text{ for } 1 \leq i \leq N\}$
- Least-squares: $\min ||Ap - v||_2^2 = \min \sum_i r_i^2$

Polynomial approximation (among others)

More penalization strategies:

- Least-squares: $\min ||Ap - v||_2^2 = \min \sum_i r_i^2$
  denoting $r_i := (Ap)_i - v_i$ (error at the ith point).
- More generally, penalization strategies: $\min \sum_i \phi(r_i)$, where $\phi : \mathbb{R} \to \mathbb{R}_+$ is a convex *penalty* function: $\phi(t) = 0$ iff $t = 0$.

What do you want?	Which ϕ can you take?
Ignore good outliers	$\phi(t) = \|t\|$, or $\phi(t) = [\|t\| - a]_+$
Avoid bad outliers	$\phi(t) = -a^2 \log(1 - t^2/a^2)$
Do something in between	$\phi(t) = t^2$.



Polynomial approximation (among others)

Let $(t_1, v_1), \ldots (t_N, v_N)$ in $\mathbb{R}^2$.
We have a set of basis functions $\phi_0(t), \ldots, \phi_d(t)$.
Find $p_0, \ldots, p_d$ such that $v_i \approx \sum_{j=0}^{d} p_j \phi_j(t_i)$.
With $[A]_{ij} := \phi_j(t_i)$, we will try to get $v \approx Ap$.

Often used in practice:

- ϕ_j's that make A diagonal (Legendre polynomials, ...)
- Piecewise linear functions: $\phi_j(t_i) = \delta_{ij}$.
- For periodic phenomenons: $\phi_j(t) = \cos(\lambda_j t + \gamma_j)$
- $\phi_j(t) = \exp(-(t - c_j)^2)$, ...

When there are two antagonistic objectives: regularization

Suppose we want to find p such that $v \approx Ap$.
The cost of a candidate p is $\sum_i |p_i| = ||p||_1$,
and we also want to make it small.

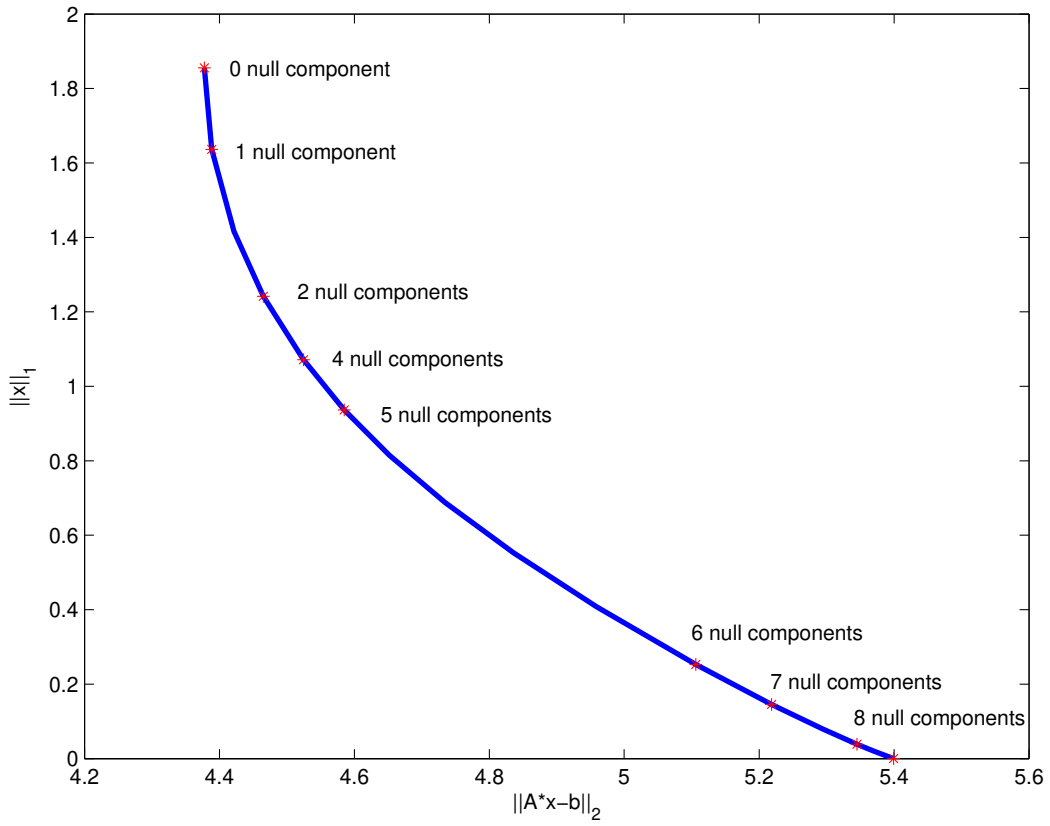
$$\text{" min}(||v - Ap||, ||p||_1)\text{"} \rightsquigarrow \min ||v - Ap|| + \gamma ||p||_1$$

for every $\gamma \geq 0$ gives the *Pareto* curve of the problem.

Interesting feature: the solution p_γ has often many zeros.

The solution has often many zeros

$A \in \mathbb{R}^{30\times 10}$, $0 \leq \gamma \leq 20$



Application:

understanding Swiss cantons correlations

In Switzerland, some laws are subject to a referendum. The results show sometimes differences between cantons.

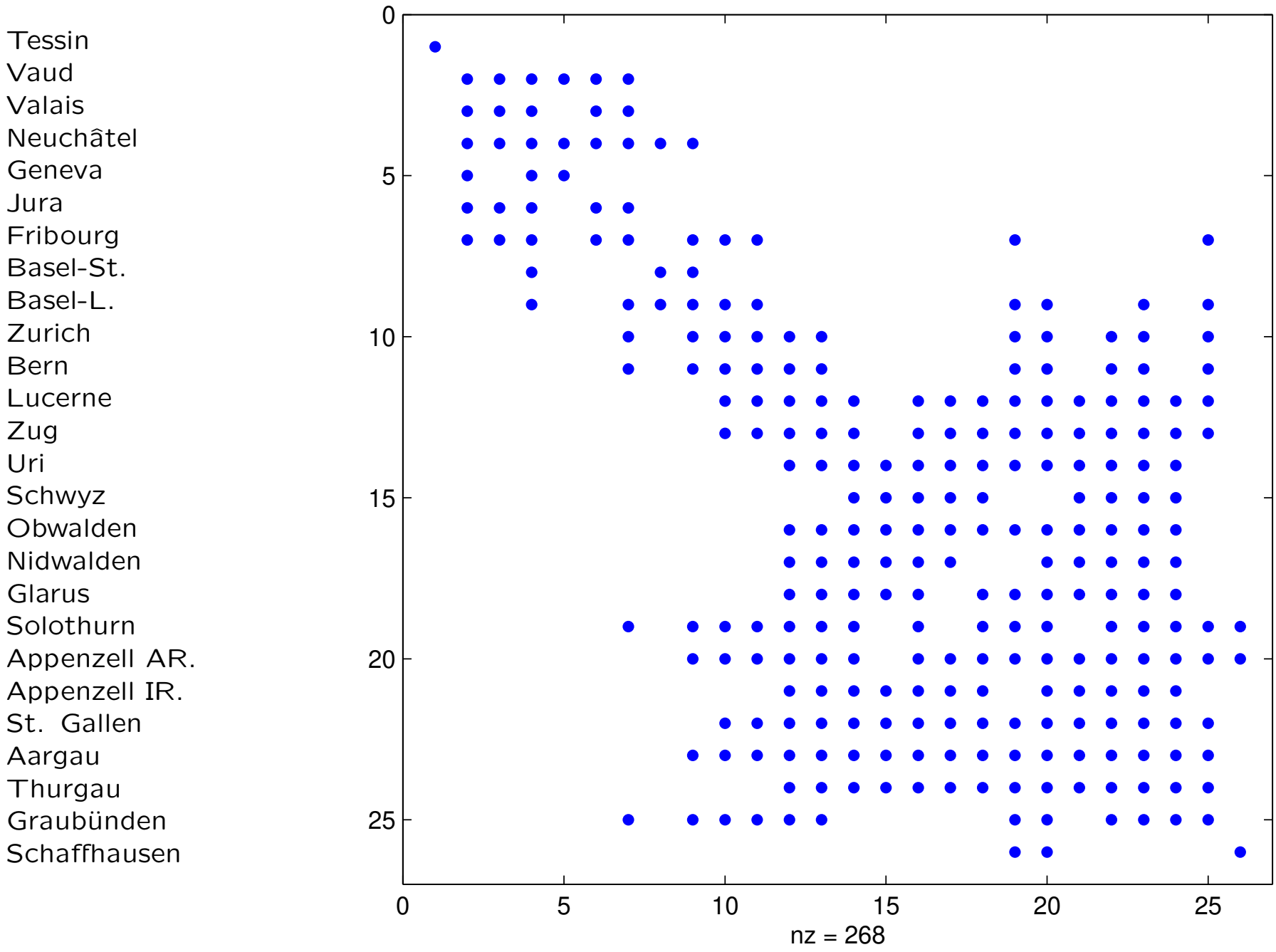
Model: Each canton is a random variable $y_i \in [-50\%, 50\%]$, approximated as $y \sim N(0, \Sigma)$, with unknown Σ.

Fact: y_i is independent of y_j iff $[\Sigma]_{ij} = 0$.

A starting point for Σ: To estimate Σ, we look at the 24 past referendums, and we build the standard correlation estimator $\widehat{\Sigma}$.

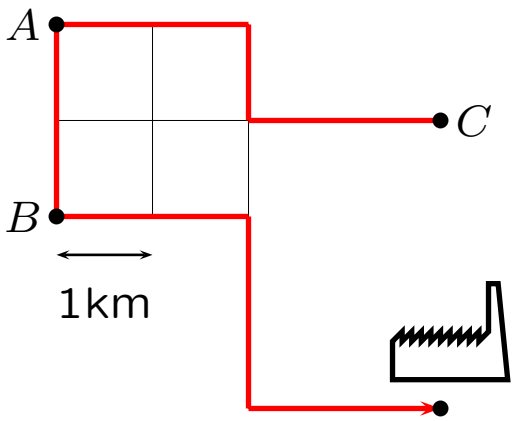
We then solve $\min\{||\Sigma - \widehat{\Sigma}||_2 + \gamma||\Sigma||_1 : \Sigma \in \mathbb{S}_+^{26}\}$.

Next picture: There is a blue dot at (i, j) iff $|\Sigma_{ij}| > 0.0018$.


Tessin
Vaud
Valais
Neuchâtel
Geneva
Jura
Fribourg
Basel-St.
Basel-L.
Zurich
Bern
Lucerne
Zug
Uri
Schwyz
Obwalden
Nidwalden
Glarus
Solothurn
Appenzell AR.
Appenzell IR.
St. Gallen
Aargau
Thurgau
Graubünden
Schaffhausen
0
5
10
15
20
25
nz = 268

II. Understanding coalitions

Carpool without fights



A, B, and C work at the factory and each have a car. Transportation cost: 1 CHF/km.

$c(P)$ denotes P's travel cost when they are all in one car. The *gain* of carpool is $v(P) := \sum_{i \in P} c(\{i\}) - c(P)$.

P	cost c	$v(P)$
$\{A\}$	8	0
$\{B\}$	6	0
$\{C\}$	7	0
$\{A, B\}$	8	6
$\{A, C\}$	12	3
$\{B, C\}$	11	4
$\{A, B, C\}$	13	8

Carpool without fights

When $P := \{A, B, C\}$, C drives and pays all the costs (13CHF).

A, B have to pay him a *fair compensation*, i.e. some money to enter in the coalition.

How much?

P	cost c	$v(P)$
$\{A\}$	8	0
$\{B\}$	6	0
$\{C\}$	7	0
$\{A, B\}$	8	6
$\{A, C\}$	12	3
$\{B, C\}$	11	4
$\{A, B, C\}$	13	8

Equivalently, how do we split the benefit (8CHF)?

Let x_i be the fraction of the benefit taken by i.

Example: if $(x_A, x_B, x_C) = (0.25, 0.25, 0.5)$, A and B take 2CHF and C takes 4CHF of the benefits. That means that A pays $8 - 2 = 6$CHF, B pays $6 - 2 = 4$CHF, and C pays $7 - 4 = 3$CHF.

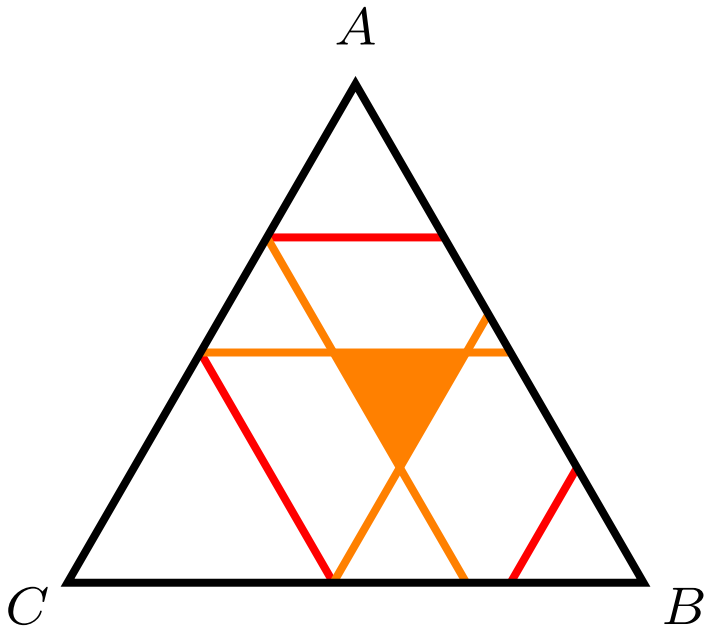
What is a fair compensation?

Suppose that we impose (x_A, x_B, x_C) to A, B, C, and that they have to decide whether they join a coalition or not.

Let $\mathcal{N} := \{A, B, C\}$ and $\bar{v} := v(\mathcal{N})$.

- If $x_i \cdot \bar{v} < v(\{i\})$, i has no incentive to join the coalition.
- More generally, if there is $P \subseteq \mathcal{N}$ for which $\sum_{i \in P} x_i \bar{v} < v(P)$, those in P will never enter the coalition.
- If $\exists i$ s.t. $\forall P \subseteq \mathcal{N}$ we have $x_i \cdot \bar{v} > v(P) - v(P \backslash \{i\})$, the others can (and will) negotiate with i to get at least $x_i \cdot \bar{v} - \max\{v(P) - v(P \backslash \{i\}) : P \subseteq \mathcal{N}\}$ more. This will not change i's incentive wrt carpool.
- For L. Shapley, a compensation $x \in \Delta_{|\mathcal{N}|}$ is *fair* iff
  $$\sum_{i \in P} x_i \cdot \bar{v} \geq v(P) \, \forall \, P \subseteq \mathcal{N}$$
  and $x_i \cdot \bar{v} \leq \max\{v(P) - v(P \backslash \{i\}) : P \subseteq \mathcal{N}\} \, \forall \, i \in \mathcal{N}$
  (iff, for some authors, x is in the *core of the cooperative game*).

The set of fair carpool



We use the carpool problem data. Points are represented in barycentric coordinates (e.g. $A = (1, 0, 0)$).

Conditions $\sum_{i \in P} x_i \cdot \bar{v} \geq v(P)$ yields $x_A \leq \frac{9}{13}$, $x_B \leq \frac{10}{13}$, $x_C \leq \frac{7}{13}$.

$x_i \cdot \bar{v} \leq \max_{P \subseteq \mathcal{N}} v(P) - v(P \backslash \{i\})$ gives $x_A \leq \frac{6}{13}$, $x_B \leq \frac{6}{13}$, $x_C \leq \frac{4}{13}$.

The intersection is non-empty, thus a coalition can occur.

When is the set of fair points nonempty?
i.e. is there a property of v implying fair points?

Convex games:

games where coalitions are possible

Let $\mathcal{N} := \{1, \ldots, n\}$. A function $v : 2^{\mathcal{N}} \to \mathbb{R}_+$ is *convex* if

$$v(P) + v(Q) \leq v(P \cap Q) + v(P \cup Q) \text{ for every } P, Q \subseteq \mathcal{N}$$

and $v(\phi) = 0$. Note that v is **discrete**.

Note 1: Definition by L. Shapley (1965) in Game Theory.
J. Edmonds (1966) called v a *supermodular function*.

Note 2: Replacing $2^{\mathcal{N}}$ by $\{0,1\}^n$ ($\mathcal{N}\backslash\{1\}$ is now $(0, 1, \ldots, 1)$),
there exists a concave (!) function $\bar{v}$ on $[0,1]^n$
for which $\bar{v}(x) = v(x)$ on $\{0,1\}^n$. **(Lovász)**

Note 3: As natural substitute for gradient in a direction,
we can set $\nabla v(P)[Q] := v(P \cup Q) - v(P \backslash Q)$.
If v is convex, then $\nabla(\nabla v(P)[Q])[Q] \geq 0$.
(Compare with $\nabla^2 \bar{v} \succeq 0$ for usual convex functions)

Convex games:
games where coalitions are possible

$v(P) + v(Q) \leq v(P \cap Q) + v(P \cup Q)$ for every $P, Q \subseteq \mathcal{N}$.

Note 4: If $j \notin P$, $v(P) + v(\{j\}) \leq v(P \cup \{j\})$:
the larger the coalition, the larger its gain.

Theorem 1 (Shapley)
If v is convex, there exists a fair compensation.

Proof: Let $\mathcal{N}_j := \{1, \ldots, j\}$. Define $x_i := (v(\mathcal{N}_i) - v(\mathcal{N}_{i-1}))/\bar{v}$ for all i.
The condition $x_i \bar{v} \leq \max_{P \subseteq \mathcal{N}} v(P) - v(P \backslash \{i\})$ is satisfied for $P := \mathcal{N}_i$.

To check the condition $\sum_{i \in P} x_i \bar{v} \geq v(P)$, let us take $P \subseteq \mathcal{N}$

and let $j := \min\{k : k \in \mathcal{N} \backslash P\}$, so that $\mathcal{N}_{j-1} \subseteq P$. By convexity,

$$v(P) + v(\mathcal{N}_j) \leq v(\mathcal{N}_{j-1}) + v(P \cup \{j\}), \text{ i.e. } \sum_{k \in P \cup \{j\}} x_k \bar{v} - v(P \cup \{j\}) \leq \sum_{k \in P} x_k \bar{v} - v(P).$$

Recursively (take $P \cup \{j\}$ instead of P above, etc...), we get:
$\sum_{k \in \mathcal{N}} x_k \bar{v} - v(\mathcal{N}) = \bar{v} - \bar{v} = 0 \leq \sum_{k \in P} x_k \bar{v} - v(P)$.
Hence, x is a fair compensation. ■

III. Statistical estimation: Maximum likelihood problems

Exploring brains (Nemirovski/Ben-Tal/Margalit)

Positron Emission Tomography (PET)
is a powerful non-invasive technique for measuring *metabolic* activity (unlike X-rays) in the human body.
Detection of tumors, visualization of blood flow,...

The patient swallows/inhales some positron emitters, which disintegrate in the body producing positrons.

Each positron decays instantly with surrounding electrons, emitting two 500 keV photons in opposite directions.

The patient is surrounded by a ring of captors.

The two photons hit simultaneously a pair of captors: we know the line where the decay happened.

Problem: reconstruct a decay density map from all these lines, given possible errors.

Maximum likelihood in action

Since $y_i \sim \mathrm{Po}((P\lambda)_i)$, we have by definition:

$$\mathbb{P}[y_i = n|\lambda] = \frac{(P\lambda)_i^n}{n!}\exp(-(P\lambda)_i)$$

We compute the λ that is the most likely to be:

$$\lambda^* := \arg\max_{\lambda\geq 0} \mathbb{P}[y|\lambda] = \arg\max_{\lambda\geq 0} \prod_{i\in A} \frac{(P\lambda)_i^{y_i}}{y_i!}\exp(-(P\lambda)_i).$$

NON CONVEX!?

No: a simple transform makes it convex.

$$\lambda^* = \arg\max_{\lambda\geq 0} \log\mathbb{P}[y|\lambda] = \arg\max_{\lambda\geq 0} \sum_{i\in A} y_i \log((P\lambda)_i) - (P\lambda)_i.$$

Maximum likelihood and convex optimization

Let p_λ be probability distributions on $Q \subseteq \mathbb{R}^n$,
indexed by $\lambda \in \Lambda \subseteq \mathbb{R}^m$.
The *likelihood function* for $y \in Q$ is $L_y(\lambda) := p_\lambda(y)$.
Given observations y, the *most likely* parameter λ is:

$$\lambda^* := \arg\max\{L_y(\lambda) : \lambda \in \Lambda\}.$$

Example where convexity pops up naturally
Suppose that $y = A\lambda + \xi$, where $A \in \mathbb{R}^{n\times m}$
and ξ_i are IID with density p. Then

$$p_\lambda(y) = \prod_i p(y_i - (A\lambda)_i) \;\Leftrightarrow\; \log L_y(\lambda) = \sum_i \log(p(y_i - (A\lambda)_i)),$$

which is concave if $t \mapsto \log(p(t))$ is concave.

Two examples where $\log(p(t))$ is concave

- **Gaussian noise:** If $\xi_i \sim N(0, \sigma^2)$,
  by definition $p(t) = \exp(-t^2/(2\sigma^2))/\sqrt{2\pi\sigma^2}$, and

$$\log L_y(\lambda) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}||y - A\lambda||^2.$$

  We get a least-square problem.

- **Uniform noise:** If $\xi_i \sim U(-a, a)$,
  by definition $p(t) = 1/(2a)$ for $t \in [-a, a]$, and

$$\log L_y(\lambda) = \sum_i \log(p(y_i - (A\lambda)_i)),$$

  which is maximized on every λ
  such that $-a \leq y_i - (A\lambda)_i \leq a$.

For next week

- Can you solve a general optimization problems?
- How can you *solve* convex optimization problems? Simple methods, and how well they can perform.