Statistics and optimization
Motivation

Sylvain Sardy

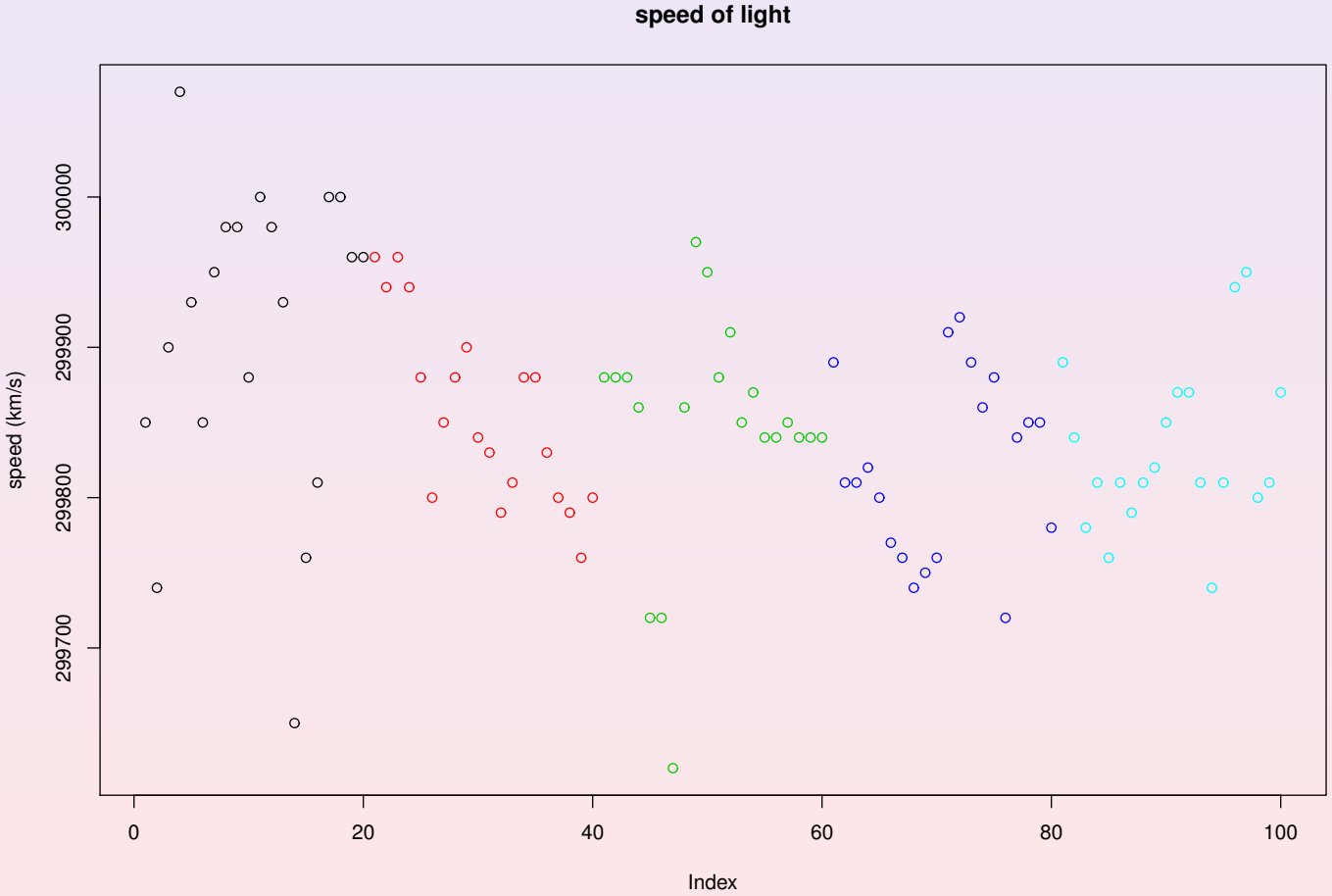
Université de Genève

1. Some applications of optimization

- decision making, management, e.g., best strategy
- engineering design, e.g., minimal cost for maximal reliability
- utilization of resources, e.g., energy (hydro-electricity, windmills, nuclear, ...)
- minimal cost, maximal profit, minimal risk
- statistics

2. Optimization in Statistics/Machine learning

Data set 1: $N = 5x20$ measurements of speed of light (Michelson, 1879)



To estimate the speed of light μ, we can think about finding a value $\hat{\mu}$ the closest to the data:

- with the ℓ_2 metric

$$\hat{\mu} = \arg\min_{\mu \in \mathbb{R}} f(\mu)$$

  with $f(\mu) = \sum_{n=1}^{N} (y_n - \mu)^2 = \|\mathbf{y} - \mu \mathbf{1}\|_2^2$.
- with the ℓ_1 metric

$$\hat{\mu} = \arg\min_{\mu \in \mathbb{R}} f(\mu)$$

  with $f(\mu) = \sum_{n=1}^{N} |y_n - \mu| = \|\mathbf{y} - \mu \mathbf{1}\|_1$.

- if we assume

$$Y_1, \ldots, Y_N \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\mu, \sigma^2),$$

then you know that $(\bar{y}, s^2) = (\hat{\mu}, \hat{\sigma}^2)^{(\mathrm{MLE})}$.

This means that

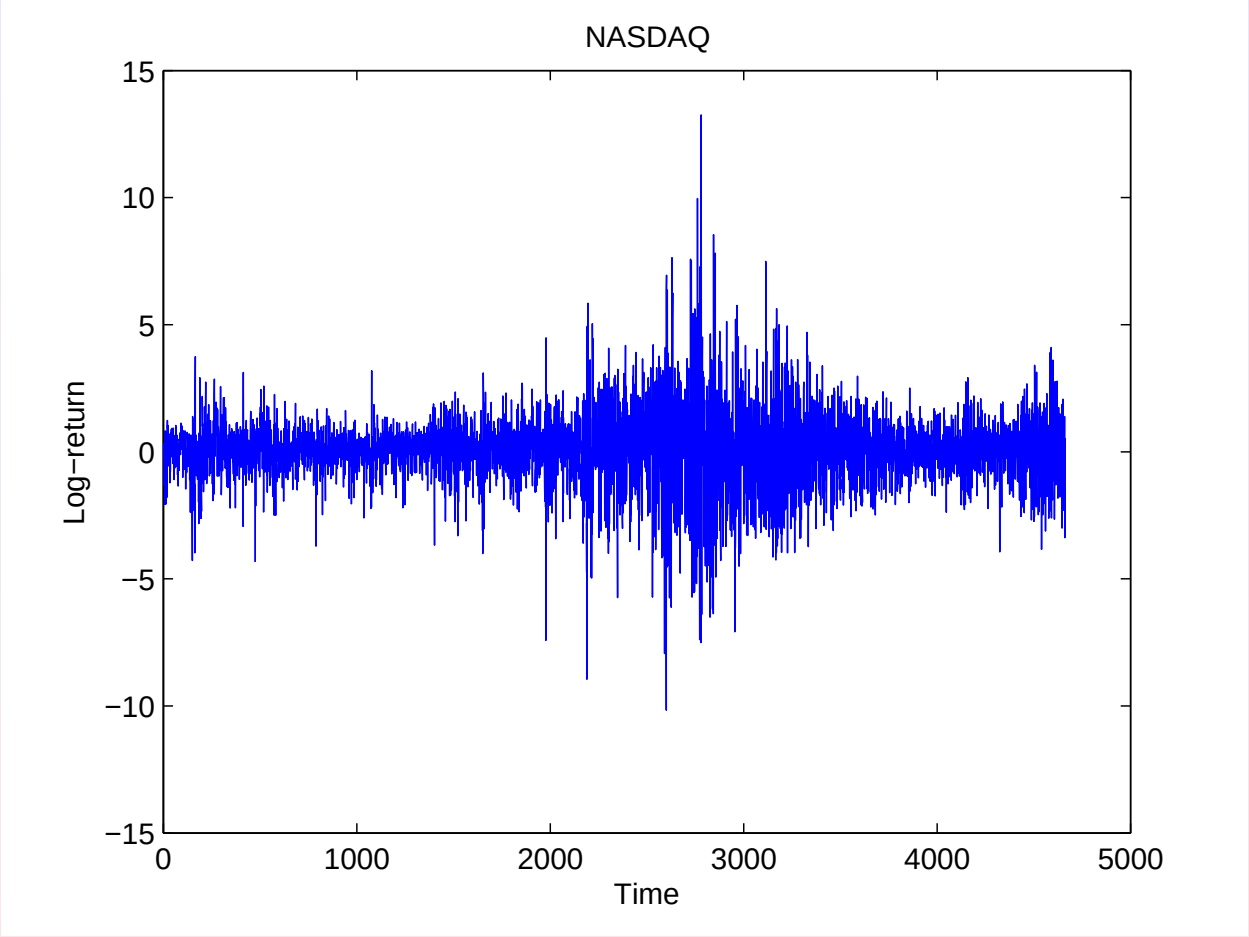
$$\begin{aligned}(\bar{y}, s^2) &= \underset{\mu \in \mathbb{R}, \sigma > 0}{\operatorname{argmax}} L(\mu, \sigma^2; y_1, \ldots, y_N) \\ &= \underset{\mu \in \mathbb{R}, \sigma > 0}{\operatorname{argmin}} f(\mu, \sigma),\end{aligned}$$

where

$$f(\mu, \sigma^2) = -\log L(\mu, \sigma^2; y_1, \ldots, y_N)$$

is the negative log-likelihood.

Data set 2: Time series $Y_t = \log(p_t/p_{t-1})$ consists of log-returns for the price p_t of the stocks at time t.



Contemporary finance seeks to control the risk/volatility.

The ARCH model (Engle, 1982) assumes

$$\begin{aligned} Y_t &= \sigma_t \epsilon_t \\ \sigma_t^2 &= \alpha_0 + \alpha_1 Y_{t-1}^2, \end{aligned} \tag{1}$$

where $\epsilon_t \overset{\text{i.i.d.}}{\sim} (0, 1)$, $\alpha_0 > 0$ and $\alpha_1 \geq 0$.
The corresponding likelihood function

$$L(\alpha_0, \alpha_1; \mathbf{y}) = \Pi_{t=2}^{T} \frac{1}{\sqrt{2\pi(\alpha_0 + \alpha_1 y_{t-1}^2)}} \exp(-\frac{y_t^2}{2(\alpha_0 + \alpha_1 y_{t-1}^2)})$$

can be maximized to get the MLE

$$(\hat{\alpha}_0, \hat{\alpha}_1)(\mathrm{MLE}) = \arg \max_{\alpha_0 > 0, \alpha_1 \geq 0} f(\alpha_0, \alpha_1),$$

where $f(\alpha_0, \alpha_1) = L(\alpha_0, \alpha_1; \mathbf{y})$.

Data set 3: estimation of density function f from

$$Y_1, \ldots, Y_N \overset{\text{i.i.d.}}{\sim} f_Y.$$

Example: SAT math score

- Histogram $\hat{f}_{Y,\lambda}$: nonparametric estimator of f_Y. Its smoothness is governed by a regularization parameter λ, called binwidth.
- Gaussian parametric model with $k = 2$ modes:

$$f_Y = p_1 \, \mathrm{N}(\mu_1, \sigma_1^2) + p_2 \, \mathrm{N}(\mu_2, \sigma_2^2).$$

The five parameters can be estimated by

$$(\hat{p}_1, \hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1, \hat{\sigma}_2) = \arg \max_{p_1 \in [0,1],\ \mu_1 \in \mathbb{R},\ \mu_2 \in \mathbb{R},\ \sigma_1 > 0,\ \sigma_2 > 0} f(p_1, \mu_1, \mu_2, \sigma_1, \sigma_2),$$

where $f(p_1, \mu_1, \mu_2, \sigma_1, \sigma_2) = L(p_1, \mu_1, \mu_2, \sigma_1, \sigma_2; \mathbf{y})$.

Data set 4: Regression for a medical prognostic.

$N = 97$ men are about to receive a prostatectomy.
One response and $P = 8$ covariates measured on each patient:

- $Y =$ `lpsa`: log(prostate specific antigen) is the *response variable*
- $X_1 =$ `lcavol`: log(cancer volume)
- $X_2 =$ `lweight`: log(prostate weight)
- $X_3 =$ `age`: age of patient
- $X_4 =$ `lbph`: log(benign prostatic hyperplasia amount)
- $X_5 =$ `svi`: seminal vesicle invasion
- $X_6 =$ `lcp`: log(capsular penetration)
- $X_7 =$ `gleason`: Gleason score
- $X_8 =$ `pgg45`: percentage Gleason scores 4 or 5.

Goal 1: predict Y based on knowledge contained in $X_1, \ldots, X_8$.
Goal 2: disregard unimportant covariates.

A standard approach assumes a linear model:

$$Y_n = \alpha_0 + \sum_{p=1}^{P} \alpha_p X_{np} + \epsilon_n$$

and estimate the coefficients $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_P)$ for a good fit based on the ℓ_2 metric:

$$\hat{\boldsymbol{\alpha}}^{(\mathrm{LS})} = \arg \min_{\boldsymbol{\alpha} \in \mathbb{R}^{P+1}} f(\boldsymbol{\alpha})$$

where $f(\boldsymbol{\alpha}) = \|\mathbf{y} - X\boldsymbol{\alpha}\|_2^2 = \sum_{n=1}^{N} (y_n - \alpha_0 - \sum_{p=1}^{P} \alpha_p X_{np})^2$.

The cost function f is differentiable, ie, $\nabla_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha})$ exists everywhere. We will see the first order optimality conditions play a key role:

$$\nabla_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha}) \equiv \mathbf{0} \Leftrightarrow X^{\mathrm{T}} X \hat{\boldsymbol{\alpha}}^{(\mathrm{LS})} := X^{\mathrm{T}} \mathbf{y}.$$

In many modern applications, e.g.,

- chemometrics
- genetics
- cancer research

we have more information/covariates than subjects, aka, the $P > N$ situation.

Consequently, the normal equations have an infinite number of solutions.
LS estimation does not lead to a unique estimate,
and the bias–variance tradeoff is poor!

Improved estimators have been proposed, including:

- **Ridge regression** (Hoerl and Kennard, 1970) regularizes least squares by solving

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^{P+1}} f(\boldsymbol{\alpha})$$

with $f(\boldsymbol{\alpha}) = \|\mathbf{y} - X\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_2^2$ for some $\lambda \geq 0$.

- **Lasso** (Tibshirani, 1996) instead solves

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\|\mathbf{y} - X\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1$$

  for some $\lambda \geq 0$.

Note: the intercept α_0 is often not penalized.

The cost function f is not differentiable, leading to the notion of subdifferential ∂f (set of subgradients).

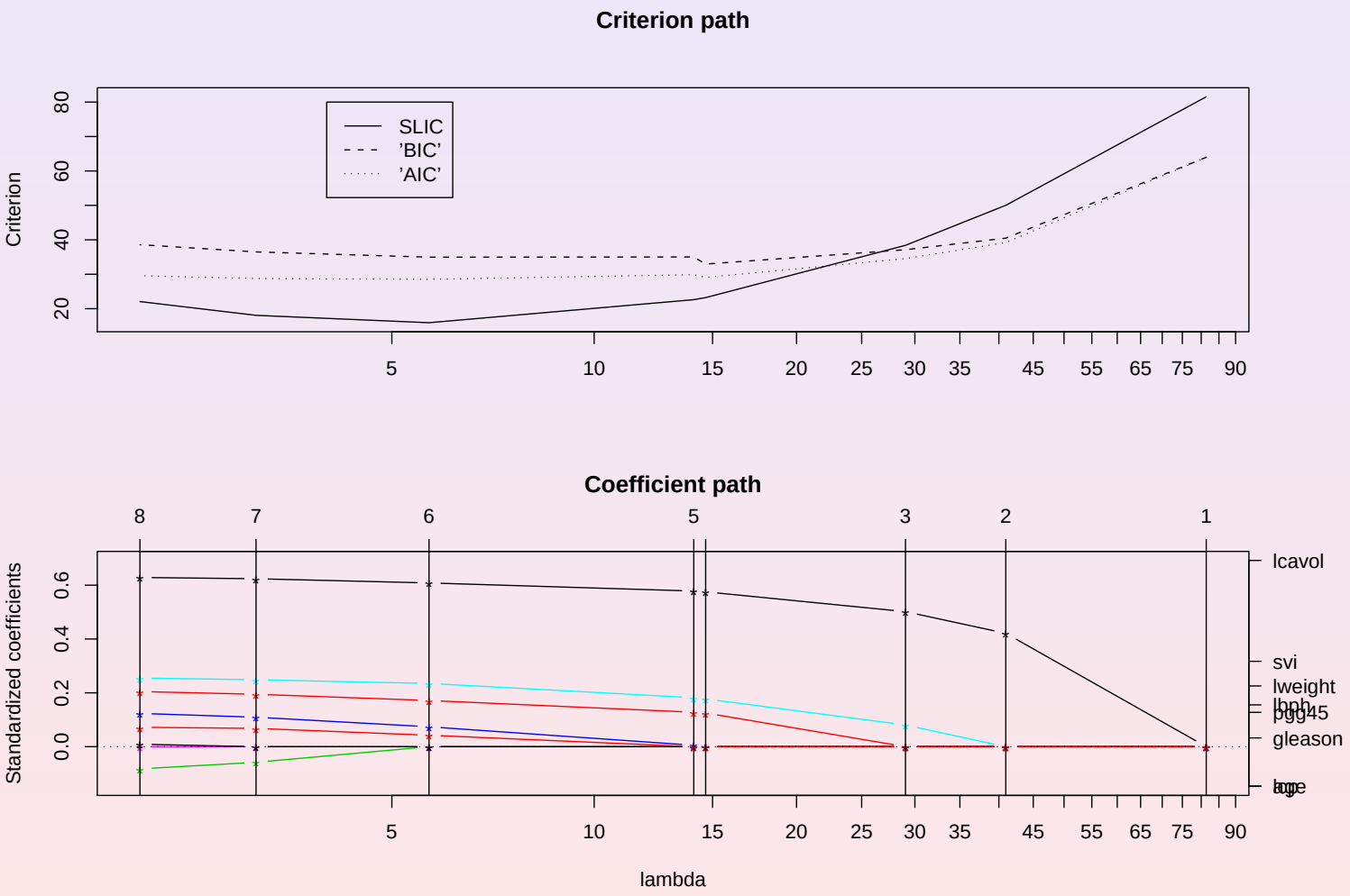
We will see the first order optimality conditions for $\hat{\boldsymbol{\alpha}}_\lambda$ to be a solution are

$$\mathbf{0} \in X^{\mathrm{T}} X \hat{\boldsymbol{\alpha}}_\lambda - X^{\mathrm{T}}\mathbf{y} + \lambda \mathbf{s} =: \partial f(\hat{\boldsymbol{\alpha}}_\lambda),$$

where $\mathbf{s} \in \partial\|\boldsymbol{\alpha}\|_1$ with $s_p = \begin{cases} \{1\} & \text{if } \alpha_p > 0 \\ \{-1\} & \text{if } \alpha_p < 0 \\ [-1,1] & \text{if } \alpha_p = 0. \end{cases}$

Note: *sparsity* property for $\lambda \geq \|X^{\mathrm{T}}\mathbf{y}\|_\infty$.

Typical lasso path:



Data set 5: $N = 173$ crabs. We have the following measurements:

- the response Y = Number of males attached to the female,
- $P = 1$ covariate: X_1 = Width of female crab

Poisson generalized linear model assumes

$$Y_n \mid x_n \sim \text{Poisson}(\eta_n) \quad \text{with} \quad \eta_n = \exp(\alpha_0 + \alpha_1 x_n),$$

where exp is the canonical inverse link function, and estimates $\boldsymbol{\alpha}$ by MLE

$$\boldsymbol{\alpha} = \arg\min_{\boldsymbol{\alpha} \in \mathbb{R}^2} f(\boldsymbol{\alpha})$$

with
$f(\boldsymbol{\alpha}) = -\log L(\boldsymbol{\alpha}; \mathbf{y}) = \sum_{n=1}^{N} \exp(\alpha_0 + \alpha_1 x_n) - y_n(\alpha_0 + \alpha_1 x_n).$

Another possible GLM model is

$$Y_n \mid x_n \sim \text{Poisson}(\eta_n) \quad \text{with} \quad \eta_n = \alpha_0 + \alpha_1 x_n,$$

and estimates $\boldsymbol{\alpha}$ by MLE

$$\boldsymbol{\alpha} = \arg \min_{\boldsymbol{\alpha} \in \Omega \subset \mathbb{R}^2} f(\boldsymbol{\alpha}), \quad \text{where} \quad \Omega = \{\boldsymbol{\alpha} \in \mathbb{R}^2 : \alpha_0 + \alpha_1 x_n > 0 \text{ for all } n\}$$

with $f(\boldsymbol{\alpha}) = -\log L(\boldsymbol{\alpha}; \mathbf{y}) = \sum_{n=1}^{N} \alpha_0 + \alpha_1 x_n - y_n \log(\alpha_0 + \alpha_1 x_n)$.

For the $P > N$ situation in GLM, Park and Hastie (*L1-Regularization Path Algorithm for GLM*, 2007) solve

$$\min_{\boldsymbol{\alpha}} - \log L(\boldsymbol{\alpha}; \mathbf{y}) + \lambda \|\boldsymbol{\alpha}\|_1.$$

with the `glmnet` library in `R` that solves

Comment: if not all entries of $\boldsymbol{\alpha}$ are penalized, the optimization may not have a solution.

Data set 6: Observations Y_t over time $t = 1, \ldots, 2000$.
We believe they come from a succession of regimes.

Model:

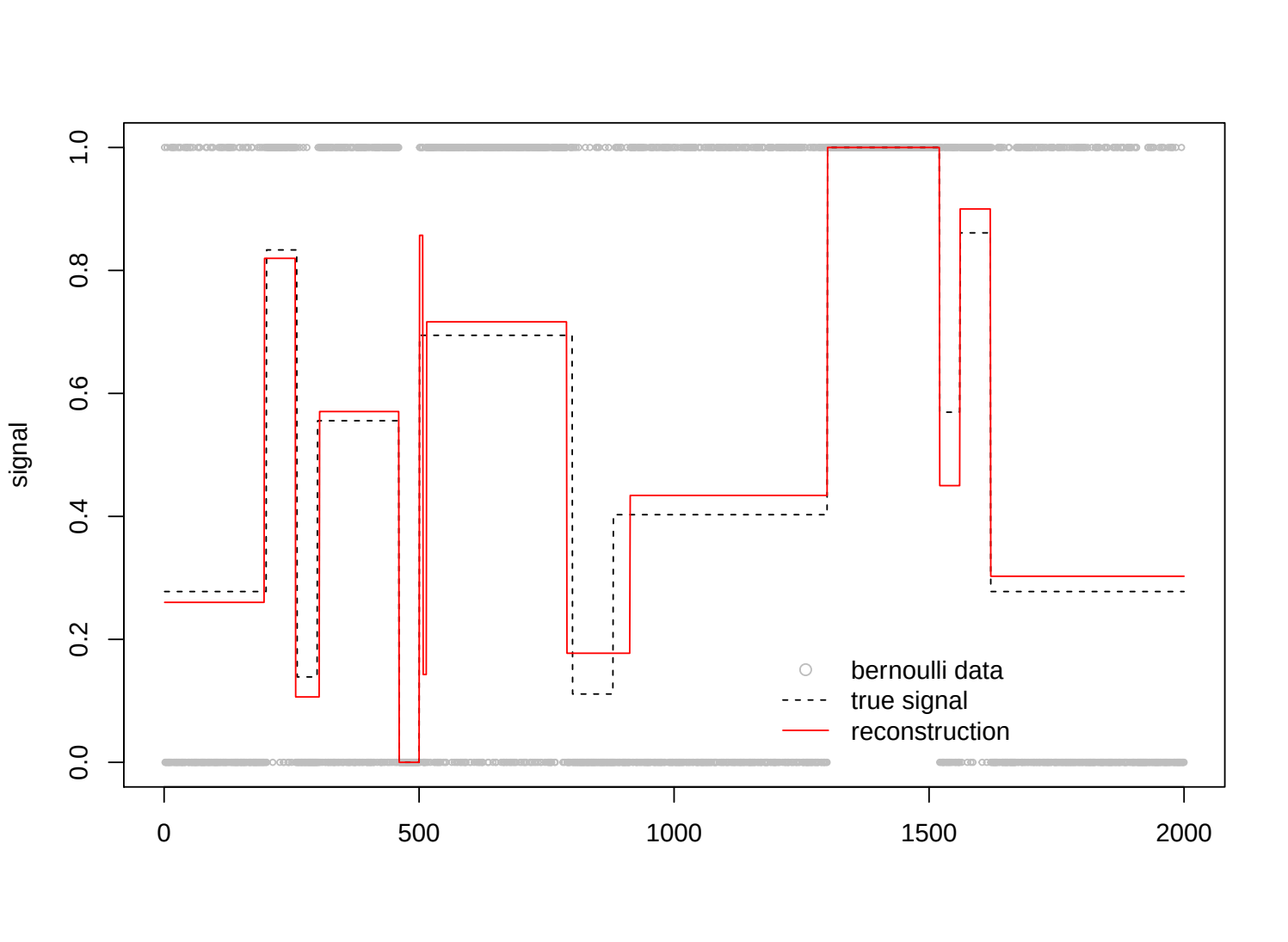
$$Y_t \sim F_{(\mu_t, \eta)} \quad \text{for} \quad t = 1, \ldots, 2000,$$

where $\mu_t = \mathbb{E}(Y_t)$ is a piecewise constant function and η a nuisance parameter.
Each constant piece is a regime.

Goal: detect the times $t_1, t_2, \ldots, t_K$ of regime switch, where K is also unknown.

This is a *segmentation* problem.

Bernoulli example: $F_{(\mu_t,\eta)}$ is Bernoulli($\mu_t = p_t$).



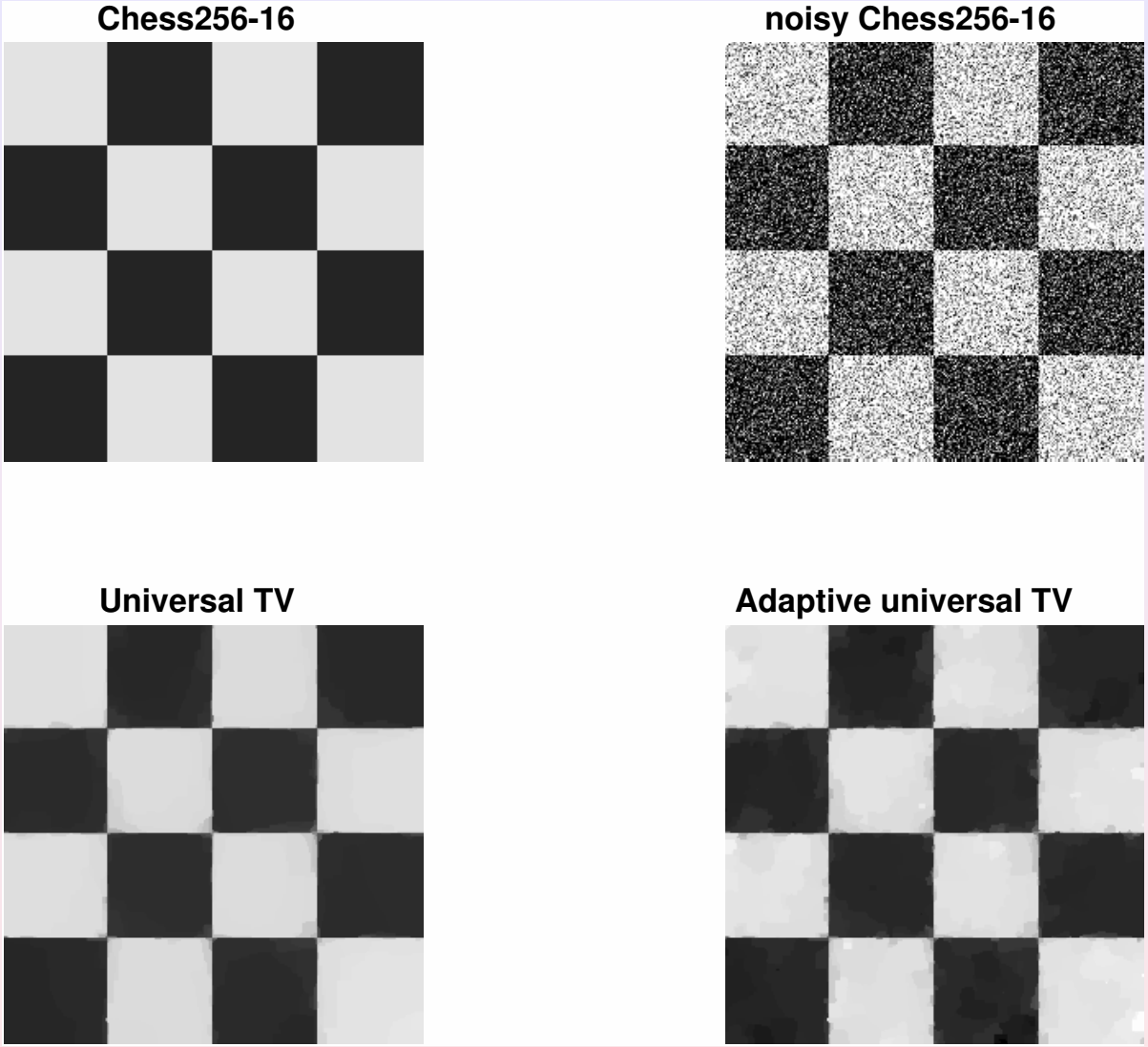
The reconstruction $\hat{\mathbf{p}} = (\hat{p}_1, \ldots, \hat{p}_{2000})$ is obtained by solving the *total variation* optimization problem

$$\min_{\mathbf{p} \in \mathbb{R}^{2000}} f(\mathbf{p}),$$

with $f(\mathbf{p}) = -\log L(\mathbf{p}; \mathbf{Y}) + \lambda \sum_{t=1}^{1999} |p_{t+1} - p_t|$.

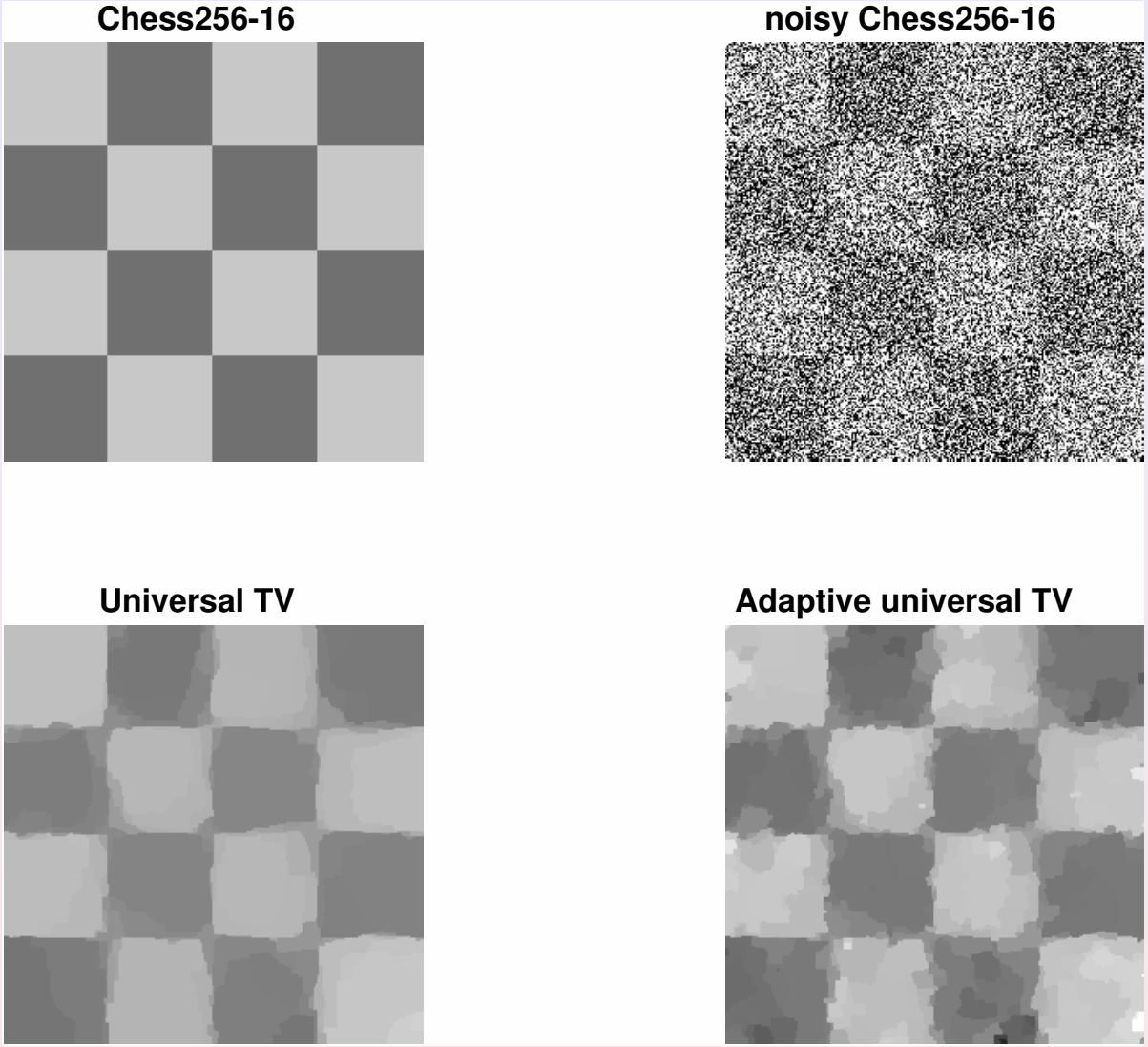
(L. Rudin , S. Osher and E. Fatemi (1992), *Nonlinear total variation based noise removal algorithms*).

Total variation is also employed in image processing



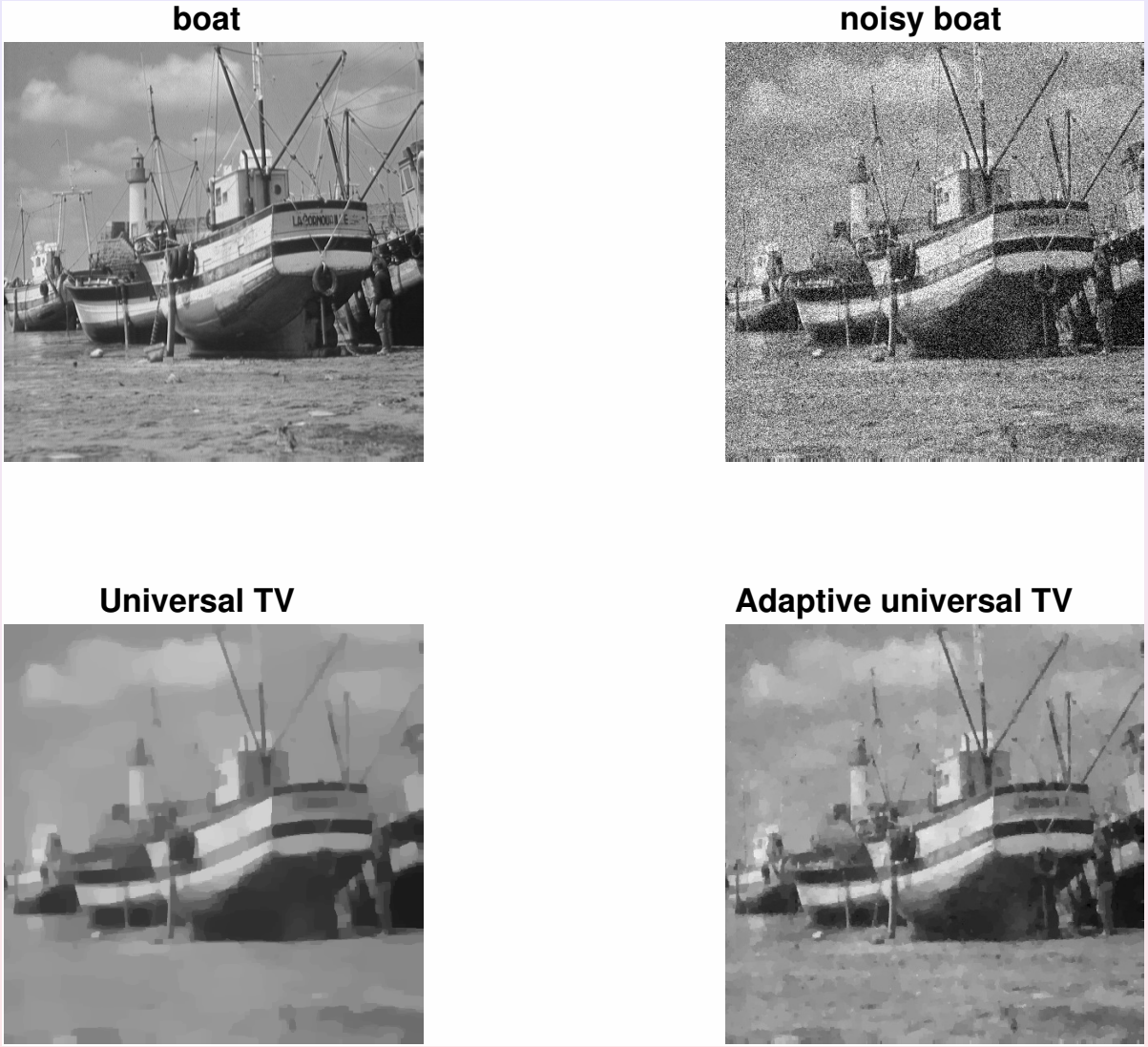
Gaussian image data $Y_{i,j} \sim \mathrm{N}(\mu_{i,j}, \eta)$.

Total variation is also employed in image processing



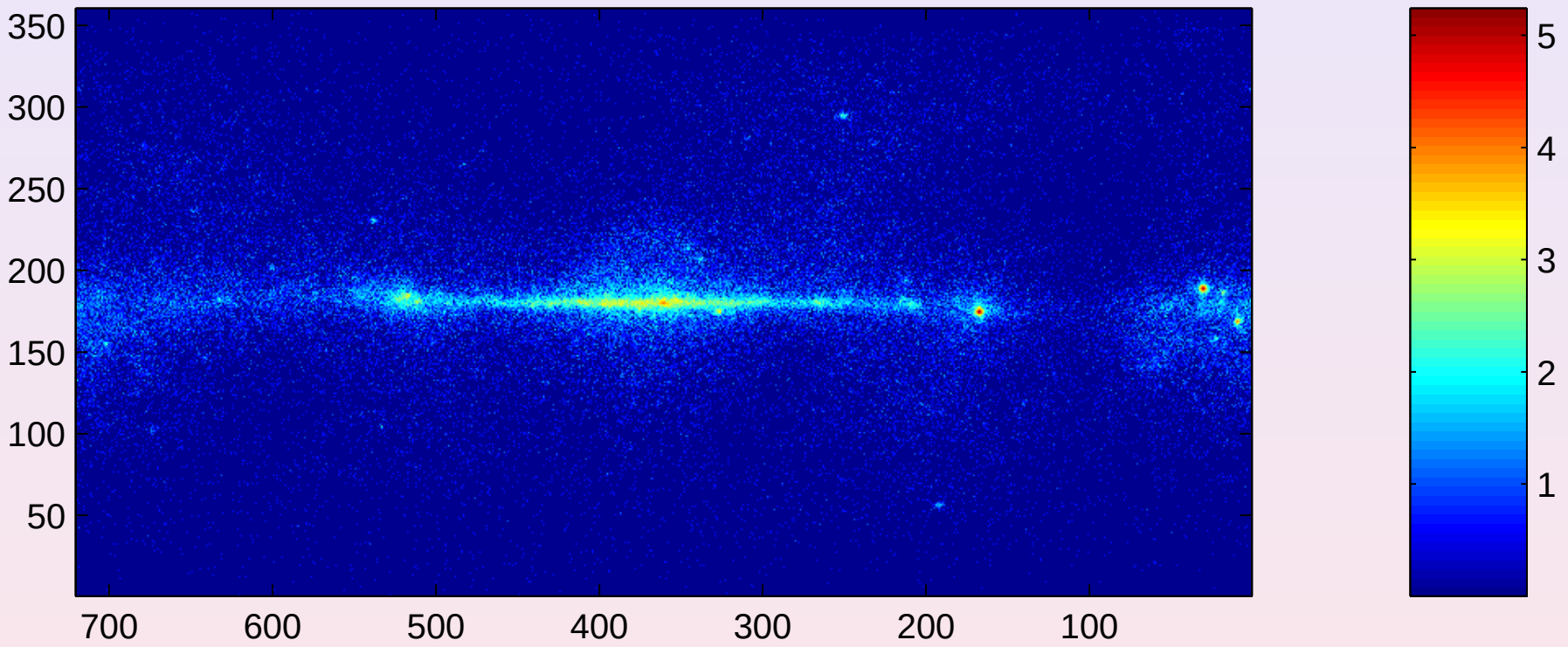
Gaussian image data $Y_{i,j} \sim \mathrm{N}(\mu_{i,j}, \eta)$.

Total variation can also be employed to non-piecewise constant images to *denoise* the noisy image.



Gaussian image data $Y_{i,j} \sim \mathrm{N}(\mu_{i,j}, \eta)$.

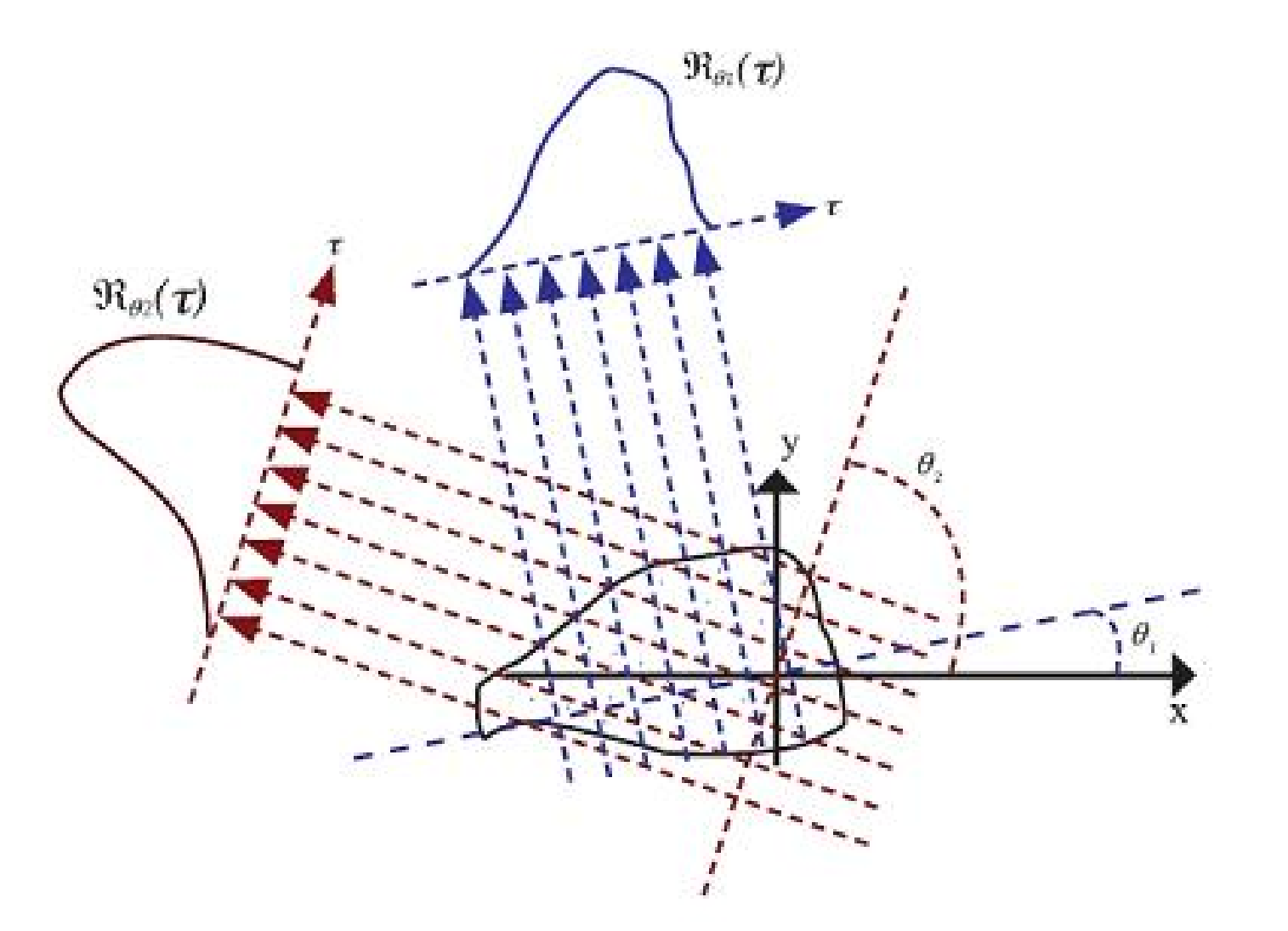
Denoising Energetic Gamma Ray Experiment Telescope data.



Poisson image data $Y_{i,j} \sim \mathrm{Poisson}(\mu_{i,j})$.

Data set 7: compressed sensing
[Candes, Donoho, Tao, Tanner (2006)]

Positron emission tomography scanner



Positron emission tomography integration lines

Mathematical model (noiseless case):

$$\mathbf{y} = X\boldsymbol{\mu},$$

where X is an $n \times N$ matrix with $n < N$.
Goal: retrieve $\boldsymbol{\mu}$ from a linear system of less equations than unknowns!

Additional assumption: $\boldsymbol{\mu}$ is *sparse* with k out of n non-zero entries.

By solving

$$\min_{\boldsymbol{\mu}} \|\boldsymbol{\mu}\|_1 \quad \text{subject to (s.t.)} \quad \mathbf{y} = X\boldsymbol{\mu},$$

one may be able to retrieve $\boldsymbol{\mu}$ exactly!

Mathematical model (noiseless case):

$$\mathbf{y} = X\boldsymbol{\mu},$$

where X is an $n \times N$ matrix with $n < N$.
Goal: retrieve $\boldsymbol{\mu}$ from a linear system of less equations than unknowns!

Additional assumption: $\boldsymbol{\mu}$ is *sparse* with k out of n non-zero entries.

By solving

$$\min_{\boldsymbol{\mu}} \|\boldsymbol{\mu}\|_1 \quad \text{subject to (s.t.)} \quad \mathbf{y} = X\boldsymbol{\mu},$$

one may be able to retrieve $\boldsymbol{\mu}$ exactly!

Data set 8: Netflix price.

Movie	1	2	3	4	...	*P*
User 1	?	?	1	?	...	
User 2	?	?	?	3	...	
User 3	1	?	?	5	...	
⋮						
User N	?	2	?	?	...	

where $N = 500'000$, $P = 20'000$, 99% ratings missing, rating$\in \{1, 2, 3, 4, 5\}$.

Goal: $1 M prize for 10% improvement in filling '?' correctly.

Proposal to recover the entire matrix X from partial observations Y_{np}:

$$\min_{X \in \mathcal{M}_{N \times P}} \sum_{\text{observed } (n,p)} (Y_{np} - X_{np})^2 \quad \text{s.t.} \quad \text{rank}(X) = r.$$

Combinatorially very hard.

Instead, convexify the problem by solving

$$\min_{X \in \mathcal{M}_{N \times P}} \sum_{\text{observed } (n,p)} (Y_{np} - X_{np})^2 \quad \text{s.t.} \quad \|X\|_* \leq \lambda,$$

where $\|X\|_* = \sum_j \lambda_j(X)$ is the nuclear norm of X (sum of singular values of X).

3. Summary

Many statistical estimators $\hat{\mathbf{x}}$ are defined as minimizer of a function

$$\begin{aligned} f: \quad X &\to \mathbb{R} \\ \mathbf{x} &\to f(\mathbf{x}) \end{aligned}$$

where:

- f is the *cost function*
- X is the *feasible set*.

Notation: $f^* = \min_{\mathbf{x} \in X} f(\mathbf{x})$ and $\hat{\mathbf{x}} = \arg\min_{\mathbf{x} \in X} f(\mathbf{x}) =: \mathbf{x}^*$.

Many statistical estimators are based on f of the form

$$f(\mathbf{x}) = \ell(\mathbf{x}; Y) + \lambda \operatorname{pen}(\mathbf{x})$$

inspired by Tikhonov (1963) regularization.
They can be seen as a *maximum a posteriori* estimator where $\lambda \operatorname{pen}(\mathbf{x})$ is the negative log-prior distribution.

Most Bayesian estimators are rather based on *posterior mean* calculated by MCMC methods.

4. What you will learn and not learn in this course

I expect you to be proactive, ask questions and have ideas.

Will learn:

- mathematical optimization issues, e.g., existence, uniqueness, convexity, optimality conditions
- descent algorithms
- non-differentiable convex optimization

Will not learn:

- discrete optimization
- non-descent algorithms (simulated annealing, genetic algorithms)
- how to prove convergence of algorithms