Convex optimisation with applications in statistics

Sylvain Sardy

3ème cycle, EPFL
July 30, 2006

1. Motivation

Maximum likelihood is a widely used estimation principle (Fisher, 1912)

$$\max_{\boldsymbol{\alpha}\in\Omega} \Pi_n f(Y_n, \boldsymbol{\alpha}).$$

Theorem: Under certain regularity conditions including:

- differentiability,
- $\boldsymbol{\alpha}_0 \in \mathrm{int}(\Omega)$,

then, with $\mathrm{P} \overset{n\to\infty}{\longrightarrow} 1$, the likelihood equations

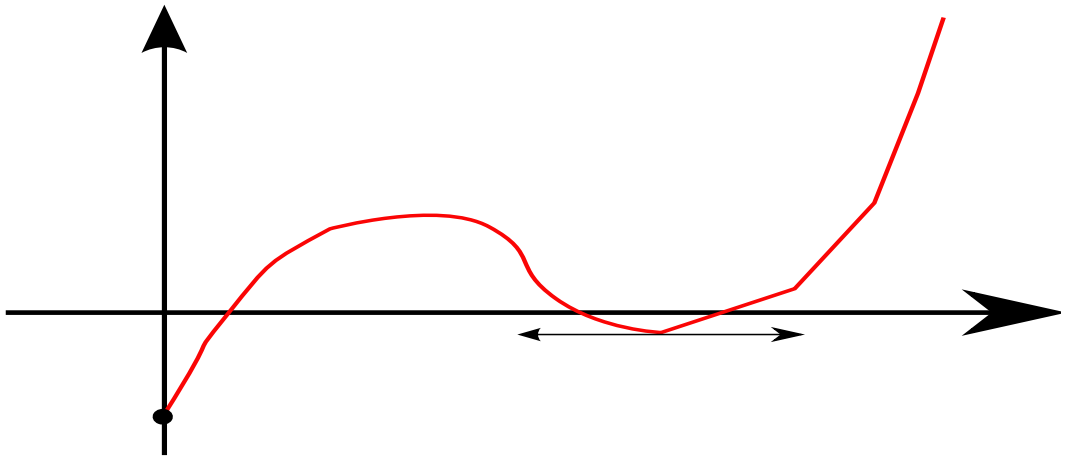
$$\nabla_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}, \mathbf{Y}) = \sum_n \nabla_{\boldsymbol{\alpha}} \log f(Y_n, \boldsymbol{\alpha}) = \mathbf{0}$$

have a root $\hat{\boldsymbol{\alpha}}_n$ (among possibly many) that is consistent.

– The theory does not tell us *which root to choose* in practice. It would be convenient if the root was *uniquely defined*.

Solution: *strict convexity*.

– Even if differentiable, the likelihood equations are *not necessary nor sufficient conditions* to define the maximum.



$\hat{\boldsymbol{\alpha}} \notin \text{int}(\Omega)$

Solution: *coercivity*.

Model selection is a recurrent Grail in regression, classification, latent variables, times series, etc.

Regression example 1.1: $N = 97$ men are about to receive a prostatectomy. Data: antigen level Y_n and $P = 8$ covariates X_{np} (e.g., age, prostate weight, ...)

Goal: predict the antigen's level with many $\hat{\alpha}_p = 0$ in

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

The maximum likelihood estimation of $\boldsymbol{\alpha}$ solves

$$\min_{\boldsymbol{\alpha}} -l(\boldsymbol{\alpha}, \mathbf{Y}) = \min_{\boldsymbol{\alpha}} \|\mathbf{Y} - X\boldsymbol{\alpha}\|_2^2,$$

but all $\hat{\alpha}_p \neq 0$.

```
> lm(y~X)

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept)  6.281e-17  7.193e-02 8.73e-16  1.00000
Xcrlcavol    6.919e-01  1.036e-01    6.677 2.11e-09 ***
Xcrlweight   2.257e-01  8.443e-02    2.673  0.00896 **
Xcrage      -1.462e-01  8.318e-02   -1.758  0.08229 .
Xcrlbph      1.553e-01  8.480e-02    1.832  0.07040 .
Xcrsvi       3.172e-01  1.011e-01    3.136  0.00233 **
Xcrlcp      -1.475e-01  1.273e-01   -1.159  0.24964
Xcrgleason   3.259e-02  1.137e-01    0.287  0.77506
Xcrpgg45     1.276e-01  1.247e-01    1.024  0.30885
```

Model selection: keep **'s and drop the others.

Drawback: ignores collinearities.

Discrete optimization: *best subset variable selection*

$$\min_{\boldsymbol{\alpha}} -l(\boldsymbol{\alpha}, \mathbf{Y}) \quad \text{s.t.} \quad k \text{ coefficients } \alpha_p = 0.$$

Equivalent to

$$\min_{\boldsymbol{\alpha}} -l(\boldsymbol{\alpha}, \mathbf{Y}) + \lambda \|\boldsymbol{\alpha}\|_0,$$

where $\|\boldsymbol{\alpha}\|_0 = \sum_p \alpha_p^0$ with the convention that $0^0 = 0$.

Drawbacks:

- Unstable.
- C_k^P MLEs must be calculated for $k = 0, 1, \ldots, P$.

Continuous optimization: ℓ_1 *penalized likelihood*

$$\min_{\boldsymbol{\alpha}} -l(\boldsymbol{\alpha}, \mathbf{Y}) + \lambda \|\boldsymbol{\alpha}\|_1,$$

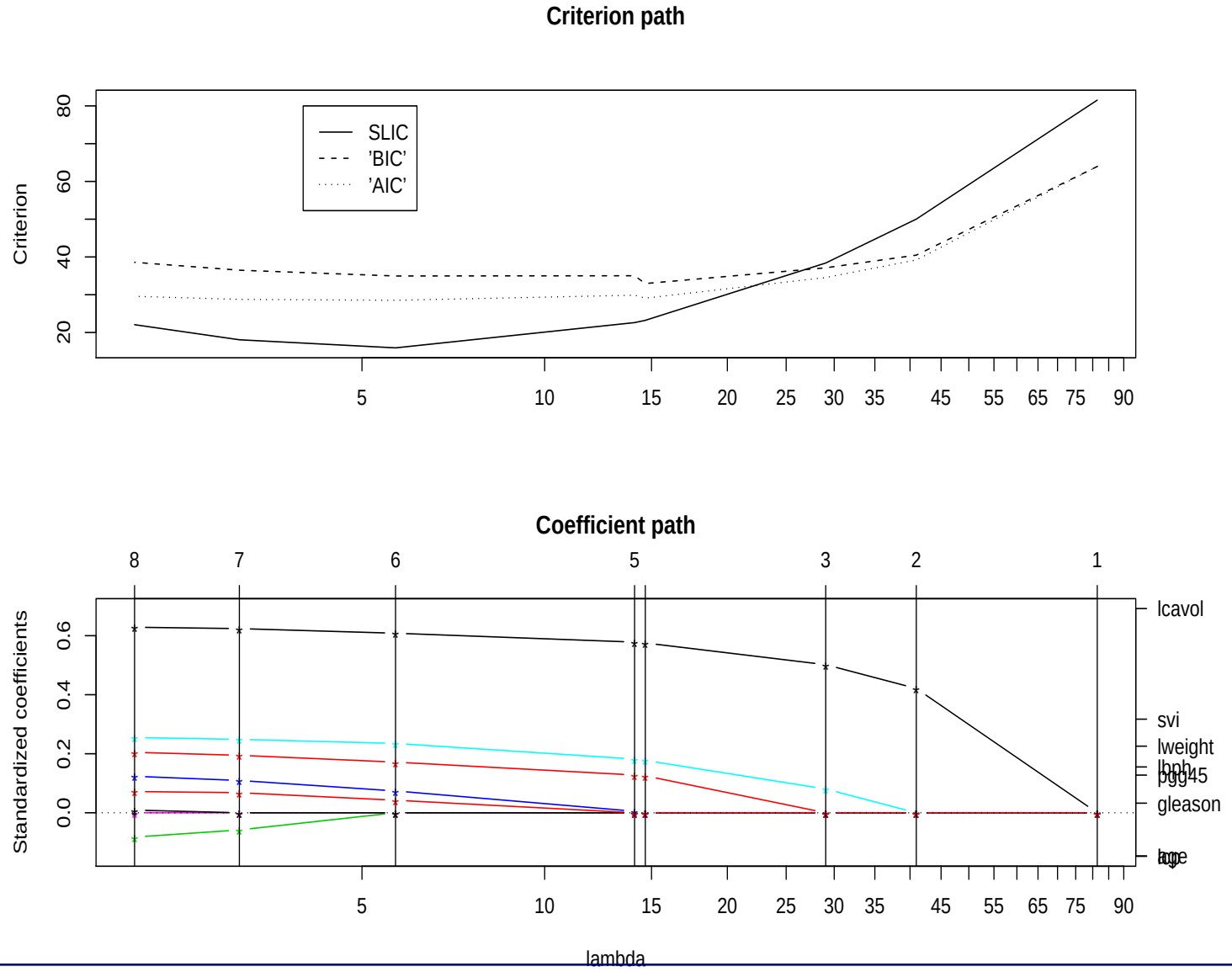
where $\|\boldsymbol{\alpha}\|_1 = \sum_p |\alpha_p|$.

Under some *strict convexity* and *coercivity* assumptions:

- model selection is guaranteed, i.e., some $\hat{\alpha}_p(\lambda) = 0$,
- existence and uniqueness are guaranteed.

Example 1.1 revisited: Lasso (Tibshirani 1996)

$$\min_{\alpha_0, \boldsymbol{\alpha}} \frac{1}{2} \|\mathbf{Y} - \alpha_0 - X\boldsymbol{\alpha}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_1.$$


Criterion path
Criterion
SLIC
'BIC'
'AIC'
20
40
60
80
5
10
15
20
25
30
35
45
55
65
75
90
Coefficient path
Standardized coefficients
0.0
0.2
0.4
0.6
8
7
6
5
3
2
1
lcavol
svi
lweight
gleason
lambda

Many estimators are defined as the solution of an optimization problem

$$\min_{\mathbf{x}} \mathrm{loss}(\mathbf{x}, \mathbf{Y}) + \mathrm{pen}(\mathbf{x}),$$

where:

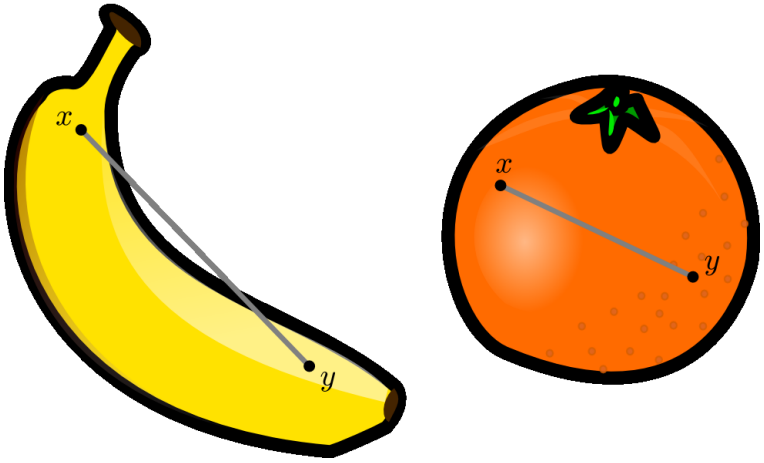
- $\mathrm{loss}()$ is a loss function/goodness-of-fit to the data,
- $\mathrm{pen}()$ is a penalty function/prior.

The mathematical properties (e.g., convexity, coercivity, differentiability) of these functions provide the corresponding estimator with different statistical properties (e.g., existence, uniqueness, model selection (thresholding)).

2. Convexity

Definition of convex set: A set $C \subset R^N$ is convex if for every $\mathbf{x} \in C$ and $\mathbf{y} \in C$, the line segment joining $\mathbf{x}$ and $\mathbf{y}$ also lies in C,
i.e., for every $\lambda \in [0, 1]$, then $\lambda\mathbf{x} + (1 - \lambda)\mathbf{y} \in C$.

Example 2.1: a banana is not convex, an orange is.



Example 2.2: Let $A_{M\times N}$ be a matrix and $\mathbf{b} \in R^M$.

Then $S = \{\mathbf{x} \in R^N : \ A\mathbf{x} \geqslant \mathbf{b}\}$ is convex.

Application: estimation of Poisson intensities $\boldsymbol{\mu} = A\mathbf{x} \geqslant \mathbf{0}$.

Property 2.1: Let C_i be convex sets.

Then the intersection $\bigcap C_i$ is also convex.

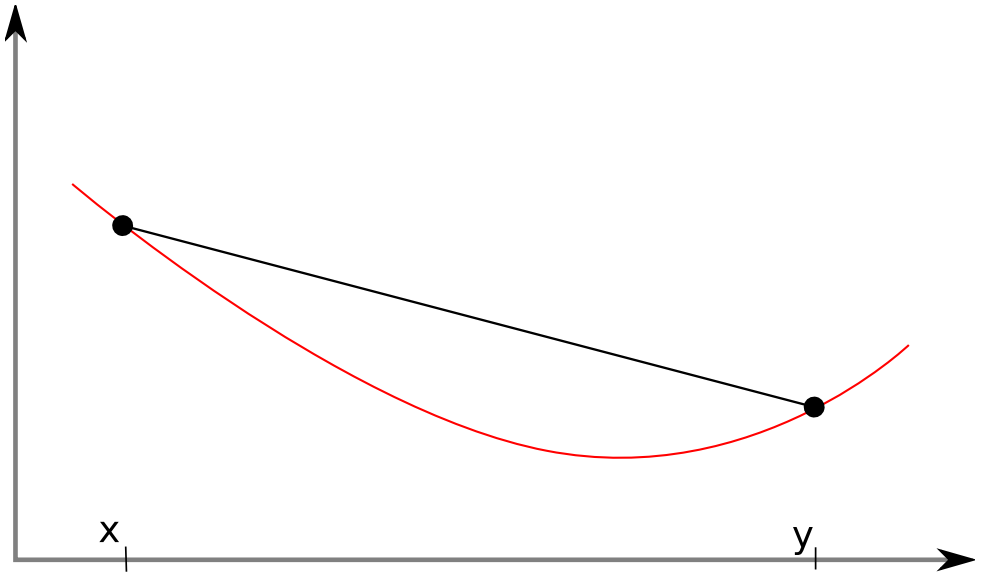
Application: estimation of a density $\mathbf{f} \geqslant \mathbf{0}$ such that $\mathbf{w}'\mathbf{f} = 1$.

Definition of convex function: A function $f(x)$ defined on a convex set C

- is strictly convex on C if

$$f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) < \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$$

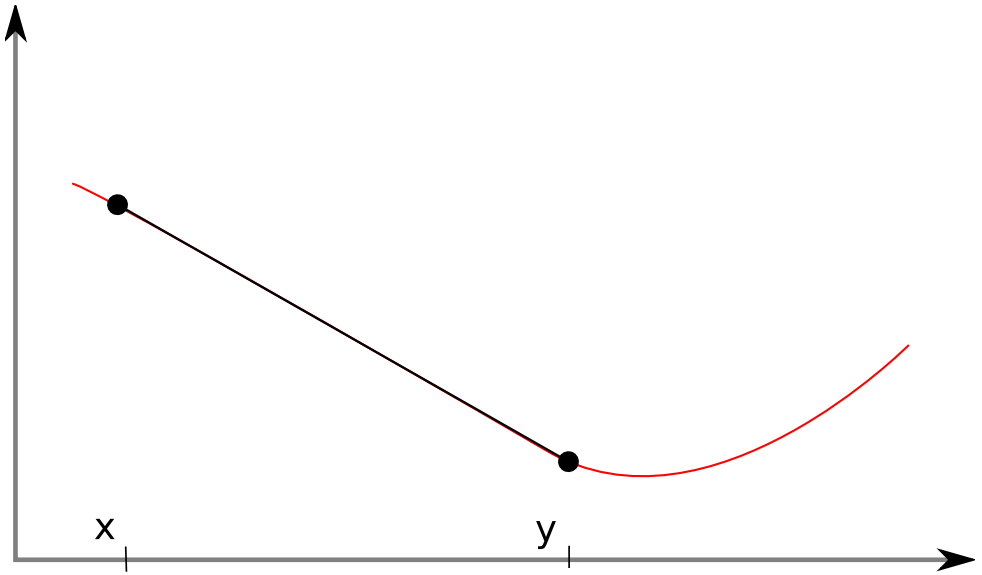
for all $\mathbf{x}$, $\mathbf{y}$ in C with $\mathbf{x} \neq \mathbf{y}$ and all $0 < \lambda < 1$.



- is convex on C if

$$f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leqslant \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$$

for all $\mathbf{x}$, $\mathbf{y}$ in C and all $0 \leqslant \lambda \leqslant 1$;



Examples 2.3 of $\mathrm{loss}(x; Y)$ and $\mathrm{pen}(x)$ functions:

- Gaussian negative log-likelihood:

$$-l(x; Y) = \frac{1}{2}(x - Y)^2$$

- ℓ_1 and ℓ_2 penalties:

$$\mathrm{pen}(x) = |x| \quad \text{and} \quad \mathrm{pen}(x) = x^2$$

- Poisson negative log-likelihood:

$$-l(x; Y) = -Y \log x + x$$

- Support vector machine regression loss:

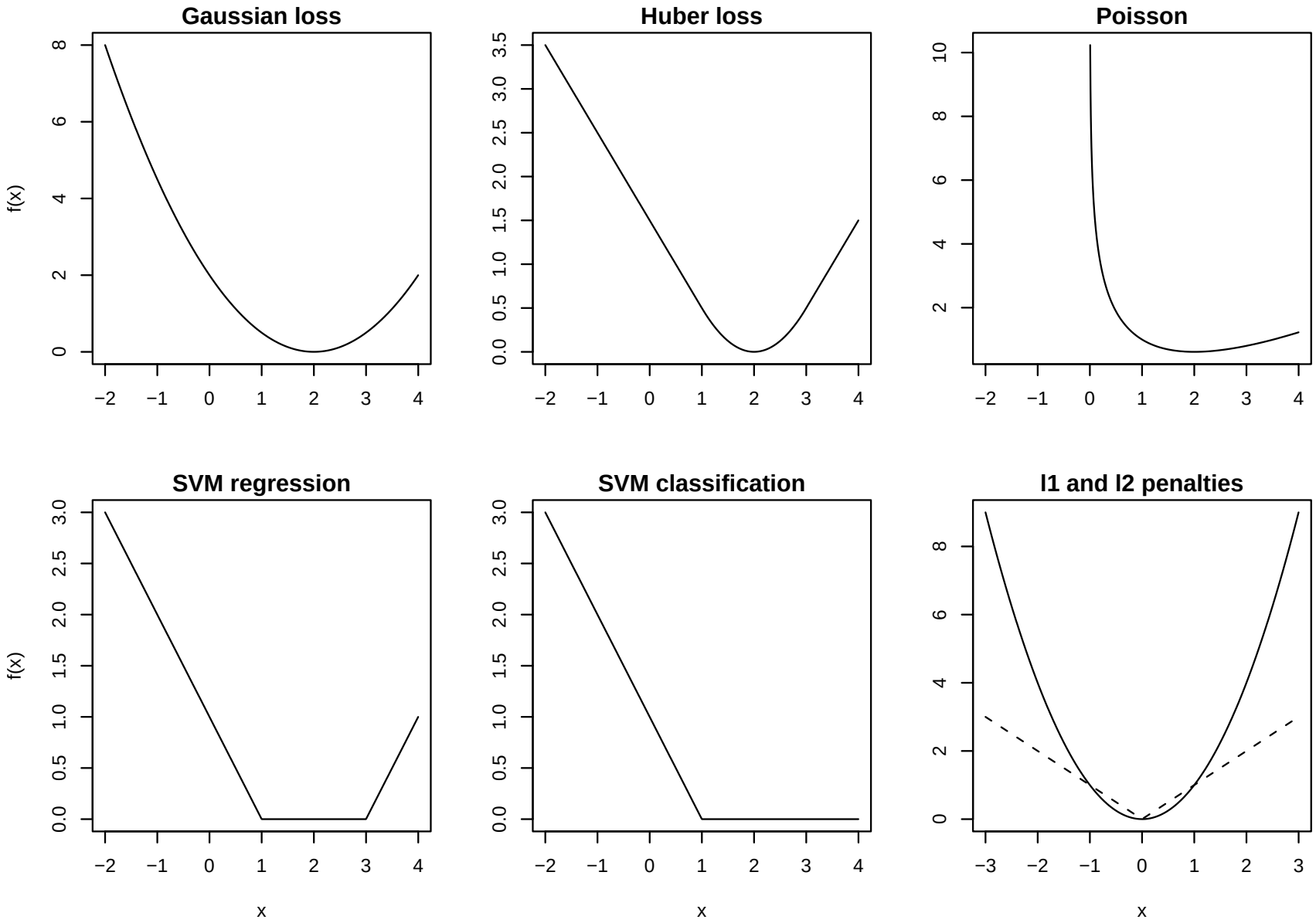
$$\text{loss}(x, Y) = \begin{cases} 0 & |x - Y| \leqslant \tau \\ |x - Y| - \tau & \text{else} \end{cases}$$

- Support vector machine classification loss:

$$\text{loss}(x, Y) = (1 - Yx)_+$$

- Robust regression with Huber loss function

$$\text{loss}(x, Y) = \begin{cases} (x - Y)^2/2 & |x - Y| \leqslant \tau \\ \tau|x - Y| - \tau^2/2 & \text{else} \end{cases}$$


Gaussian loss
Huber loss
Poisson
SVM regression
SVM classification
l1 and l2 penalties
f(x)
x

Property 2.2: Suppose $\partial^2 f / \partial x_i x_j$ are continuous on C open and convex. If the Hessian of f is $Hf(\mathbf{x}) \geqslant 0$ (resp. $Hf(\mathbf{x}) > 0$) on C, then f is $convex$ (resp. $strictly\ convex$) on C.

Note: the reciprocal is not true for strict convexity, e.g., $f(x) = x^4$.

Property 2.3: Every convex function on C open and convex is continuous.

Property 2.4: If $f_1(\mathbf{x}), \ldots, f_k(\mathbf{x})$ are convex functions on C convex, then

$$f(\mathbf{x}) = f_1(\mathbf{x}) + \cdots + f_k(\mathbf{x})$$

is convex. Moreover, if at least one $f_i(\mathbf{x})$ is strictly convex, then so is $f(\mathbf{x})$.

Application: Poisson ℓ_1-penalized likelihood

$$f(\mathbf{x}) = \sum_n -Y_n \log x_n + x_n + \lambda \|\mathbf{x}\|_1$$

is strictly convex provided that at least one $Y_n \neq 0$.

Property 2.5: Let f be convex and A be a matrix. Then $f \circ A$ is convex.

If the convexity is strict and $\mathrm{Ker}(A) = \{\mathbf{0}\}$, then $f \circ A$ is strictly convex.

Application: Generalized linear models.

Theorem 2.1: Any local minimizer of a $convex$ function $f(\mathbf{x})$ defined on a $convex$ set $C \in R^n$ is also a $global$ minimizer.

If the convexity is $strict$, then the global minimizer is $unique$.

Theorem 2.2: If f is convex and $\partial f/\partial x_i$ are continuous on C convex, then any critical point of f is a global minimizer.

Application: if it exists, the solution to the likelihood equations

$$\nabla_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}, \mathbf{Y}) = \sum_n \nabla_{\boldsymbol{\alpha}} \log f(Y_n, \boldsymbol{\alpha}) = \mathbf{0}$$

provides a maximum likelihood estimate, unique if the convexity is strict.

Warning: Convexity is more the exception than the rule.

But you can maybe choose the likelihood, the link function and the priors such that (parts of) the posterior likelihood is convex.

3. Coercivity

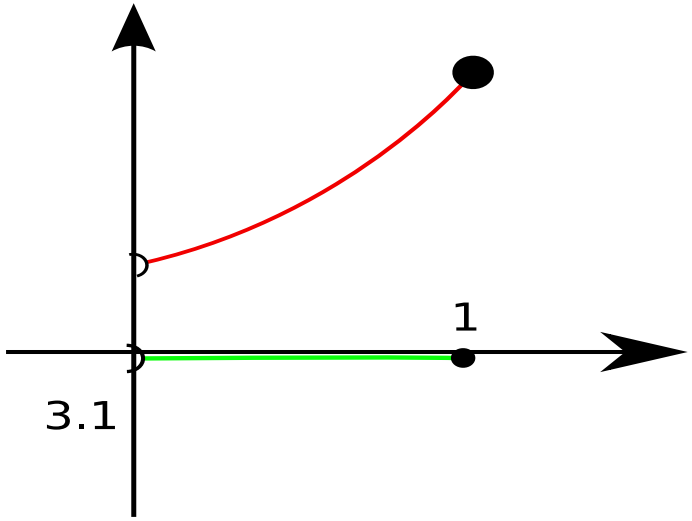
Convexity does not tell us whether a solution/estimate exists, but rather tells the conditions and properties it may satisfy, if it exists.

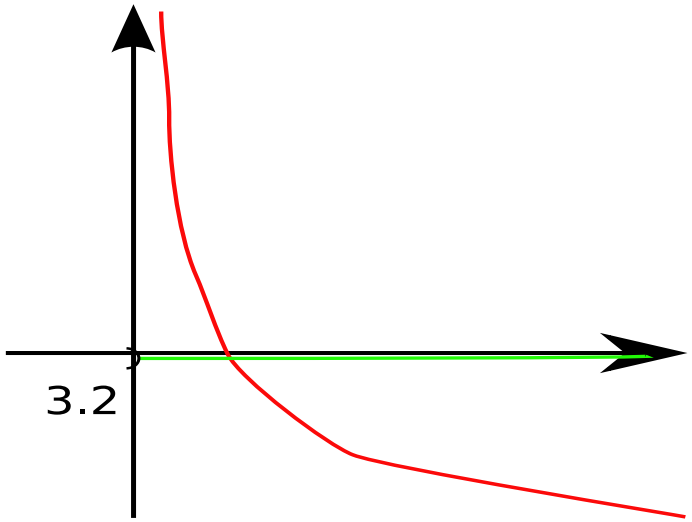
Weierstrass Theorem: let D be a compact subset of R^P. If f is continuous on D, then f has a global maximizer and minimizer on D.

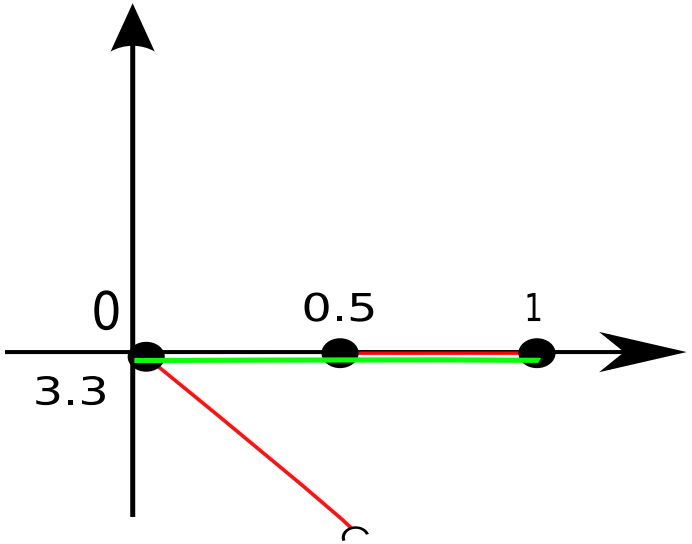
Counter-example 3.1: $f(x) = \exp(x)$ on $D = (0, 1]$.

Counter-example 3.2: $f(x) = -\log(x)$ on $D = R^+$.

Counter-example 3.3: $f(x) = -x \cdot 1(x < 0.5)$ on $D = [0, 1]$.


1
3.1


3.2


0
0.5
1
3.3

In statistics, the domain is rarely compact, e.g., $D = R$, $D = (0, \infty)$.

Definition of coercive function: a continuous f on R^N is *coercive* if

$$\lim_{\|\mathbf{x}\| \to \infty} f(\mathbf{x}) = +\infty.$$

Example 3.1: The Gaussian loss $f(x) = \frac{1}{2}(x - Y)^2$ is coercive.

Example 3.2: The Poisson loss $f(x) = -Y \log x + x$ is coercive on $D = (0, \infty)$ provided $Y \in \{1, 2, 3, \ldots\}$.

Theorem: let f be continuous on R^N. If f is coercive, then f has a global minimizer.

Moreover, coercivity implies that $\hat{\boldsymbol{\alpha}} \in \text{int}(\Omega)$.

4. Statistical estimators

Least squares estimate: $\min_{\boldsymbol{\alpha}} \|\mathbf{Y} - X\boldsymbol{\alpha}\|_2^2$.

Existence: f is coercive if $\mathrm{Ker}(X) = \{\mathbf{0}\}$, and continue.

Uniqueness: The cost function is convex, and strictly convex if $\mathrm{Ker}(X) = \{\mathbf{0}\}$.

Strict convexity is lost if $\mathrm{Ker}(X) \neq \{\mathbf{0}\}$ (e.g., microarray applications).

Why: take $\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 + \epsilon\mathbf{k}$ with $\mathbf{k} \in \mathrm{Ker}(X)$.
Since $X\boldsymbol{\alpha}_1 = X\boldsymbol{\alpha}_2$, then $f(\lambda\boldsymbol{\alpha}_1 + (1-\lambda)\boldsymbol{\alpha}_2) = f(\boldsymbol{\alpha}_1) = f(\boldsymbol{\alpha}_2)$ for all $\lambda \in (0, 1)$.

The gradient vector is continuous and R^P is convex, so any solution to the normal equations

$$\nabla_{\boldsymbol{\alpha}} l(\boldsymbol{\alpha}, \mathbf{Y}) = -X'\mathbf{Y} + X'X\boldsymbol{\alpha} = \mathbf{0}$$

is a least squares solution, unique if $\mathrm{Ker}(X) = \{\mathbf{0}\}$.

If $\mathrm{Ker}(X) \neq \{\mathbf{0}\}$, a unique solution can be defined by choosing the one of minimum ℓ_2 norm: the Moore-Penrose solution.

Generalized linear model: $\min_{\boldsymbol{\alpha}\in\Omega} -l(\boldsymbol{\mu} = g^{-1}(X\boldsymbol{\alpha}); \mathbf{Y})$,
where g is the link function.

Existence: $-l \circ g^{-1}$ coercive and continue.

Uniqueness: $-l \circ g^{-1}$ strictly convex, Ω convex and $\mathrm{Ker}(X) = \{\mathbf{0}\}$.

Example 4.1: Poisson GLM and $\mathrm{Ker}(X) = \{\mathbf{0}\}$

- log-link: $-l \circ g^{-1}(\boldsymbol{\eta}) = \sum_n -Y_n\eta_n + \exp(\eta_n)$ is strictly convex and $\Omega = R^P$ is convex.
- identity link: $-l \circ g^{-1}(\boldsymbol{\mu}) = \sum_n -Y_n \log \mu_n + \mu_n$ is strictly convex and $X\boldsymbol{\alpha} \geqslant 0$ is convex.

Ridge regression: $\min_{\boldsymbol{\alpha}} \|\mathbf{Y} - X\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_2^2$, $\lambda > 0$.

Existence: by coercivity and continuity.

Uniqueness: by strict convexity. In fact $\hat{\boldsymbol{\alpha}}_\lambda = (X'X + \lambda I)^{-1}X'\mathbf{Y}$.

Lasso: $\min_{\boldsymbol{\alpha}} \|\mathbf{Y} - X\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1$, $\lambda > 0$.

Existence: by coercivity and continuity

Uniqueness: guaranteed if $\mathrm{Ker}(X) = \{\mathbf{0}\}$. Otherwise $\mathrm{Ker}(X)$ may intersect the faces of the ℓ_1 ball.

Thresholding: for a *finite* $\lambda_{\mathbf{Y}} < \infty$, then $\hat{\boldsymbol{\alpha}}_{\lambda_{\mathbf{Y}}} = \mathbf{0}$ is the global minimum. ℓ_1 penalized likelihood does variable selection.

Proof: let $\mathbf{g} = \nabla_{\boldsymbol{\alpha}} - l \circ X = -X'\mathbf{Y} + X'X\boldsymbol{\alpha}$ be the gradient of the likelihood. The ℓ_1 penalized likelihood is not differentiable, but its *generalized* gradient can be defined everywhere as

$$r_p(\boldsymbol{\alpha}) = \begin{cases} |g_p(\boldsymbol{\alpha}) + \lambda \frac{\alpha_p}{|\alpha_p|}| & \text{if} \quad |\alpha_p| \neq 0; \\ \min_{0 \leqslant |\eta| \leqslant \lambda} |g_p(\boldsymbol{\alpha}) + \eta| & \text{if} \quad |\alpha_p| = 0. \end{cases}$$

A necessary and sufficient condition for $\hat{\boldsymbol{\alpha}} = \mathbf{0}$ to be a global minimizer is that $\mathbf{r}(\mathbf{0}) = \mathbf{0}$, i.e., $|g_p(\mathbf{0})| \leqslant \lambda$ for all p, i.e., $\lambda \geqslant \lambda_{\mathbf{Y}} = \|X'\mathbf{Y}\|_\infty$.

SVM regression: For the regression model

$$\mathbf{Y} = X\boldsymbol{\alpha} + \boldsymbol{\epsilon},$$

the SVM estimate of $\boldsymbol{\alpha}$ solves

$$\min_{\boldsymbol{\alpha}} \sum_{n=1}^{N} (|Y_n - \mathbf{x}_n'\boldsymbol{\alpha}| - \tau)_+ + \lambda \|\boldsymbol{\alpha}\|_2^2.$$

Existence: by coercivity and continuity.

Uniqueness: like ridge regression, the estimate is unique regardless of $\mathrm{Ker}(X)$.

Recently, Hastie proposed:

$$\min_{\boldsymbol{\alpha}} \sum_{n=1}^{N} (|Y_n - \mathbf{x}_n'\boldsymbol{\alpha}| - \tau)_+ + \lambda \|\boldsymbol{\alpha}\|_1.$$

ℓ_1-**Markov random field smoothing**. For the regression model

$$Y_n = f(x_n) + \epsilon_n,$$

a Markov random field assumption on $\mathbf{f}$ and Bayes theorem lead to solving

$$\min_{\mathbf{f}} \frac{1}{2}\|\mathbf{Y} - \mathbf{f}\|_2^2 + \lambda \sum_{n=2}^{N} |f_{n-1} - f_n|.$$

Existence and uniqueness.

5. Iterative algorithms

Gradient methods: Given a current iterate $\mathbf{x}_k$, get a new one with

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \delta_k D_k \nabla f(\mathbf{x}_k), \quad \delta_k \geqslant 0,$$

where:

- $\nabla f(\mathbf{x}_k)$ is the gradient
- $D_k > 0$ is a positive definite symmetric matrix
- δ_k is the stepsize

Example 5.1: steepest descent chooses $D_k = I$

Properties: simple, no overhead calculations, but slow.

Example 5.2: Pure Newton's method chooses $\delta_k = 1$ and $D_k = \left(\nabla^2 f(\mathbf{x}_k)\right)^{-1}$.

The idea of Newton's method is to minimize at each iteration the quadratic approximation of f around the current point x^k

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

Properties: converges very fast, but requires $\nabla^2 f(\mathbf{x})$ and $\left(\nabla^2 f(\mathbf{x})\right)^{-1}$.

Application 5.1: Poisson GLM with log-link. The gradient and Hessian are

$$\begin{aligned}
\nabla_{\boldsymbol{\alpha}} f(\boldsymbol{\alpha}) &= \sum_n -y_n \mathbf{x}_n + \exp(\mathbf{x}_n' \boldsymbol{\alpha}) \mathbf{x}_n \\
&= X'(\mathbf{y} - \exp(X\boldsymbol{\alpha})) \\
&= X'(\mathbf{y} - \mu^k) \\
\nabla^2 f(x^k) &= \sum_n \exp(\mathbf{x}_n' \boldsymbol{\alpha}) \mathbf{x}_n \mathbf{x}_n' \\
&= X'WX, \quad \text{with} \quad W = \operatorname{diag}(\exp(\mathbf{x}_n' \boldsymbol{\alpha})) > 0.
\end{aligned}$$

So

$$\boldsymbol{\alpha}^{k+1} = \boldsymbol{\alpha}^k - (XWX')^{-1}(X'(\mathbf{y} - \mu^k)).$$

This is the so-called Iterated Reweighted Least Squares.

Example 5.3: Discretized Newton's method chooses $D_k = \left(\hat{\nabla}^2 f(\mathbf{x}_k)\right)^{-1}$.

Example 5.4: Quasi-Newton's methods updates $D_{k+1} = D_k + \frac{\mathbf{p}_k \mathbf{p}_k'}{\mathbf{p}_k' \mathbf{q}_k} - \ldots$ where $\mathbf{p}_k = \mathbf{x}_{k+1} - \mathbf{x}_k$ and $\mathbf{q}_k = (\nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k))$.

In `R`: option "BFGS" (Broyden, Fletcher, Goldfarb and Shanno) in `optim`. Properties: best general purpose quasi-Newton method known, but requires storage of D_k and matrix-vector multiplication.

In `R`: option "Nelder-Mead" in `optim` is not a q-N method. Properties: not a gradient method, weak theoretical convergence, works reasonable well in small dimension.

In `R`: option "CG" (conjugate gradient) in `optim` is not a q-N method. Properties: no storage of D_k, but fragile (conjugacy progressively lost).

Newton's methods' drawbacks:

- f must be twice differentiable.
  E.g.: $f(x) = \frac{2}{3}|x|^{3/2}$ is strictly convex, but $x_k = (-1)^k$ does not converge.
- difficult to adapt to constrained optimization
- Problems when $\nabla^2 f(\mathbf{x}) \not\succ 0$.
- Requires:
  - calculating or estimating $\nabla_{\mathbf{x}} f(\mathbf{x})$
  - calculating or estimating $\nabla^2 f(\mathbf{x}_k)$ and $\left(\nabla^2 f(\mathbf{x}_k)\right)^{-1}$
  - storing them.

Coordinate descent methods: to solve

$$\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{x} \in X = X_1 \times \cdots \times X_P$$

solve a succession of easy sub-problems

$$\min_{\xi \in X_i} f(x_1, \ldots, X_{i-1}, \xi, x_{i+1}, \ldots, x_P)$$

Properties:

- easy to implement
- makes practical sense if closed-form and separability
- convergence properties similar to steepest descent
- can handle certain constraints and certain non-differentiability

Application 5.2: Lasso solves

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

$$= \min_{\boldsymbol{\alpha}_{-i},\alpha_i} \frac{1}{2}\|\mathbf{Y} - X_{-i}\boldsymbol{\alpha}_{-i} - \mathbf{x}_i\alpha_i\|_2^2 + \lambda\|\boldsymbol{\alpha}_{-i}\|_1 + \lambda|\alpha_i|$$

Choose i, let $\mathbf{r} = \mathbf{Y} - X_{-i}\boldsymbol{\alpha}_{-i}$ and $\tilde{y} = \frac{\mathbf{r}'\mathbf{x}_i}{\mathbf{x}_i'\mathbf{x}_i}$. The sub-problem in $\xi = \alpha_i$ is

$$\min_{\xi} \frac{1}{2}(\xi - \tilde{y})^2 + \frac{\lambda}{\mathbf{x}_i'\mathbf{x}_i}|\xi|.$$

The closed form solution is $\xi = \begin{cases} \tilde{y} - \lambda/(\mathbf{x}_i'\mathbf{x}_i) & \tilde{y} > \lambda/(\mathbf{x}_i'\mathbf{x}_i) \\ 0 & |\tilde{y}| \leqslant \lambda/(\mathbf{x}_i'\mathbf{x}_i) \\ \tilde{y} + \lambda/(\mathbf{x}_i'\mathbf{x}_i) & \tilde{y}_n < -\lambda/(\mathbf{x}_i'\mathbf{x}_i) \end{cases}$.

Theorem: Convergence to the global minimum by convexity and separability of non-differentiable part.

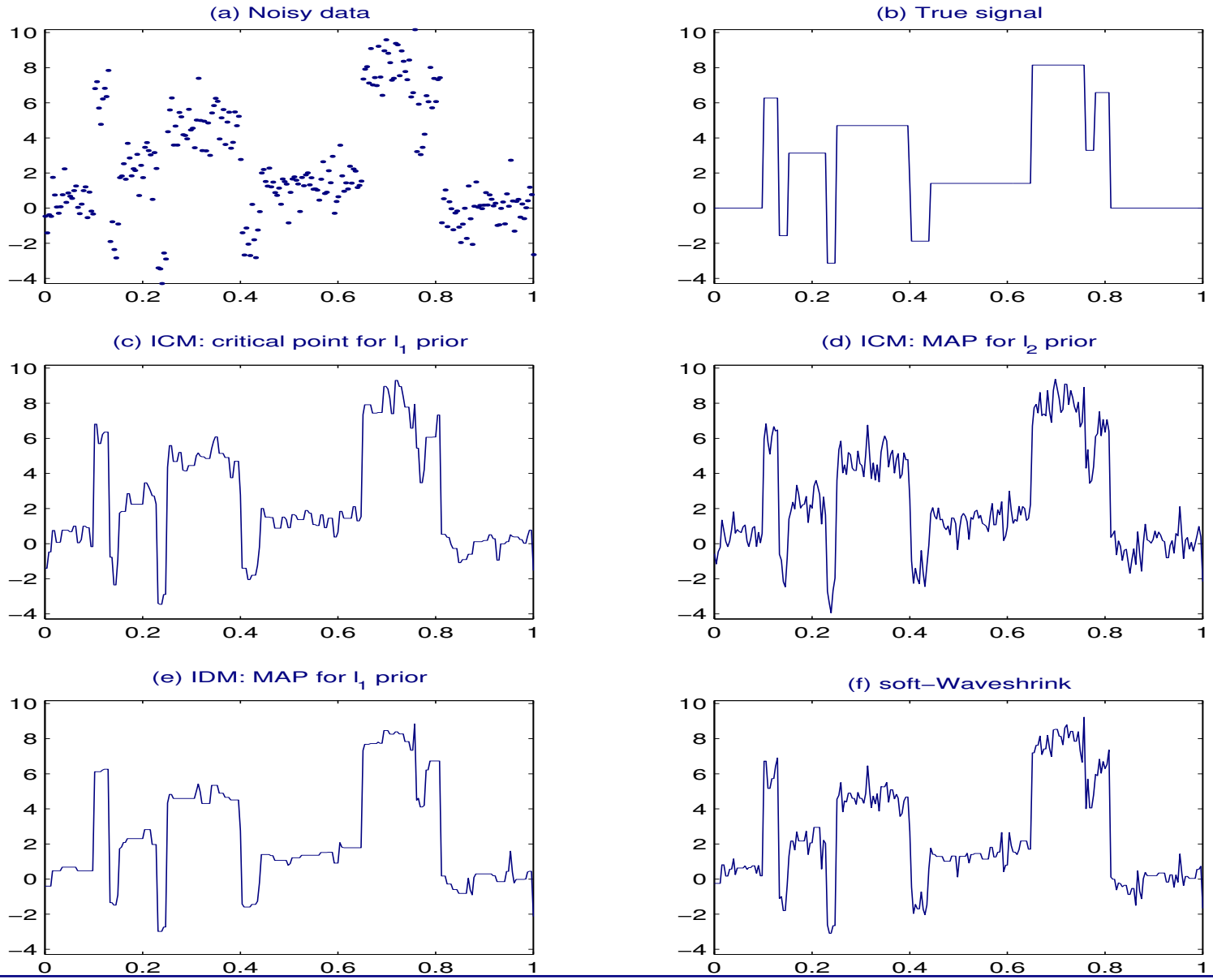
Application 5.3: ℓ_1 Markov random field smoother solves

$$\min_{\mathbf{f}} \frac{1}{2}\|\mathbf{Y}-\mathbf{f}\|_2^2 + \lambda \sum_{n=2}^{N} |f_{n-1} - f_n|.$$

Hans Künsch found that ICM (Besag, 1986), which is a coordinate descent method, does not converge to the global minimum, but to a critical point, despite strict convexity.

Problem: non-separability of the non-differential part.

Solution: use duality theory of convex programming to derive an equivalent optimization problem (the dual) to which we can apply coordinate descent methods.


(a) Noisy data
(b) True signal
(c) ICM: critical point for l_1 prior
(d) ICM: MAP for l_2 prior
(e) IDM: MAP for l_1 prior
(f) soft–Waveshrink

Convex program:

The estimators we saw are solutions to convex programs

$$\min_{\mathbf{x}} g(A\mathbf{x}) + \lambda \|B\mathbf{x}\| \quad \text{s.t.} \quad \begin{cases} C\mathbf{x} \geqslant \mathbf{0} \\ \mathbf{w}'\mathbf{x} = 1 \\ \mathbf{x} \in \Omega \end{cases}$$

with linear constraints.

Primal: suppose f convex in

$$(P) \qquad \min_{\mathbf{x}} f(\mathbf{x}) \quad \text{s.t.} \quad \begin{cases} E\mathbf{x} = \mathbf{e} \\ A\mathbf{x} \leqslant \mathbf{b} \\ \mathbf{x} \in \Omega \end{cases}$$

Define the Lagrangian function $L(\mathbf{x}, \mathbf{w}, \boldsymbol{\mu}) = f(\mathbf{x}) + \mathbf{w}'(\mathbf{E}\mathbf{x} - \mathbf{e}) + \boldsymbol{\mu}'(\mathbf{A}\mathbf{x} - \mathbf{b})$, and the dual function: $q(\mathbf{w}, \boldsymbol{\mu}) = \inf_{\mathbf{x} \in \Omega} L(\mathbf{x}, \mathbf{w}, \boldsymbol{\mu})$.

Dual:

$$(D) \qquad \max_{\mathbf{w}, \boldsymbol{\mu} \geqslant \mathbf{0}} q(\mathbf{w}, \boldsymbol{\mu})$$

Theorem: if (P) has a minimum at $\mathbf{x}^\star$, then (D) has a maximum at $(\mathbf{w}^\star, \boldsymbol{\mu}^\star)$ and the duality gap $f(\mathbf{x}^\star) - q(\mathbf{w}^\star, \boldsymbol{\mu}^\star) = 0$ (always $\geqslant 0$ in general).

Application 5.3 (rev.): ℓ_1 Markov random field smoother solves

- Gaussian case:

$$(P) \qquad \min_{\mathbf{f}} \frac{1}{2}\|\mathbf{Y}-\mathbf{f}\|_2^2 + \lambda \sum_{n=2}^{N} |f_{n-1}-f_n| \quad = \quad \min_{\mathbf{f}} \frac{1}{2}\|\mathbf{Y}-\mathbf{f}\|_2^2 + \lambda \|B\mathbf{f}\|_1,$$

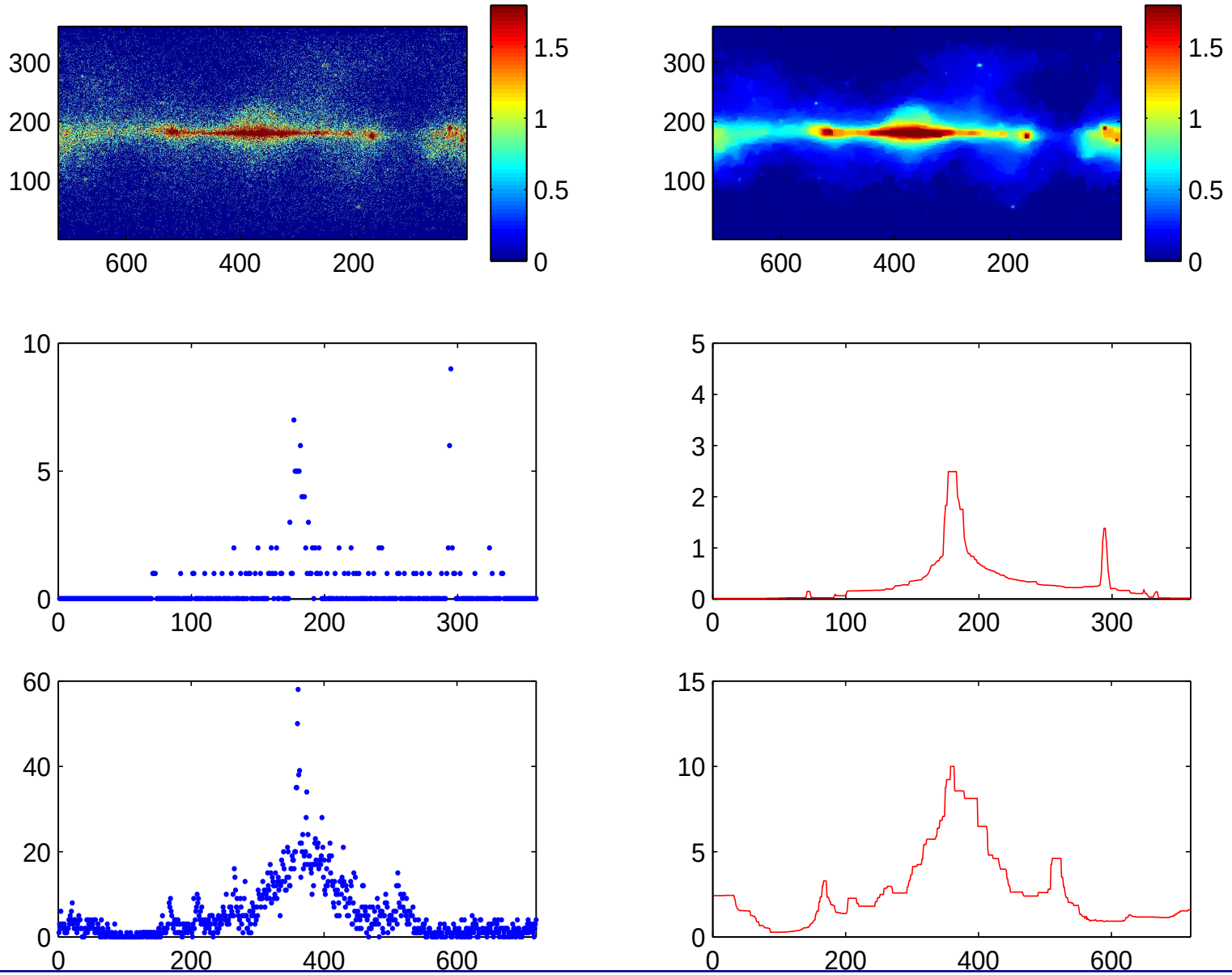
which has zero duality gap with

$$(D) \qquad \max_{\mathbf{w}} \mathbf{s}' B' \mathbf{w} + \frac{1}{2}\|B'\mathbf{w}\|_2^2 + \sum_n \begin{cases} 0 & \text{if } |w_n| \leqslant \lambda \\ \infty & \text{otherwise} \end{cases}.$$

- Poisson case:

$$(D) \qquad \max_{\|\mathbf{w}\|_\infty \leqslant \lambda} \mathbf{s}' \log(1 + B'\mathbf{w}).$$

A coordinate descent method converges quickly in R to the unique minimum.


300
200
100
600
400
200
1.5
1
0.5
0
10
5
0
0
100
200
300
5
4
3
2
1
60
40
20
0
200
400
600
15
10

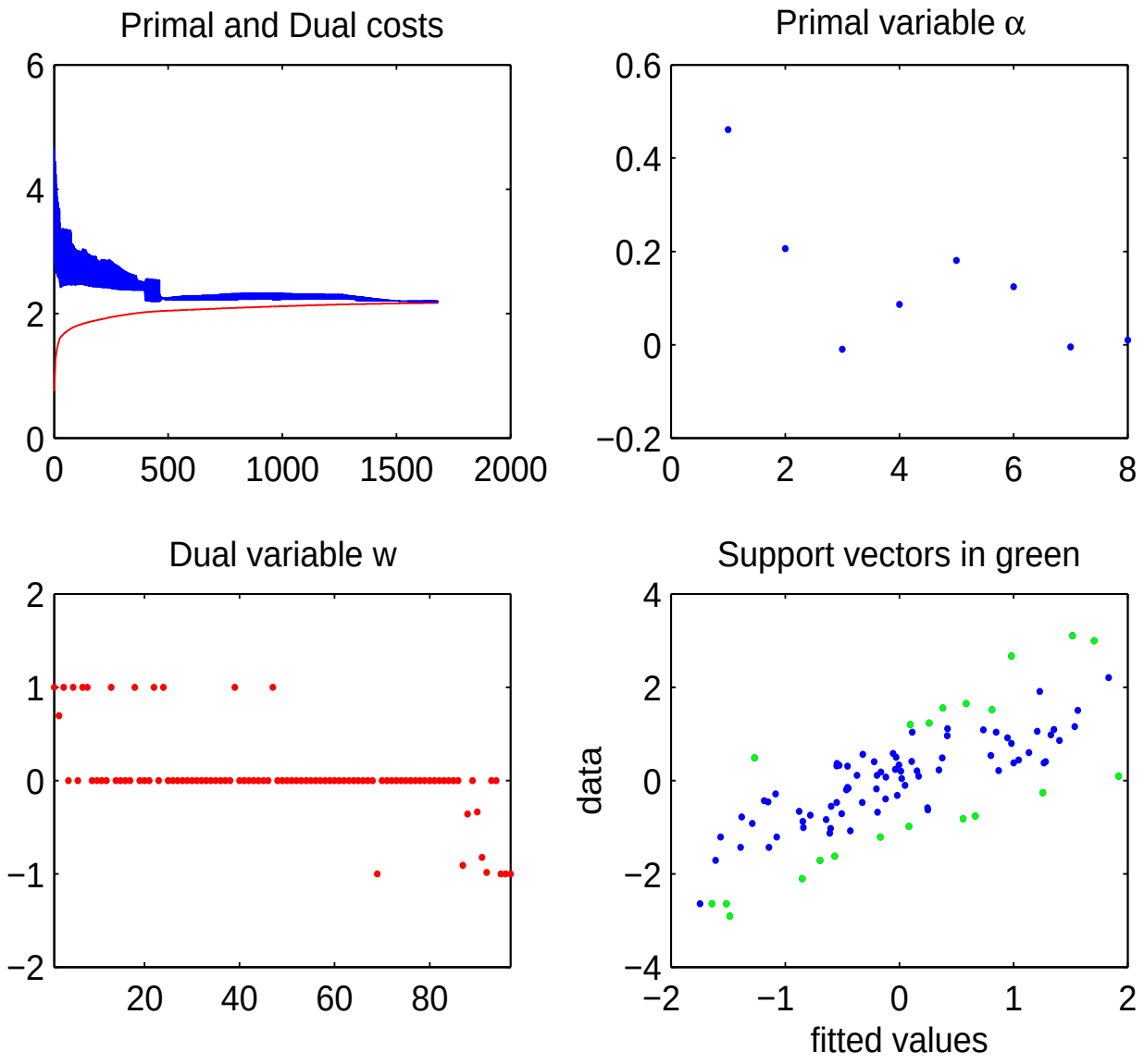
Application 5.4: SVM regression for prostate cancer data solves

$$\min_{\boldsymbol{\alpha}} \sum_{n=1}^{N} (|Y_n - \mathbf{x}_n' \boldsymbol{\alpha}| - \tau)_+ + \lambda \|\boldsymbol{\alpha}\|_2^2.$$

which has zero duality gap with

$$\min_{\mathbf{w}} \|X'\mathbf{w}\|_2^2/(4\lambda) + \mathbf{w}'\mathbf{s} + \tau \sum_n \begin{cases} |w_n| & \text{if } |w_n| \leq 1; \\ \infty & \text{otherwise.} \end{cases}$$

A coordinate descent method converges quickly in `R` to the unique minimum.


Primal and Dual costs
6
4
2
0
0
500
1000
1500
2000
Primal variable α
0.6
0.4
0.2
0
−0.2
0
2
4
6
8
Dual variable w
2
1
0
−1
−2
20
40
60
80
Support vectors in green
4
2
0
−2
−4
data
−2
−1
0
1
2
fitted values

6. Conclusions

Check for existence and uniqueness. Look for convexity.

Think of coordinate relaxation algorithms.

Dual problems may be easier to solve.

All methods have their limitations.

Nonlinear Programming, D. P. Bertsekas, Athena Scientific.

The Mathematics of Nonlinear Programming, Peressini, Sullivan, Uhl, Springer.

(2006) A Coordinate Gradient Descent Method for Nonsmooth Separable Minimization, Tseng, P. and Yun, S. preprint.

Nonsmooth, Nonconvex Optimization: Theory, Algorithms and Applications, Overton, A.L., 11 July 2006, EPFL 11am in MA 12.