Statistics and optimization
Unconstrained univariate optimization

Sylvain Sardy

Université de Genève

1. Problem statement

We want to minimize a function

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

Solving univariate problems is the corner stone of many multivariate optimization algorithms.

Motivation 1: many algorithms seek a minimum to

$$f : X \subset \mathbb{R}^{\mathrm{N}} \to \mathbb{R}$$

by solving successions of line search/univariate optimization problems.

Given a current feasible point $\mathbf{x}^{(i)}$ at iteration i and given a descent direction $\mathbf{d}^{(i)}$, they search for a solution α^* to

$$\min_{\alpha > 0} f(\mathbf{x} + \alpha \mathbf{d})$$

to get to the next point $\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} + \alpha^* \mathbf{d}^{(i)}$.

The term "line search" was coined because we search along the line in direction $\mathbf{d}$.

Motivation 2: many estimators require the selection of a regularization/smoothing parameter λ.

λ controls the bias/variance trade-off. Selection of λ is difficult.

Existing methods:

- Cross validation for essentially any estimator and noise, and generalized cross validation [Golub, Heath and Wahba 1979] for linear estimators and Gaussian noise:

$$\min_{\lambda \geq 0} \mathrm{GCV}(\lambda)$$

where:

$$\mathrm{GCV}(\lambda) = \frac{1}{N}\|\mathbf{y} - H(\lambda)\mathbf{y}\|_2^2 / [\frac{1}{N}\mathrm{tr}\{I_N - H(\lambda)\}]^2.$$

E.g., $H(\lambda) = X(X^{\mathrm{T}}X + \lambda I_N)^{-1}X^{\mathrm{T}}$ for ridge regression.

- Stein unbiased risk estimation (SURE) [Stein, 1981] for weakly differentiable estimators (e.g., lasso) and Gaussian noise:

$$\min_{\lambda \geq 0} \mathrm{SURE}(\lambda)$$

where:

$$\mathrm{SURE}(\lambda) = -N\sigma^2 + \|\mathbf{y} - \hat{\boldsymbol{\mu}}(\lambda)\|_2^2 + 2\sigma^2 \sum_{n=1}^{N} \frac{\partial \hat{\mu}_n}{\partial y_n}(\lambda).$$

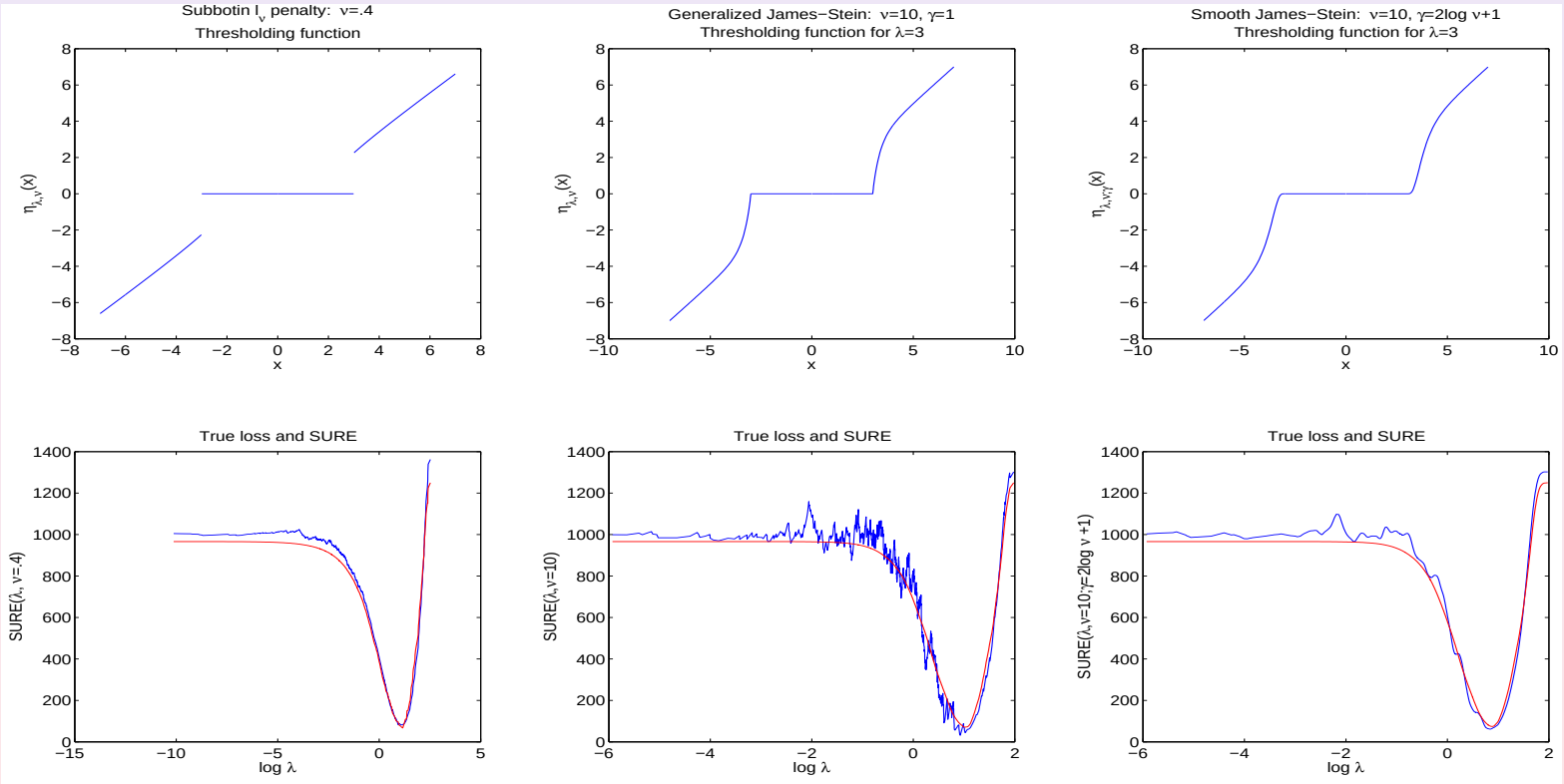
The goal of CV, GCV, SURE, etc. is to estimate a goodness-of-fit measure between the estimate $\hat{\boldsymbol{\mu}}$ and the unknown estimand $\boldsymbol{\mu}$.

For Gaussian noise, the ℓ_2-loss and corresponding risk/mean squared error are often the goodness-of-fit measures:

$$\mathrm{Loss}(\hat{\boldsymbol{\mu}}_\lambda, \boldsymbol{\mu}) = \|\hat{\boldsymbol{\mu}}_\lambda - \boldsymbol{\mu}\|_2^2 \quad \text{and} \quad \mathrm{Risk}(\hat{\boldsymbol{\mu}}_\lambda, \boldsymbol{\mu}) = \mathbb{E}\left(\mathrm{Loss}(\hat{\boldsymbol{\mu}}_\lambda, \boldsymbol{\mu})\right)$$

and

$$\mathbb{E}(\mathrm{SURE}(\lambda)) = \mathrm{Risk}(\lambda).$$



Goal: find a λ that minimizes the blue line.

2. Closed form expression

Before employing an iterative algorithm, analyze the mathematical properties of your function. For instance check existence, uniqueness, or *closed form expression* of a solution.

The following univariate problems have a closed form solution:

- $\min_x \frac{1}{2}(y-x)^2 + \lambda x^2$
- $\min_x \frac{1}{2}(y-x)^2 + \lambda |x|$

The following multivariate problems have a closed form solution:

- $\min_{\boldsymbol{\alpha}} \frac{1}{2}\|\mathbf{y} - X\boldsymbol{\alpha}\|_2^2 + \lambda \boldsymbol{\alpha}^{\mathrm{T}} Q \boldsymbol{\alpha}$, where Q is a symmetric positive definite matrix
- $\min_{\boldsymbol{\alpha}} \frac{1}{2}\|\mathbf{y} - X\boldsymbol{\alpha}\|_2^2 + \lambda \|\boldsymbol{\alpha}\|_1$, where X has orthogonal columns.

for all $\lambda \geq 0$.

3. Iterative methods

Many univariate or one-dimensional methods exist:

- Golden section (requires f)
- Pure Newton method (requires f, f' and f'')

3.1 Golden section (Kiefer, 1953)

This method does not require knowledge of f'.

It is based on *bracketing a minimum*:

Algorithm (from Bertsekas):

1. Calculate

$$\begin{aligned} x_k &= a_k + (b_k - a_k); \quad g(x_k) \\ y_k &= b_k - (b_k - a_k); \quad g(y_k) \end{aligned}$$

2. Three cases:
   - If $g(x_k) < g(y_k)$, set

$$\begin{aligned} a_{k+1} &= a_k;\ b_{k+1} = x_k \quad \text{if} \quad g(a_k) \leq g(x_k) \\ a_{k+1} &= a_k;\ b_{k+1} = y_k \quad \text{if} \quad g(a_k) > g(x_k) \end{aligned}$$

   - If $g(x_k) > g(y_k)$, set

$$\begin{aligned} a_{k+1} &= y_k;\ b_{k+1} = b_k \quad \text{if} \quad g(y_k) \geq g(b_k) \\ a_{k+1} &= x_k;\ b_{k+1} = b_k \quad \text{if} \quad g(y_k) < g(b_k) \end{aligned}$$

   - If $g(x_k) = g(y_k)$, set

$$a_{k+1} = x_k;\ b_{k+1} = y_k$$

3. If $|b_k - a_k| >$, then $k = k + 1$, goto 1;
   Else stop and return $(a_k + b_k)$=2, or $[a_k; b_k]$, or both.

Choose $= (3 - \sqrt{5})$=2.

3.4 Pure Newton method

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

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

By setting the derivative of f^k to zero

$$f'(x^{(k)}) + f''(x^{(k)})(x - x^{(k)}) = 0,$$

we obtain the next iterate $x^{(k+1)}$ as the minimum of $f^{(k)}$:

$$x^{(k+1)} = x^{(k)} - f'(x^{(k)})/f''(x^{(k)}),$$

assuming $f''(x^{(k)}) \neq 0$.
Newton's method typically converges very fast.
But convergence to a global minimum is not guaranteed.

Property: Let $\alpha^* \in D_f$ be a local minimum of a function f. If $f \in C^2$ and if $f''(\alpha) > 0$, then there exists a neighborhood of α^* such that for all initial value x_0 in the neighborhood, the sequence $\{x_n\}$ of Newton iterates converges to α.

4. Black boxes

Many softwares (`R`, `MatLab`) offer functions to do line search:

- `R`: `optim`, `nlminb`, `optimize`, ...
- `MatLab`: `srchgol`, `fminbnd`, `fminunc`, `fmincon`, `fminsearch`, ...

Do NOT apply them blindly. Read a description of the function to understand what method it uses.

There is no universal method that works uniformly best.
The method you use depends on the optimization problem you have to solve.