

ℓ^1 -regularization

Fallstudien der mathematischen Modellbildung, Teil 2

20.10.2023 - 21.11.2023,

paul.catala@tum.de

- Enforcing structure helps with ill-conditioning and under-determined systems.
A popular prior is **sparsity**, i.e. assuming the solution has only a few non-zero entries

The diagram illustrates the equation $Ax = y$ using a grid-based representation. Matrix A is a 6x10 grid of squares, some of which are shaded gray. Vector x is a 10x1 column of squares, with 5 squares shaded orange. Vector y is a 6x1 column of squares, with 4 squares shaded gray. An equals sign is placed between the vectors x and y .

- **Rationale:** signals/data are often sparse in some basis / living on low-complexity domain.

- If $c_i, i = 1, \dots, n$ denotes the columns of A , the system rewrites

$$y = \sum_{i=1}^n x_i c_i$$

(c_i) is an over-complete basis (or dictionary), and the goal is to select a subset of this basis that is sufficient to express $y \rightarrow$ **regressor selection**, or **variable selection**.

- If $c_i, i = 1, \dots, n$ denotes the columns of A , the system rewrites

$$y = \sum_{i=1}^n x_i c_i$$

(c_i) is an over-complete basis (or dictionary), and the goal is to select a subset of this basis that is sufficient to express $y \rightarrow$ **regressor selection**, or **variable selection**.

- A natural candidate to promote sparsity of solutions is the ℓ^0 -norm

$$\|x\|_0 = \# \{i \in \{1, \dots, n\} ; x_i \neq 0\}$$

! It is actually not a norm !

- If $c_i, i = 1, \dots, n$ denotes the columns of A , the system rewrites

$$y = \sum_{i=1}^n x_i c_i$$

(c_i) is an over-complete basis (or dictionary), and the goal is to select a subset of this basis that is sufficient to express $y \rightarrow$ **regressor selection**, or **variable selection**.

- A natural candidate to promote sparsity of solutions is the ℓ^0 -norm

$$\|x\|_0 = \# \{i \in \{1, \dots, n\} ; x_i \neq 0\}$$

! It is actually not a norm !

- The corresponding regularized problem is

$$\min \|Ax - y\|^2 + \lambda \|x\|_0$$

and in the noiseless case

$$\min \|x\|_0 \quad \text{s.t.} \quad Ax = y$$

- Remember that the penalized form is always equivalent to a constrained form with adequate parameter, *i.e.*

$$\min_{x \in \mathbb{R}^n} \|Ax - y\|^2 \quad \text{s.t.} \quad \|x\|_0 \leq \tau \quad (1)$$

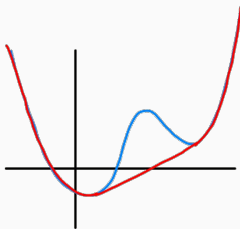
- NP-hard** combinatorial, **non-convex** problem. Direct strategy: check every possible sparsity pattern, *i.e.* fix subsets J of non-zero entries in x and solve the least-squares

$$\min_{\tilde{x} \in \mathbb{R}^n} \|A_J \tilde{x} - y_J\|^2$$

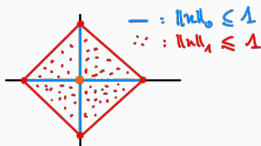
There are $\binom{n}{k}$ possible supports for each sparsity level \rightarrow infeasible for large n

- Possible approximations of the problem:
 - Greedy algorithms (e.g. orthogonal matching pursuit)
 - **Convex relaxation**

- **Definition (Convex envelope).** The convex envelope of a function $I(x)$ is the largest convex $J(x)$ such that $J(x) \leq I(x)$.



- **Theorem.** The convex envelope of $\|x\|_0$ for x restricted to $\|x\|_\infty \leq \alpha$ is $\|x\|_1/\alpha$



- Relax ℓ^0 -penalty into ℓ^1 -penalty

$$\min \|Ax - y\|_2^2 + \lambda \|x\|_1 \quad (\text{LASSO})$$

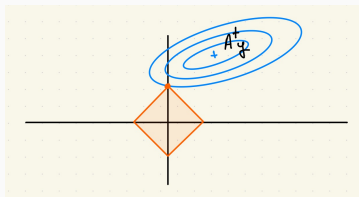
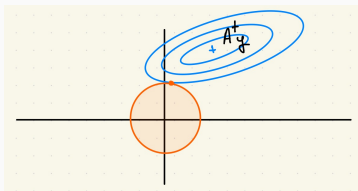
Called Lasso¹ (Least Absolute Shrinkage and Selection Operator) or basis pursuit denoising. When $\lambda = 0$, we obtain the basis pursuit² problem

$$\min \|x\|_1 \quad \text{s.t.} \quad Ax = y \quad (\text{BP})$$

- Main properties are:

Shrinkage: like Tikhonov regularization, Lasso penalizes large coefficients

Selection: unlike Tikhonov, Lasso produces sparse estimates



¹Tibshirani, 1996

²Donoho, early 1990's

LAGRANGE DUAL FUNCTION FOR LASSO

- We can reformulate the problem under a constrained form

$$\min \frac{1}{2} \|z - y\|^2 + \lambda \|x\|_1 \quad \text{s.t.} \quad z = Ax$$

and deduce the Lagrangian:

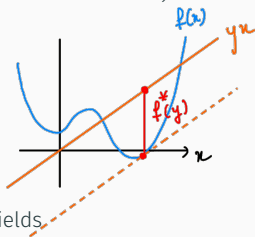
$$\mathcal{L}(x, z, \nu) = \frac{1}{2} \|z - y\|^2 + \lambda \|x\|_1 + \nu^\top (z - Ax)$$

- Minimization over z yields $\tilde{z} = y - \nu$. Minimization over x on the other hand is less obvious, since we have lost differentiability

$$\inf_x \lambda \|x\|_1 - \langle A^\top \nu, x \rangle = - \left(\sup_x \langle A^\top \nu, x \rangle - \lambda \|x\|_1 \right)$$

Definition (Conjugate function). The convex conjugate of $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is

$$f^*(y) = \sup_x \langle y, x \rangle - f(x)$$



With $J(x) := \lambda \|x\|_1$, the minimization over x and z yields

$$\mathcal{L}(\tilde{x}, \tilde{z}, \nu) = \nu^\top y - \frac{1}{2} \|\nu\|^2 - J^*(A^\top \nu)$$

- **Definition (Dual norm).** Given a norm $\|\cdot\|$ on \mathbb{R}^n , the associated dual norm is

$$\|y\|_* = \sup \left\{ y^\top x ; \|x\| \leq 1 \right\}$$

Example. $\|\cdot\|_1$ and $\|\cdot\|_\infty$ are dual to each other.

- **Proposition.** The conjugate function of $\|x\|$ is

$$f^*(y) = \begin{cases} 0 & \text{if } \|y\|_* \leq 1 \\ \infty & \text{otherwise} \end{cases}$$

Proof.¹ If $\|y\|_* > 1$, then by definition there exists $w \in \mathbb{R}^n$ such that $\|w\| \leq 1$ and $y^\top w > 1$. Taking $x = tw$ and letting $t \rightarrow \infty$ we obtain

$$y^\top x - \|x\| = t(y^\top w - \|w\|) \rightarrow \infty,$$

hence $f^*(y) = \infty$. If $\|y\|_* \leq 1$, since $y^\top x \leq \|x\|\|y\|_*$ for all x , then $y^\top x - \|x\| \leq 0$, and $x = 0$ is the maximizer.

¹Boyd, Vandenberghe, *Convex Optimization*, Example 3.26

- If $J(x) = \lambda \|x\|_1$, then $J^*(y)$ is the indicator of $\{\|y\|_\infty \leq \lambda\}$.
- Altogether, we obtain

$$\mathcal{L}(\tilde{x}, \tilde{z}, \nu) = \nu^\top y - \frac{1}{2} \|\nu\|^2 - i_{\{\nu: \|\nu\|_\infty \leq \lambda\}}(A^\top \nu)$$

where we denote i_C the indicator function of the set C . Hence the Lasso dual problem reads

$$\max \nu^\top y - \frac{1}{2} \|\nu\|^2 \quad \text{s.t.} \quad \|A^\top \nu\|_\infty \leq \lambda$$

$\|\cdot\|_1$ is convex but not differentiable at 0. How to derive optimality conditions?

- Recall the standard inequality for convex functions

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$$

- **Definition (Sub-differential).** The sub-differential of f at x is

$$\partial f(x) = \{v \in \mathbb{R}^n ; \forall y \in \mathbb{R}^n, f(y) \geq f(x) + \langle v, y - x \rangle\}$$

Note that $\partial f(x)$ is convex. If f is differentiable, then $\partial f(x) = \{\nabla f(x)\}$.

- **Proposition.** For any function f ,

$$x_* = \operatorname{argmin}_x f(x) \iff 0 \in \partial f(x)$$

Proof. x_* minimizer of $f \iff \forall x, f(x) \geq f(x_*) = f(x_*) + \langle 0, x - x_* \rangle \iff 0 \in \partial f(x)$.

Some basic rules

- $\partial f(x) = \{\nabla f(x)\}$ if f is differentiable at x
- $\partial(\alpha f) = \alpha \partial f$ if $\alpha > 0$
- $\partial(f_1 + f_2)(x) = \partial f_1(x) + \partial f_2(x)$
- if $g(x) = f(Ax + b)$ where f is convex, then $\partial g(x) = A^\top \partial f(Ax + b)$

- $|x|$ is differentiable at any $x \neq 0$ with derivative ± 1 . At 0,

$$(\forall z \in \mathbb{R}, |z| \geq yz) \iff y \in [-1, 1]$$

so $\partial|0| = [-1, 1]$.

- Generalization:

$$v \in \partial\|x\|_1 \iff \begin{cases} v_i = \text{sign}(x_i) & \text{if } x_i \neq 0 \\ v_i \in [-1, 1] & \text{if } x_i = 0 \end{cases}$$