Internship proposal: “Benchopt for MCP and other sparse regression formulations”

Chair of research and teaching in artificial intelligence

Keywords: *Optimization, coding, sparse regression, MCP*

`Benchopt` presentation

Optimization has a central role in machine learning (ML). Though theory has come handy to assert the effectiveness (or ineffectiveness) of many algorithms, practical considerations are often not taken as seriously by the ML community.

Choosing the best algorithm to solve an optimization problem often depends on

- The data **scale, conditionning**
- The objective parameters **regularisation**
- The implementation **complexity, language**

An impartial selection requires a time-consuming **benchmark**!

The goal of the package `benchopt`[1] is to make this step as easy as possible, and allow practitioners to evaluate a wide diversity of algorithm on their own dataset.

Internship description

Within this project, the goal of the intern will be to improve the `benchopt` package and website in general. On the mid/long-term, this would lead to a publication in a major conference in ML, once the package has become strong enough, and could cover a wide diversity of optimization problems.

A precise task for the intern would be to start providing a benchmark for the Minimax Concave Penalty (MCP) regression problem [10].

Let us remind the definition of the MCP estimator. First, for some parameters $\gamma > 1$ and $\lambda \geq 0$, we define the 1D penalty for any $t \in \mathbb{R}$ as follows:

$$p_{\lambda,\gamma}^{\text{MCP}}(t) = \begin{cases} \lambda|t| - \frac{t^2}{2\gamma}, & \text{if } |t| \leq \gamma\lambda, \\ \frac{1}{2}\gamma\lambda^2, & \text{if } |t| > \gamma\lambda \ . \end{cases} \tag{1}$$

[1] https://github.com/benchopt/benchOpt

The proximity operator[2] of $p_{\lambda,\gamma}$ for parameters $\lambda > 0$ and $\gamma > 1$ is defined as follows [2, Sec. 2.1]:

$$\mathrm{prox}^{\mathrm{MCP}}_{\lambda,\gamma}(t) = \begin{cases} \frac{\mathrm{ST}(t,\lambda)}{1-\frac{1}{\gamma}} & \text{if } |t| \leq \gamma\lambda \\ t & \text{if } |t| > \gamma\lambda \end{cases} , \tag{2}$$

where $\mathrm{ST}(t,\lambda) = \mathrm{sign}(t) \cdot (|t| - \lambda)_+$ for any $t \in \mathbb{R}$ and $\lambda \geq 0$, and similarly $\mathrm{HT}(t,\lambda) = t \cdot \mathbb{1}_{\{|t|>\lambda\}}$. The proximal operator defined in Eq.(2) is also known as the Firm Shrinkage [5]. The MCP estimator is then given by:

$$\hat{\beta}^{(\lambda,\gamma)}(y) \triangleq \mathop{\arg\min}_{\beta \in \mathbb{R}^p} \frac{1}{2n} \|y - X\beta\|_2^2 + \sum_{j=1}^{p} p^{\mathrm{MCP}}_{\lambda,\gamma}(\beta_j) \quad , \tag{3}$$

where $X \in \mathbb{R}^{n \times p}$ is the design matrix and $y \in \mathbb{R}^n$ the observed signal.

Illustration of the penalty and proximal operator are provided in Fig.1 for some parameters, including limiting behavior.

The optimization solvers of interest that will be investigated are:

- Coordinate descent [9, 2]
- Proximal gradient descent [1]
- Difference of convex (DC) programming [6]
- Proximal bundle methods [3]

More recent algorithms could be investigated and added to the package, see for instance [8] or [7], or variants using an interesting formulation of the MCP penalty as a bi-convex program, see [4]. Upon success, a new algorithm to handle MCP could also then be considered.

Skills required

- `Python`
- `Git`
- `R` (not mandatory, but could come handy)

Supervision Team

- Joseph Salmon: joseph.salmon@umontpellier.fr
- Cássio Fraga Dantas.: cassiofragadantas@gmail.com

Salary

Gross monthly salary: approx 550 Euros.
This work will be funded by the ANR CaMeLOt ANR-20-CHIA-0001-01.

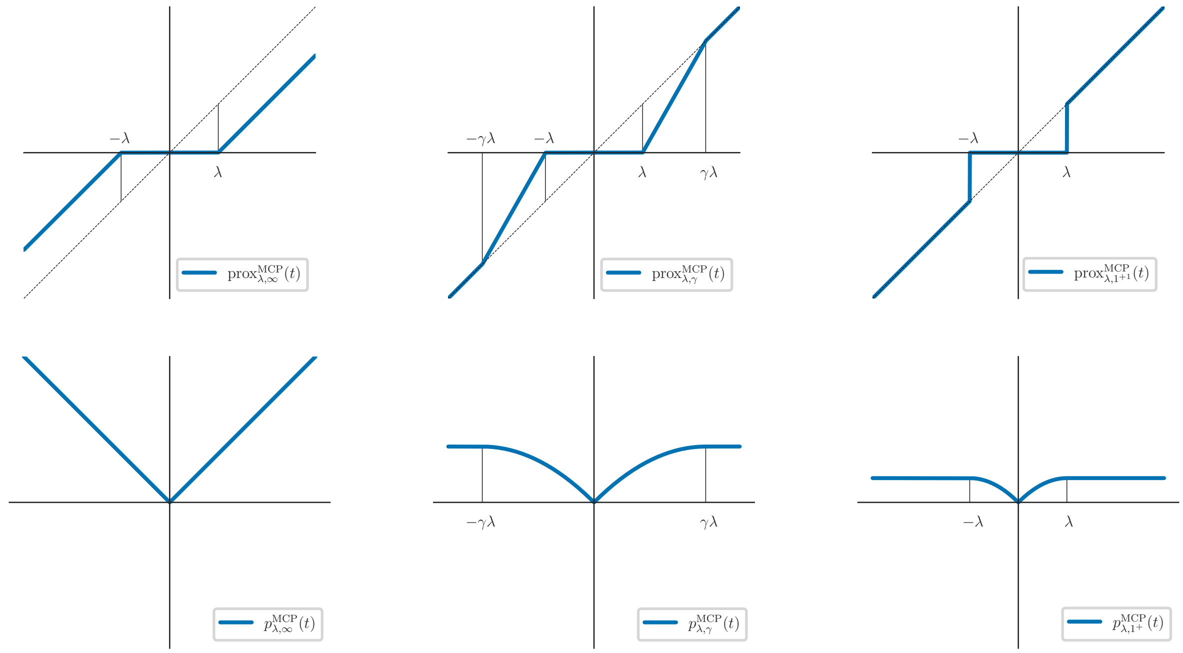
Duration

The internship could last from 4 to 6 months.

Location

The internship will be located in Montpellier (Univ. Montpellier), within the mathematics department (IMAG).

[2] For a function p this operator is defined by $\mathrm{prox}(x) = \arg\min_{x'} p(x') + \|x - x'\|^2/2$.



(a) $\gamma \to +\infty$: recovers $\mathrm{ST}(\cdot, \lambda)$ (b) $\gamma = 2.3$ (c) $\gamma \to 1^+$: recovers $\mathrm{HT}(\cdot, \lambda)$

Figure 1: Penalty (bottom) and associated proximal operator (top) for a fixed λ, with $\gamma \to +\infty$ (a), for $\gamma = 2.3$ (b) and $\gamma \to 1+$ (c).

References

[1] A. Beck and M. Teboulle. "A fast iterative shrinkage-thresholding algorithm for linear inverse problems". *SIAM J. Imaging Sci.* 2.1 (2009), pp. 183–202.

[2] P. Breheny and J. Huang. "Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection". *Ann. Appl. Stat.* 5.1 (2011), p. 232.

[3] W. de Oliveira. "Proximal bundle methods for nonsmooth DC programming". *Journal of Global Optimization* 75.2 (2019), pp. 523–563.

[4] M. Fornasier and H. Rauhut. "Iterative thresholding algorithms". *Applied and Computational Harmonic Analysis* 25.2 (2008), pp. 187–208.

[5] H.-Y. Gao and A. G. Bruce. "WaveShrink with firm shrinkage". *Statist. Sinica* (1997), pp. 855–874.

[6] G. Gasso, A. Rakotomamonjy, and S. Canu. "Recovering sparse signals with non-convex penalties and DC programming". *IEEE Trans. Signal Process.* 57.12 (2009), pp. 4686–4698.

[7] A. Rakotomamonjy, R. Flamary, G. Gasso, and J. Salmon. *Provably Convergent Working Set Algorithm for Non-Convex Regularized Regression*. Tech. rep. 2020.

[8] A. Rakotomamonjy, G. Gasso, and J. Salmon. "Screening Rules for Lasso with Non-Convex Sparse Regularizers". *ICML*. Vol. 97. 2019, pp. 5341–5350.

[9] T. T. Wu and K. Lange. "Coordinate descent algorithms for lasso penalized regression". *Ann. Appl. Stat.* (2008), pp. 224–244.

[10] C.-H. Zhang. "Nearly unbiased variable selection under minimax concave penalty". *Ann. Statist.* 38.2 (2010), pp. 894–942.