Bilevel optimization for hyperparameter tuning in structured sparse models

Supervision:

- Alexandre Gramfort, Parietal, Inria Saclay Ile-de-France, (alexandre.gramfort@inria.fr)
- Joseph Salmon, Université de Montpellier (joseph.salmon@umontpellier.fr)
- Quentin Bertrand, Parietal, Inria Saclay Ile-de-France, (quentin.bertrand@inria.fr)

Context A whole variety of sparse estimators have been developed over the last decades. In particular it is possible to finely model brain responses using refined sparse linear models. Such models take into account the spatial aspects as well as the temporal aspects of the brain responses [4]. These estimators are based on the resolution of non-smooth optimization problems, which can now be solved in a very efficient way, e.g., using block coordinate descent techniques [5]. However, these estimators rely on hyperparameters, namely, regularization parameters in this context. Setting regularization parameters for sparse estimators is notoriously difficult, though crucial in practice: the quality of the estimator can dramatically varies for different values of these regularization parameters. This calibration challenge can prevent practitioners from using these refined models in practice.

For this internship, the motivation is to develop tools based on differential programming [1] to *automatically* set such hyperparameters for time-frequency models [4], and provide practitioners with automatically calibrated method for brain signals analysis. If one denotes $Y \in \mathbb{R}^{n \times q}$ the data (EEG and MEG), $X \in \mathbb{R}^{n \times p}$ the design matrix, $\Phi \in \mathbb{C}^{q \times m}$ a time-frequency transform (Short-time Fourier Transform), $B \in \mathbb{R}^{p \times q}$ the unknown the brain activity and its time-frequency transform $Z = B\Phi \in \mathbb{C}^{p \times m}$ the problem reads:

$$\min_{Z \in \mathbb{C}^{p \times m}} J(Z, \lambda, \mu) = \min_{Z \in \mathbb{C}^{p \times m}} \frac{1}{2} \|Y - XZ\Phi^\top\|^2 + \sum_{j=1}^{p} \lambda_j \|Z_{j,\cdot}\|^2 + \sum_{j=1}^{p} \sum_{k=1}^{m} \mu_{jk} |Z_{j,k}| \quad ,$$

where $\lambda \in \mathbb{R}_+^p$ and $\mu \in \mathbb{R}_+^{p,m}$ are hyperparameters we aim to estimate.

Methods The most popular hyperparameter optimization approach is grid-search. Grid-search however requires to choose a predefined grid for each parameter, which scales exponentially in the number of parameters. Given the dimension of λ and μ this is not an option. In the same vein random-search approaches [2], which are typical to adress higher dimensional problems, can be slow when the number of hyperparameters is very large. Another approach is to cast hyperparameter optimization as a bilevel optimization problem [3] one can solve by gradient descent. The key challenge for these methods is the estimation of the gradient w.r.t. the hyperparameters [1], which is here a challenge due to the nonsmooth optimization problem considered.

Environment The internship will take place at the Inria Saclay, in the Parietal team. This is a large team working focused on mathematical methods for statistical modeling of brain function using neuroimaging data (fMRI, MEG, EEG). Particular topics of interest include machine learning techniques, numerical and parallel optimization, differentiable programming, applications to human cognitive neuroscience, and scientific software development.

Requirements

- Strong mathematical background. Knowledge in numerical optimization is a plus.
- Good programming skills in Python. Knowledge of a deep-learning library is a plus.

References

[1] Atılım Günes Baydin, Barak A Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: a survey. *The Journal of Machine Learning Research*, 18(1):5595–5637, 2017.

[2] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13(1):281–305, 2012.

[3] Quentin Bertrand, Quentin Klopfenstein, Mathieu Blondel, Samuel Vaiter, Alexandre Gramfort, and Joseph Salmon. Implicit differentiation of lasso-type models for hyperparameter optimization. In *Proc. ICML 2020*, July 2020.

[4] Alexandre Gramfort, Daniel Strohmeier, Jens Haueisen, Matti S Hämäläinen, and Matthieu Kowalski. Time-frequency mixed-norm estimates: Sparse m/eeg imaging with non-stationary source activations. *NeuroImage*, 70:410–422, 2013.

[5] M. Massias, A. Gramfort, and J. Salmon. Celer: a fast solver for the lasso with dual extrapolation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 3321–3330, 2018.