PhD offer: properties of proximal operators-learning neural networks

Organization

- start date: October 2023
- supervision: Mathurin Massias (CR, Inria Lyon), Rémi Gribonval (DR, Inria Lyon), Emmanuel Soubiès (CR CNRS, IRIT, Toulouse)
- location: the PhD will take place in the Inria OCKHAM team, hosted by the Laboratoire d'Informatique du Parallélisme (46 allée d'Italie, 69007 Lyon)

Context

Optimization problems of the form $x^* = \mathrm{argmin}_x f(x) + g(x)$ are ubiquitous in machine learning, imaging and signal processing, where their composite nature allows enforcing two properties on the solution simultaneously. The *datafit* f ensures that x^* correctly fits the observed data, while the *regularizer* g imposes that it exhibits a low level of complexity, e.g. is sparse.

Since the 2000s, proximal methods have been applied with tremendous success to composite optimization. They are designed to solve such problems for a wide variety of regularizers g, the prototypical being the ISTA algorithm for the Lasso [1], corresponding to $g = \|\cdot\|_1$. In the most simple proximal algorithm, iterates x_k are obtained by first performing a gradient descent step on f, followed by the application of the so-called proximal operator of g:

$$x_{k+1} = \mathrm{prox}_{\gamma g}(x_k - \gamma \nabla f(x_k)) \ , \tag{1}$$

where the proximal operator is defined as the solution of an optimization problem:

$$\mathrm{prox}_{\gamma g}(x) = \underset{y}{\mathrm{argmin}}\, \gamma g(y) + \frac{1}{2}\|x - y\|^2 \ .$$

The convergence guarantees of such methods have been thoroughly studied and are now rigorously established, including a wealth of variants [2].

Drawing inspiration from the form of iterations (1), Plug-and-Play (PNP) methods [3] aim at providing better solutions for the original task, by replacing the proximal operator in proximal methods by a learned operator H_σ. For example, (1) becomes in the PNP framework

$$x_{k+1} = H_\sigma(x_k - \gamma \nabla f(x_k)) \ .$$

The strength of this method is that learning the operator H_σ (hence the name *learning based* approach) provides a greater flexibility than using fixed proximal operators (*model based* approach). Beyond proximal gradient, PNP methods have been successfully applied to many proximal algorithms such as ADMM or Douglas-Rachford. However, they lack the theoretical guarantees of model-based methods in terms of convergence.

Goals of the PhD

One popular way to learn the operator in PNP methods is through neural networks [4]. The PhD will explore two sets of questions related to learning-based proximal methods with neural networks.

Guaranteed learning of proximal operators The first one is how to learn operators H_σ that are *guaranteed* to be proximal operators. Such a property is highly desirable, as it would grant learning-based methods the strong theoretical guarantees of model-based ones, in particular regarding convergence. As related works, Ryu et al. [5] proposed to constraint the Lipschitz constant of the network to be close to one, and relied on fixed-point theory to prove convergence of PNP algorithms. Yet this solution is not satisfying: first, bounding the Lipschitz constant of a network precisely is not an easy task, and current proposals are very crude. In addition, it restricts by too much the class of functions that can be learned; in particular, proximal operators of non convex functions can have arbitrarily large Lipschitz constants in the non convex case. Other alternatives involve implicit networks directly modelling the Jacobian [6], but it requires numerical integration procedures that prevent its applications to deep neural networks.
On the contrary, we propose to use the characterization of proximal operators of [7], that must correspond to gradients of convex functions. In particular, this yields promising study directions on the necessary and sufficient positive semidefiniteness of the Jacobian of such networks. Beyond proximal operators, other applications are foreseeable in other domains where learning functions that are gradients of convex functions is relevant, such as optimal transport [8]. After having proposed new neural architectures to provably implement proximal operators, it will be crucial to characterize the kind of functions that such architectures are able to approximate, and at which cost in term of size, in order to devise practical methods with theoretical grounding. While this question has already been asked for other types of constraints neural networks in the field of neural network compression [9], it remains unanswered in the framework of PNP networks.

Optimal learned proximal operators for sparse recovery Second, in the context of sparse learning, the PhD will investigate the properties of learned operators compared to model-based ones. For explicit sparse penalties, there exists a vast literature on the optimal choice of penalty amongst specific classes of functions. In particular, it is well-known that in some sense the L1 norm is the tightest convex relaxation of the L0 pseudonorm. In the realm of non-convex sparse regularizers, MCP and CEL0 [10] are also optimal with respect to other criteria in terms of continuous relaxations of the L0 pseudo-norm. The PhD will study the type of penalties that correspond to proximal operators that can be implemented by neural networks, and their relationship with hand-crafted ones. In particular, much attention has been devoted to unrolling algorithms, e.g. to model the ISTA iterations for the Lasso:

$$x_{k+1} = \texttt{soft\_thresholding}((\mathrm{Id} - \gamma A^\top A)x_k - A^\top b)$$

as the action of a layer of a neural network: matrix multiplication, bias addition, and non linearity application: $x_{k+1} = \sigma(Wx_k + b)$. This direction has been studied in [11] using the popular ReLU non linearity, which corresponds to a soft-thresholding. However, using learned proximal operators in the non linearities may boost the performance of such unrolled networks, by going beyond the limited L1 norm [12]. After studying the practical properties of such unrolled algorithm, the PhD will focus on the theoretical analysis of the learned proximal operators, and their potential optimality as continuous relaxations of the L0 penalty.

Together with these theoretical directions, the PhD work will also develop efficient open source implementations of the new architectures, relying on pytorch, and interacting with other libraries developed in the OCKHAM team, such as `pyfaust` and `celer`. Such developments will favor practical applications within the environment of ENS Lyon, in particular with the team of Nelly Pustelnik (CR CNRS, Laboratoire de physique).

References

[1] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(11):1413–1457, 2004.

[2] H. Bauschke and P. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

[3] S. Venkatakrishnan, C. Bouman, and B. Wohlberg. Plug-and-play priors for model based reconstruction. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 945–948. IEEE, 2013.

[4] T. Meinhardt, M. Moller, C. Hazirbas, and D. Cremers. Learning proximal operators: Using denoising networks for regularizing inverse imaging problems. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1781–1790, 2017.

[5] E. Ryu, J. Liu, S. Wang, X. Chen, Z. Wang, and W. Yin. Plug-and-play methods provably converge with properly trained denoisers. In *International Conference on Machine Learning*, pages 5546–5557. PMLR, 2019.

[6] J. Richter-Powell, J. Lorraine, and B. Amos. Input convex gradient networks. *arXiv preprint arXiv:2111.12187*, 2021.

[7] R. Gribonval and M. Nikolova. A characterization of proximity operators. *Journal of Mathematical Imaging and Vision*, 62(6-7):773–789, 2020.

[8] A. Korotin, L. Li, A. Genevay, J. Solomon, A. Filippov, and E. Burnaev. Do neural optimal transport solvers work? a continuous wasserstein-2 benchmark. *NeurIPS*, 34:14593–14605, 2021.

[9] A. Gonon, N. Brisebarre, R. Gribonval, and E. Riccietti. Approximation speed of quantized vs. unquantized relu neural networks and beyond. *IEEE Transactions on Information Theory*, 2023.

[10] E. Soubies, L. Blanc-Féraud, and G. Aubert. A continuous exact \ell_0 penalty (cel0) for least squares regularized problem. *SIAM Journal on Imaging Sciences*, 8 (3):1607–1639, 2015.

[11] P. Ablin, T. Moreau, M. Massias, and A. Gramfort. Learning step sizes for unfolded sparse coding. *NeurIPS*, 32, 2019.

[12] C. Yang, Y. Gu, B. Chen, H. Ma, and H. So. Learning proximal operator methods for nonconvex sparse recovery with theoretical guarantee. *IEEE Transactions on Signal Processing*, 68:5244–5259, 2020.