Identification et exploitation explicite de structure en apprentissage régularisé

- **Franck Iutzeler** (Univ. Grenoble Alpes)

Mots-clé : optimisation non-lisse, opérateur proximal, régularisation, lasso.

Résumé : L'optimisation mathématique tient une place de plus en plus importante en science des données. Ceci est dû en partie à la difficulté croissante des taches d'apprentissage mais aussi à la structure particulière des problèmes de minimisation associés. En effet, face à l'augmentation de la taille des données (et en particulier du nombre de mesures par échantillon), des méthodes de réduction de dimension ont été proposées. La plus connue d'entre elles est certainement le lasso, qui cherche un régresseur linéaire performant mais aussi parcimonieux [1]. Ce problème se pose naturellement comme la somme d'un terme de perte quadratique et d'un terme de régularisation non-différentiable : la norme 1. Afin de le minimiser, il a tout de suite été suggéré d'utiliser l'opération de seuillage doux, qui correspond à l'opérateur proximal associé à la régularisation par la norme 1 [2]. Cette opération est aussi bien utile à la minimisation de l'objectif qu'à l'identification des coordonnées nulles du régresseur. Le lien entre la quête de structures de faible dimension en apprentissage et les algorithmes proximaux en optimisation est donc naturel et s'étend bien au delà de cet exemple.

Dans cet exposé, je m'intéresserai à la caractérisation mathématique de la structure sous-jacente des problèmes régularisés. Nous donnerons donc quelques résultats sur les opérateurs proximaux qui permettent d'examiner la structure de non-différentiabilité du problème. Afin d'observer (numériquement) cette structure, nous considérerons uniquement les opérateurs proximaux explicites comme la norme 1, la semi-norme 0, ou la norme nucléaire. Ainsi, nous verrons que les résultats ci-dessus impliquent une identification de structure en temps fini pour une grande classe d'algorithmes proximaux [3].

Ensuite, je donnerai quelques exemples d'exploitation algorithmique de ce phénomène. Une fois la structure locale du problème identifiée, la vitesse de convergence est en général améliorée (voir par exemple [4] dans le cas du gradient proximal). Pour aller encore plus loin, il est souvent possible de restreindre notre objectif à une variété (représentant la non-différentiabilité locale). Nous nous trouvons alors en face d'un problème d'optimisation lisse sur variété. Malheureusement, il est en général impossible de savoir si la structure identifiée est optimale ou non. Nous présenterons donc un algorithme qui alterne entre une étape de gradient proximal et une étape de Newton sur variété [5]. Sous des conditions de stabilité de structure, cette méthode converge super-linéairement, montrant ainsi l'intérêt de l'exploitation de structure en pratique. Finalement, nous illustrerons l'intérêt et les limites de cette méthode sur des problèmes d'apprentissage régularisés.

Références :

[1] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[2] D. Donoho, J. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455,1994.

[3] F. Iutzeler, J. Malick. Nonsmoothness in Machine Learning: specific structure, proximal identification, and applications. *Set-Valued and Variational Analysis*, 28(4):661–678, 2020.

[4] J. Liang, J. Fadili, G. Peyré. Activity Identification and Local Linear Convergence of Forward–Backward-type Methods. *SIAM Journal on Optimization*, 27(1)408-437, 2017.

[5] G. Bareilles, F. Iutzeler, J. Malick. Newton acceleration on manifolds identified by proximal-gradient methods. *preprint arXiv:2012.12936*, 2020.