Journal de la Société Française de Statistique
Vol. 160 No. 3 (2019)

Discussion: Concentration for ordered smoothers

Titre: Discussion : Concentration pour des lisseurs ordonnés

Pierre Bellec[1]

This discussion focuses on concentration results such as Proposition 1 in Arlot (2019) used by the author to study the Slope heuristics and minimal penalties. I will restrict my discussion to the case of the normal mean model with Ridge regression estimates, to highlight a phenomenon surprisingly different from Arlot (2019, Proposition 1).

Consider the following setting: $y \sim N(\mu, I_n)$ is observed for an unknown mean $\mu \in \mathbb{R}^n$. A design matrix $X \in \mathbb{R}^{n\times p}$ is available and the practitioner wishes to fit Ridge regression estimates

$$\hat\beta = \operatorname*{arg\,min}_{b\in\mathbb{R}^p} \|y - Xb\|^2 + \lambda\|b\|^2.$$

It is well known that the above estimate is linear, $X\hat\beta = A_\lambda y$ for a deterministic matrix $A_\lambda = X(X^TX + \lambda I_p)^{-1}X^T$. The practitioner chooses a grid of tuning parameters $\lambda_1 < \ldots < \lambda_M$, and selects one using, e.g., Mallows (1973) criterion $C_p(A) = \|y - Ay\|^2 + 2\,\mathrm{trace}(A)$ or other selection/aggregation methods. In order to study the performance of C_p-tuned Ridge regression, or its variant based on Q-aggregation (Rigollet, 2012; Dai et al., 2012, 2014) as explained in Bellec (2018), one needs to study the one-sided concentration of random variables of the form

$$C_p(A) - C_p(A^*) - \mathbb{E}[C_p(A) - C_p(A^*)] - c\|(A - A^*)y\|^2,$$

for some absolute constant $c > 0$ and uniformly over all matrices A, A^* in the model set $\mathscr{M} = \{A_\lambda, \lambda = \lambda_1, \ldots, \lambda_M\}$. It is common to bound the above random variable first for fixed matrices A, A^*, obtain exponential probability bounds, and finally use the union bound over all matrices $A \in \mathscr{M}$. This union bound induces a uniform upper bound on the previous display of order $\log M$, where $M = |\mathscr{M}|$ is the cardinality of the model set, e.g., as in Proposition 3.1 of the work Arlot (2019) discussed here. This induces oracle inequalities that grow with $\log M$.

Surprisingly, in the case of ordered smoothers Kneip (1994) such as the above Ridge regression setting, chaining arguments (e.g., Adamczak (2015); Dirksen (2015)) lead to the bound

$$\mathbb{P}\Big\{ \sup_{\lambda\ge 0} \big(C_p(A_\lambda) - C_p(A^*) - \mathbb{E}[C_p(A_\lambda) - C_p(A^*)] - c\|(A_\lambda - A^*)y\|^2\big) \le Cx \Big\} \ge 1 - Ce^{-x}.$$

for some absolute constant $C > 0$ and any $x \ge 1$, cf. Bellec and Yang (2019). In particular, the above deviation inequality is independent of both the dimension and the cardinality of the model

[1] Department of Statistics
Busch Campus, Rutgers University
Piscataway, NJ 08854, USA.

 ISSN: 2102-6238

set $\mathscr{M}$, in striking contrast to union bound arguments used in Proposition 3.1 of Arlot (2019) or Dai et al. (2014); Bellec (2018). The major consequence of such uniform deviation bound is that ordered linear smoothers can be optimally tuned, no matter how many tuning parameters are considered or how coarse the grid is, at no statistical cost: The procedure $\hat{y}$ in Bellec and Yang (2019) that leverages the above uniform deviation inequality enjoys the oracle inequality

$$\mathbb{E}\big[\|\hat{y}-\mu\|^2 - \min_{j=1,\dots,M} \|A_{\lambda_j} y - \mu\|^2\big] \le C\sigma^2.$$

I am wondering if the above uniform deviation inequality has consequences for the Slope heuristics or the minimal penalty phenomenon, and I wish to congratulate the author for this insightful survey.

References

Adamczak, R. (2015). A note on the hanson-wright inequality for random vectors with dependencies. *Electronic Communications in Probability*, 20.

Arlot, S. (2019). Minimal penalties and the slope heuristics: a survey. *arXiv preprint arXiv:1901.07277*.

Bellec, P. C. (2018). Optimal bounds for aggregation of affine estimators. *Ann. Statist.*, 46(1):30–59.

Bellec, P. C. and Yang, D. (2019). The cost-free nature of optimally tuning tikhonov regularizers and other ordered smoothers. *preprint*.

Dai, D., Rigollet, P., L., X., and T., Z. (2014). Aggregation of affine estimators. *Electon. J. Stat.*, 8:302–327.

Dai, D., Rigollet, P., and Zhang, T. (2012). Deviation optimal learning using greedy Q-aggregation. *The Annals of Statistics*, 40(3):1878–1905.

Dirksen, S. (2015). Tail bounds via generic chaining. *Electronic Journal of Probability*, 20.

Kneip, A. (1994). Ordered linear smoothers. *The Annals of Statistics*, 22(2):835–866.

Mallows, C. L. (1973). Some comments on c p. *Technometrics*, 15(4):661–675.

Rigollet, P. (2012). Kullback–Leibler aggregation and misspecified generalized linear models. *Ann. Statist.*, 40(2):639–665.

 ISSN: 2102-6238