Sara van de Geer
ETH Zürich

Concentration in high-dimensional statistics

High-dimensional statistics is about models where one has more parameters p than observations n. The theory contains a high concentration of challenging mathematical issues. We will encounter convex analysis, random matrix theory, approximation theory and, last but not least, concentration of measure. We illustrate this for linear regression and also briefly discuss other models.

Suppose one observes an n-dimensional Gaussian vector Y of the form

$$Y = f^0 + \epsilon,$$

with f^0 an unknown mean vector and ϵ standard Gaussian noise. Consider a given $n \times p$ design matrix X, with $p > n$, and the estimator

$$\hat{\beta} := \arg\min_{\beta \in \mathbb{R}^p} \Big\{ \underbrace{\|Y - X\beta\|_2^2}_{\text{least squares loss}} + \underbrace{2\lambda\|\beta\|_1}_{\text{regularization penalty}} \Big\}$$

where $\lambda > 0$ is a given tuning parameter. This estimator, called "Lasso" (Tibshirani (1996)), is extremely popular in high-dimensional regression. The penalty $\beta \mapsto 2\lambda\|\beta\|_1$ regularizes the problem: it ensures that certain entries in the vector $\hat{\beta}$ are set to zero.

The theoretical properties of the Lasso are well understood. We will present some recent further refinements of this theory.

Consider the minimizer of the noiseless problem

$$\beta^* := \arg\min_{\beta \in \mathbb{R}^p} \Big\{ \|f^0 - X\beta\|_2^2 + 2\lambda\|\beta\|_1 \Big\}.$$

Let $\hat{f} := X\hat{\beta}$ and $f^* := X\beta^*$.

We are interested in the behaviour of the "approximation error" $\|f^* - f^0\|_2$ and "estimation error" $\|\hat{f} - f^*\|_2$. Note that the approximation error is a deterministic quantity that can be studied using approximation theory. Clearly, if the tuning parameter λ is small this error will be small too. The estimation error is random and is typically large for small λ.

To deal with the estimation error one can apply concentration of measure. It is shown in van de Geer and Wainwright (2016) and Bellec and Tsybakov (2016) that $\epsilon \mapsto \|\hat{f} - f^*\|_2$ is Lipschitz. Hence by concentration of measure, $\|\hat{f} - f^*\|_2$ concentrates around its median, m^* say. A bound for the median m^* is as follows. The subdifferential of $\beta \mapsto \|\beta\|_1$ is equal to

$$\partial\|\beta\|_1 := \{z \in \mathbb{R}^p : \ \|z\|_\infty \le 1, \ \beta^T z = \|\beta\|_1\}.$$

The vector β^* satisfies the Karush-Kuhn-Tucker conditions

$$X^T(X\beta^* - f^0) + \lambda z^* = 0$$

where $z^* \in \partial \|\beta^*\|_1$. For any $S \subset \{1, \ldots, p\}$ let z^*_{-S} be the vector z^* restricted to the complement of the set S.

Theorem *Let $S \subset \{1, \ldots, p\}$ be a subset of the variables such that $\|z^*_{-S}\|_\infty < 1$. Let $\lambda(1 - \|z^*_{-S}\|_\infty) \geq \sqrt{2n \log(4p)} + \sqrt{2n \log(2)}$. Then have*

$$m^* \leq \sqrt{|S|} + \sqrt{2 \log(2)}.$$

The flavour of this result is that the squared estimation error is roughly the number of variables (columns of X) needed to approximate the signal f^0. In other words, the estimator adapts to the sparsity of the approximation f^* of f^0.

Moreover, under certain conditions on the design X one can show that $\sqrt{|S|}$ is of small order $\|f^* - f^0\|_2$. Thus, typically, the approximation error dominates the estimation error. Applying random matrix theory, the design conditions are met with large probability when X is a random matrix with independent rows from an appropriate distribution.

A bound for the approximation error $\|f^* - f^0\|_2$ can be established using convex analysis. Here occurs a geometric quantity that can be thought of as an ℓ_1 version of canonical correlation. We show in some examples that the obtained bound is tight.

Finally, we present some extensions to other norms and loss functions for non-Gaussian models.

References:

P.C. Bellec and A.B. Tsybakov. Bounds on the prediction error of penalized least squares estimators with convex penalty, 2016. arXiv:1609.06675v1.

R. Tibshirani. Regression analysis and selection via the Lasso. *Journal of the Royal Statistical Society Series B*, 58:267–288, 1996.

S. van de Geer and M. Wainwright. On concentration for (regularized) empirical risk minimization, 2016. arXiv:1512.00677v2, to appear in *Sankhya*.