Approximation of cross-validation error for linear regression with piecewise continuous nonconvex penalties

Ayaka Sakata[1] and Tomoyuki Obuchi[2]

1 Department of Statistical Inference & Mathematics, Institute of Statistical Mathematics

2 Department of Mathematical & Computing Science, Tokyo Institute of Technology

We investigate the signal reconstruction performance of sparse linear regression in the presence of noise when piecewise continuous nonconvex penalties are used. Among such penalties, we focus on the smoothly clipped absolute deviation (SCAD) penalty. The contributions of this study are three-fold:

We first present a theoretical analysis of a typical reconstruction performance, using the replica method, under the assumption that each component of the design matrix is given as an independent and identically distributed (i.i.d.) Gaussian variable. This clarifies the superiority of the SCAD estimator compared with ℓ_1 in a wide parameter range, although the nonconvex nature of the penalty tends to lead to solution multiplicity in certain regions. This multiplicity is shown to be connected to replica symmetry breaking in the spin-glass theory, and associated phase diagrams are given. We also show that the global minimum of the mean square error between the estimator and the true signal is located in the replica symmetric phase.

Second, we develop an approximate formula efficiently computing the cross-validation error without actually conducting the cross-validation, which is also applicable to the non-i.i.d. design matrices. It is shown that this formula is only applicable to the unique solution region and tends to be unstable in the multiple solution region. We implement instability detection procedures, which allows the approximate formula to stand alone and resultantly enables us to draw phase diagrams for any specific dataset.

Third, we propose an annealing procedure, called nonconvexity annealing, to obtain the solution path efficiently. Numerical simulations are conducted on datasets to examine these results to verify the consistency of the theoretical results and the efficiency of the approximate formula and nonconvexity annealing. The characteristic behaviour of the annealed solution in the multiple solution region is addressed. Another numerical experiment on a real-world dataset of Type Ia supernovae is conducted; its results are consistent with those of earlier studies using the ℓ_0 formulation.