Response to Associate Editor

January 20, 2017

We appreciate the helpful feedback from the reviewer. We have addressed your questions and comments. Below we give a point-by-point response to each of the questions posed by the reviewer:

1. The authors provided insufficient justification for using a large number of regularization parameters

   We have updated Section 1 with more examples of problems with multiple regularization parameters.

2. Some important details have been omitted from the empirical results. Full reproducibility is expected

   We have included more details in Section 3 regarding the simulation studies.

3. The empirical results cover a relatively small range of scenarios

   We've added a new example in the paper on low-rank matrix completion. Due to its significantly different data structure, we needed to modify our method to calculate the gradient of the validation loss. In particular, we needed to add another step to obtain the gradient optimality conditions.

4. The technical conditions seem quite restrictive from a practical point of view, and need further explanation/justification (or weakening)

   As noted by Reviewer 2, our paper does not need the strict convexity assumption that we have mistakenly included. This assumption has been removed. Our technical conditions as written are now applicable to many popular penalized regression settings.

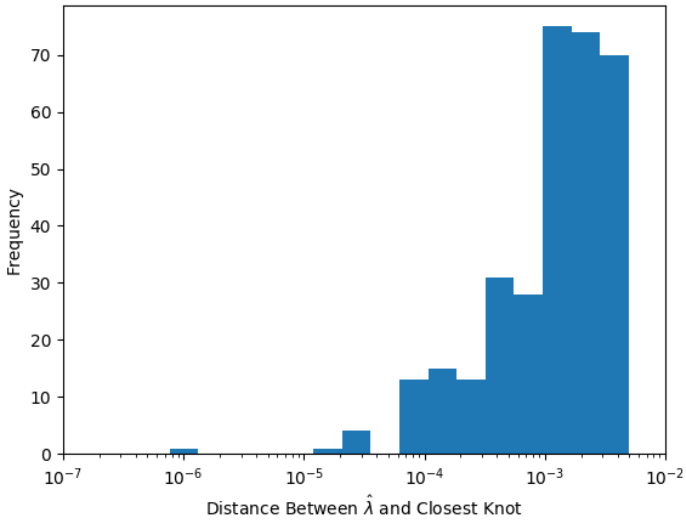
5. Make sure to provide all code for all experiments

   It is.

6. They state in the abstract and in the paper: "For many penalized regression problems, the validation loss is actually smooth almost-everywhere with respect to the penalty parameters." I assume that almost everywhere means "almost everywhere with respect to Lebesgue measure." But of course, this same statement is true of the objective itself, for which gradient descent cannot be used. The relevant condition seems to be to be whether the loss is smooth almost everywhere with respect to the probability measure induced by the true sampling model, which is not the case for e.g. lasso/group lasso/etc. Can the authors please clarify and elaborate on this point?

   The editor is correct in that the condition of interest is whether the loss is smooth almost everywhere with respect to the probability measure induced by the true sampling model. However we believe that even under the true sampling model, the probability measure is

Figure 1: Histogram of the closest distance between the $\hat{\lambda}$ and a knot along the lasso path



the same except over the penalty parameter knots where a set of measure zero where the validation loss is not smooth with respect to the penalty parameters. Our implementation of gradient descent searches the penalty parameter space agnostic to the fact that the training criterion is not smooth. Therefore it is very unlikely for gradient descent to end up exactly at a knot location. In addition, we believe that it is very unlikely for the penalty parameter that minimizes the validation loss to be where the validation loss is non-smooth. For a small perturbation in the validation dataset, it would seem that the penalty parameter would also perturb a small amount. It seems unlikely that the penalty parameter would jump to another value where the validation loss is non-smooth.

We performed a simulation study to see if our believe holds true. We considered a simple case of the lasso with 50 covariates and plotted the distance between the lasso parameter that minimizes the validation loss to the closest knot along the lasso path. In the simulations, lasso parameter that zeroed out all the features was on average around 1. The response was generated data from the model

$$y = X\beta + \sigma\epsilon$$

where $\beta = (1, 1, 1, 0, ..., 0)$ and ϵ and X were generated from a standard Gaussian distribution. We then found the lasso path for the training criterion

$$\hat{\beta}(\lambda) = \arg\min_{\beta} \|y - X\beta\|_T^2 + \lambda\|\beta\|_1$$

In addition, we calculated $\hat{\lambda}$ that minimizes the validation error using grid search:

$$\hat{\lambda} = \arg\min_{\lambda} \|y - X\hat{\beta}(\lambda)\|_V^2$$

As shown in the Figure 6, the regularization parameter at each iteration is rarely exactly at the knots in the lasso path.