Response to critiques for article EJS2002-023RA0

Tun Lee Ng and Michael A. Newton

February 2021

We thank the associate editor and referee for extremely useful comments that have allowed us to elaborate our work and fully restructure the manuscript. Briefly, our response entails merging and reorganizing previous theorems/proofs into three more inclusive results about the random weighting solutions: one on conditional model-selection consistency (Theorem 3.1), one on convergence (Theorem 3.2) and one on asymptotic conditional distribution (Theorem 3.3). Within each of these, various conditions on weight distributions and penalty structure have been consolidated compared to the original manuscript, thanks to your suggestions. Furthermore, we have in the interim established that a novel two-step modification of random weighting allows for a unification of statements about uncertainty in the selection process and distributions within a selected model. As summarized more clearly below, the new Theorem 3.4 finds a common convergence rate for the penalty parameters λ_n assuring appropriate selection and distributional properties.

In addition to the substantial reorganization and extension of the main theoretical contributions, the revision addresses two other major concerns from the original paper. We make a more well-informed discussion of the relationship between random weighting methods, Bayesian approximation, and bootstrap sampling theory. We also include a section with numerical studies. Guided by simulation designs in the recent literature, we present a thorough numerical study of the two-step weighting method and we also demonstrate its performance in a benchmark data set. We hope you find the quality and significance of the revised work to meet the EJS standards.

In the following we provide responses to specific points, numbering comments as **ReviewerID.CommentNumber**.

1 Associate Editor

In this paper, the authors consider the sparse linear regression problem and study asymptotic properties of the weighted likelihood bootstrap for the lasso.

(AE.1) I think to make the paper sensible, it should be concentrating on the results that are interesting in the modern context. The fixed p case is not much interesting and there is no need to separately state and prove this case (but may be absorbed inside a general statement and proof). Section 5 gives the impression that the results are valid for a lot more general scheme of weights. If that is fully

true, I hardly see any point in doing the special case of exponential first, and then repeat the results for the general case (with appropriate conditions on the weights). The results should be unified, but if this is not possible, the results may be just restricted to the exponential case. The same remark applies to the weights on the penalty. A general form allowing flexibility in weights in both loss and penalty is preferable, from which all individual cases may be concluded. In summary, the paper at its present form is not attractive, but there are potentially interesting results. It appears that the paper may be substantially condensed by focusing on the setup of interest only and results using general weights.

Reply: Thank you for your helpful suggestion that leads to a major revision of the original manuscript. In response, we make a concerted effort in condensing and generalizing the original results. Specifically, the original Theorems 3.3 and 3.4 – about conditional model selection consistency for fixed p and growing p_n settings – are now combined together into a single theorem (Theorem 3.1 in the revised manuscript) that handles growing p_n setting. The fixed p setting could then be inferred as a special case of this new theorem.

Moreover, we completely remove Section 5 in the original manuscript that contained all corollaries that dealt with different random-weight distributions and different weighting schemes. Instead, all theorems in the revised manuscript are now generalized to allow flexibility in random-weight distributions, and the exponential weights could then be inferred as a special case. Besides that, these theorems also cover results for all three different weighting schemes (no penalty weights, a common penalty weight, and different penalty weights) that we consider in our paper. It is interesting to note that: even though the different-penalty-weight case seems like a general case for the other two weighting schemes, it actually involves more complicated expressions that necessitate markedly different constraints. This warrants separate consideration for these different weighting schemes in Proposition 3.1 and Theorem 3.1 of the revised manuscript.

Furthermore, we completely restructure all our proofs in the Appendix to align with these consolidated theorems. The preceding lemmas in the Appendix section serve as building blocks for the proofs of the theorems in the main text. The proofs are now leaner, more organized and easier to follow than the ones in the original manuscript.

(AE.2) Theorem 3.1 about a randomly weighted estimator's convergence is not useful for inference because no one would be willing to use this as an estimator (far more complicated than the lasso without offering any additional benefit). The distributional approximation in Theorem 3.2 in the present form is not attractive, but it would be if p is allowed to grow and a sparse normal approximation is established. Theorem 3.4 is the only result of interest apart from a possibly upgraded Theorem 3.2.

Reply: Thank you for your constructive feedback that leads to a substantial extension with new results in the revised manuscript. First, the theorem about conditional model selection consistency in growing dimensional setting (originally Theorem 3.4 in the previous manuscript) – an important result in our manuscript – is now presented first as Theorem 3.1 in the revised manuscript. Next, we use the theorems about conditional consistency and conditional asymptotic normality (when true $\boldsymbol{\beta}_0$ is not sparse) in fixed p setting (orginally Theorems 3.1 and 3.2 in the previous manuscript) to illustrate the conflicting demands on regularization λ_n in terms of estimation-accuracy and model-selection. Consequently, even under fixed p setting, there is no single λ_n that enables the random-weighting samples to have conditional sparse normality (or conditional oracle property), i.e. simultaneously achieving conditional model selection consistency and attaining conditional asymptotic normality on the true support of $\boldsymbol{\beta}$. All these results are now covered in Section 3.1 of the revised manuscript.

Subsequently, in Section 3.2, we introduce new results by considering an extension to our random-weighting approach. We retain the framework of repeatedly drawing random weights and optimize the weighted LASSO objective function, but now the optimization consists of two steps: the first step serves as a variable-selection step, and the second step involves a weighted least-squares estimation (with the same weights that we use in the first step) on the selected variables. This could be seen as the random-weighting version (i.e., weighted bootstrap) of Liu and Yu (2013)'s LASSO+LS estimator (LASSO variable selection followed by Least-squares estimation for selected variable). We proceed to establish that the random-weighting samples obtained from this extended framework achieve conditional sparse normality under growing p_n setting with a common λ_n in Theorem 3.4, which is an improvement to the results we have in Section 3.1.

(AE.3) The centering at the least square estimator is misfit too, as centering at the lasso estimator appears to be more appropriate in this context.

Reply: We generalize our theorem on fixed-dimensional conditional asymptotic distribution (now Theorem 3.3; previously Theorem 3.2) such that centering could be done at any strongly consistent estimator that satisfies certain regularity condition. Two potential candidates for centering are the least square estimator, as well as the lasso estimator with regularization parameter $\lambda_n = o(\sqrt{n})$. We acknowledge that centering at the lasso estimator lies more closely to a frequentist's interpretation about bootstrapped samples, which we mention in Section 4.2 of the revised manuscript.

(AE.4) Although the authors put a Bayesian emphasis in the paper including the title, to me, the results appear to be about exchangeable bootstrap for the lasso, at least in the way it is presented in its current form. Quite contrary to the title, the procedure has not been shown to approximate a posterior distribution in some well-defined setting. I don't find the discussion in Section 4.1 convincing. For a sparse prior, the posterior distribution is not approximated by a non-singular

normal, but by a mixture of different dimensional normal (or under a stronger condition, by a single lower-dimensional normal); see Castillo, Schmidt-Hieber and van der Vaart (2015, Annals of Statistics, Vol 43, pp 1986-2018 — this paper should be cited). The case of fixed p is not much interesting. This is not to say that the weighted bootstrap methodology is not useful for inference, but its connection with a conventional posterior distribution is questionable. The obtained results are thus about the bootstrap, and should be viewed in that way. There are some results on bootstrap for the lasso (some negative, some positive), and a thorough comparison with those results is essential.

Reply: Our work is motivated by recent developments in random weighting (e.g., Newton et al. (2020), Fong et al. (2019), Bissiri et al. (2016)) with a focus on approximate posterior inference, but we recognize limitations of our theoretical findings on this matter. Hence, we de-emphasize the notion of approximate posterior inference in our revised manuscript by removing it from the title of our paper and by restructuring the Introduction and Discussion to provide a more well-informed context for the new findings (Sections 1.1 and 4.1).

We thank the Associate Editor for pointing us to the work of Castillo et al. (2015) that advances our understanding of asymptotic posterior distributions, which we have cited accordingly in the revised manuscript. We further improve our discussion in Section 4.1, by mentioning that the traditional Bernstein-von-Mises limit is attained only if true $\boldsymbol{\beta}_0$ is not sparse. Meanwhile, in the revised manuscript, we include new results on the extension of the random-weighting framework with a two-step procedure. The newly added Theorem 3.4 affirms that the conditional distribution of the resulting random-weighting samples amasses around the true support of $\boldsymbol{\beta}$, and these samples attain asymptotic Gaussian distributional behavior on this true support. Theorem 3.4 is therefore comparable to Corollary 2 in Castillo et al. (2015), although different techniques are deployed; for instance we consider almost sure weak conditional convergence, whereas Castillo et al. (2015) considers sample average total-variation distance convergence, and we have no explicit prior structure. Yet the basic message of both is that the mass of the posterior distribution, on the one hand, and the random-weighting distribution, on the other, are similarly concentrating on the correct model subset according to the same Gaussian law.

In addition, more discussions are added in this revised manuscript to bring out the message that whilst our work was motivated from a Bayesian perspective, it has also found some resemblance (and contributed) to the bootstrap literature. Specifically, we dedicate Section 1.2 to go through the bootstrap literature and to compare them with our random-weighting approach. We argue that our random-weighting approach (both original framework and the extended version) has meaningful sampling theory interpretation in Section 4.2.

(AE.5) Numerical performance should be reported and a thorough comparison with comparable results from the bootstrap and the Bayesian literature should be presented.

Reply: In Section 5 of the revised manuscript, we include new simulation studies to compare random-weighting samples (obtained from the extended framework) with samples obtained from the Bayesian LASSO method as well as samples from a standard LASSO residual bootstrap (which is a commonly used and easily implementable bootstrap approach). In particular, our simulation results illustrate the potential of the random-weighting approach to give surrogate Bayesian samples. Besides that, in certain simulation cases, the random-weighting approach outperforms the residual bootstrap approach in terms of higher coverage probabilities for two-sided 90% confidence interval with comparable average interval widths. A benchmark data example is also included to illustrate our extended random-weighting methodology.

(AE.6) The initial part of Appendix A is too familiar, and may be dropped.

Reply: Yes, we have removed that part from our current revision.

2 Referee

(R1.1) The article does not seem to have any numerical results, which I find a bit surprising. Can you elaborate? Your aim is to use this WBB scheme to approximate certain posterior laws, which itself can be found using more direct approaches (Gibbs sampling?). So, it seems reasonable to provide some numerical results.

Reply: Thank you for bringing this matter to our attention. In fact, this issue is partially covered in our response to **(AE.4)**. In our revised manuscript, we cover in Section 1.1 about our original motivation that stems from a Bayesian perspective, which encompasses some discussion about the reasons behind the advent of various random-weighting algorithms in the literature – including the need for alternative methods to obtain posterior samples. In Section 5 of the revised manuscript, we also include new simulation studies that compare the Bayesian LASSO (which is a fully Bayesian model) samples and our random-weighting samples. In particular, our simulation results illustrate the potential of the random-weighting approach to provide surrogate Bayesian samples.

(R1.2) If I consider the estimator in (1.4), and repeatedly compute it B times, it is likely that the each time the sign vector of $\widehat{\boldsymbol{\beta}}_n^w$ will vary. So, as an user, how do I interpret Theorem 3.3? Which sign vector does the user decide to use? It does not seem, that you aim to average out the B different estimators, as is done in bootstrapping. Please clarify this issue.

I can imagine that in case of the asymptotic distributional result (Theorem 3.2), you can use averaging over B samples, to approximate the posterior limit law mentioned in Section 4.1. Is there a similar interpretation about model selection?

I think, you should more clearly state the implications of random weighting in Theorem 3.3 and the earlier Proposition.

It would be also helpful if you discuss the use of the consistency result (Theorem 3.1) in context of how it would be usable for posterior inference? I assume there is no fixed β_0 in the Bayesian approach, so what are you trying to capture?

Reply: Thank you for bringing up this matter for clarification. Since our work was motivated from a Bayesian perspective, we are interested in the posterior probability of selecting a variable (i.e. whether the sign of $\widehat{\boldsymbol{\beta}}^w_{n,j}$ is equal to zero or not) – hence the name of the theorem "Conditional Model Selection Consistency". This matter is illustrated in the newly added simulation studies in our revised manuscript, where we track the (posterior) probability of whether the j^{th} variable is selected into the model by the particular method.

By "averaging over B samples", I am assuming you are referring to obtaining certain point estimates based on the sampled marginal posterior distribution such as the posterior mean. In this case, getting this "point estimate" from the posterior probability of selecting a variable (or more precisely, posterior distribution of signs) is akin to doing a variable selection, where the selected model consists of variables which has a higher posterior probability of getting selected than getting discarded.

Thank you for your helpful suggestions. Theorems 3.3 and 3.4 have already been consolidated together and presented as Theorem 3.1 in the revised manuscript. We have included more explanations about the implications of the theorems in the revised manuscript. In particular, for theorem concerning conditional model selection consistency, we can interpret the result as concentration of the conditional distribution of the signs of $\widehat{\boldsymbol{\beta}}^w_n$ around the neighborhood of the true signs of $\boldsymbol{\beta}$ as $n \to \infty$.

Bayesian asymptotics do assume existence of true parameter $\boldsymbol{\beta}_0$, and posterior consistency results indicate concentration of posterior distribution around the neighborhood of $\boldsymbol{\beta}_0$. This issue is discussed in more detail in our response to **(R1.9)**.

(R1.3) The Introduction is very complicated and does not speak of the main goal in simple language.

Reply: Thank you for your constructive feedback. In our revised manuscript, we rewrite our Introduction section to highlight the main point of our work – random-weighting in LASSO regression. We then dedicate two subsections to cover our motivation that stems from approximate posterior inference and how our work relates to existing frequentists' bootstrap literature.

(R1.4) There are some papers on perturbation bootstrap for penalized regression estimators (mainly Lasso and variants), by Das and Lahiri (2019) and Das et al. (2019). You may want to compare your methodology with theirs and see what is new in your case.

Reply: Thank you for suggesting the helpful references on the bootstrap literature. These two articles are an integral part of our citations in Section 1.2 of our revised manuscript that deals with literature review about LASSO-related bootstrap algorithms. Furthermore, we also argue in Section 4.2 of the revised manuscript that the theoretical results concerning our random-weighting approach have meaningful sampling theory interpretations.

(R1.5) In Theorem 3.2, why do you center at the OLS? I expect to see centering at $\boldsymbol{\beta}_0$. Your result mimics a bootstrap distributional consistency result, but, the target here is $\boldsymbol{\beta}_0$ itself.

Reply: This issue is partially covered in our response to **(AE.3)**. We generalize the fixed-dimensional conditional asymptotic normality result (previously Theorem 3.2; now Theorem 3.3 in the revised manuscript) such that centering could be done at any strongly consistent estimator that satisfies certain regularity condition. Indeed, our result mimics a bootstrap distributional consistency result, which we discuss at length in Section 4.2 of the revised manuscript. Since we are interested in the conditional distribution (given data) of random-weighting samples, much like a frequentist's interest in the conditional distribution (given data) of bootstrapped samples given data, centering at a LASSO estimator (with appropriate rate of regularization λ_n) would be appropriate as a comparison. On the other hand, centering at $\boldsymbol{\beta}_0$ results in additional terms that depend on the realized sample path, thus convergence in conditional distribution could not be established for almost every data set. We include Remark A.1 in the Appendix to lay out the mathematical details of this different choice of centering.

(R1.6) It is not clear to me how you choose W_i's? Obviously, I see $E(W_i) = 1$ is a requirement while proving results, but other than that it seems any positive random variable with certain particular type of moments should do the job? Is it true? One needs a more refined analysis of the WBB procedure to judge its true benefits.

Reply: This issue is partially covered in our response to **(AE.1)**. In our revised manuscript, we have generalized all our theorems to allow flexibility in the choice of random weight distribution. Specifically, we could draw any i.i.d. positive random weights with finite fourth moment to achieve all the asymptotic results that we establish in Section 3 (Main Results), except for Theorem 3.4 which requires $E(W_i) = 1$.

(R1.7) I do not understand the utility of introducing weights on the penalty term? Neither, I get the reason for writing Section 5 separately? The results in earlier sections can be a special case of the latter?

Reply: Indeed, this issue is covered in our response to **(AE.1)**. In our revised manuscript, we have consolidated all our theorems in Section 3 to allow flexibility in random-weight distributions and to handle different penalty weight structures, where

exponential-random-weight case could then be inferred as a special case. Thus, we completely remove Section 5 in the original manuscript that contains all the corollaries that deal with the aforementioned generalizations.

The utility of introducing weights on the penalty term was discussed in Newton et al. (2020). In particular, they found that introducing weights at the penalty term help to improve certain numerical performances of the random-weighting approach. There is also meaningful statistical interpretation for introducing penalty weights. In particular, we mention in the beginning of Section 1 of the revised manuscript that the LASSO objective function corresponds to the log posterior density from a Gaussian likelihood and a double Exponential prior. Introducing penalty weights is seen as assigning weights on the individual prior terms, just like how we assign weights on the likelihood components.

(R1.8) I have not checked the proofs, but while looking into them, I have some questions. In the middle of page 21, in the proof of Theorem 3.1, why do you write, Conditional on the data? The term on the next line $X'\epsilon/n$ does not involve weights. Similarly, in the penultimate convergence statement on that page, it seems that quantity is a joint function of both $(\epsilon_i\ ;\ W_i)$ and hence you need a convergence under the joint probability distribution, not a conditional convergence statement. Could you please check.

Reply: Thank you for bringing up this matter for clarification. Indeed, there are two sources of randomness in the random-weighting approach, namely the error terms $\boldsymbol{\epsilon}$ and the user-defined weights $\boldsymbol{W}$. For this, we consider a common probability space with common probability measure $P = P_D \times P_W$, where P_D is the probability measure of the observed data $Y_1, Y_2, \cdots$, and P_W is the probability measure of the triangular array of random weights (as mentioned in both the original and the revised versions of the manuscript). Our asymptotic results then focus on the convergence of conditional probabilities given data, that is, given the sigma-field $\mathcal{F}_n := \sigma(Y_1, \ldots, Y_n) = \sigma(\epsilon_1, \ldots, \epsilon_n)$. In particular, we are interested in convergence of $P(\cdot\,|\mathcal{F}_n)$ almost surely P_D, i.e. convergence of $P(\cdot\,|\mathcal{F}_n)$ for almost every data set. The study of convergence of $P(\cdot\,|\mathcal{F}_n)$ under a weighted bootstrap framework is not new; it is found in papers such as Mason and Newton (1992) and Lyddon et al. (2019), which we cite in the revised paper.

Going back to your question on the proof (that segment of the proof has since been reorganized as Lemma A.8 of the revised manuscript), the term $X'\boldsymbol{\epsilon}/n$ converges almost surely under P_D to zero by the Strong Law of Large Number. Subsequently, using the Weak Law of Large Numbers, we show that for any $\xi > 0$,

$$P\left(\left\|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i(W_i - \mu_W)\boldsymbol{x}_i\right\| > \xi \middle| \mathcal{F}_n\right)$$

converges to zero almost surely under P_D, or equivalently,

$$\frac{1}{n}X'(D_n - \mu_W I_n)\boldsymbol{\epsilon} \xrightarrow{c.p.} \mathbf{0} \quad a.s.\ P_D.$$

An alternative proof (without invoking the Weak Law of Large Numbers) would be: For all $j = 1, \cdots, p$ and for any $\xi > 0$, by Chebyshev's inequality,

$$\begin{aligned} P\left(\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i x_{ji}(W_i - \mu_W)\right| > \xi \middle| \mathcal{F}_n\right) &\leq \frac{1}{\xi^2} Var\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i x_{ji} W_i \middle| \mathcal{F}_n\right) \\ &= \frac{\sigma_W^2}{\xi^2 n^2}\sum_{i=1}^{n}\epsilon_i^2 x_{ji}^2 \end{aligned}$$

which converges to zero almost surely under P_D because

$$\frac{\sigma_W^2}{\xi^2 n} \to 0$$

whereas

$$\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 x_{ji}^2 = \mathcal{O}(1)$$

almost surely under P_D due to existence of second moment of $\boldsymbol{\epsilon}$ and x_{ji} is bounded for all i, j.

We have also removed the term "conditional on data" from our proof since it sounds confusing and does not aid the reader's understanding.

(R1.9) Finally, the entire goal of the article is to approximate certain posterior laws as you mention briefly (in Section 4). So, your asymptotic results should be based on the same framework, and that means not assuming a $\boldsymbol{\beta}_0$ parameter in (1.2). Neither you can assume classical sparsity settings? Probably, this issue needs explanation in depth.

Reply: This is a good point and we have restructured the paper to make our contribution and its connection to existing work more clear. Regarding the goal of the article, the revision explains how our work originated from a Bayesian perspective (Section 1.1), especially recent results from nonparametric Bayesian analysis. The work also has strong connections to frequentists' bootstrap methods, which we discuss more fully in the revised Section 1.2. After presenting the main results, we return to the issue in the discussion (Section 4), arguing that random weighting has meaningful Bayesian and sampling theory interpretations, in the framework where a true parameter $\boldsymbol{\beta}_0$ is natural in both perspectives. We recall classical Bayesian asymptotic results, such as the Bernstein-von-Mises Theorem that guarantees asymptotic normality of posterior measures. We also discuss connections to more recent work involving sparsity. Castillo et al. (2015) assumed the existence of sparse $\boldsymbol{\beta}_0$, and established Bayesian asymptotic results for a high-dimensional Bayesian linear regression model. Others, such as

Narisetty and He (2014), also assumed classical sparsity settings when they established posterior model selection consistency for a Bayesian model. In Section 4.1 we relate random weighting results to these Bayesian asymptotic results. See also our response to **(AE.4)**.

References

Pier Giovanni Bissiri, Christopher C. Holmes, and Stephen G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 78(5):1103–1130, 2016.

Ismael Castillo, Johannes Schmidt-Hieber, and Aad van der Vaart. Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986–2018, 2015.

Debraj Das and S. N. Lahiri. Distributional consistency of the lasso by perturbation bootstrap. *Biometrika*, 106(4):957–964, 2019.

Debraj Das, Karl Gregory, and S. N. Lahiri. Perturbation bootstrap in adaptive lasso. *The Annals of Statistics*, 47(4):2080–2116, 2019.

Edwin Fong, Simon Lyddon, and Christopher C. Holmes. Scalable nonparametric sampling from multimodal posteriors with the posterior bootstrap. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

Hanzhong Liu and Bin Yu. Asymptotic properties of lasso+mls and lasso+ridge in sparse high-dimensional linear regression. *Electronic Journal of Statistics*, 7:3124–3169, 2013.

S. P. Lyddon, C. C. Holmes, and S. G. Walker. General bayesian updating and the loss-likelihood bootstrap. *Biometrika*, 106(2):465–478, 2019.

David M. Mason and Michael A. Newton. A rank statistics approach to the consistency of a general bootstrap. *The Annals of Statistics*, 20(3):1611–1624, 1992.

Naveen Naidu Narisetty and Xuming He. Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, 42(2):789–817, 2014.

Michael Newton, Nicholas G. Polson, and Jianeng Xu. Weighted bayesian bootstrap for scalable posterior distributions. *The Canadian Journal of Statistics*, 2020.