We are grateful to the reviewers for their helpful feedback. Recall that, in this paper, we observe that existing approximate cross-validation (ACV) methods may be slow and inaccurate in GLM problems with high data dimension (D). To address this issue, we provide a new ACV method, which we show is both fast *and* accurate for approximately low-rank (ALR) data. And we provide an efficiently computable upper bound on the error of our ACV method.

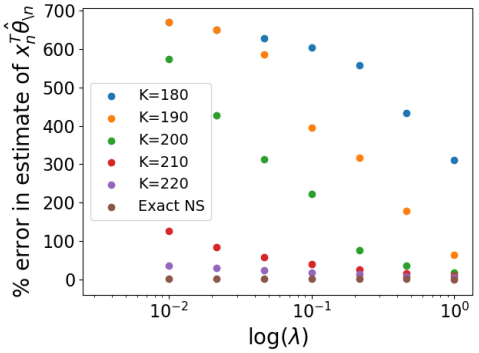
We agree with **R1** that we do not focus on asymptotic analysis. We see our focus on finite-sample bounds as a strength rather than a deficit. By considering the data at hand rather than an imagined infinite population (which may often be subject to misspecification concerns), we provide computable bounds that users can apply to their own data, with accuracy guarantees for their particular application. That being said, we do provide some conditions under which our bounds asymptotically go to zero (and so must be tight) in Corollaries 1 and 2. We believe a full treatment of bound asymptotics would be a major undertaking and out of scope of the current paper.

R1 suggests using principal components analysis (PCA) to reduce dimensionality of the covariate matrix. We see two potential interpretations of this suggestion. (1) We care about the full high-dimensional GLM, but we could use PCA as an approximation within existing ACV methods — as an alternative to our approximation. (2) Every problem is already low-dimensional because practitioners always apply dimensionality reduction (e.g. PCA) first. The issue in both cases is that PCA is task agnostic; e.g. the covariates associated with a response need not be in the first components of PCA [Jolliffe 1982]. So, for (2), there is real interest in the full-covariate GLM. For (1), we note that our method does encode task, namely by reducing the Hessian (a function of both covariates and predictors) via its associated quadratic form (lines 91–92). We will illustrate with an empirical comparison to the proposal in (1) in our revision.

R1 suggests citing "Nouridine to illustrate the importance of leave-one-out estimators for prediction risk estimation." We will provide more discussion of the benefits of LOOCV (vs K-fold CV) as shown in e.g. Figure 1 of Rad and Maleki [2020]. We searched for relevant papers by (e.g.) Noureddine El Karoui but did not find one. If R1 could provide an exact citation, we would be happy to include any appropriate references.

We agree with **R3** that having a hyperparameter K is undesirable. However, we note that K is different from most "tuning parameters," where too-large or too-small values are both bad. For K we instead recommend that practitioners use the largest K allowed by their computational budget. Crucially, our cheaply computable error bounds allow a practitioner to check if this K allows sufficient accuracy.

We agree with **R3 and R4** that we should be more clear what is meant by ALR data. In all rigorous mathematical statements, we use the typical definition of ALR corresponding to having only a few large singular values (e.g. Corollary 1). We will add a more exact statement and discussion early in the paper. As **R4** notes, the issue is compounded by our citation of Udell and Townsend [2019], who use a different ALR definition. Despite the difference, we still believe their work helps motivate why many matrices are ALR in the spectral sense (cf. the success of our method on real datasets). We will clarify this point in the paper. **R3** also suggests we may need additional assumptions to estimate Q_n well with our method. While the example given by R3 shows that using the top eigenvectors and eigenvalues can lead to poor estimates of some Q_n, we note that Prop. 3 implies this issue cannot happen *on average*. As our interest is in computing Q_n for all n, we believe that Prop. 3 does show that H being ALR is sufficient for the success of our method.



R4 asks [Q2] about Prop. 4 in App. E.1. Recall γ_d are the eigenvalues of $\sum_n b_n b_n^T$; so $\gamma_D = \sum_n \langle b_n, v_D \rangle^2$. If $b_n \propto v_D$ for some n, we have $\gamma_D \geq \langle b_n, v_D \rangle^2 = \|b_n\|_2^2$. We will clarify this logic.

R4 observes [Q3] that our experiments use $\lambda \geq 1$ and wonders about $\lambda < 1$. We chose λ to make the optimization problems sufficiently regular, so exact CV runtime would be reasonable for the larger experiments. Still, we appreciate R4's caution. To address dependence on λ, we ran a $N = 600, D = 400$ synthetic logistic regression task of approximate rank $R = 200$ (the singular values follow $\sigma_{200} \approx 10.0, \sigma_{201} \approx 1.0$) for a range of λ values. In the figure, we show percent error in the estimate of exact CV, $x_n^T \hat{\theta}_{\setminus n}$, via the NS approximation across different settings of K and λ. The errors' change with λ is minimal, except for values of K near R. We will include real-data experiments for the "low-λ" regime in our revision.

R4 asks [Q1]: what are the failure modes of the approximation in Algorithm 2, line 5? This approximation has two parts: (1) using a single iteration of the subspace iteration method and (2) a diagonal approximation to the Hessian. For (1), the subspace iteration method converges in the fewest iterations when there is a sharp dropoff after the first K singular values. For (2), the Hessian is the least diagonal when the covariates are all linearly dependent and completely diagonal when they are orthogonal. We will expand on these points with experiments in an updated version of the paper.

■ *References*: I. T. Jolliffe. A note on the use of principal components in regression. Applied Statistics. 1982.