We thank the reviewers for their helpful comments. First, we briefly recap our contributions. We provide the first computationally efficient and accurate method for approximate cross-validation (ACV) in the following setting: structured models fit with MLE or MAP, where the task is a structured prediction. In so doing, we show (theoretically and empirically) that IJ-style ACV ideas can be applied at an inexact (MLE or MAP) optimum; the latter is a new result even for traditional ACV. By contrast, previous IJ approximations (e.g. Giordano et al. [2019]) do not apply to structured prediction and also assume an exact optimum; Newton-step (NS) methods (e.g. Rad and Maleki [2020]) *could* be applied to structured prediction, but we show (lines 118–121, Appendix G) they are not computationally efficient; and Bürkner et al. [2020] focus on leave-future-out for time series (but not other forms of structured prediction) within a fully Bayesian (rather than MLE or MAP) framework.

Novelty: (A) R2 is concerned that we are applying a "general framework for ACV from a previous paper." But note that Giordano et al. [2019], Beirami et al. [2017], Koh & Liang [2017], Koh et al. [2019], Stephenson & Broderick [2020], and Wilson et al. [2020] all work in the framework of Eq. (1) of Giordano et al. [2019], whose additive form does not allow for the structured tasks we consider here. Observe that Bürkner et al. [2020] extend Bayesian ACV methods to leave-future-out CV for time series models; that work demonstrates the non-trivial nature of extensions to structured tasks. (B) R2 also describes our new theoretical error bounds as applying to HMMs and CRFs. But note that our error bound (Prop. 2) applies much more broadly — both to all of the structured tasks we consider here, including more general MRFs — as well the Giordano et al. [2019] framework with inexact optima (a result which was not previously established but is more practically relevant than results requiring exact optima). (C) R2 is concerned that our methods apply only to models with latent processes. While we believe models with latent structure represent a widely used and interesting class of models, we note that we do present methodology (Sections E, F, and Algorithm 4) and experiments (Section 5, lines 265–291) for CRFs, which contain no latent variables. And our inexact optimum theory applies beyond structured models.

Additional experiments: R2 asks for comparisons against pre-existing IJ or NS ACV methods. Unfortunately, there are no existing IJ methods that apply to the tasks we consider. To the best of our knowledge, NS has not previously been applied to these problems, but we do consider it. In lines 118–121 and Appendix G, we discuss the computational challenges of NS methods, which arise from computing and inverting a new Hessian for each CV fold. E.g., for the time-varying Poisson process problem ($T = 50{,}400$), IJ computation across all $1{,}000$ folds took about 12 minutes. By contrast, the NS approximation here would take roughly *113 hours*. On sufficiently small datasets, though, the NS and our IJ can be much closer in performance. We will include NS timing across datasets, and discussion, in a revision of our main text.

Assumptions: We fully agree with R1 that our assumptions could be stated more clearly. We will collect them in the text, and we note them here for clarity. We require (1) a model fit via optimization, (2) twice differentiability of the model objective and invertibility of the Hessian matrix at the initial model fit $\hat{\Theta}$, and (3) the ability to write the model fits across CV folds, $\hat{\Theta}^{\backslash \mathbf{o}}$, as optima of the same weighted objective for all folds $\mathbf{o}$. These conditions are satisfied by a broad class of empirical/regularized risk minimization problems, including widely used probabilistic models of structured data fit via MLE or MAP. We further clarify that *any* optimization method (stochastic or exact; with or without early stopping; etc.) may be used to obtain the initial fit $\hat{\Theta}$ for input to Algorithm 1. When computing the actual derivatives in Algorithm 1, we assume that the gradients are computed exactly at $\hat{\Theta}$. Computing these gradients involves a single pass through the dataset before approximating CV for all folds.

Structured tasks: R1 notes that we need to clarify that our novelty here is for the combination of structured models *paired with* structured tasks. We completely agree and will be sure to make this point very early in a revised manuscript.

Number of folds: R4 is concerned that 500 folds in our neural CRF experiment is "too large to use in practice." Fig. 1 of Rad and Maleki [2020] show leave-one-out CV (with hundreds of folds) substantially outperforming $\{3, 5, 10\}$-fold CV at estimating out-of-sample error. A goal of modern ACV methods is to allow a larger number of folds, and therefore more accurate final estimate overall, in practice.

Quantitative results: R4 is concerned that our paper "lacks quantitative results." We interpret this concern to mean that the reviewer would like to see approximation error and timing reported in numbers, in addition to being plotted in the figures we have included in our submission. We will be sure to include both numbers and figures in a revision.

Inexact initial fit: R4 is would like us to clarify how our method provides a solution to the inexact initial fit problem. We prove (Section 4) that the IJ approximation error increases smoothly with the error in the initial fit. Proposition 2 explicitly bounds the approximation error and suggests that we can use "good enough" initial fits without substantively sacrificing ACV accuracy. Our neural CRF experiments in Section 5 provide empirical confirmation.

References (beyond original paper) ○ P. W. Koh, K. S. Ang, H. Teo, P. Liang. NeurIPS'19.
○ A. Wilson, M. Kasy, L. Mackey. AISTATS'20.