We thank the reviewers for their time and effort reviewing our paper. We are pleased that you found our work to "solve an important problem" (R1), our method to be "elegant" (R2), and our paper to be "well written" (R3). Below we respond to the key comments and issues in order of apperance.

R1: "Diverse mini-batch Active Learning" and open-source code We will discuss the paper and consider it as an additional baseline. At the latest, source code will be made publicly available upon publication.

R2: Batch active learning by imputing labels When we write "greedy," we are referring to the sequential setting in which we alternate between querying a single data point and updating the model. When we say "naive batch" we mean ranking and taking the top b points. The proposed imputation approach could be considered as an alternative batch construction method not currently discussed in the paper. The key issue with this method is indeed the cost of performing a model update after each label is imputed. While a linear increase in runtime due to repeated model updates may be acceptable for constructing smaller-sized batches, in the large-scale settings we consider it simply becomes infeasible. Interestingly, there is a close relationship between the imputation approach and our method: in fact, our method also uses the expected (i.e., imputed) labels under the model to construct the batch in a principled way; however, importantly, it does not require making repeated predictions over the entire pool set or updating the model after each point is added to the batch. We will add a thorough discussion on this topic to a future version of the paper.

R2: Logistic vs. probit regression We agree with R2 that the discussion on logistic regression is clearer by considering probit regression to begin with. We will update the section accordingly.

R2: Extending entropy-based methods to the batch setting is not necessarily difficult We agree the wording is misleading. The main issue is how to do this efficiently for complex, non-linear models. We will clarify this point.

R2: Runtime discussion In general, constructing the batches has negligible cost (cf. Section 5 for a discussion on computational complexity) compared to updating the model, so increasing the batch size allows to decrease overall computational cost. We will add a more detailed comparison with other methods to the experimental section.

R3: "Nonmyopic active learning of Gaussian processes: an exploration-exploitation approach" While GPs suffer from scalability issues and cannot be applied to many of the domains we considered, we agree that batch active learning (AL) with GPs is under-discussed in the paper. Krause & Guestrin (2007) consider mutual information (MI; also known as BALD) as an acquisition criterion, which makes this work related to ours. However, they immediately reduce the batch formulation to the sequential greedy case. We highlight similarities and differences to sequential greedy MI/BALD in Sections 4 and 6, and will add a discussion on extending sequential greedy methods to the batch setting (cf. *R2: Extending entropy-based methods to the batch setting is not necessarily difficult*). Less relevant for our work are GP-specific details (as we focus on BNNs) and proofs relying on submodularity (as the diminishing returns property does not necessarily hold when performing approximate inference and stochastic optimization). We will expand the related work section accordingly.

R3: Clarifying first equality in eq. (4) The first equality is achieved by applying Bayes' rule to the posterior (eq. (1) in the main paper), taking the logarithm, and applying linearity of expectations:

$$\begin{aligned}\mathbb{E}_{\mathcal{Y}_p|\mathcal{X}_p,\mathcal{D}_0}[\log p(\boldsymbol{\theta}|\mathcal{D}_0 \cup (\mathcal{X}_p,\mathcal{Y}_p))] &= \mathbb{E}_{\mathcal{Y}_p|\mathcal{X}_p,\mathcal{D}_0}\left[\log p(\boldsymbol{\theta}|\mathcal{D}_0) + \log p(\mathcal{Y}_p|\mathcal{X}_p,\boldsymbol{\theta}) - \log p(\mathcal{Y}_p|\mathcal{X}_p,\mathcal{D}_0)\right] \\ &= \log p(\boldsymbol{\theta}|\mathcal{D}_0) + \mathbb{E}_{\mathcal{Y}_p|\mathcal{X}_p,\mathcal{D}_0}\left[\log p(\mathcal{Y}_p|\mathcal{X}_p,\boldsymbol{\theta})\right] + \mathbb{H}[\mathcal{Y}_p|\mathcal{X}_p,\mathcal{D}_0],\end{aligned} \tag{1}$$

where we used $\mathbb{E}_{\mathcal{Y}_p}\left[-\log p(\mathcal{Y}_p|\mathcal{X}_p,\mathcal{D}_0)\right] = \mathbb{H}[\mathcal{Y}_p|\mathcal{X}_p,\mathcal{D}_0]$. We will make this derivation clearer in the next version.

R3: Motivation for the batch AL setting can be improved This scenario is practical in a number of real-world applications, particularly when the cost of acquiring labels is high but can be parallelized. Examples include crowd-sourcing a complex labeling task, leveraging parallel simulations on a compute cluster, or performing experiments that require resources with time-limited availability (e.g. a wet-lab in natural sciences). In all these cases, being able to generate a quality batch of query points without having to wait for previous labels to be acquired can be extremely advantageous. The importance of the scenario is further evidenced by the volume of existing literature and ongoing research on this topic. See for example Hoi et al. (2006), Guo & Schuurmans (2008), Wei et al. (2015), Sener & Savarese (2018), Kirsch et al. (2019), and many more references therein, all of which are concerned with the batch AL setting. We thank R3 for making this point, and will add content motivating the batch AL setting to the paper.

R3: Evaluated datasets As a methodology paper, our goal is to demonstrate the usefulness of our proposed method in a broad range of scenarios. We performed experiments on several small- and large-scale regression and classification datasets. While we agree with R3 that the datasets evaluated in the experimental section do not necessarily reflect real-world scenarios, the experimental protocols we used resemble benchmarks from multiple important (batch) AL papers (e.g., Hernandez-Lobato & Adams, 2015; Gal et al., 2017; Sener & Savarese, 2018). In fact, our work goes beyond what is typically tractable with Bayesian approaches (e.g., *cifar10* and *year*), demonstrating the usefulness and scalability of our method. Since the performance of the method is independent of labelling cost, we argue it is sufficient to use real-world settings as motivating examples (see discussion above) and leave specific applications to future work.