Thanks to all of the reviewers for their time and effort, and both constructive and critical comments.

▷ **All Reviewers.** General Comment 1 - code and reproducibility Due to file size limits we included our code as an anonymous downloadable link in Appendix J, but we need to make this more obvious, thanks. We gladly commit to making the code publicly available on paper acceptance. General Comment 2.1 - Marginal Likelihood (ML) approximation Our ML estimate follows the standard approach in EP, which is to assume that the site approximations are fixed. Proving that "at convergence, the parameters of the approximate factors can be considered fixed" may require analyzing the energy function of our model and a good idea for future work, thanks. General Comment 2.2 - estimate of ML approximation We use the ML of the approximate GP in our work and also in EP. A previous study on GP classification [Assessing Approximate Inference for Binary Gaussian Process Classification, Kuss & Rasmussen] thoroughly compared various ML approximations, and found that the ML approximation we use matches the predictive accuracy very well. General Comment 3 - novelty We thank Reviewer 1 and Reviewer 3 for acknowledging the non-triviality of the locality property and practical efficiency of our work. Locality is our central result, and guarantees that the complex nested optimization of the Wasserstein distance reduces to a relatively simple and efficient 1-d update. Reviewer 1 puts it well, that "replacing it [KL] by the L2 Wasserstein should strike the majority of researchers as an obvious desirable improvement", and moreover that "Wasserstein distance is very hard to calculate. It is even harder to do approximate inference with it. A general procedure for approximate Bayesian inference by minimising some sort of Wasserstein distance to the posterior would be a large boon for the field." General Comment 4 - numerical stability If the CDF is accessible, Equation (5), which forms the basis of our lookup tables, is stable and it avoids divergence. Intuitively, this is because the Gaussian function in the integration is bounded by 1 and when the integration variable is very large or small, the Gaussian function approaches 0 rapidly. If the CDF isn't accessible, there are double integrations which can be computed in one pass. In such cases, the numerical stability may be problematic; investigating this would fairly deserve another paper.

▷ **Reviewer 1.** We thank reviewer 1 for their supportive comments and helpful suggestions on *e.g.* the broader impact, which we will incorporate in the final version. "Found analytically for fewer distributions than EP" While we agree that the number of analytically tractable cases for QP is probably less than that of EP, various numerical schemes may be employed and in many cases made efficient using lookup tables; see also General Comment 4.

▷ **Reviewer 2.** "No code is given" Please see General Comment 1. "not clear if the proposed method is worth it" Please see General Comment 3. "not clear how the estimate of the marginal likelihood is optimized"; "in EP one can show that at convergence, the parameters of the approximate factors can be considered fixed"; "the accuracy of the estimate of the marginal likelihood". Please see General Comments 2.1 and 2.2.

▷ **Reviewer 3.** "empirical advantages of the method are not demonstrated" We respectfully remind that in 8 out of 10 tasks, our method achieves consistently better test log-likelihoods on repeated experiments, than the strong baseline that is EP. Besides, Figure 1.a and 1.b illustrate the effectiveness of our method in alleviating the over-estimation of variances of EP. "primary motivation and reason for pursuing QP over EP" Please see General Comment 3. "No analysis of fixed points is given" Theoretical analysis of fixed points is an interesting area for future work, thanks. At present we offer an empirical analysis of fixed points as given in our experiment section. "any modifications for future work for which you could prove convergence and analyze the fixed points?" Changes such as the double loop EP [Expectation Consistent Approximate Inference, Opper & Winther] and proving the property pointed out by Reviewer 2 that "at convergence, the parameters of the approximate factors can be considered fixed" are an interesting direction for future work, thanks. However, we cannot yet rule out that our method is *already* provably convergent under appropriate assumptions. "Is there any guidance for constructing the look-up tables" Please see General Comment 1.

▷ **Reviewer 4.** "a lot of references to the Appendix . . . the paper a little hard to digest" We will take the suggestion about bringing part of the supplement to the main paper, *e.g.* with regards to more explanation and discussion on WD. We will also use the extra page in the final version for this. "I'm not sure whether the page on the locality property is enlightening and really surprising" We respectfully disagree and are deeply disappointed that the reviewer places this comment in the list of 'weaknesses'. Locality is central to our contribution, and much harder to show here than for EP. We ask the reviewer to kindly consider the broader relevance outlined in General Comment 3. "Note sure whether the authors intend to release code" Please see General Comment 1. "The degree of novelty is pretty small" Please see General Comment 3. "marginal likelihood and its accuracy" Please see General Comments 2.1 and 2.2. "numerical aspects of the L2 Wasserstein distance computations" Please see General Comment 4. "A discussion why values for $p$ different from 2 are not interesting to consider." We briefly mention $p$ different from 2 in *e.g.* line 72–73, 188–189 and Appendix B. These cases are interesting but also *even more* challenging to handle. "does not provide evidence . . . better suited in cases where EP has 'deficiencies' " The specific shortcoming of EP is over-estimation of variances as pointed out on *e.g.* lines 13, 29-31; we will clarify this even further in the final version, thanks. We give both theoretical analysis of local updates mitigating over-estimation of variances (sec. 4.3) and empirical evidence that at convergence our method predicts better and with lower predictive variances.