We thank the reviewers for the valuable comments and suggestions made. We will focus on addressing the main remarks regarding baseline, scalability, complexity and the full batch setting in the following paragraphs.

Baseline and models. The reviewers' main concern is the lack of baseline besides MCVI and the absence of a more complex model. We recognize that this gap makes the evaluation of the method difficult. Our main goal was to demonstrate that the use of OQ offers theoretical guarantee and can be efficiently used for inference. Taking into account the suggestions made by all reviewers, the new version includes three baselines: Monte Carlo Variational Inference (MCVI), Quasi Monte Carlo Variational Inference (QMCVI), Randomized Quasi Monte Carlo Variational Inference (RQMCVI). Notably, QVI converges faster on almost all experiments except on the Poisson GLM experiment where similar performance with QMC is observed (Figure 1, second column displays the result for the Forest experiment). In addition, following [7,24] and the suggestions of **reviewers 1,2,3**, we included a more challenging Bayesian Neural Network (BNN) experiment with a larger dataset. The network consists of a Multi Layer Perceptron (30 neurons ReLU activated) with normal prior weights and inverse Gamma hyperprior on mean and variance. The dimension of the latent space is $K = 62$ and regression was performed on the UCI Metro dataset with $L = 48204$ data points (see Figure 1. RQVI procedure led to computational instability). Notably, it exhibits quick convergence for QVI with a bias of 5%. Increasing the number of neurons (and thus the posterior dimension) beyond this setup results in a too large bias of 20% in the ELBO estimation. Altogether the experiment section amounts to five methods, three baselines, three models (Bayesian Linear Regression (BLR), Poisson GLM, BNN) and five datasets (Boston, Fires, Life Expect., Frisk and Metro) with learning rate analysis.

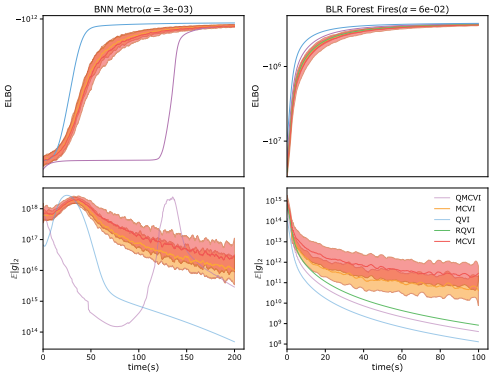

Figure 1: ELBO (first row, log scale) and expect gradient norm (second row, log scale) during the optimization procedure for the BNN model (Metro dataset, left) and the BLR model (Forest dataset, right) as function of time. Variance for MC (red area) and QMC (orange area) estimator is obtained by 20 re-run for each experiment.

Scalability. Questions were raised by **reviewers 1,2,3** about the scalability of the method with respect to dataset size and dimension. We do not claim that this method is suitable for high dimensional posteriors. We considered it to be the main limitation of the approach as it is for local HPV [24]. For a MC sample size N, when considering the d-dimensional variational distribution $X^{\Gamma_N,\lambda}$ in place for X^{λ}, we introduce a bias in $\mathcal{O}(N^{-\frac{\alpha}{d}})$ for the ELBO estimation. The number of data points L is not a bottle-neck since the complexity associated with computing the MC and QVI estimators are similar (see Eq. 7 for the cubature formula). Active research is underway to reduce bias in higher dimensions [22,26,28].

Complexity. To address the question of **Reviewer 1**, the complexity of getting an approximation of the optimal quantizer $X^{\Gamma_N,\lambda}$ is in $\mathcal{O}(N \log N)$ [33] but only needs to be constructed once and can be used throughout the inference since optimality is preserved for the variational family considered. Consequently, the construction of the optimal quantizer is not a limiting factor. It is accurate that the method will not be viable without this property.

RP gradient in the full batch setting. Reviewers 2,3 pointed out the lack of sufficient discussion about the importance of gradient variance in the full batch setting for the ELBO minimization problem. Gradient variance and CV methods are discussed more thoroughly in [7,24,5] (see L35-L42) and it is an essential issue in stochastic optimization in general. To **reviewer 2**, the gradient variance is displayed in all experiments as red shaded area on even rows (description of gradient variance evolution has been clarified as it can lead to confusion) and is computed on 20 re-reruns.

A relevant point is raised by **reviewer 2** about the full-batch setting. It is true that we consider only the variance associated with sampling from the variational family while in mini-batch sampling, the dominant term would be in $O\left(S^{-1}\right)$ for S-sized batches. We underline that i) it would not exhibit significant variance reduction except on large datasets; ii) even though it would reduce MCVI RP gradient variance, it would also reduce its norm, making it difficult to assess the relative gain for the MCVI method; iii) choosing the batch size S can be difficult, depends on other hyperparameters and is currently beyond our analysis scope. The chosen framework was motivated by extending previous studies [7,24] with full batch RP gradient to deterministic sampling. The new version includes the motivation for the choice of the full batch setting and the comparative performance of control variate and alternative sampling ([7] shows that RQMC outperforms HPV control variate [24] in a similar setting).

Other comments. As underlined by **reviewer 2**, the explanation about how to use this method for model checking can be confusing. Put simply, since QVI converges in fewer epochs, we can estimate $\mathcal{L}(\lambda)$ with its quantized counterpart $\widehat{\mathcal{L}}^N_{OQ}(\lambda)$ with a precision given by theorem 1. As pointed out by **Reviewer 4**, we agree that this approach could be better suited for IWAE/DReG/Jackknife VI. However, our derivations rely on the optimal quantizer's technical properties, and it is quite challenging to use it for these gradient estimators (more precisely, it is likely that consistency is not preserved).