We thank reviewers for their constructive feedback on our work. We are happy to see that the problem of data summarization for scalable and privacy-preserving Bayesian inference in high-dimensions is recognized as important for the community, and our approach was found technically sound and usable in real-world applications.

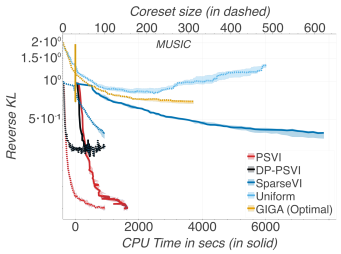
R2 (novelty): Although pseudodata-based sparsifications for VI are not new in ML, this idea is novel and nontrivial in the context of summarization. It also provides three key benefits that are specific to this setting, namely: it enables (1) summarization in high-dimensions, (2) private release of summarizations, and (3) batch construction (reducing complexity). Further, our derivations take advantage of the particular form of our objective/gradients for efficient computation.

R3 (connections with GP literature): Sparse methods in GPs have been indeed inspiring for delevoping PSVI. Similarly to (Titsias, 09) and in contrast to (Seeger et al., 03; Snelson and Ghahramani, 06), (1) we learn pseudopoints via continuous optimization on the KL between the exact and approximate posterior, avoiding overfitting by construction, and (2) do not modify the model prior, but introduce pseudopoints as variational parameters. Note that our method is applicable in both supervised and unsupervised learning settings. We will expand on these connections in our revision.

R3 (applications to broader likelihood functions): Our method is agnostic to the particular form of data likelihood functions and can be readily applied to classification problems (see e.g. the Bayesian logistic regression experiment in Section 4 and Supplement C.3.1). We will emphasize that PSVI maintains the unifying sparse exponential family interpretation for any statistical model it is applied to, with pseudopoints weights and likelihood terms corresponding respectively to the natural parameters and sufficient statistics for the family of pseudocoreset posterior approximations.

R2 (clarity of Proposition 1): In Gaussian mean inference, under the standard coresets formulation, for fixed dataset size and data dimensionality, the minimum required coreset size to reduce KL below a given threshold is bounded per Proposition 1. This bound depends on the dimension and increases when summarising a dataset of larger dimensionality, as empirically demonstrated in the difference between the KL plots of baseline methods in 200 and 500 dimensions (Fig. 2(a) and (b) of the paper)—implying impractical summary sizes for good KL. Pseudocoresets are not constrained by this bound and can achieve arbitrary KL reduction by a single pseudopoint, regardless of data dimensionality.

R5 (pseudodata and posterior quality): Learned pseudodata are explicitly optimized to approximate (in the KL sense) the exact posterior for a given statistical model, forming "approximate sufficient statistics" of the full data. Ongoing experiments showed us that pseudocoreset posteriors can be successfully applied in predictive analysis offering improvements in test accuracy/rmse. Though Hilbert coresets and uniform sampling might eventually achieve higher KL reduction for (often prohibitively) large coreset sizes, we are primarily interested in small coresets, where PSVI is outperfoming baselines in the tradeoff of KL reduction, coreset size and CPU time (required for both summary construction *and* subsequent inference); in contrast, Hilbert coresets are fundamentally constrained in this regime both due to data dimensionality (as is SparseVI as well), and information-geometric limitations (see (Campbell & Beronov, 19) and plot shown on the right).



R3 (pseudodata weights): The varational parameters size in PSVI is dominated by pseudopoints in high-dimensions. Weights seem to be a natural ingredient for data summarization, that can account for coreset points multiplicity, hence enabling more expressive sparse posteriors, without having a significant bearing on the computational cost and the robustness of optimization. Importantly weights can differentiate posterior approximations among datasets of different size. For example, removing the variational parameters w in the Gaussian mean inference experiment (Section 4), won't allow correctly adjusting the covariance of the pseudocoreset posterior, which is not a function of pseudopoints location.

R2,3,5 (private scheme): A major desideratum in Bayesian coresets is maximising the automation of inference. Using the subsampled Gaussian mechanism is a decisive step towards pursuing this goal in DP extensions of coresets: our privatisation method removes requirements on computing sensitivities for noise calibration, enables adaptive clipping of gradients guided by private statistics on pseudopoints potentials, and gives tight estimates of the accumulated privacy cost via moments accounting—the latter allows many gradient steps under DP leading to good convergence in KL in practice, even when pseudodata are initialised from an uninformed prior, potentially far from true observations (Section 4). On the other hand, privatising via noise addition in the first place requires strong public knowledge/assumptions on the (typically infinite) data likelihood sensitivities. Moreover, exponential mechanism based private selection for incremental schemes of summarization would not allow tight composition of privacy over a large number of iterations.

R4 (privacy evaluation and related work): We kept δ parameter fixed to $1/N$ over all experiments on private inference, as this allows reasonable relaxations of pure DP quarantees. Fig. 4 of the paper presents the achieved posterior approximation quality over a range of values for the ε parameter for both our method and the baseline, profiling methods behavior over the regime of strong and weak privacy guarantees. DP schemes for coresets applicable in computational geometry already exist (Feldman et al., 09; 17), whilst the idea of releasing private dataset compressions has been also pursued in kernel methods (Balog et al., 18), sparse regression (Zhou et al., 07), and compressive learning (Schellekens et al.19); however, none of these approaches is directly applicable to summarising for general-purpose Bayesian inference, which led us to the decision of comparing against a standard private VI method.

R2,3 (clarity of presentation, minor comments): We will address all typos, fix inconsistent notation, adapt sections length and clarity according to your suggestions, and expand on the noted references. Thank you for pointing these out.