We thank the reviewers for their careful evaluation of our manuscript, and are glad to see that all reviewers appreciated the clear writing and value of our new unifying theoretical framework of coreset construction.

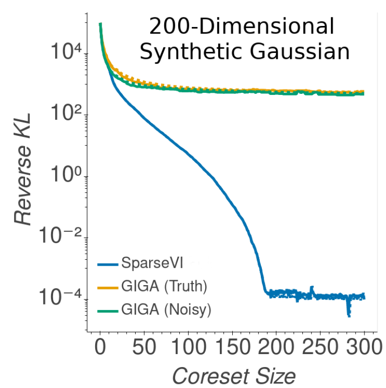
R1 (# samples/potentials): Two answers—one theoretical and one practical. On the theoretical side, we have new work that employs standard concentration inequalities to obtain the desired finite sample approximation error guarantees. This result will provide guidance on tuning S in a follow-up paper, but is outside the scope of the present work. The practical answer is simpler: use as many features as is computationally feasible. We will add a note to the final draft regarding this point. Note that Props 1 & 2 are with respect to the true norms, and do not rely on any particular approximation/sampling scheme.

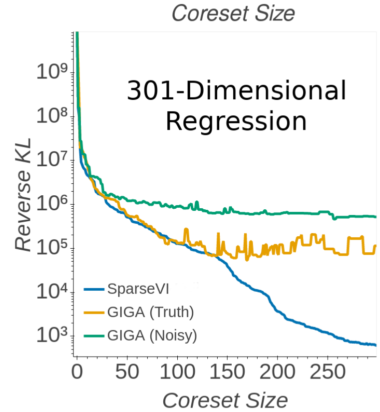
R1 (step size): γ_t was not discussed adequately in the current draft; thanks for pointing this out. As is the case with many optimization algorithms, γ_t is just a tuning parameter. Generally, it should decay such that the effect of tangent space approximation noise is eventually removed. We decided against developing Robbins-Monro-like theory for this due to both space constraints and the fact that in practice, decay rates other than $1/t$ worked best (see appendix D).

R2 (novelty/significance): We disagree that the work is incremental in view of past Bayesian coresets work, but we appreciate the point you raise, and will provide a more detailed discussion of the following points in the final draft. In particular, all prior Bayesian coreset constructions need a weighting distribution $\hat{\pi}$. Note that this is a very severe limitation; $\hat{\pi}$ isn't just a single tuning parameter, it's an entire distribution, and the fact that it is constant fundamentally limits the coreset construction (see results and discussion below right). Our first main contribution—which we believe is quite significant in the coresets literature—is to remove this bottleneck entirely. This demonstrates for the first time that fully automated, statistically rigorous coreset construction is possible. Further, past work provided no guidance on the meaning of $\hat{\pi}$; the second main contribution of the manuscript is a unifying info-geometric formulation that clarifies that $\hat{\pi}$ serves as the "anchor point" of a tangent space on the coreset manifold. We expect (and are already finding in our own ongoing work) that our new unifying info-geometric theory will open the door to many new Bayesian coreset-based inference methods. Naturally, given that our algorithm is the first instantiation of the new approach, we pay a computational price; we believe this price is well-worth it as a first foray into fully-automated coreset construction.

R1&3 (unclear plots): Thank you for your comments; in short, we agree completely. We tried to present results for numerous models / datasets / metrics despite limited space by combining results, but in hindsight "compressed" a bit too much. The two main metrics we want to use for comparison are computation time (i.e., construction cost) and coreset size (i.e., downstream inference cost). Thus we will (1) remove the iteration # plot (Fig. 3a), as it conflates these two metrics, and (2) split/format the remaining plots so that each dataset / model is clearly identifiable.

R1,2,3 (experiments): Although the underlying motivation for Bayesian coresets research is large-scale inference, the current work does not aim to explore the limits of data dimension and size in the new sparse VI formulation; this is a complex topic for which a statistically comprehensive treatment would merit a separate paper in itself (cf. Lucic et al 2018, "training Gaussian mixtures at scale via coresets" vs. many preceding papers on mixture coresets). Note that this is not unusual for a subfield still in its early exploratory phase (cf. the history of stochastic methods for regression or VI), where contributions tend to be foundational in nature, rather than computational. We focus here on the foundational problem of removing the weighting distribution $\hat{\pi}$ of past coreset methods; we verify that this is sound by demonstrating a reasonable level of performance on simple illustrative problems to avoid the confounding difficulties of more complex models. We will make these points clear in the final draft. However, to demonstrate feasibility in higher dimensions, we have increased the dimension of the synthetic Gaussian example to 200, and also applied our method to a new 301-dimensional regression problem with 300 Gaussian basis functions plus 1 constant function on 10,000 house sale records from the 2018 UK land registry dataset (both results shown right). This illustrates a key strength of our new methodology: previous Hilbert coresets eventually reach a performance limit due to using a single tangent space approximation, with quality depending on the choice of $\hat{\pi}$ (green,orange), while our method (blue) is "manifold aware" and continues to improve. To capture this in the final draft we will include the new regression experiment and increased-dimension Gaussian experiment, and increase the iteration count for all tests to clearly show the performance limit of Hilbert coresets in each. We will also highlight this limitation of Hilbert coresets in the text. Note: after submission we found that the Poisson regression log likelihood was numerically unstable in a rare circumstance that was triggered in one of the datasets (biketrips), leading to the anomalous result. We have fixed this instability.





R2,3 (minor comments): π_1 is defined in footnote 1, but we will emphasize next to Eq. (2). The Bregman comment will be removed as it is not used directly. We will address all typos. Thank you both very kindly for the careful edits!