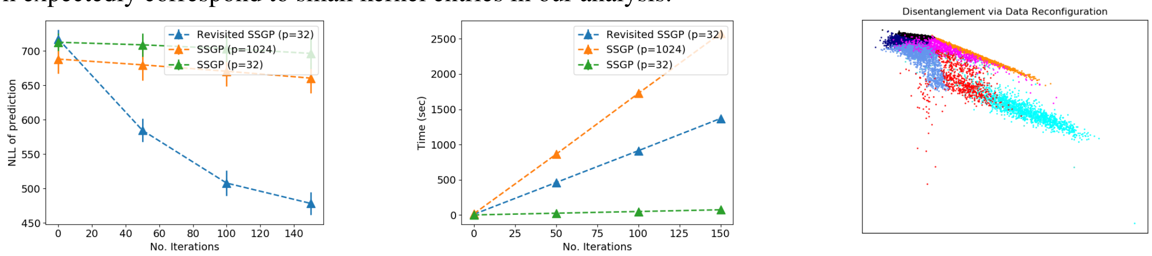
To all reviewers: We thank all the reviewers for providing valuable feedback. We have included additional experiments (on the GAS dataset with 10k data points) below to compare the scalability and performance of Revisited SSGP (rSSGP) with $p = 32$ samples to that of SSGPs with $p = 32$ and $p = 1024$ samples. Fig. 1(a) shows that rSSGP outperforms SSGP in terms of NLL of prediction (and RMSE, which is omitted here due to space constraints and will be included in the revised version) even when SSGP is using many more samples. Fig. 1(b) shows the runtime tradeoff incurred by the overhead of VAE training. While rSSGP expectedly runs slower than SSGP with the same number of samples, it outperforms and runs faster than SSGP with 1024 samples, which implies that rSSGP is more scalable than SSGP with the same performance (i.e., $p > 1024$). Fig. 1(c) shows the distribution of embedded data fitted to a Gaussian mixture model with 8 components (the 2D projection onto the span of the first two eigenvectors). The latent inputs form clusters with varying density (we plot this for convenience because each cluster covariance is an 18×18 matrix). The mixture weights are $[0.046, 0.063, 0.070, 0.130, 0.153, 0.155, 0.167, 0.215]$ which roughly corresponds to the 2^{-i} values as required by our practical conditions. As a consequence, we observe 85% of latent input pairs to be cross-cluster pairs, which expectedly correspond to small kernel entries in our analysis.



(a) NLL of prediction (b) Training time (c) Latent input distribution

Figure 1: Comparative performance of Revisited SSGP ($p = 32$) and SSGP ($p = 32$ and 1024) on GAS (10k) dataset.

R1: VAE training: Both VAE's reconstruction loss and SSGP's NLL are combined additively into one loss function, which enables end-to-end training of both VAE's and SSGP's parameters. The dimensions of x and z are not necessarily the same (e.g., $|x| = 8$ and $|z| = 4$). Compressing x into z (with a Gaussian mixture prior on the VAE) allows us to reconfigure the data onto a latent space conditioned to exhibit the disentanglement that enables our theoretical analysis (please refer to Appendix D1 and our visualization above). **Analysis for other kernels:** Extending our current analysis to other kernels is feasible since the spectral theorem applies generally to many shift-invariant GP kernels [1, 2]. In fact, we can modify Lemmas 3 and 4 in Appendix A (Equations 13-17) to make our analysis applicable to a broader range of exponential kernels. We will include this. **Sample complexity result:** We will state it in Theorem 2 as you suggested.

R2: Suggested literature: We have discussed (Rahimi and Recht) in lines 96-104 and in footnote 3. We will state this explicitly in the introduction and include the other works in our discussion. While these works generate bounds on SSGP that only hold for a fixed parameter configuration, ours hold universally on the entire parameter space. **Restrictiveness of practical conditions:** Our analysis will hold (approximately) if we can reconfigure and embed the data onto a latent space that exhibits such conditions. As such, in 3.3.2, we adopt a VAE with a Gaussian Mixture prior to embed the raw data (please see Appendix D1 and the extra plots above) such that our conditions are likely to hold.

R3: Practicality of conditions: Please see the above response for R2. **Training complexity**: The VAE's complexity per update iteration is $\mathcal{O}(b \cdot \mathrm{poly}(p))$ where p is the number of VAE parameters and b is the batch size. For large datasets, $n \gg p$ so the overhead is very mild with respect to n. To the best of our knowledge, both VAE and GP have no guarantee for convergence in terms of total no. iterations. We show empirically (see above plot) that the resulting scheme will not be slower than normal SSGP with sufficiently many spectral samples to produce similar performance and surely faster than full GP with a per-iteration update cost of $\mathcal{O}(n^3)$. In addition, for full GP prediction, the $\mathcal{O}(n^2)$ memory cost is a bottleneck, whereas our scheme allows batch training and does not incur this expensive memory cost.

R4: Thank you for recognizing our theoretical contribution. **Rigor of VAE:** The proposed VAE is a practical measure to achieve the conditions that enable our analysis and is more than a baseline attempt to cluster data. We will put a remark to clarify as suggested. **Analysis for other kernels:** Please refer to our response for R1. **Computation of the additional term:** The model is optimized via updating its parameters along the direction of the gradient. While the exact gradient is not tractable, its unbiased stochastic estimate can be computed using the reparameterization trick as described in [22] (a standard practice in many VAE works). **Large-scale experiments:** In Appendix D2, we show results on a large dataset with 0.5 million data points, on which full GP is already infeasible. We will include extra results to showcase our performance. **Unused decoder:** While the reconstructed data is not used for prediction, it is useful as an auxiliary training objective to obtain our practical conditions. It would be ideal if there is an alternative approach to ensure this happens in a more direct manner and reduce the difficulty of the learning problem, which would in turn allow us to tighten the sample complexity further. This is an interesting direction to take for future research.

[1]: Fourier Feature Approximations for Periodic Kernels in Time-Series Modelling (Tompkins and Ramos, 2018).
[2]: Generalized Spectral Kernels (Samo and Roberts, 2015).