We thank all reviewers for their careful reading and useful comments. Due to space limit, we focus our response on the main comments. In particular, we notice Assumption 3.1 and Assumption 3.3 drew most attentions. *We'd first like to state that despite these assumptions, to the best of our knowledge, this paper presents the first successful try at proving SGD converges (and recovers the true noise) in correlated settings. Open problems still exist, some assumptions can be further relaxed, yet we believe this paper presents the first crack at this challenging topic.*

1. **Explanation on Assumption 3.1:** We would like to point out that the exponential eigendecay assumption (Assumption 3.1) is satisfied by a wide range of kernels including the RBF kernel (the Gaussian kernel), see Section 4.3.1 in Rasmussen, C. E. (2003), "Gaussian processes in machine learning", which is commonly seen in the GP literature. Other kernels with non-exponential eigendecay mostly decay at a polynomial rate, e.g., Matern kernel, see section 2.3 in Bach, F. (2017), "On the equivalence between kernel quadrature rules and random feature expansions". We do have additional theoretical guarantees for kernels with polynomial eigendecay where the optimization error is still $O(\frac{1}{K})$ and the statistical error is $O(m^{-\frac{1}{2}+\varepsilon})$ where $0 < \varepsilon < \frac{1}{2}$ depends on the particular decay rate. We will include these results in the camera-ready version if accepted.
2. **Explanation on Assumption 3.3:** When $M = 1$, the considered model is the standard GP formulation and satisfies Assumption 3.3 directly. Meanwhile, in response to reviewer 3's question, both signal variance and noise variance need to be estimated when $M = 1$, and there is no closed form solution for them together. When extending $M = 1$ to $M > 1$, the already challenging proof becomes extremely challenging without Assumption 3.3. We also have simulation results which were not included due to space constraints, suggesting that SGD works well for $M > 1$. Providing theoretical guarantees for $M > 1$ without Assumption 3.3 remains an open and challenging question.
3. **Estimation for lengthscale:** To clarify, we only claim the convergence guarantee for estimating signal and noise variances. The extension to convergence guarantees for the lengthscale parameter in RBF kernel is extremely challenging: both the proof for Lemma 4.1 and 4.2 presents additional challenges if looking at the lengthscale since it lies inside the exponential as a denominator. However, the case studies suggest that SGD may still be used for estimating the lengthscale in practice. Meanwhile, we also have additional simulation experiments suggesting that SGD also recovers the true lengthscale up to some statistical error, which will be added if space permits.
4. **Theoretical contribution:** We also want to emphasize the technical challenges even when we consider estimating only the signal and noise variances.

   First we explain the role of "correlation": here we mean the correlation of $\{\mathbf{y}_{\xi_k}\}_{k=1}^K$ conditioning on $\mathbf{X}$, which results in the correlation of $\{g(\boldsymbol{\theta}^{(k-1)})\}_{k=1}^K$. The statistical error is the deviation of $g(\boldsymbol{\theta}^{(k-1)})$ from $\mathbb{E}(g(\boldsymbol{\theta}^{(k-1)})|\mathbf{X}_{\xi_k})$, averaged over each iteration $1 \le k \le K$, which is less concentrated (larger variance) if the correlation among $g(\boldsymbol{\theta}^{(k-1)})$ is strong. Therefore the statistical error only converges to 0 when the minibatch size m tends to ∞, thus m is assumed to be large instead of being a constant.

   In order to have a lower bound on the approximate curvature (see Lemma 4.2) that is independent of m, we establish novel upper and lower bounds on $\sum_{j=1}^m \lambda_{lj}\lambda_{l'j}(\sum_{i=1}^M \theta_i^{(k)}\lambda_{ij} + \theta_{M+1}^{(k)})^{-2}$ with high probability when m is large, where λ_{lj} is the jth eigenvalue of $\mathbf{K}_{f,n}^{(l)}$.

   The proof for Lemma 4.1 is also non-trivial since the error bounds holds uniformly for all $\boldsymbol{\theta} \in [\theta_{\min}, \theta_{\max}]^{M+1}$, and the gradient involves the trace of an $m \times m$ high-dimensional matrix whose entries all depend on $\boldsymbol{\theta}$ in a non-linear way. We apply Taylor's expansion and a novel truncation technique to avoid the difficulty caused by the high non-linearity.
5. **Model selection v.s. model inference/prediction:** The model selection (estimation of hyperparameter) is indeed our core task in the paper, which is also an important problem in GPs. The case studies presented in this paper demonstrate that SGD helps us find better hyperparameters and thus has better prediction performance. The proposed SGD technique cannot directly carry over to the prediction task, but existing methods on prediction can all be applied after the model selection. One should note that for prediction we only need to invert the kernel matrix once, but for model selection, each update of hyperparameters requires one inversion, which is more time-consuming. The speed-up of training process provided by SGD still leads to huge overall computational savings.
6. **Comparison with EGP (fairness):** For data size of 10^6, EGP took a few days to fit with 8 GPUs while our model took 30 minutes to train on 1 CPU. We note that kernel matrix partitioning done via GPU in EGP is for acceleration and not enhanced performance over exact inference. Numerical instabilities are addressed via the kernel matrix preconditioner–a task that we also do when comparing with EGP. In addition, although SGD in correlated setting is a sequential operation, for minibatch size over 10^4, GPU acceleration can be achieved in similar fashion as EGP via kernel matrix partitioning.