We would like to thank you and appreciate your clear review, important points and invested time. Before going through your comments, we want to inform you that we have added the results of another high dimensional experiment, visual odometry task. Please note that these results were almost ready at the time of the submission, but due to the page limit we decided to cover two high dimensional experiments, single-double pendulum, and two low dimensional experiments, Lorenz and NCLT. But, because of the concerns of some reviewers about higher dimensional experiment results, we have merged single-double pendulum experiment and added visual odometry results in the rebuttal version.

We try to clarify point by point based on your comments and feedback.

Paper completely ignores methodology of system identification, e.g., no mention of EM algorithm, considering they are talking about Kalman smoothing filtering etc. (Which forms the E part of EM methods)

R1.

Solving a system identification problem from EM point of view is now added to the paper section 3. We elaborate more on the importance of EM algorithm in the system identification tasks and consider our GIN and proposed variational-based system identifier as convoluted types of EM algorithm. Please refer to lines 75-80 and 88-102.

The network doesn't enforce regularity or restrictions on the output, as it outputs a covariance matrices or Kalman gains etc which result in covariance matrices

R2.

In the paper, for calculating the KG, we conduct the following procedure in the GRU^{KG} :

1- Calculate $\mathbf{f}(\mathbf{\Sigma}_t^-)$, where $\mathbf{f}$ is convolution for high dimensional data and identity for low dimensional data.

2- Pass $\mathbf{f}(\mathbf{\Sigma}_t^-)$ and $\mathbf{R}_t$ to the GRU cell and propose a DNN on the output. The proposed DNN outputs a lower triangular matrix L_t with positive diagonal elements, e.g. Elu+1 activation function, and then model $[\mathbf{H}_t\mathbf{\Sigma}_t^-\mathbf{H}_t^T + \mathbf{R}_t]^{-1} = L_t L_t^T$

3- Construct the KG as: $\mathbf{\Sigma}_t^-\mathbf{H}_t^T L_t L_t^T$

By this approach, Cholesky factor consideration, we ensure the positive definiteness of $[\mathbf{H}_t\mathbf{\Sigma}_t^-\mathbf{H}_t^T + \mathbf{R}_t]^{-1}$ and provide a KG which does not affect the positive definiteness of resulted covariance matrices. (The procedure for the smoothing gain is similar)

Gaussian state approximation makes this method same as UKF/EKF etc. Gaussian mapped to a Gaussian is a linear transform you can create a complex mechanism to estimate linear coefficients, but the map is linear

R3.

We agree that the map is linear in the latent space, but it is worth noting two points here:

1- Non-linear functions, $f(.)$ and $h(.)$ in the paper, stands for non-linearity handling, meaning that the more complex $f(.)$ and $h(.)$ could handle more non-linearity and if the length between time-steps of observations is sufficiently small, then the constructed latent model is linear or can be closely approximated by a good linear approximator, which is the case in our experiments.

2- By considering a dynamic system of the form: $\mathbf{x}_t = f(\mathbf{x}_{t-1}) + q_t$ and $\mathbf{w}_t = h(\mathbf{x}_t) + r_t$ and applying linearization, the system can be written as $\mathbf{x}_t = \mathbf{F}_t(\mathbf{x}_{t-1})\mathbf{x}_{t-1} + q_t$ and $\mathbf{w}_t = \mathbf{H}_t(\mathbf{x}_{t-1})\mathbf{x}_t + r_t$. By conducting filtering parameterization as $p(\mathbf{x}_t|\mathbf{w}_{1:t})$, then the posteriors are obtained as $(\mathbf{x}_t^- + \mathbf{K}(\mathbf{H_t})_t[\mathbf{w}_t - \mathbf{H}_t\mathbf{x}_t^-], \mathbf{\Sigma}_t^- - \mathbf{K}(\mathbf{H_t})_t[\mathbf{H}_t\mathbf{\Sigma}_t^-\mathbf{H}_t^T + \mathbf{R}_t]\mathbf{K}(\mathbf{H_t})_t^T)$. Despite regular LGSSM, where $\mathbf{F}_t(\mathbf{x}_{t-1}) = \mathbf{F}_t$, $\mathbf{H}_t(\mathbf{x}_{t-1}) = \mathbf{H}_t$ and $\mathbf{K}(\mathbf{H_t})_t = \mathbf{K}_t$, we propose 1) K basic state transition and emission matrices, $\mathbf{F}^k$ and $\mathbf{H}^k$ and interpolate them based on the information from the latent state $\mathbf{x}$ which is modeled by a DNN and constructs more accurate $(\mathbf{F}_t, \mathbf{H}_t)$. 2) We model $\mathbf{K}(\mathbf{H_t})_t$ with further non-linearity which may give us a better estimation of $\mathbf{K}(\mathbf{H_t})_t$ (although the map is linear).

Based on these two points and the paper results, the GIN is outperforming LGSSM and SIN in the all settings, where the difference between the GIN and LGSSM is that the GIN performs linearization more accurately than the LGSSM possibly because of the further non-linearity. The GIN also outperforms the SIN that does not perform linearization over the system $\mathbf{x}_t = f(\mathbf{x}_{t-1}) + q_t$ and $\mathbf{w}_t = h(\mathbf{x}_t) + r_t$, but approximate f and h directly with DNN.

Kalman smoother was proposed by Rauch, Tung and Striebel, fair to call it Kalman as it uses joint Gaussian conditioning to get the gain formulas, but RTS needs to be cited at least once in the paper.

R4.

Thanks for the important related paper and your recommendation, we have cited this paper in the related works section.

I am still curious how the model avoids mode collapse, a phenomenon where jointly estimating hidden state and the transition model often results in very conservative estimates for both.

R5.

To prevent the model being stuck into mode collapse, we provided three solutions:

1- By introducing k sets of $\mathbf{F}^k, \mathbf{H}^k$, where each set of $\mathbf{F}^k, \mathbf{H}^k$ models different dynamics, we introduce a loss term with a small constant factor which tries to increase the distance of each pair of $\mathbf{F}^k, \mathbf{H}^k$ set. Intuitively, the presence of different dynamics can easily modify the states in each update. We found this method as a potential solution to prevent the model go through the mode collapse.

2- Considering the negative distance of consecutive pairs of states as additional loss term with a small constant factor (the distance can be considered as euclidean difference of mean or KL of two consecutive states). Intuitively, this solution is forcing the states to not have overlap with each other and impose them to change in each update step.

3- In the first few epochs, we only learn auto-encoder(MLPs) and $\mathbf{F}^{(k)}$ and $\mathbf{H}^{(k)}$, but not *Dynamics Network* parameters $\alpha_t(\mathbf{x}_{t-1})$. All the parameters are jointly learned, afterwards. This allows the system to learn good embedding and various meaningful dynamics at first, then learns how to employ K different dynamics variables.

In the simulation results, we have used the third option.

How do you ensure that Kalman Gain and Smoothing gain you compute results in positive semidefinite covariance for Gaussian state you estimate.

R6.

This concern is addressed in R2.

why not use classical EM or with complexities proposed variational EM as a baseline

R7.

In our high dimensional experiments, single-double pendulum and visual odometry implementations, we compare the results of the GIN with EKVAE and KVAE. Where the proposed algorithms in both of them may address your concern (Both are EM-based variational inference). We briefly summarize the proposed algorithm of the KVAE for instance:

In KVAE, $p(\mathbf{s})$ is parameterized as: $\log \int p(\mathbf{s}, \mathbf{w}, \mathbf{x}) d\mathbf{w} d\mathbf{x} \geq \mathbb{E}_{q_\phi(\mathbf{x},\mathbf{w}|\mathbf{s})}\big[\log \frac{p_\theta(\mathbf{s}|\mathbf{w}) p_\gamma(\mathbf{w}|\mathbf{x}) p_\gamma(\mathbf{x})}{q_\phi(\mathbf{w},\mathbf{x}|\mathbf{s})}\big] = \mathbb{E}_{q_\phi(\mathbf{w}|\mathbf{s})}\Big[\log \frac{p_\theta(\mathbf{s}|\mathbf{w})}{q_\phi(\mathbf{w}|\mathbf{s})} + \mathbb{E}_{p_\gamma(\mathbf{x}|\mathbf{w})}\big[\log \frac{p_\gamma(\mathbf{w}|\mathbf{x}) p_\gamma(\mathbf{x})}{p_\gamma(\mathbf{x}|\mathbf{w})}\big]\Big] = \mathcal{L}_{\theta,\gamma,\phi}(\mathbf{s})$, where they use $q_\phi(\mathbf{s}|\mathbf{x})$ as an approximation of the posterior $p_\theta(\mathbf{s}|\mathbf{x})$, and γ is dynamics parameters for filtering/smoothing. In the E step, with fixed θ, ϕ and γ parameters they sample $\mathbf{w}$ and $\mathbf{x}$ from $q_\phi(\mathbf{w}|\mathbf{s})$ and $p_\gamma(\mathbf{x}|\mathbf{w})$ and estimate $\hat{\mathcal{L}}_{\theta,\gamma,\phi}$. Then, in the M step, they optimize the parameters such that $\hat{\mathcal{L}}_{\theta,\gamma,\phi}$ is maximized.

The proposed structure in the EKVAE is more complex than KVAE and outperforms it. These two methods are EM variational-based baselines in our paper.

Why not include Log-likelihoods of the estimates, Mean squared error on its own is a not a good metric to judge estimators performance. The reason it is perfectly feasible to have a very bad model that gives you high MSE but estimates a bad posterior/State probability

R8.

In all of the experiments, both high-low dimensional, we are maximizing the log-likelihood and the log-likelihood is reported in the high dimensional results. However, in the low dimensional results, all of the baselines reported their performance in term of MSE so that we also provide MSE results for the sake of comparison. We can include the log-likelihood results of the low dimensional experiments, either in the main paper or in the supplementary, if you still find it necessary.