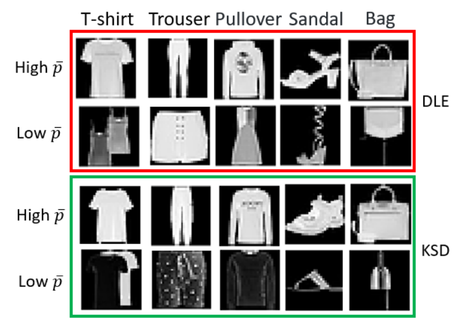
We thank reviewers for their insightful comments and here we would like to address some questions raised in the review.

R1: "Consistency results are given but these assume the parameter space is compact (and other not so simple to check assumptions)..." The compactness condition is only used to define a domain on which Assumption 2 and 4 hold. If Assumption 2 and 4 are defined over a neighbourhood region of the true parameter, we can remove the compactness condition by adding an extra proof which shows $(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\theta}})$ eventually fall in such a neighbourhood, but doing so would introduce further technical complications. The compactness condition is among a set of conditions commonly used in classic consistency proofs (see e.g., Wald's Consistency Proof, 5.2.1, van der Vaart, 1998). It is possible to derive weaker conditions given specific choices of $\boldsymbol{f}$ or $p(\boldsymbol{x}; \boldsymbol{\theta})$. However, in the current manuscript, we only focus on more *generic* settings and conditions that would give rise to estimation consistency and useful asymptotic theories.

R1: "... though this further assumes a (fairly strong) condition of uniform convergence ..." The uniform convergence on Hessian is needed to control the residual of the asymptotic expansion (eq. 25, 26) and is a slight modification of a classical regularity condition on the uniformly bounded third order derivative (5.3, van der Vaart, 1998). Again, this assumption may be weakened given specific choices of $\boldsymbol{f}$ and $p(\boldsymbol{x}; \boldsymbol{\theta})$ but we focus on investigating generic settings where specific choices of $\boldsymbol{f}$ and $p(\boldsymbol{x}; \boldsymbol{\theta})$ are not available.

"R1: it would be good to compare DLE to for example KSD on a complex model. R2: more empirical examples on non-toy datasets..."

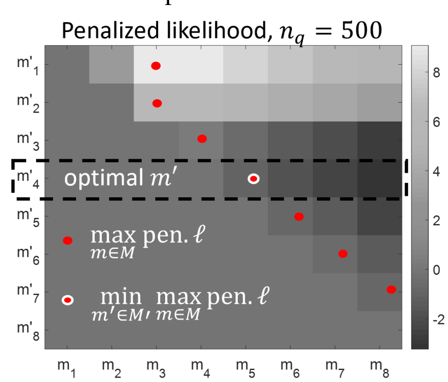


We run the same typical/outlier image detection task in Section 6.2 on Fashion MNIST dataset and compare DLE and KSD (see the figure). Both methods work well and seem to assign high likelihood to similar images. However, the tails of the fitted densities seem different, judging from low likelihood images. We observe similar results on MNIST dataset and will provide analysis into the differences of tail behaviors in the revision.

"R2: ... but it didn't compare to other methods like Contrastive Divergence or Noise Contrastive Estimation (NCE)."

Contrastive divergence was commonly used for restricted Boltzman Machine (RBM) where Gibbs sampling can be efficiently done. However, we consider a much wider family of models. In our MNIST experiment, the model is more complicated than an RBM and Gibbs sampling is hard. Other MCMC methods such as Metropolis-Hasting are unlikely to succeed as it is also difficult to design a proposal distribution in a 784-dimensional space. Due to the difficulties of applying MCMC in high dimensional tasks, we restrict our discussion on sampling-free methods for their computational efficiency and reliability in those applications. We tried NCE in the MNIST experiment but cannot find a good noise distribution which would give comparable results to DLE and KSD. Those results will be presented in revision.

R2:"... how a practitioner could select the Stein features to use?" R3: "...some guidelines or heuristics for how to select the feature.."



Section 4.3 provides an information-criterion based model selection method. Suppose M is a set of different choices of Stein features. Given a parameter θ, one should select the features $\hat{m}(\boldsymbol{\theta}) := \arg\max_{m \in M} \mathbb{E}_q[\ell(\boldsymbol{\theta}, \hat{\boldsymbol{\delta}}(m))]$ as this choice would minimize $\mathrm{KL}[q \| r_{\hat{\boldsymbol{\delta}}} p_{\boldsymbol{\theta}}]$ (see eq. 2).

If we have a set of candidate density models M', we can jointly select density model and Stein feature at the same time: $(\hat{m}, \hat{m}') := \arg\min_{m' \in M'} \max_{m \in M} \mathbb{E}_q[\ell(\hat{\boldsymbol{\theta}}(m'), \hat{\boldsymbol{\delta}}(m))]$, where $(\hat{\boldsymbol{\theta}}(m'), \hat{\boldsymbol{\delta}}(m))$ are estimated parameters under the model choice (m', m). Replacing $\mathbb{E}_q[\ell(\boldsymbol{\theta}, \hat{\boldsymbol{\delta}}(m))]$ with the penalized likelihood derived in Section 4.3, we can get a practical model selection method. We create a numerical experiment and plot the calculated penalized likelihood using scaled colors with respect to both M and M' (see the figure on the right). It shows that our information criterion can indeed select the optimal density model. We will explain this procedure in our revision.

R2: "Given these conflicting forces, how does one choose the Stein features?" Yes, there *can* be a trade-off between efficiency and overfitting. This happens in classic settings too: MLE is an efficient estimator, but suffers from overfitting when the dataset is small. Given a small number of samples, we may have to settle for a less efficient estimator to avoid overfitting. However, the aforementioned information criterion can be used to select Stein features in this setting.

R3: "...may imply that that the estimated density $p(x, \boldsymbol{\theta})$ is not positive everywhere..." The unnormalized density model $\bar{p}(\boldsymbol{x}; \boldsymbol{\theta})$ in our problem, by definition, should be non-negative everywhere for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, thus the estimated density $\bar{p}(\boldsymbol{x}; \hat{\boldsymbol{\theta}})$ is also non-negative. The estimated *density ratio* is guaranteed to be positive only within X_q, but the estimated density is guaranteed to be positive everywhere by definition. We will clarify this in our revision.