Response to Reviewer 1

We sincerely thank Reviewer 1 for referring us to four relevant papers [1-4]. We will include a detailed discussion on these papers in our revised manuscript, but would like to emphasize that our work differs from, and is complementary to this literature. This is further explained below:

Paper [1] provides a very interesting relationship between Fisher divergence and Stein's operator, whereby Hyvarinen score and its extensions are cast as specific Stein's operators. Paper [4] establishes a more general result than that of S. Lyu.

In contrast, our work focuses on utilizing Hyvarinen score (and its extensions) to develop information quantities for both interpretability and computational benefits (e.g. in our example of fast tree approximation). That said the computational motivation still remains to be the getting rid of normalizing constants.

Regarding your first technical question: It follows from an application of weak law of large number that by minimizing the Hyvranin score the parametric density converges to the shadow of the true data generating distribution under mild assumptions. If the model class is well-specified then convergence to the data generating distribution is guaranteed.

Regarding your second technical question raised: The Fisher entropy is intimately related to the Cramer-Rao Bound in parameter estimation. Fano's inequality also gives lower bounds of model selection/message decoding error (so larger bound implies more complexity for 'description'). We understand that Reviewer 1 may have a different interpretation of our wording, and we will clarify this in the revision.

[1] Liu, Qiang, Jason Lee, and Michael Jordan. 'A kernelized Stein discrepancy for goodness-of-fit tests.' ICML 2016.

[2] Chwialkowski, Kacper, Heiko Strathmann, and Arthur Gretton. 'A kernel test of goodness of fit.' ICML 2016.

[3] Gorham, Jackson, and Lester Mackey. 'Measuring sample quality with kernels.' ICML 2017.

[4] Liu, Qiang, and Dilin Wang. 'Stein variational gradient descent: A general purpose bayesian inference algorithm.' NIPS 2016.

Response to Reviewer 2

We appreciate Reviewer 2's comments and recommendations.

We will do another proof reading and remove typos. Due to page limit we could not include more extensive experiments, but we plan to do so in subsequent works.

Reviewer 2 raised an excellent point about the extension from pairs of variables to groups of variables. We have some recent results on this (not reported due to page limit), and hope to expand on them and present them in future work.

Regarding the two questions raised by the Reviewer 2, the values of p ranges from 2 to 11 and the joint density is determined by the conditional exponential family distribution in (12) and the discovered tree structure; the second term in (2) is q(y), and p(y) has been implicitly included in the expectation.

Response to Reviewer 3

We appreciate Reviewer 3's comments and recommendations.

Reviewer 3 raised an excellent point about defining cross entropy. We think it is doable and it would be another analogy to the classical cross entropy. Despite theoretical interest, we have not yet found a real data application for this. We will include a discussion of this in the revision.

Like the point made in (Marsh 2013), it is also true in our case that the relative entropy between distributions is more interpretable than entropy on its own. We have not found an application of gradient mutual information in the settings of information bottleneck. But we believe this direction is extremely interesting. As suggested by Reviewer 3, we will do another proof reading and remove the typos.