Reviewers 1 + 3:

> *‘Hard to see how TRE was relevant for energy-based modelling . . . [the] description of NCE was insufficient’*

Thank you for pointing this out; we will remedy this by strengthening the NCE connection from earlier on in the paper.

> *‘AIS connections could be better’*

We agree, and will add further discussion. However, we presently think the connection is more conceptual than technical.

> *‘MNIST is rather easy. Does it work on harder datasets?’*

> *‘Would higher dimension problems require considerably more intermediate distributions? If so, how bad is it?’*

While too preliminary to be included in the paper, we have some evidence from ongoing work showing that TRE can work on higher-dimensional image datasets using only modestly more intermediate distributions.

> *‘Is it easy to say if the limitations in [Stratos & McAllester] apply to your method and, if not, why not?’*

We believe the answer is no, since Stratos & McAllester prove strong limitations on “high-confidence lower bounds on mutual information” and our MI estimates are not lower bounds (nor upper bounds). Finally, we thank you for the relevant additional references.

Reviewer 2:

> *‘hard to assess if the claimed severity of the proposed "density-chasm" is valid. . . [the chasm] does not seem to hinder state-of-the-art results in many applications.’*

> *‘No comparisons to models other than "single ratio" estimations are given. It raises the question, why this problem was never addressed before, or if it was addressed before why there is no comparison.’*

As noted by reviewers 1 & 3, the “density-chasm problem” is a frequently occurring issue for practitioners using density-ratio estimation. Despite this fact, we are not aware of any prior work has clearly labelled the issue and provided a general-purpose solution. This gap in the literature motivated our paper in the first place.

In specific applications, it is often possible to design ratio-estimation tasks that avoid the density-chasm problem. As we show in our MNIST experiments, single-ratio estimation (i.e. NCE) can work very well if the noise distribution is sufficiently close to the data distribution, as is the case for the RQ-NSF model. A similar strategy (of learning a powerful noise distribution) has been used many times in the literature, and could be viewed as one of the core motivations behind GANs. However, learning *both* a ratio-estimator and complex noise distribution, in order to reduce the chasm, can be challenging/impractical, and hence it is useful for practitioners to have another method at their disposal which can work with simple noise distributions (as illustrated on MNIST in Figure 5).

The strategy of learning a good noise distribution only applies in the context of energy-based modelling. It does not apply to the more general problem setting of estimating a density-ratio p/q, where p and q are *fixed* in advance. In this general setting (which includes our MI & representation learning experiments), very few viable approaches exist. We think that the single density-ratio baselines we compare to—which include results from a 2019 Neurips paper—are representative of the state-of-the-art.

> *‘With little analytical guarantees and only few experiments (most of them involving variants of Gaussian noise) it is hard to assess how generalizable the results are.’*

The correctness of TRE is straightforwardly derived from that of single-density ratio estimation, for which there are extensive analytical guarantees (which we cite in Sec 2). Experiments in 4.1 & 4.2 are intentionally simple illustrations of the method using synthetic Gaussian data. Experiments in 4.3 & 4.4 use significantly more complex, non-Gaussian data. Whilst the absolute number of datasets used is modest, we believe the substantial performance increase of TRE over single ratio methods in a diverse set of applications (MI estimation, representation learning & energy-based modelling) is strong evidence of its generalizability.

> *‘The introduced [waymark] mechanisms all appear very ad hoc. How did you decide between the one or another?’*

As discussed in the conclusion, we agree further research on the waymark mechanisms would be valuable despite the good empirical performance of those presented, which were motivated by simplicity. As you noted, across all applications of TRE we used ‘dim-wise mixing’ for discrete data and the ‘linear combinations’ for continuous data.

> *‘What did you use as free parameter θ (line 88 and Fig.1)? ... [in Fig.1] The axes scaling/cut makes p(x) invisible... Can you elaborate on [line 133]? In Fig. 1(a) p(x) has sigma=1e-6 and q(x) has sigma=1 ... Could you provide a derivation of Eq. (5) from equal variances (e.g. in the appendix)?’*

θ is given in Sec 3, Eq 7 of the Appendix; we will clarify this. p(x) is missing in Fig.1 because it perfectly overlaps with the blue curve; we will amend the caption. Line 133 is slightly wrong: all of our experiments *except* 4.1 (i.e. Fig 1) preserve variance; we will amend it and add a one-line derivation of the variance property. Thank you for the comments.

Reviewer 4

> *‘The main weakness is that the authors never reveal which energy-based model they use in the experiments section.’*

We state in Sec 3.4 that the model has the form $r(\mathbf{x};\boldsymbol{\theta})q(\mathbf{x})$, where $r(\mathbf{x};\boldsymbol{\theta})$ is the product of all the bridges (Eq. 4) and $q(\mathbf{x})$ is a noise distribution. In Sec 4.4, we state how each bridge is parameterised, and the different choices of noise distribution. Sec 1 & 5 of the appendix give all the architectural/training details necessary to reproduce our results.

> *‘The relationship to AIS could be made a bit more prominent’*

We agree, and will include more discussion in the final paper (please see also our response to Revs 1+3).