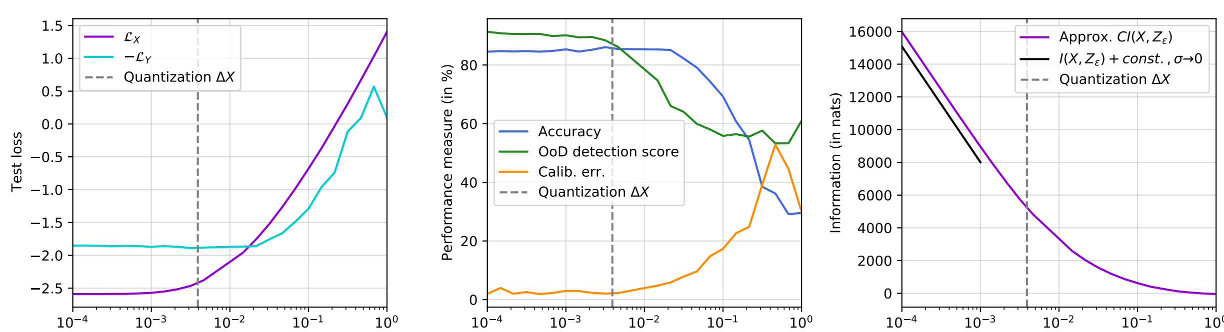
We sincerely thank all reviewers for their time and constructive feedback. We will add all minor clarifications and corrections to the final version (R2, R4, R5), as well as additional generated samples (R1). We are also thankful for the idea of R1 on how to extend our method using perturbations on Z, and will investigate this in the future. We address the main questions and criticisms in the following:

$CI(Y,Z)$ as a lower bound (R1). Thank you for this comment, we realized that we did not mention this in the paper, even though the answer is straight forward and enlightening: $CI(Y,Z)$ is in fact a lower bound of $I(Y,Z)$. This can be readily seen from Eq. 6: The first term is the entropy $h(Y)$, because the label distribution $p(Y)$ is known exactly. The second term can be rewritten as the negative cross-entropy $-h_q(Y \mid Z)$. For $I(Y,Z)$, we have the negative entropy $-h(Y \mid Z)$ as the second term instead. Because $h_q(Y \mid Z) \geq h(Y \mid Z)$ (Gibbs' inequality), we have $CI(Y,Z) \leq I(Y,Z)$. This is essentially the the same as the variational bound originally proposed by Barber & Agakov (2003): Their Eq. 3 corresponds to our Eq. 6, noting that their x is our Y, and their y is our Z. This is a bound that only works in this specific case, as the label distribution $p(Y)$ must be known for it to apply. We will add this to the final version.

Mathematical assumptions about the network g_θ (R4). We agree that the assumptions should be added to the text and propositions more explicitly, and we will rectify this for the final version. However, we do not see the assumptions as a 'fundamental technical difficulty': For all INN architectures used in practice, they are fulfilled by construction. This includes GLOW, RealNVP, NICE, i-ResNet, and more. In none of these cases, there is any need for any additional constraints, i.e. the assumptions are fulfilled per default. We refer to works such as Virmaux & Scaman (2018); Behrmann et al. (2020) for further details.

Strengthening Prop. 1 and properties of CI (R4). We also think that the CI is of great interest in general and should be further investigated in future. In our case, it is only used in a very specific way, so we did not consider strengthening or extending Prop. 1. Instead, we would like to refer to Xu et al. (2020), who derive various further theoretical results and insights concerning CI in general.

Effect of hyperparameter σ (R5). In line with this suggestion, we will add some more experiments to the appendix concerning the effect of σ. As a first step, the following figure shows the behaviour for 25 different models trained with σ between 10^{-4} and 10^{0} (x-axis), and fixed $\gamma = 0.2$. We find that the loss values (left) and performance characteristics (middle) do not depend on σ below a threshold that is comparable to the qantization step size ΔX. The models performance does not decrease even when σ is 50 times smaller than ΔX. Detrimental effects might occur more easily if the quantization steps are larger, e.g. $\Delta X = 1/32$ as used by Kingma & Dhariwal (2018). The rightmost plot compares our approximation of $CI(X, Z_\varepsilon)$ with the asymptotic $I(X, Z_\varepsilon) + const.$ for $\sigma \to 0$, where the constant is unknown. The slope of the approximation agrees well for small σ, but breaks down for larger values. This, and further experiments concerning the role of σ will be added to the final version.



References.

Barber, D. and Agakov, F. V. The im algorithm: a variational approach to information maximization. In *Advances in neural information processing systems*, pp. None, 2003.

Behrmann, J., Vicol, P., Wang, K.-C., Grosse, R., and Jacobsen, J.-H. Understanding and mitigating exploding inverses in invertible neural networks. *arXiv preprint arXiv:2006.09347*, 2020.

Kingma, D. P. and Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in neural information processing systems*, pp. 10215–10224, 2018.

Virmaux, A. and Scaman, K. Lipschitz regularity of deep neural networks: analysis and efficient estimation. In *Advances in Neural Information Processing Systems*, pp. 3835–3844, 2018.

Xu, Y., Zhao, S., Song, J., Stewart, R., and Ermon, S. A theory of usable information under computational constraints. *arXiv preprint arXiv:2002.10689*, 2020.