Lossy Compression For Lossless Prediction

Yann Dubois
Vector Institute
yanndubois96@gmail.com

Benjamin Bloem-Reddy
The University of British Columbia
benbr@stat.ubc.ca

Karen Ullrich
Facebook AI Research
karenu@fb.com

Chris J. Maddison
Vector Institute and University of Toronto
cmaddis@cs.toronto.edu

Abstract

Most data is automatically collected and only ever "seen" by algorithms. Yet, data compressors preserve perceptual fidelity rather than just the information needed by algorithms performing downstream tasks. In this paper, we characterize the minimum bit-rate required to ensure high performance on all predictive tasks that are invariant under a set of transformations, such as data augmentations. Based on our theory, we design unsupervised objectives for training neural compressors. Using these objectives, we train a generic compressor that achieves substantial rate savings (more than $1000\times$ on Imagenet) compared to JPEG on 8 datasets, without decreasing downstream classification performance.

1 Introduction

Zetabytes (10^{21}) of data are collected every year (Reinsel et al., 2017), which is too much for humans to process. As a result, most of this data will only be processed by algorithms performing a task. So, there is a growing need for compression methods that retain only the information necessary to ensure high performance on downstream tasks.

Existing compression methods assume either that all information is important or that the goal is perceptual fidelity. However, much of that information is not useful for downstream tasks. For example, image classification is often invariant under small rescalings or rotations, but a standard compressor will faithfully reconstruct this information. If we care only about predictive performance, we should be able to improve compression by discarding such information, as seen in Fig. 1. Our goal is to quantify the bit-rate gains that come from removing this unnecessary information and to learn compressors that discards it.

The minimum bit-rate required for high performance on a supervised task corresponds to compressing the labels. Achieving this rate requires access to the labels, and caring only about a single task. Instead, we want a compressed representation that ensures good performance on *any* future tasks of interest. Importantly, this set will rarely be known at compression time or might be too large to even enumerate. We overcome these challenges by focusing on tasks that are *invariant* under user-defined transformations.

In this paper, we characterize the minimum bit-rate needed to ensure predictability of all invariant tasks. The key is that we construct a worst-case task, which bounds your performance on any invariant tasks. As a result, the bit-

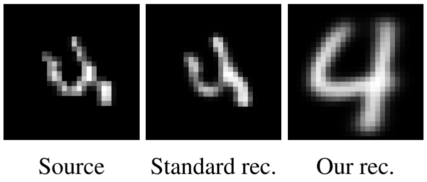


Figure 1: Our unsupervised coder achieves better compression by keeping only the information that is necessary for desired tasks. (left) source augmented MNIST digit; (center) a neural compressor optimized for perceptual similarity achieves a 130 bit-rate; (right) our invariant compressor achieves a 48 bit-rate.

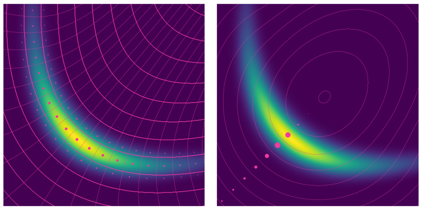


Figure 2: Compression rates of a Banana source can be decreased when downstream tasks are rotation invariant. (left) a neural transform coder achieves a 5.4 bit-rate; (right) our learned invariant transform coder achieves a 2.5 bit-rate. Pink lines are quantization boundaries, dots are code vectors.

rate required to perform well on *all* invariant tasks is exactly the rate required to compress the labels of the worst-case task. Intuitively, this task is to recognize which examples are transformed version of one another, and the rate savings come from discarding that information, as seen in Fig. 2.

In addition to establishing optimal rates, we provide unsupervised neural compression methods that approximate these rates. In particular, we design two training objectives. One is a modified variational autoencoder (VAE) (Kingma & Welling, 2014) that reconstructs prototypical examples, as shown in Fig. 1. The second is a simple modification of InfoNCE (Oord et al., 2019), used in self-supervised learning (SSL), which allows us to convert pre-trained SSL models into powerful compressors.

2 Minimum Bit-Rates for High Predictive Performance

The goal of lossy compression is to find the number of bits (the *bit-rate*) required to store a r.v X so that it can be reconstructed to within a certain tolerance. Specifically, given a distribution $p(Z|X)$ mapping X to a r.v. Z (the *representation*) and a distortion measure $\mathrm{D}[X, Z]$, Shannon's (1959) rate-distortion (RD) theory, characterizes the minimal achievable rate for a distortion threshold δ by

$$\mathit{Rate}(\delta) = \min_{p(Z|X) \text{ s.t. } \mathrm{D}[X,Z] \leq \delta} \mathrm{I}[X; Z] \ . \tag{1}$$

Our aim is different: to find the minimal bit-rate required to store X so that we can still achieve high performance on a set of downstream tasks. To do this, we define a distortion that ensures that predicting from the compressed Z is approximately as good as from X. All proofs are in Appx. C.

Let us represent a downstream tasks by a set of r.v.s $\mathcal{T} = \{Y_1, Y_2, \ldots\}$, which are jointly distributed with X and represent all variables that we may be interested in predicting. For example, $Y_1 \in \mathcal{T}$ might indicate whether X is an image of a dog, while $Y_2 \in \mathcal{T}$ might indicate whether X is hand-drawn. Let $\mathrm{R}[Y \mid X] = \inf_q \mathrm{E}_{p(X,Y)}[-\log q(Y|X)]$ be the best log-loss risk when predicting Y from X.

We would like to find a compressed representation Z such that the increase in risk due to predicting $Y \in \mathcal{T}$ from Z (as opposed to X) is bounded, i.e., $\mathrm{R}[Y \mid Z] - \mathrm{R}[Y \mid X] \leq \delta$, for all tasks. This suggests using the following as a distortion measure for our rate-distortion theory,

$$\mathrm{D}_{\mathcal{T}}[X, Z] := \sup_{Y \in \mathcal{T}} \quad \mathrm{R}[Y \mid Z] - \mathrm{R}[Y \mid X] \,, \tag{2}$$

Unfortunately, working with Eq. (2) assumes access to all downstream tasks of interest $\mathcal{T}$ during compression and the ability to optimize over them, which is unrealistic in practice. However, sets of tasks that we care about are oriented to human goals, which suggests that there may be exploitable structure in realistic task sets. For example, image classification often relates to the concepts present in the image, rather than fine-grain details about brightness and object positions. These tasks are thus invariant to mild augmentation such as brightness changes and translations. In such case, the maximization in Eq. (2) is achieved by certain "hardest" invariant tasks. Intuitively, these hard tasks consist of retaining only the information which you should *not* be invariant to. This can essentially be done by predicting a prototypical version of the input X, denoted as $M(X)$.

Formally, we assume that the conditional distribution of each $Y \in \mathcal{T}$ is invariant to an equivalence relation $\sim$ on X's sample space .[1] That is, $x \sim x' \implies p(Y \mid x) = p(Y \mid x')$. Under weak regularity conditions, we prove that there exists a *maximal invariant*[2] task $M(X) \in \mathcal{T}$ that simplifies Eq. (2) to

$$\mathrm{D}_{\sim}[X, Z] = \mathrm{R}[M(X) \mid Z] \,. \tag{3}$$

Our "Rate-Invariance" theorem uses this simplification to characterize the minimal bit-rate required to ensure high predictive performance on invariant tasks $\mathcal{T}$. Let $\mathrm{H}[\cdot]$ denote the (discrete or differential) entropy. Intuitively, Eq. (3) suggests that the minimal bit-rate is related to compressing $M(X)$, which requires $\mathrm{H}[M(X)]$ bits. We formalize this by incorporating our distortion in RD theory.

Theorem 1 (Rate Invariance)**.** For $\delta \geq 0$, let $\mathit{Rate}(\delta)$ denote the minimum achievable bit-rate for transmitting Z such that for any invariant $Y \in \mathcal{T}$ we have $\mathrm{R}[Y \mid Z] - \mathrm{R}[Y \mid X] \leq \delta$. Then $\mathit{Rate}(\delta)$ is 0 if $\delta \geq \mathrm{H}[M(X)]$ and otherwise it is finite and given by

$$\mathit{Rate}(\delta) = \mathrm{H}[M(X)] - \delta = \mathrm{H}[X] - \mathrm{H}[X \mid M(X)] - \delta. \tag{4}$$

[1] As a reminder, $\sim$ is an equivalence relation iff for all $x, x', x'' \in \mathcal{X}$: (reflexivity) $x \sim x$, (symmetry) $x \sim x' \iff x' \sim x$, and (transitivity) $x \sim x'$ and $x' \sim x'' \implies x \sim x''$. Note that invariances w.r.t. $\sim$ essentially subsume all notion of invariance (e.g., groups, semigroups, functions).

[2] M is any function such that $x \sim x' \iff M(x) = M(x')$. See Appx. B for examples.

Theorem 1 relates compression and statistical learning theory by showing that allowing a δ decrease in log-loss performance can save *exactly* δ bits during compression. To illustrate Thm. 1, consider a scenario where we require no loss in predictive performance, i.e., $Rate(0) = \mathrm{H}[X] - \mathrm{H}[X \mid M(X)]$, and contrast this rate to the standard lossless compression rate, $\mathrm{H}[X]$. That is, we can remove any information in X that is not in $M(X)$ while ensuring lossless prediction of all invariant tasks.

Consider compressing a sequence of n coin flips. If one is only interested in predicting labels that are permutation invariant, then instead of compressing the entire sequence, one could simply compress the number of heads. The number of heads is a maximal invariant for permutation invariance. In this case, the sequence can be compressed to a bit-rate that grows as $\mathcal{O}(\log n)$, as opposed to $\mathcal{O}(n)$ for the lossless compression case. In we show how our result recovers other previous results (i) lossless compression; (ii) unlabeled graph compression (Rashevsky, 1955); (iii) multiset compression (Varshney & Goyal, 2007); (iv) the information bottleneck (Tishby et al., 2000).

3 Learning Invariant Compression using Data Augmentations

In this section, we design practical loss functions for training neural compressors that approximate optimal rates. We do so by optimizing an unconstrained (equivalent[3]) formulation of Eq. (1), i.e.,

$$Rate(\delta) = \min_{p(Z|X)} \mathrm{I}[X;Z] + \beta(\delta)\,\mathrm{R}[M(X) \mid Z]\,. \tag{5}$$

Both terms in the minimization of Eq. (5) are challenging to compute. Here we address how to approximate them. See Appx. D for derivations and the resulting training algorithms.

For $\mathrm{I}[X;Z]$, we rely on Ballé et al.'s (2017) Variational Compressors (VC), which uses a specific entropy model $q_\theta(Z)$ and the bound $\mathrm{I}[Z;X] \leq \min_\theta \mathrm{E}_{p(X)p_\varphi(Z|X)}[-\log q_\theta(Z)]$. The challenge with our term $\mathrm{R}[M(X) \mid Z]$ is that maximal invariants are typically inaccessible. To overcome this we make two assumptions, both of which are implicit in SSL. First, we assume that we can sample data augmentations A to which our tasks should be invariant, e.g., image shearing. This enables sampling of equivalent points x, i.e., $x \sim X$. Second, we assume that no two observations $X, X' \in \mathcal{D}$ in our dataset are equivalent $X \not\sim X'$, i.e., they are not augmented versions of one another. This assumption lets us use each sample in $\mathcal{D}$ as the maximal invariant of its own equivalence class, i.e., $M(x) = X$ for all $x \sim X$. The following two variational upper bounds on our invariance distortion encourage networks to solve the task (explicitly or implicitly) of mapping $A(X)$ back to X.

Variational Invariant Compressor (VIC). Our first loss is closely related to the VC and the VAE. The model has an encoder $p_\varphi(Z|X)$, an entropy model $q_\theta(Z)$, and a decoder $q_\phi(X|Z)$. Given a data point X, we apply an augmentation $A(X)$, pass it through the encoder to get a representation Z. The decoder then attempts to reconstruct the unaugmented X from $A(X)$. This leads to the objective,

$$\mathcal{L}_{\text{VIC}}(\phi, \theta, \varphi) := -\sum_{X \in \mathcal{D}} \mathrm{E}_{p(A)p_\varphi(Z|A(X))}[\log q_\theta(Z) + \beta \cdot \log q_\phi(X \mid Z)]\,. \tag{6}$$

Bottleneck InfoNCE (BINCE). Our second loss is based on InfoNCE (Oord et al., 2019). For every X, we sample a sequence of random points $\boldsymbol{X} = (X_+, X_1, \ldots, X_n)$, where $X_+ = A(X)$ is an augmentation of X ("positive") and each X_i are non equivalent examples $X_i \not\sim X$ ("negatives"). Let $\boldsymbol{Z} = (Z_+, Z_1, \ldots, Z_n)$, be the corresponding representations given by $p_\varphi(Z \mid X)$. Let f_ψ be a discriminator that is optimized to score the equivalence of two representation. The final loss is:

$$\mathcal{L}_{\text{BINCE}}(\phi, \theta, \psi) := -\sum_{X \in \mathcal{D}} \mathrm{E}_{p(A)p_\varphi(Z, \boldsymbol{Z}|X,A)}\left[\log q_\theta(Z) + \beta \log \frac{\exp f_\psi(Z_+, Z)}{\sum_{Z' \in \mathbf{Z}} \exp f_\psi(Z', Z)}\right]. \tag{7}$$

Eq. (7) is the standard SSL loss with an additional entropy bottleneck.[4] Instead of directly predicting X as in VIC, BINCE retains information about X by classifying (as seen by the softmax) which Z is associated with an equivalent example $x \sim X$. This has the advantage of not requiring a high dimensional decoder but can require many negative examples n. Both VIC and BINCE give rise to efficient compressors by passing X through $p_\varphi(Z|X)$ and entropy coding using $q_\theta(Z)$. In theory they can recover the $\delta = 0$ optimal bit-rate, i.e., $\mathrm{H}[M(X)]$, in the limit of infinite samples ($|\mathcal{D}|$,n) and unconstrained variational families.

[3] Any $\beta \geq 0$ corresponds to a $\delta \geq 0$, so we can find different solutions on the RD curve by sweeping over β.

[4] The bottleneck arises from our desire to increase compression, but this SSL objective can be interested in its own right as such bottleneck provably improve generalization of downstream predictors (Dubois et al., 2020).

4 Experiments and Discussion

Let us evaluate our learned compressors. See Appx. F for experimental details and Appx. G for additional results.

Banana. We compress samples of the Banana distribution (Ballé et al., 2020) assuming downstream tasks are rotation invariant w.r.t. the origin. Our method (Fig. 1, right) learns disk-shaped quantization bins to retain only information about the norm of x, $M(x) = \|x\|_2$, and disregard angular information. As a result, the bitrate is 2.5 as opposed to 5.4 for a compressor with standard distortion (Fig. 1, left).

MNIST. Next, we evaluate the effect of standard geometrical augmentations on MNIST compression performance. We contrast our VIC with a standard VC. We assume that we know the downstream tasks to be invariant to the set of augmentations, which we enforce by augmenting the test set. To reach at least 99% downstream accuracy our compressor requires only 48 bits compared to 130 for the VC (reconstructions from both models can be seen in Fig. 2). Similar gains can be seen for any level of downstream accuracy as seen in Fig. 3.

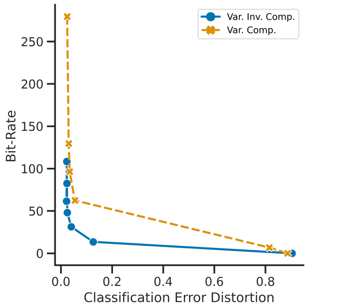


Figure 3: By considering invariances, our unsupervised compressors (in blue) improves MNIST compression (as in Fig. 2.) compared to standard compressors (in orange) for different distortion thresholds. The x-axis is the test error of a ResNet18 that classifies MNIST using reconstructions.

Table 1: Converting a pretrained SSL model into a zero-shot compressor achieving substantial bit-rate gains while allowing accuracy similar to supervised models predicting from raw images.

	Imagenet	STL10	PCam	Cars196	CIFAR10	Food101
Bit-rate gains vs JPEG	1030×	200×	63×	600×	18×	252×
Our MLP probe (Acc.)	75.3%	98.6%	82.4%	80.7%	95.2%	88.1%
Supervised (Acc.)	76.1%	99.0%	82.6%	49.1%	96.7%	81.8%

Zero-shot compressor using SSL. As previously pointed out, BINCE consists in a standard SSL loss with an additional entropy bottleneck. Given the availability and impressive results from pretrained SSL models, a natural question is whether we can take advantage of pre-trained SSL methods to give rise to powerful (invariant) compressors. To investigate this question, we add and train an entropy bottleneck on top of a SSL model. Specifically, we first download a SOTA SSL model, CLIP (Radford et al., 2021), and freeze it. Then we add an entropy bottleneck and train it on the CLIP's output on a small dataset, MSCOCO (Lin et al., 2015). [5] Finally, we evaluate our resulting compressor on various datasets (different tasks and shapes) which were never seen during training.

Table 1 shows that this simple method gives rise to a powerful and generic compressor, which achieves more than $1000\times$ bit-rate gains compared to JPEG. The bit-rate gains (1st row) are significant across all datasets, [6] even on PCAM (Veeling et al., 2018) which consists of biological tissues. Importantly, these gains do not come at the cost of removing information needed for desired tasks. The second row shows the accuracy of a simple multi-layer perceptron trained on the output of our compressed representation. The last row shows Radford et al.'s (2021) baselines from a near SOTA fully supervised model trained on the uncompressed images.

Discussion. Given the exponentially increasing amount of collected data and the prevalence of task-specific algorithms that analyze that data, it is urgent to rethink our current task-agnostic compression paradigm. To the best our knowledge we derive the first theoretical framework for task-centric compression (see Appx. E for related work). Furthermore, our theory and experimental results shows that the recent advancements and open-sourcing of pretrained self-supervised networks can be exploited to decreases current compression rate by orders of magnitude.

[5] This takes less than one hour on a single GPU.

[6] The large variance in bit-rate gains come from the variance in shapes of the raw images. For example, CIFAR10 has 32×32 bits so there is less information to gain compared to JPEG.

Acknowledgments

We would like to thank Emile Mathieu for helpful feedback. Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute. BBR acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC): RGPIN-2020-04995, RGPAS-2020-00095, DGECR-2020-00343.

References

Eirikur Agustsson and Lucas Theis. Universally Quantized Neural Compression. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/92049debbe566ca5782a3045cf300a3c-Abstract.html`.

Johannes Ballé, Valero Laparra, and Eero P. Simoncelli. End-to-end Optimized Image Compression. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL `https://openreview.net/forum?id=rJxdQ3jeg`.

Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL `https://openreview.net/forum?id=rkcQFMZRb`.

Johannes Ballé, Philip A. Chou, David Minnen, Saurabh Singh, Nick Johnston, Eirikur Agustsson, Sung Jin Hwang, and George Toderici. Nonlinear Transform Coding. *CoRR*, abs/2007.03034, 2020. URL `https://arxiv.org/abs/2007.03034`. _eprint: 2007.03034.

Toby Berger. Rate distortion theory for sources with abstract alphabets and memory. *Information and Control*, 13(3):254–273, September 1968. ISSN 00199958. doi: 10.1016/S0019-9958(68)91123-6. URL `https://linkinghub.elsevier.com/retrieve/pii/S0019995868911236`.

Benjamin Bloem-Reddy and Yee Whye Teh. Probabilistic Symmetries and Invariant Neural Networks. *J. Mach. Learn. Res.*, 21:90:1–90:61, 2020. URL `http://jmlr.org/papers/v21/19-322.html`.

Thomas M. Cover and Joy A. Thomas. *Elements of information theory (2. ed.)*. Wiley, 2006. ISBN 978-0-471-24195-9. URL `http://www.elementsofinformationtheory.com/`.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, June 2009. doi: 10.1109/CVPR.2009.5206848. ISSN: 1063-6919.

Yann Dubois, Douwe Kiela, David J. Schwab, and Ramakrishna Vedantam. Learning Optimal Representations with the Decodable Information Bottleneck. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/d8ea5f53c1b1eb087ac2e356253395d8-Abstract.html`.

Jarek Duda. Asymmetric numeral systems. *CoRR*, abs/0902.0271, 2009. URL `http://arxiv.org/abs/0902.0271`. _eprint: 0902.0271.

Morris L. Eaton. Group Invariance Applications in Statistics. *Regional Conference Series in Probability and Statistics*, 1:i–133, 1989. ISSN 1935-5912. URL `https://www.jstor.org/stable/4153172`. Publisher: Institute of Mathematical Statistics.

Ian S. Fischer. The Conditional Entropy Bottleneck. *Entropy*, 22(9):999, 2020. doi: 10.3390/e22090999. URL https://doi.org/10.3390/e22090999.

Gergely Flamich, Marton Havasi, and José Miguel Hernández-Lobato. Compressing Images by Encoding Their Latent Representations with Relative Entropy Coding. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/ba053350fe56ed93e64b3e769062b680-Abstract.html.

Robert G. Gallager. *Information Theory and Reliable Communication*. John Wiley & Sons, Inc., USA, 1968. ISBN 978-0-471-29048-3.

Tilmann Gneiting and Adrian E Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, March 2007. ISSN 0162-1459, 1537-274X. doi: 10.1198/016214506000001437. URL http://www.tandfonline.com/doi/abs/10.1198/016214506000001437.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. *arXiv:1406.2661 [cs, stat]*, June 2014. URL http://arxiv.org/abs/1406.2661. arXiv: 1406.2661.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 1026–1034. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.123. URL https://doi.org/10.1109/ICCV.2015.123.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, June 2016. IEEE. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.90. URL http://ieeexplore.ieee.org/document/7780459/.

Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Francis R. Bach and David M. Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 448–456. JMLR.org, 2015. URL http://proceedings.mlr.press/v37/ioffe15.html.

Nick Johnston, Damien Vincent, David Minnen, Michele Covell, Saurabh Singh, Troy Chinen, Sung Jin Hwang, Joel Shor, and George Toderici. Improved lossy image compression with priming and spatially adaptive bit rates for recurrent networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4385–4393, 2018.

Nick Johnston, Elad Eban, Ariel Gordon, and Johannes Ballé. Computationally efficient neural image compression. Technical report, Google Research, 2019. URL https://arxiv.org/pdf/1912.08771.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6114.

Alexander Kraskov, Harald Stoegbauer, and Peter Grassberger. Estimating Mutual Information. *Physical Review E*, 69(6):066138, June 2004. ISSN 1539-3755, 1550-2376. doi: 10.1103/PhysRevE.69.066138. URL http://arxiv.org/abs/cond-mat/0305641. arXiv: cond-mat/0305641.

Saunders Mac Lane and Garrett Birkhoff. *Algebra*. American Mathematical Soc., 1999. ISBN 978-0-8218-1646-2. Google-Books-ID: L6FENd8GHIUC.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. Publisher: Ieee.

Jooyoung Lee, Seunghyun Cho, and Seung-Kwon Beack. Context-adaptive entropy model for end-to-end optimized image compression. *ICLR*, 2019.

Erich L. Lehmann and Joseph P. Romano. *Testing statistical hypotheses*. Springer texts in statistics. Springer, New York, 3rd ed edition, 2005. ISBN 978-0-387-98864-1.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context. *arXiv:1405.0312 [cs]*, February 2015. URL `http://arxiv.org/abs/1405.0312`. arXiv: 1405.0312.

David J. C. MacKay. Bayesian Model Comparison and Backprop Nets. In J. E. Moody, S. J. Hanson, and R. P. Lippmann (eds.), *Advances in Neural Information Processing Systems 4*, pp. 839–846. Morgan-Kaufmann, 1992. URL `http://papers.nips.cc/paper/488-bayesian-model-comparison-and-backprop-nets.pdf`.

James L. Massey. On the entropy of integer-valued random variables. In *Int. Workshop on Inf. Theory*, 1988.

David McAllester and Karl Stratos. Formal Limitations on the Measurement of Mutual Information. In Silvia Chiappa and Roberto Calandra (eds.), *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pp. 875–884. PMLR, 2020. URL `http://proceedings.mlr.press/v108/mcallester20a.html`.

Fabian Mentzer, George Toderici, Michael Tschannen, and Eirikur Agustsson. High-Fidelity Generative Image Compression. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/8a50bae297807da9e97722a0b3fd8f27-Abstract.html`.

David Minnen, Johannes Ballé, and George D Toderici. Joint autoregressive and hierarchical priors for learned image compression. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL `https://proceedings.neurips.cc/paper/2018/file/53edebc543333dfbf7c5933af792c9c4-Paper.pdf`.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. *arXiv:1807.03748 [cs, stat]*, January 2019. URL `http://arxiv.org/abs/1807.03748`. arXiv: 1807.03748.

Liam Paninski. Estimation of Entropy and Mutual Information. *Neural Computation*, 15(6):1191–1253, June 2003. ISSN 0899-7667, 1530-888X. doi: 10.1162/089976603321780272. URL `https://www.mitpressjournals.org/doi/abs/10.1162/089976603321780272`.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 8024–8035, 2019. URL `https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html`.

John Turner Pinkston. Encoding independent sample information sources. 1967. Publisher: MIT Research Laboratory of Electronics.

Ben Poole, Sherjil Ozair, Aäron van den Oord, Alex Alemi, and George Tucker. On Variational Bounds of Mutual Information. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5171–5180. PMLR, 2019. URL http://proceedings.mlr.press/v97/poole19a.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. *arXiv:2103.00020 [cs]*, February 2021. URL http://arxiv.org/abs/2103.00020. arXiv: 2103.00020.

Nicolas Rashevsky. Life, information theory, and topology. *The bulletin of mathematical biophysics*, 17(3):229–235, September 1955. ISSN 1522-9602. doi: 10.1007/BF02477860. URL https://doi.org/10.1007/BF02477860.

David Reinsel, John Gantz, and John Rydning. Data age 2025: The evolution of data to life-critical. *Don't Focus on Big Data*, 2017.

Jorma J. Rissanen. Generalized Kraft Inequality and Arithmetic Coding. *IBM Journal of Research and Development*, 20(3):198–203, May 1976. ISSN 0018-8646. doi: 10.1147/rd.203.0198. Conference Name: IBM Journal of Research and Development.

John Schulman. Sending Samples Without Bits-Back, 2020. URL http://joschu.net/blog/sending-samples.html.

Claude E Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec*, 4(142-163):1, 1959.

Saurabh Singh, Sami Abu-El-Haija, Nick Johnston, Johannes Ballé, Abhinav Shrivastava, and George Toderici. End-to-End Learning of Compressible Features. In *IEEE International Conference on Image Processing, ICIP 2020, Abu Dhabi, United Arab Emirates, October 25-28, 2020*, pp. 3349–3353. IEEE, 2020. doi: 10.1109/ICIP40778.2020.9190860. URL https://doi.org/10.1109/ICIP40778.2020.9190860.

Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszár. Lossy Image Compression with Compressive Autoencoders. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=rJiNwv9gg.

Naftali Tishby, Fernando C. N. Pereira, and William Bialek. The information bottleneck method. *CoRR*, physics/0004057, 2000. URL http://arxiv.org/abs/physics/0004057.

Lav R. Varshney and Vivek K. Goyal. Benefiting from Disorder: Source Coding for Unordered Data. *arXiv:0708.2310 [cs, math]*, August 2007. URL http://arxiv.org/abs/0708.2310. arXiv: 0708.2310.

Bastiaan S. Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation Equivariant CNNs for Digital Pathology. *arXiv:1806.03962 [cs, stat]*, June 2018. URL http://arxiv.org/abs/1806.03962. arXiv: 1806.03962.

Chris S. Wallace. Classification by Minimum-Message-Length Inference. In Selim G. Akl, Frantisek Fiala, and Waldemar W. Koczkodaj (eds.), *Advances in Computing and Information - ICCI'90, International Conference on Computing and Information, Niagara Falls, Canada, May 23-26, 1990, Proceedings*, volume 468 of *Lecture Notes in Computer Science*, pp. 72–81. Springer, 1990. doi: 10.1007/3-540-53504-7_63. URL https://doi.org/10.1007/3-540-53504-7_63.

Tailin Wu, Ian S. Fischer, Isaac L. Chuang, and Max Tegmark. Learnability for the Information Bottleneck. *Entropy*, 21(10):924, 2019. doi: 10.3390/e21100924. URL https://doi.org/10.3390/e21100924.

Yibo Yang, Robert Bamler, and Stephan Mandt. Improving inference for neural image compression. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 573–584. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/066f182b787111ed4cb65ed437f0855b-Paper.pdf.

Yongwook Choi and Wojciech Szpankowski. Compression of graphical structures. In *2009 IEEE International Symposium on Information Theory*, pp. 364–368, Seoul, South Korea, June 2009. IEEE. ISBN 978-1-4244-4312-3. doi: 10.1109/ISIT.2009.5205736. URL `http://ieeexplore.ieee.org/document/5205736/`.

A Notation and Assumptions

A.1 Notation

Letters that are upper-case X, calligraphic $\mathcal{X}$, and lower-case x, represent, respectively, a random variable (r.v.), its associated sample space, and a realization of it. We denote the probability density function as $p(X)$, which we suppose always exist. $\{x_i\} \overset{\text{i.i.d.}}{\sim} p(X)$ denotes independent and identical samples from $p(X)$. $X \overset{\text{d}}{\sim} \mathcal{N}(0, 1)$ denotes that X has a certain distribution (here Gaussian). The composition of a function f with a r.v. X is denoted $f(X)$. Expectations are written as: $\mathrm{E}_{p(X)}[X] := \int xp(x)\,\mathrm{d}x$, while their Monte Carlo approximations have a hat $\hat{\mathrm{E}}_{\mathrm{p(X)}}[X] = \frac{1}{|\{x_i\}|}\sum_i x_i$ for $\{x_i\} \overset{\text{i.i.d.}}{\sim} p(X)$. The KL divergence is denoted as $\mathrm{D}_{\mathrm{KL}}[p(X)\|q(X)] := \int \log p(X)\frac{p(X)}{q(X)}\,\mathrm{d}x$. The mutual information $\mathrm{I}[X; Z] := \mathrm{D}_{\mathrm{KL}}[p(X, Z)\|p(X)p(Z)]$. The (differential or discrete) entropy of a r.v. is $\mathrm{H}[X] := \mathrm{E}_{p(X)}[-\log p(X)]$, while the conditional (differential) entropy is $\mathrm{H}[X \,|\, Z] := \mathrm{E}_{p(X,Z)}[-\log p(X|Z)]$. Independence between two r.v.s is denoted with $\cdot \perp \cdot$. The cardinality of a set is denoted by $|\cdot|$. $\cdot \circ \cdot$ denotes a composition of two functions. $x \sim x'$ denotes that x and x' are equivalent w.r.t. an equivalence relation on $\mathcal{X}$ (the exact relation being implicit). The equivalence class of x under $\sim$ consist of all elements that are equivalent to x, i.e. $[x] := \{x' \in \mathcal{X} \,|\, x' \sim x\}$. We will often use variational optimization over probability distribution, when the variational family is not made explicit it means that the optimiztion is over all possible densities, e.g. $\min_{q(Y \,|\, X)}$ means that that the optimization is done over the collection of all conditional probability densities on $\mathcal{Y}$ given $x \in \mathcal{X}$.

Letters X,Z,Y respectively refer to the input, representation and target of a predictive task.

A.2 Assumptions

Our results require some assumptions and we make additional mild assumptions for clarity, we discuss those in this section. Extending our framework to different losses (dropping Assumption 1) should be investigated in future work. **All other assumptions should hold in most practical scenarios.**

Assumption 1 (Log Loss)**.** We restrict ourselves to the Bayes risk w.r.t. log loss. Using the log loss makes the link between information theory and Bayes risk more natural as seen in Lemma 3. This is the standard loss to train neural classifiers. Although our theory also holds for regression tasks (continuous Y) it is more common to work with the squared error in that case. This assumption is necessary in the current version of Thm. 1, but we hope to generalize the result to other losses in future work.

Assumption 2 (Existence of Densities)**.** We restrict ourselves to cases where the probability mass/density function exist, i.e., to probability distributions that are absolutely continuous w.r.t. to the underlying measure. This is not a necessary assumption but it simplifies the notation, and ensures that the differential entropy of r.v.s is well defined.

Assumption 3 (Bounded Bayes Risk)**.** We restrict ourselves to tasks Y s.t. the Bayes risk is always bounded, i.e. $\forall Z$ we have $|\,\mathrm{R}[Y \,|\, Z]\,| < \infty$. This ensures that taking differences of Bayes risks as in Def. 3 is well defined. We could also directly assume that the latter difference is well defined, but that would require dealing with limits which would unnecessarily complicate the proofs. For the case of discrete Y this is for example the case when $\mathrm{Var}[Y] < \infty$. [7] For continuous Y the same is true (as $\mathrm{H}[Y \,|\, Z] \leq \mathrm{H}[Y] \leq \frac{1}{2}\log(2\pi e(\mathrm{Var}[Y])) < \infty$) as long as the the (differential) entropy is not $-\infty$ which essentially happens if there are no "singularities".

Assumption 4 (Measurability of Functions)**.** We assume that all functions are measurable. We particularly require (i) the measurability of $M(\cdot)$ which implies that $M(X)$ is a r.v.. (ii) the measurability

[7] For the case where $\mathcal{Y} = \mathbb{Z}$ we have $\mathrm{H}[Y \,|\, Z] \leq \mathrm{H}[Y] \leq \frac{1}{2}\log(2\pi e(\mathrm{Var}[Y]) + \frac{1}{12}) < \infty$ where the second inequality comes from (Massey, 1988) and the last inequality comes from our assumption of finite $\mathrm{Var}[Y]$. This can be generalized to any countable $\mathcal{Y}$ by realizing that Y can always be rewritten as a bijection of Y' where $\mathcal{Y}' = \mathbb{Z}$ which has finite entropy due to the previous proof, and so Y also does.

of the projection $\pi : \mathcal{X} \rightarrow \mathcal{X}/\sim$ which implies that there always exists a maximal invariant as π is one of them. This assumption essentially holds for all practical purposes.

Assumption 5 (Existence of regular conditional probabilities)**.** We restrict probability spaces (e.g. Radon spaces) that satisfy the regular conditional probability property, so that all considered random variables admit a regular conditional probability. This is necessary to ensure the existence of probability kernel in Lemma 2. This assumption essentially holds for all practical purposes.

Assumption 6 (Countably Many Equivalence Classes)**.** We restrict our discussion to equivalences $\sim$ on $\mathcal{X}$ s.t. the quotient set $\mathcal{X}/\sim := \{[x] \mid x \in \mathcal{X}\}$ is countable. This ensures that $M(X)$ is a discrete r.v. thereby ensuring that our invariance distortion $\mathrm{D}_{\sim}[X, Z]$ is independent of the choice of maximal invariant M as the conditional entropy is invariant to bijections. As currently written our results (not only the proofs) do not hold without that assumption, but they can probably be extended.

Note that this assumption holds when $\mathcal{X}$ is countable which always happens in practice due to floating point arithmetics, i.e. every real number has to be rounded to the closest 64 bits number. Another perspective is to say that $\mathcal{X}$ is actually uncountable, but that all tasks we care about are always invariant to rounding to the nearest 64 bits number due to floating point arithmetics. As a result, the maximal invariant is the usual maximal invariant rounded to the closest floating point. For example, if X is a 2D Gaussian we cannot work directly with translations on the y-axis (which gives uncountably many $[x]$, one for each real number on the x-axis), but can work with y-axis invariance combined with invariance to rounding on the x-axis (e.g. closest 64 bits number).

Assumption 7 (Finite $\mathrm{H}[M(X)]$)**.** We restrict our discussion to X and equivalences on $(\mathcal{X}, \sim)$ s.t. if there exists a maximal invariant M then at least one has finite discrete entropy $\mathrm{H}[M(X)] < \infty$. This is a necessary condition for $M(X)$ to be in the tasks $\mathcal{T}$ of interest as we restricted our tasks of interests to the ones with bounded Bayes risk (Assumption 3). We require this in the proof of Prop. 1 to ensure that $\mathrm{D}_{\sim}[X, Z] = \mathrm{R}[M(X) \mid Z]$ is achievable. As discussed in Assumption 3, it is sufficient (but not necessary) to assume that $M(X)$ has finite variance which is implied from X having finite variance. as is often the case.

B Formal Definitions

In the main paper we were relatively informal in our definitions, here we restate our main definitions more formally.

First let us define the notion of valid distortion, which is necessary for applying the original rate distortion theorem (Shannon, 1959; Cover & Thomas, 2006). [8]

Definition 1 (Valid Distortion)**.** Let X and Z be two r.v.s that respectively t.v.i. $\mathcal{X}$ and $\mathcal{Z}$. Then an (expected) distortion D is said *valid* w.r.t. X, Z if there exists a point wise distortion $d : \mathcal{X} \times \mathcal{Z} \rightarrow \mathbb{R}^+$ s.t. for any x, z s.t. $p(x, z) > 0$ we have [9] $d(x, z) \leq d_{max} < \infty$ and

$$\mathrm{D} := \mathrm{E}_{p(X,Z)}[d(X, Z)] \tag{8}$$

Let us define the notion of representation Z which arises by encoding X using $p(Z \mid X)$ independently of any task Y.

Definition 2 (Representation)**.** Let X, Z be two r.v.s and $\mathcal{T}$ be a set of r.v.s. Z is a *representation* of X for $\mathcal{T}$ iff $\forall Y \in \mathcal{T}$ we have the pairwise conditional independence $Y \perp Z \mid X$.

Let us now define the tasks of interest.

[8] The rate distortion theorem has been extended to more general cases (Pinkston, 1967; Gallager, 1968; Berger, 1968) but we preferred using the original theorem for clarity.

[9] Note that Cover & Thomas's (2006) use the stronger assumption of $\max_{x \in \mathcal{X}, z \in \mathcal{Z}} d(x, z) \leq d_{max} < \infty$ but they only require that for any x, z s.t. $p(x, z) > 0$ we have $d(x, z) \leq d_{max} < \infty$ to bound the worst case distortion between two sequences (achievability proof of the rate distortion theorem on page 321 and the achievability proof on page 327).

Definition 3 (Invariant Tasks of Interest)**.** Let X and Y be two r.v.s that respectively t.v.i. $\mathcal{X}$ and $\mathcal{Y}$. Let $\sim$ denote an equivalence relation on $\mathcal{X}$ satisfying Assumption 6 and s.t. X satisfies Assumption 7. We say that Y is an *invariant task of interest* w.r.t. $(\mathcal{X}, \sim)$, iff:

- for any r.v. Z the Bayes risk is bounded $\mathrm{R}[Y \mid Z] < \infty$ (Assumption 3).
- the function $x \mapsto p(Y \mid x)$ is invariant w.r.t. $\sim$, i.e. $\forall x, x' \in \mathcal{X}$ we have

$$x \sim x' \implies p(Y \mid x) = p(Y \mid x') \tag{9}$$

Furthermore, we denote as $\mathcal{T}$ all such invariant tasks of interest ($\mathcal{X}$ and $\sim$ are implicit in the notation).

Let us now recall our desired distortion $\mathrm{D}_{\mathcal{T}}$.

Definition 4 (Invariance Distortion)**.** Let X and Z be two r.v.s. Let $\mathcal{T}$ be the invariant tasks of interest as in Def. 3. The *invariance distortion* $\mathrm{D}_{\mathcal{T}}$ is defined as:

$$\mathrm{D}_{\sim}[X, Z] := \sup_{Y \in \mathcal{T}} \quad \mathrm{R}[Y \mid Z] - \mathrm{R}[Y \mid X] \tag{10}$$

Finally, let us define the notion of maximal invariant.

Definition 5 (Maximal Invariant)**.** Let $\sim$ denote an equivalence relation on $\mathcal{X}$. We say that a measurable (Assumption 4) function $M : \mathcal{X} \to \mathcal{M}$ is a *maximal invariant* w.r.t. $(\mathcal{X}, \sim)$ iff

$$\forall x, x' \in \mathcal{X} \quad x \sim x' \iff M(x) = M(x') \tag{11}$$

Note that our notion of maximal invariants generalizes the notion of maximal invariants in probabilistic group theory (Eaton, 1989). Refer to Lehmann & Romano (2005) for many more examples in the group case. Important examples of maximal invariants can be summarized informally as follows:

Translations If $\mathcal{X} = \mathbb{R}^2$ and the examples are invariant to translations of the second coordinate: $x \mapsto x + [0, t]^T$; then a maximal invariant is the first coordinate $M : x \mapsto x_1$.

Scaling If $\mathcal{X} = \mathbb{R}^n$ and the examples are invariant to scalar scaling : $x \mapsto c \cdot x$; then a maximal invariant is given by rescaling by the last coordinate $M : x \mapsto x/x_n$.

Rotations If $\mathcal{X} = \mathbb{R}^n$ and the examples are invariant to rotations : $x \mapsto \mathrm{Rot}(\theta) \cdot x$; then a maximal invariant is given by the Euclidean norm $M : x \mapsto ||x||_2$.

Permutations If $\mathcal{X} = \mathbb{R}^n$ and the examples are invariant to permutations of the coordinates : $x \mapsto (x_{\pi(1)}, \ldots, x_{\pi(n)})$; then a maximal invariant is the empirical measure (the type class, or histogram).

Graph Isomorphisms If $\mathcal{X}$ is the set of all graphs, and the examples are invariant to graph isomorphisms; then a maximal invariant is the graph canonization.

C Proofs: Optimal Bit-Rate

In this section we prove all our claims in Sec. 2.

C.1 Reformulating and Validating Def. 4

In this section we prove the equivalence between Def. 4 and the nicer $\mathrm{H}[M(X) \mid Z]$ which is the core of our work.

The main steps in the proof are the following:

1. We show that if Y is an invariant tasks then $Y - M(X) - X$ forms a Markov chain.
2. Using the strict properness of the log loss, we relate the Bayes risk to the differential entropy:

$$\mathrm{R}[Y \mid Z] = \mathrm{H}[Y \mid Z] \tag{12}$$

3. Using (1) and (2), the chain rule and the data processing inequality we show that the supremum is achieved by $M(X)$:

$$\sup_{Y \in \mathcal{T}} \mathrm{R}[Y \mid Z] - \mathrm{R}[Y \mid X] = \mathrm{H}[M(X) \mid Z] - \mathrm{H}[M(X) \mid X] \tag{13}$$

4. As M is a (deterministic) function and $M(X)$ is discrete we have $\mathrm{H}[M(X) \mid X] = 0$

$$\sup_{Y \in \mathcal{T}} \mathrm{R}[Y \mid Z] - \mathrm{R}[Y \mid X] = \mathrm{H}[M(X) \mid Z] \tag{14}$$

The first step, consists in showing that for invariant tasks we have $Y \perp\!\!\!\perp X \mid M(X)$. Specifically, we prove that that any conditionally invariant r.v. can be decomposed as a function of a maximal invariant and independent noise. This can be seen as a probabilistic extension of the theorem on projections (Theorem 19 and its corollary in Lane & Birkhoff (1999)) . This can also be seen as a generalization of an important probabilistic group theoretical results (Theorem 4.4 in Eaton (1989), Theorem 7 in Bloem-Reddy & Teh (2020)), to any equivalences (rather than only groups) and without making the assumption of (marginal) invariance of $p(X)$ to $\sim$.

The following intermediate result is needed.

Lemma 1. A measurable function $f \colon \mathcal{X} \to \mathcal{F}$ is invariant with respect to $(\mathcal{X}, \sim)$ if and only if there exists a measurable function $g \colon \mathcal{M} \to \mathcal{F}$ such that $f(x) = (g \circ M)(x)$ for all $x \in \mathcal{X}$.

Proof. Clearly, if $f(x) = (g \circ M)(x) = g(M(x))$ then f is $(\mathcal{X}, \sim)$-invariant because M is.

Conversely, assume that f is $(\mathcal{X}, \sim)$-invariant. Let $F_0 \in \mathcal{F}$ denote an arbitrary fixed element. Then we can construct g as

$$g(m) = \begin{cases} f(x) & \text{if } m \text{ is in the range of } M; \\ F_0 & \text{otherwise} \end{cases} .$$

□

Lemma 2. Let X and Y be 2 r.v., and $M : \mathcal{X} \to \mathcal{M}$ be a maximal invariant w.r.t. $(\mathcal{X}, \sim)$ as in Def. 5. Then Y is (conditionally) invariant w.r.t. $(\mathcal{X}, \sim)$ as in Eq. (9) if and only if $Y \perp\!\!\!\perp X \mid M(X)$.

Proof. Assume that Y is (conditionally) invariant w.r.t. $(\mathcal{X}, \sim)$ as in Eq. (9). By standard results, there exists a probability kernel $K(A, x)$ such that for all measurable sets A, $x \mapsto K(A, x)$ is a measurable function mapping $\mathcal{X} \to \mathbb{R}_+$. Conditional invariance means that $x \sim x' \Rightarrow K(A, x) = K(A, x')$ for all x, x'. That is, as a function of x, $K(A, \bullet)$ is invariant w.r.t. $(\mathcal{X}, \sim)$. By Lemma 1, $x \mapsto K(A, x)$ can be expressed as a measurable function of M, i.e., $K(A, x) = K'(A, M(x))$, for another probability kernel K'. This implies that $P(Y|X) = P(Y|M(X))$ almost surely. Because $M(X)$ is a function of X, that further implies that $P(Y|X, M(X)) = P(Y|M(X))$, i.e., $Y \perp\!\!\!\perp X \mid M(X)$. □

We now relate the Bayes risk and conditional entropy. This is a simple lemma that directly comes from the fact that the conditional distribution $p(Y \mid Z)$ is the Bayes predictor.

Lemma 3. Let Y, X be r.v.s with bounded $\mathrm{R}[Y \mid X]$ then the log loss Bayes irsk is equal to the conditional (differential) entropy:

$$\mathrm{R}[Y \mid X] = \mathrm{H}[Y \mid X] \tag{15}$$

Proof.

$$\begin{aligned} \mathrm{R}[Y \mid X] &= \inf_{q(Y \mid X)} \mathrm{E}_{p(X,X)}[-\log q(Y|X)] && \text{Definition} && (16) \\ &= \mathrm{E}_{p(X,Z)}[-\log p(Y|X)] && \text{Strict Proper.} && (17) \\ &= \mathrm{H}[Y \mid X] && \text{Definition} && (18) \end{aligned}$$

Where Eq. (17) uses the strict properness of the logarithmic scoring function rule (Gneiting & Raftery, 2007). □

In the rest of this section we will often be working with the entropy $\mathrm{H}[M(X)]$ and conditional entropies such as $\mathrm{H}[M(X) \mid Z]$. Importantly, we would like our results to be independent of the choice of maximal invariant M. We now prove that this will indeed be the case as all these (conditional) entropy terms are independent of the choice of M. We only prove it for the marginal entropy $\mathrm{H}[M(X)]$ but the same proof holds for conditional entropies.

Lemma 4. Let X be a r.v. that t.v.i $\mathcal{X}$. Let $\sim$ denote an equivalence relation on $\mathcal{X}$ satisfying Assumption 6. Let M and M' two different maximal invariants w.r.t. $(\mathcal{X}, \sim)$ as in Def. 5, then $\mathrm{H}[M(X)] = \mathrm{H}[M'(X)]$.

Proof. First notice that if M and M' are both maximal invariant then [10] there exists a bijective function $f : \mathcal{M} \to \mathcal{M}$ s.t. $M' = f \circ M$. Indeed, from the projection theorem (Theorem 19 in Lane & Birkhoff (1999)) we know that M is a maximal invariant if and only if there is a bijective function $g : \mathcal{X}/\sim \to \mathcal{M}$ s.t. the maximal invariant is the composition of g and the projection onto equivalence classes, i.e. $\forall x \in \mathcal{X}$ we have $M(x) = g([x])$. Let g' be the corresponding bijection for M'. Then we have $M' = f \circ M$ with $f := g' \circ g^{-1}$ which is indeed bijective $f^{-1} := g \circ g'^{-1}$.

Due to Assumption 6, $M(X)$ is a discrete r.v. and so $\mathrm{H}[M(X)]$ is the discrete entropy, which is invariant to bijective functions indeed $\mathrm{H}[M(X)] = \mathrm{I}[M(X); M(X)] = \mathrm{I}[f(M(X)), f(M(X))] = \mathrm{H}[f(M(X))]$ where we used the invariance of mutual information to bijections (Kraskov et al., 2004). We thus conclude that $\mathrm{H}[M(X)] = \mathrm{H}[f(M(X))] = \mathrm{H}[M'(X)]$ as desired. □

We now prove that all $M(X)$ are always in the set of downstream tasks $\mathcal{T}$.

Lemma 5. Let X be a r.v. Let $\mathcal{T}$ be the invariant tasks of interest as in Def. 3. Then any maximal invariant (Def. 5) M is such that $M(X) \in \mathcal{T}$ and there exists at least one maximal invariant.

Proof. First, we have to prove that a maximal invariant always exists. We do so by construction. By definition equivalent elements have the same equivalence class and so $x \sim x' \iff [x] = [x']$. We thus have that the projection map $\pi : x \mapsto [x]$ satisfies Eq. (11). Due to Assumption 4 the projection map is measurable and so it is a maximal invariant.

As there exists (at least) one maximal invariant we have that there exists $M(X)$ s.t. $\forall Z$ we have $\mathrm{R}[M(X) \mid Z] = \mathrm{H}[M(X) \mid Z] \leq \mathrm{H}[M(X)] < \infty$, where the first equality comes from Lemma 3, and second inequality from the fact that conditioning decreases entropy, and the final inequality comes from Assumption 7. Due to Lemma 4 we further have that all maximal invariants M_i have a bounded $\mathrm{R}[M_i(X) \mid Z]$ which is necessary for all $M_i(X)$ to be in $\mathcal{T}$. As all maximal invariants satisfy by definition the invariance in Eq. (9), we conclude that they are all in $\mathcal{T}$ as desired. □

We are now ready to prove the desired proposition.

Proposition 1 (Nicer $\mathrm{D}_{\mathcal{T}}$)**.** Let X be a r.v. Let $\mathcal{T}$ be the invariant tasks of interest as in Def. 3, M be any maximal invariant as in Def. 5, and Z be a representation of X as in Def. 2. Let $\mathrm{D}_{\mathcal{T}}$ be as in Def. 4. Then $\mathrm{D}_{\mathcal{T}}$ is a valid distortion(Def. 1) and

$$\mathrm{D}_{\sim}[X, Z] = \mathrm{H}[M(X) \mid Z] = \mathrm{R}[M(X) \mid Z] \tag{19}$$

Proof. First let us prove that $\mathrm{D}_{\sim}[X, Z] = \mathrm{H}[M(X) \mid Z]$

$$\mathrm{D}_{\sim}[X, Z] := \sup_{Y \in \mathcal{T}} \mathrm{R}[Y \mid Z] - \mathrm{R}[Y \mid X] \qquad \text{Def. 4} \tag{20}$$

$$= \sup_{Y \in \mathcal{T}} \mathrm{H}[Y \mid Z] - \mathrm{H}[Y \mid X] \qquad \text{Lemma 3} \tag{21}$$

$$= \sup_{Y \in \mathcal{T}} \mathrm{H}[Y \mid Z] - \mathrm{H}[Y \mid X, M(X)] \qquad Y \perp M(X) | X \tag{22}$$

$$= \sup_{Y \in \mathcal{T}} \mathrm{H}[Y \mid Z] - \mathrm{H}[Y \mid M(X)] \qquad \text{Lemma 2: } Y \perp X | M(X) \tag{23}$$

$$= \sup_{Y \in \mathcal{T}} \mathrm{H}[Y \mid Z] - \mathrm{H}[Y \mid M(X), Z] \qquad \text{Def. 2: } Y \perp Z | M(X) \tag{24}$$

$$= \sup_{Y \in \mathcal{T}} \mathrm{I}[Y; M(X) | Z] \qquad \text{Def.} \tag{25}$$

$$= \sup_{Y \in \mathcal{T}} \mathrm{I}[M(X); Y | Z] \qquad \text{Symmetry} \tag{26}$$

$$= \sup_{Y \in \mathcal{T}} \mathrm{H}[M(X) \mid Z] - \mathrm{H}[M(X) \mid Y, Z] \qquad \text{Def.} \tag{27}$$

[10] It can easily be shown that this is an if and only if.

$$
\begin{aligned}
&= \mathrm{H}[M(X) \,|\, Z] - \inf_{Y \in \mathcal{T}} \mathrm{H}[M(X) \,|\, Y, Z] && (28)\\
&= \mathrm{H}[M(X) \,|\, Z] - \mathrm{H}[M(X) \,|\, M(X)] \\
&\quad + (\mathrm{H}[M(X) \,|\, M(X)] - \inf_{Y \in \mathcal{T}} \mathrm{H}[M(X) \,|\, Y, Z]) && (29)\\
&= \mathrm{H}[M(X) \,|\, Z] - \mathrm{H}[M(X) \,|\, M(X)] && \text{DPI} \quad (30)\\
&= \mathrm{H}[M(X) \,|\, Z] - 0 && \text{Discrete entropy} \quad (31)\\
&= \mathrm{R}[M(X) \,|\, Z] && \text{Lemma 3} \quad (32)
\end{aligned}
$$

To go from Eq. (24) to Eq. (27) we use the the symmetry and definition of conditional mutual information, which essentially corresponds to using the chain rule. Eq. (30) uses the fact that $\mathrm{H}[M(X) \,|\, M(X)] \leq \inf_{Y \in \mathcal{T}} \mathrm{H}[M(X) \,|\, Y, Z]$ because of the data processing inequality and the trivial fact that $M(X) - M(X) - (Z, Y)$ forms a Markov Chain. As $M(X) \in \mathcal{T}$ (Lemma 5) we also have the other inequality $\inf_{Y \in \mathcal{T}} \mathrm{H}[M(X) \,|\, Y, Z] \leq \mathrm{H}[M(X) \,|\, M(X), Z] \leq \mathrm{H}[M(X) \,|\, M(X)]$, and so we conclude that the equality $\mathrm{H}[M(X) \,|\, M(X)] = \inf_{Y \in \mathcal{T}} \mathrm{H}[M(X) \,|\, Y, Z]$ holds. Finally, Eq. (31) uses the fact that $M(X)$ is a discrete r.v. due to Assumption 6 and so the discrete conditional entropy $\mathrm{H}[M(X) \,|\, M(X)] = 0$.

It is now easy to see that $\mathrm{D}_{\mathcal{T}}$ is valid as $\mathrm{D}_{\sim}[X, Z] = \mathrm{H}[M(X) \,|\, Z] = \mathrm{E}_{p(X,Z)}[d(X, Z)]$ with $d(x, z) := -\log p(M(x) \,|\, z)$ which due to the discreteness of $M(X)$ (Assumption 6) is a function whose codomain is $\mathbb{R}^+$ as desired. As conditioning decreases entropy we also have $\mathrm{H}[M(X) \,|\, Z] \leq \mathrm{H}[M(X)] < \infty$ where the last inequality comes from Assumption 7. As $M(X)$ is discrete we conclude that $\forall x, z$ s.t. $p(x, z) > 0$ there exists d_{max} s.t. $d(x, z) \leq d_{max} < \infty$ and so $\mathrm{D}_{\mathcal{T}}$ is valid. □

C.2 Proofs for Theorem 1

Our main theoretical contribution is to characterize the minimal achievable rate to bound the Bayes risk of any invariant task. The result follows from Shannon's (1959) rate distortion theorem, the definition of $\mathrm{D}_{\mathcal{T}}$ in terms of worst Bayes risk, the fact that $\mathrm{D}_{\mathcal{T}}$ is valid, and our characterization of $\mathrm{D}_{\mathcal{T}}$ in terms of entropy (Prop. 1).

First let us restate the well known rate distortion theorem. Here we use the statement as given in Cover & Thomas (2006). Note that achievability is usually defined for deterministic encoders (e.g. on 306 of Cover & Thomas (2006)) but the proof holds for stochastic encoders (as noted on p.316 of Cover & Thomas (2006)) which we use in our work.

Lemma 6. (Theorem 10.2.1 in Cover & Thomas (2006)) Let $\mathrm{D}[X; Z]$ be a valid distortion as in Def. 1. The minimum achievable bit-rate for transmitting an i.i.d. source X with expected distortion less than $\delta \geq 0$ is given by the rate-distortion function:

$$
R(\delta) = \min_{p(Z|X) \text{ s.t. } \mathrm{D}[X;Z] \leq \delta} \mathrm{I}[X; Z] \tag{33}
$$

We can now state our rate invariance theorem.

Theorem 1 (Rate Invariance)**.** Let X be a r.v. and $\delta \geq 0$. Let $\mathcal{T}$ be the invariant tasks of interest as in Def. 3, M be any maximal invariant as in Def. 5, and Z be a representation of X as in Def. 2. Let $Rate(\delta)$ denote the minimum achievable bit-rate for transmitting an i.i.d. source of Z s.t. for any $Y \in \mathcal{T}$ we have $\mathrm{R}[Y \,|\, Z] \leq \delta + \mathrm{R}[Y \,|\, X]$. Then $Rate(\delta)$ is finite and given by

$$
\begin{aligned}
Rate(\delta) &= \max(0,\ \mathrm{H}[M(X)] - \delta) && (34)\\
&= \max(0,\ \mathrm{H}[X] - \mathrm{H}[X \,|\, M(X)] - \delta) && (35)
\end{aligned}
$$

Proof. In the following we first prove that $Rate(\delta) \leq \mathrm{H}[M(X)] - \delta$. We then prove that the rate $\max(0,\ \mathrm{H}[M(X)] - \delta)$ is achievable from which we conclude that $Rate(\delta) = \max(0,\ \mathrm{H}[M(X)] - \delta)$. Finally, we conclude by proving $\mathrm{H}[M(X)] = \mathrm{H}[X] - \mathrm{H}[X \,|\, M(X)]$ to get our result.

We want to transmit Z s.t. $\forall Y \in \mathcal{T}$ we have $\mathrm{R}[Y \,|\, Z] \leq \delta + \mathrm{R}[Y \,|\, X]$, in other words we would like $\sup_{Y \in \mathcal{T}} \mathrm{R}[Y \,|\, Z] - \mathrm{R}[Y \,|\, X] =: \mathrm{D}_{\sim}[X, Z] < \delta$. We thus need to compute the minimal achievable bit-rate for transmitting an i.i.d. source of Z s.t. $\mathrm{D}_{\sim}[X, Z] \leq \delta$. As $\mathrm{D}_{\mathcal{T}}$ is valid (Prop. 1) we can directly apply the rate distortion theorem (Lemma 6):

$$\begin{aligned}
\mathit{Rate}(\delta) &= \min_{p(Z|X) \text{ s.t. } \mathrm{D}_{\sim}[X,Z] \leq \delta} \mathrm{I}[X; Z] && \text{Lemma 6 and Prop. 1} \quad (36)\\
&= \min_{p(Z|X) \text{ s.t. } \mathrm{D}_{\sim}[X,Z] \leq \delta} \mathrm{I}[X, M(X); Z] && \text{Bijection} \quad (37)\\
&= \min_{p(Z|X) \text{ s.t. } \mathrm{D}_{\sim}[X,Z] \leq \delta} \mathrm{I}[M(X); Z] + \mathrm{I}[X; Z \,|\, M(X)] && \text{Chain Rule} \quad (38)\\
&\geq \min_{p(Z|X) \text{ s.t. } \mathrm{D}_{\sim}[X,Z] \leq \delta} \mathrm{I}[M(X); Z] && \text{Positivity} \quad (39)\\
&= \min_{p(Z|X) \text{ s.t. } \mathrm{D}_{\sim}[X,Z] \leq \delta} \mathrm{H}[M(X)] - \mathrm{H}[M(X) \,|\, Z] && (40)\\
&= \min_{p(Z|X) \text{ s.t. } \mathrm{D}_{\sim}[X,Z] \leq \delta} \mathrm{H}[M(X)] - \mathrm{D}_{\sim}[X, Z] && \text{Prop. 1} \quad (41)\\
&\geq \min_{p(Z|X) \text{ s.t. } \mathrm{D}_{\sim}[X,Z] \leq \delta} \mathrm{H}[M(X)] - \delta && (42)\\
&= \mathrm{H}[M(X)] - \delta && \text{No } Z \quad (43)
\end{aligned}$$

Where Eq. (37) uses the invariance of mutual information to bijections which here is $X \mapsto X, M(X)$. As the rate is always non-negative we have $\mathit{Rate}(\delta) \geq \max(0, \mathrm{H}[M(X)] - \delta)$.

We now prove that $\max(0, \mathrm{H}[M(X)] - \delta)$ is attainable and so $\mathit{Rate}(\delta) = \max(0, \mathrm{H}[M(X)] - \delta)$. Specifically we need to find a representation Z of X s.t.

$$\mathit{Rate}(\delta) = \begin{cases} 0 & \text{If } \delta \geq \mathrm{H}[M(X)] \\ \mathrm{H}[M(X)] - \delta & \text{Else} \end{cases} \tag{44}$$

The first case is trivial: set Z to be independent of $M(X)$ and X, e.g. a constant. Then, $\mathrm{D}_{\sim}[X, Z] = \mathrm{H}[M(X) \,|\, Z] = \mathrm{H}[M(X)] \leq \delta$ and $\mathit{Rate}(\delta) = \mathrm{I}[Z; X] = 0$.

For the second case we need $\mathit{Rate}(\delta) \geq \mathrm{H}[M(X)] - \delta$ to be an equality when $\delta > \mathrm{H}[M(X)]$. This happens iff there exists a Z s.t. inequalities Eq. (39) and Eq. (42) are equalities, i.e. iff $\mathrm{I}[X; Z \,|\, M(X)] = 0$ and $\mathrm{D}_{\sim}[X, Z] = \delta$. To get that we start from $Z = M(X)$ (which satisfies $\mathrm{I}[X; Z \,|\, M(X)] = 0$) and "modify" it s.t. $\mathrm{D}_{\sim}[X, Z] = \delta$. We can do so by "erasing" a fraction of bits by mapping all $m \to \epsilon$ with a constant probability α, similarly to binary erasure channels. Specifically, let $\mathcal{Z} := \mathcal{M} \cup \{\epsilon\}$ and Z be a r.v. that t.v.i. $\mathcal{Z}$ and whose conditional density parametrized by $\alpha \in [0, 1[$ is:

$$\forall z \in \mathcal{Z}, \forall m \in \mathcal{M}, \quad p(z \,|\, m) = \begin{cases} 1 - \alpha & \text{if } z = m \\ \alpha & \text{if } z = \epsilon \\ 0 & \text{else} \end{cases} \tag{45}$$

A simple computation then gives $\mathrm{D}_{\sim}[X, Z] = \mathrm{H}[M(X) \,|\, Z] = (1 - \alpha)\, \mathrm{H}[M(X) \,|\, Z = M(X)] + \alpha\, \mathrm{H}[M(X) \,|\, Z = \epsilon] = \alpha\, \mathrm{H}[M(X)]$. To have $\mathrm{D}_{\sim}[X, Z] = \delta$ we thus need to set $\alpha = \frac{\delta}{\mathrm{H}[M(X)]}$. Note that we do not divide by zero as if $\mathrm{H}[M(X)] = 0$ would be in the first case of Eq. (44).

We thus proved that $\max(0,\ \mathrm{H}[M(X)] - \delta)$ is obtainable and that $\mathit{Rate}(\delta) \geq \max(0,\ \mathrm{H}[M(X)] - \delta)$. From which we conclude that the best achievable bit-rate is $\mathit{Rate}(\delta) = \max(0,\ \mathrm{H}[M(X)] - \delta)$. Eq. (35), follows from $\mathrm{H}[M(X)] = \mathrm{I}[M(X); X] = \mathrm{H}[X] - \mathrm{H}[X \,|\, M(X)]$. The finiteness of $\mathit{Rate}(\delta)$ comes from the fact that $\mathit{Rate}(\delta) \leq \mathrm{H}[M(X)] < \infty$ due to Assumption 7. □

By setting $\delta = 0$ we directly get the best achievable rate for the lossless prediction but lossy compression setting.

Corollary 1 (Invariant Source Coding)**.** Let X be a r.v. and $\delta \geq 0$. Let $\mathcal{T}$ be the invariant tasks of interest as in Def. 3, M be any maximal invariant as in Def. 5, and Z be a representation of X as in Def. 2. Let $\mathit{Rate}(\delta)$ denote the minimum achievable bit-rate for transmitting an i.i.d. source of Z s.t. for any $Y \in \mathcal{T}$ we have $\mathrm{R}[Y \,|\, Z] = \mathrm{R}[Y \,|\, X]$. Then $\mathit{Rate}(\delta)$ is finite and given by

$$\begin{aligned}
\mathit{Rate}(\delta) &= \mathrm{H}[M(X)] && (46)\\
&= \mathrm{H}[X] - \mathrm{H}[X \,|\, M(X)] && (47)
\end{aligned}$$

C.3 Recovering Subcases

Corollary 1 recovers many previous results in the literature:

Unlabeled Graphs Let us consider the task of compressing unlabeled graphs, here we consider tasks that are invariant to graph isomorphisms. A possible maximal invariant is the graph canonization and $\mathrm{H}[M(X)]$ becomes the well known *structural entropy* (Rashevsky, 1955; Yongwook Choi & Szpankowski, 2009). [11] If all isomorphic graphs are permissible and equiprobable, Yongwook Choi & Szpankowski (2009) show that the structural entropy is $\mathrm{H}[S] = \mathrm{H}[X] - \mathrm{E}_{x \sim p(X)}\left[\log \frac{n!}{|\mathrm{Aut}_{\mathcal{G}}(x)|}\right]$. This is Eq. (47), with the second term corresponds to $\mathrm{H}[X \mid M(X)]$ with a uniform distribution on isomorphic graphs.

Multisets Let us derive the best achievable bit-rate for compressing multisets. Let X be any sequence and $\mathcal{T}$ be invariant to permutations of that sequence. One possible maximal invariant in that case is the empirical measure (also called type), i.e., the counts $K_1, \ldots, K_n$ of each of the n elements that are present in the sequence X. Lossless compression of multisets thus requires $\mathrm{H}[M(X)] = \mathrm{H}[K_1, \ldots, K_n]$, as discussed in (Varshney & Goyal, 2007). Using Eq. (47) we can also characterize the bits gains that you obtain by considering the invariance, namely, $\mathrm{H}[X|M(X)]$. This recovers theorem 1 of (Varshney & Goyal, 2007), where $\mathrm{H}[X|M(X)]$ is called the "order entropy". Note that similarly to our example in the text about i.i.d. coin flips, the amount of bits needed to losslessly compress the multiset grows as $\theta(\log n)$ (Varshney & Goyal, 2007).

Information Bottleneck (IB) Suppose you are interested in predicting a single task $Y = t(X)$, where t is a (deterministic) "target function". The task is invariant to the labeling so the maximal invariant is $t(\cdot)$ and the distortion becomes $\mathrm{H}[T(X) \mid Z] = \mathrm{H}[Y \mid Z]$. Then Eq. (1) becomes the information bottleneck (IB) (Tishby et al., 2000). Using Corollary 1 we see that for lossless predictions the optimal rate is $Rate(0) = \mathrm{H}[Y] = \mathrm{H}[X] - \mathrm{H}[X \mid Y] = \mathrm{I}[X; Y]$ as shown in (Wu et al., 2019; Fischer, 2020). From a compression stand point this is nevertheless not very useful as $Rate(0) = \mathrm{H}[Y]$, so IB for deterministic labels tells you to entropy code Y.

Lossless Let X be discrete. Every task will always be invariant to equality "=". In this case the maximal invariant is the identity function, and we recover Shannon's source coding theorem $Rate(0) = \mathrm{H}[M(X)] = \mathrm{H}[X]$.

D Variational Objectives

In this section we will derive the variational bounds for estimating the rate and the distortion. Recall that the optimal bit-rate is simply the Rate Distortion function wusing our invariance distortion (Eq. (36)), so pareto optimal encoder (for $delta$) can be obtained by using the following arg minimum:

$$Rate(\delta) = \min_{p(Z|X) \text{ s.t. } \mathrm{D}_{\sim}[X,Z] \leq \delta} \mathrm{I}[X; Z] \tag{48}$$

As optimization in machine learning is typically unconstrained, we can use the Lagragian relaxation instead

$$Rate(\beta) = \min_{p(Z|X)} \mathrm{I}[X; Z] + \beta(\delta) \cdot \mathrm{R}[M(X) \mid Z] \tag{49}$$

Both terms $\mathrm{I}[X; Z]$ and $\mathrm{R}[M(X) \mid Z]$ are hard to estimate from samples, so the rest of the section is devoted to deriving variational upper bounds on them.

D.1 Variational Rate $\mathrm{I}[X; Z]$

Let us discuss how to approximate the rate term $\mathrm{I}[X; Z]$. The mutual information is well known to be hard to estimate from samples (Paninski, 2003; McAllester & Stratos, 2020), but fortunaltely many

[11] Also called topological information content.

variational bounds have previously proposed (see (Poole et al., 2019)). In the follwoing we denote a family of variational distributions over Z (priors or entropy models) as $\mathcal{Q} := \{q(Z)\}$.

Mutual Information Bottleneck. The first bound that we consider is the standard upper bound on $\mathrm{I}[X; Z]$, e.g. in VAE or VIB. Specifically:

$$\begin{aligned}
\mathrm{I}[Z; X] &:= \mathrm{H}[Z] - \mathrm{H}[Z \mid X] && (50)\\
&= \mathrm{E}_{p(Z)}[-\log p(Z)] - \mathrm{H}[Z \mid X] && (51)\\
&= \min_{q \in \mathcal{Q}} \mathrm{E}_{p(X)p(Z|X)}\left[-\log \frac{p(Z)q(Z)}{q(Z)}\right] - \mathrm{H}[Z \mid X] && (52)\\
&= \min_{q \in \mathcal{Q}} \mathrm{E}_{p(X)p(Z|X)}[-\log q(Z)] - \mathrm{E}_{p(X)p(Z|X)}\left[\log \frac{p(Z)}{q(Z)}\right] - \mathrm{H}[Z \mid X] && (53)\\
&= \min_{q \in \mathcal{Q}} \mathrm{E}_{p(X)p(Z|X)}[-\log q(Z)] - \mathrm{D}_{\mathrm{KL}}[p(Z) \| q(Z)] - \mathrm{H}[Z \mid X] && (54)\\
&\leq \min_{q \in \mathcal{Q}} \mathrm{E}_{p(X)p(Z|X)}[-\log q(Z)] - \mathrm{H}[Z \mid X] && (55)
\end{aligned}$$

We call this the variational mutual information bottleneck bound. The approximation gap is then $\min_{q \in \mathcal{Q}} \mathrm{D}_{\mathrm{KL}}[p(Z) \| q(Z)]$. This bound has the advantage that if $p(Z) \in \mathcal{Q}$ then bound is tight, which is of course the case when the variational family $\mathcal{Q}$ is unconstrained.

The major issue with the mutual information bottleneck, is that no efficient compressors can in general achieve the rate given by it (Agustsson & Theis, 2020). [12]

Entropy Bottleneck. To have efficient compressors we would like a bound such that the given rate can be achieved by an entropy coder. Indeed, entropy coders have been developped for years and are now very efficient (Rissanen, 1976; Duda, 2009). To dervie such bounds it suffices to realize that the mutual information of two r.v. is upperbounded by the entropy of each of those r.v.s., specifically, $\mathrm{I}[Z; X] = \mathrm{H}[Z] - \mathrm{H}[Z \mid X] \leq HZ$ and the bound is tight for Z that are deterministic transformations of X. So Eq. (55) becomes

$$\mathrm{I}[Z; X] = \mathrm{H}[Z] \leq \min_{q \in \mathcal{Q}} \mathrm{E}_{p(Z,X)}[-\log q(Z)] \tag{56}$$

This is the standard bound used in neural compressors (Ballé et al., 2017; Theis et al., 2017). We call this the variational entropy bottleneck. Note that achieving the rate can be done efficiently if Z is discrete by entropy entropy coding using the trained $q(Z)$. This advantage comes with two main downsides of Eq. (56):

- It is generally not true that any (for any δ) optimal rate can be achieved by a discrete and deterministic Z. For teh specific case of $\delta = 0$ it is the case, as we can simply set $Z = M(X)$.
- Eq. (56) is unfortunately not suitable for gradient based optimization w.r.t. to the encoder (due to the discreteness of Z) so we typically have to add noise during training (Ballé et al., 2017) which can cause a mismatch between training and testing (Agustsson & Theis, 2020).

Despite these issues we will mostly use the entropy bottleneck bound in experiments as we want our method to give rise to practical compressors.

D.2 Variational Distortion $\mathrm{R}[M(X) \mid Z]$

Let us now consider variational upper-bounds on the distortion $\mathrm{R}[M(X) \mid Z]$.

Direct Distortion. The obvious variational bound on the conditional entropy is the standard cross entropy loss as used for the distortion of VAE, VIB, and standard VC. Let $\mathcal{Q}'$ denote a family of regular conditional distributions (decoders), then:

$$\begin{aligned}
\mathrm{R}[M(X) \mid Z] &= \mathrm{H}[M(X) \mid Z] && \text{Lemma 3} \quad (57)\\
&= \mathrm{E}_{p(Z,X)}[-\log p(M(X) \mid Z)] && (58)
\end{aligned}$$

[12] See Flamich et al. (2020) or Schulman (2020) for an $\mathcal{O}(\exp(\mathrm{I}[Z; X]))$ algorithm. Bits-back coding (Wallace, 1990) can efficiently reach the desired bit-rate only because it is in the lossless setting.

$$= \min_{q' \in \mathcal{Q}'} \mathrm{E}_{p(Z,X)}\left[-\log \frac{p(M(X) \mid Z) q'(M(X) \mid Z) p(Z)}{q'(M(X) \mid Z) p(Z)}\right] \tag{59}$$

$$= \min_{q' \in \mathcal{Q}'} \mathrm{E}_{p(Z,X)}[-\log q'(M(X) \mid Z)]$$

$$- \mathrm{D}_{\mathrm{KL}}\left[p(M(X), Z) \| q'(M(X) \mid Z) p(Z)\right] \tag{60}$$

$$\leq \min_{q' \in \mathcal{Q}'} \mathrm{E}_{p(Z,X)}[-\log q'(M(X) \mid Z)] \tag{61}$$

We call this the variational direct distortion, as we directly try to predict / reconstruct $M(X)$ using a decoder $q'(M(X) \mid Z)$. The approximation gap here is $\min_{q' \in \mathcal{Q}'} \mathrm{D}_{\mathrm{KL}}\left[p(M(X), Z) \| q'(M(X) \mid Z) p(Z)\right]$. This direct distortion is simple and well understood but requires M which is usually not known and will thus have to be approximated. Furthermore, if $\mathcal{M}$ is in high dimension (such as when reconstructing unaugmented images) the trained decoder $q'(M(X) \mid Z)$ can often underfit.

Contrastive Distortion. We now consider a bound that does not directly require M, by considering a contrastive estimator. Suppose that for every X we can sample a sequence $\boldsymbol{X} := (X^+, X_1^-, \ldots, X_n^-)$ s.t. $X^+ \sim x$ (it is "postive") and each X_i^- are *not* ($X_i^- \not\sim x$, they are "negatives"). Then, we can use the InfoNCE (Oord et al., 2019) bound on information, that is standard in self-supervised learning. Let $\boldsymbol{Z}$, be the sequence of representations that are sampled by passing $\mathbf{x}$ through $p(Z \mid X)$. Let $\mathcal{F} := \{f : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}\}$ be a family of discriminators that scores how equivalent two representations are, then:

$$\mathrm{R}[M(X) \mid Z] = \mathrm{H}[M(X) \mid Z] \tag{62}$$

$$= \mathrm{H}[M(X)] - \mathrm{I}[M(X); Z] \tag{63}$$

$$\leq \mathrm{H}[M(X)] - \min_{f \in \mathcal{F}} \mathrm{E}_{p(X)p(Z,\boldsymbol{Z}|X)}[\mathrm{I}_{\mathrm{NCE}}] \qquad \text{InfoNCE} \tag{64}$$

$$= \mathrm{H}[M(X)] - \min_{f \in \mathcal{F}} \mathrm{E}_{p(X)p(Z,\boldsymbol{Z}|X)}\left[\log n + \log \frac{\exp f(Z^+, Z)}{\sum_{Z' \in \mathbf{Z}} \exp f(Z', Z)}\right] \qquad \text{Def.} \tag{65}$$

$$= \min_{f \in \mathcal{F}} \mathrm{E}_{p(X)p(Z,\boldsymbol{Z}|X)}\left[-\log \frac{\exp f(Z^+, Z)}{\sum_{Z' \in \mathbf{Z}} \exp f(Z', Z)}\right] + (const) \tag{66}$$

Eq. (64) uses the fact that InfoNCE is a valid lower bound on mutual information (Poole et al., 2019). The last equation removes constant w.r.t. Z and shows that we are only left with a log softmax term which essentially tries to classify which of all the sampled representations is positive. Note that the contrastive distortion has the advantage of not having to reconstruct high dimensional data (e.g. for images), but it suffers from bias in the case where the number of negatives n is small (Poole et al., 2019).

D.3 Practical Losses

In practice we will use parametrized neural network for the different variational families ($q_\varphi(X|Z), q_\theta(Z), q_\phi(X \mid Z)$) as well as for the encoder ($p_\varphi(X|Z)$). As discussed in the main text we usually assume access to samples from $M(X)$ and to a random generator of augmentation A.

The four previous bounds can be mixed and matched to provide four different practical losses to learn invariant compressors. As the mutual information bounds do not give rise to practical compressors, we focus below on the entropy bottleneck bound for the rate, which gives rise to the VIC and BINCE loss discussed in the main text.

- The **VIC** loss is recovered using the variational entropy bottleneck and the variational direct distortion bounds. The final criterion for training a neural network is summarized in Algorithm 1. As previously discussed the encoder is stochastic during training to enable backpropagation but deterministic at test time, for details see (Ballé et al., 2017).
- The **BINCE** loss is recovered using the variational entropy bottleneck and the variational contrative distortion bound. The final criterion for training a neural network is summarized in Algorithm 2. Similarly to VIC the encoder is stochastic during training but deterministic at test time.

Algorithm 1 Variational Invariant Compressor (VIC) Forward Pass Single Sample

```
Require: Encoder $p_\varphi(Z|X)$, Entropy Model $q_\theta(Z)$, Decoder $q_\phi(X|Z)$
Require: Dataset $\mathcal{D}$, Random augmentation generator $A$, $\beta$
$x \leftarrow \text{select}(\mathcal{D})$                                    ▷ Sample
$\tilde{x} \leftarrow A(x)$                                                   ▷ Augment
$z \leftarrow \text{sample}(p_\varphi(Z|\tilde{x}))$                          ▷ Encode
$\text{rate\_loss} \leftarrow -\log q_\theta(z)$                              ▷ Entropy Bottleneck
$\text{distortion\_loss} \leftarrow -\log q_\phi(x|z)$                        ▷ Direct Distortion
Return $\text{rate\_loss} + \beta \cdot \text{distortion\_loss}$
```

Algorithm 2 Bottleneck InfoNCE (BINCE) Forward Pass Single Sample

```
Require: Encoder $p_\varphi(Z|X)$, Entropy Model $q_\theta(Z)$, discriminator $f_\psi$,
Require: Dataset $\mathcal{D}$, Random augmentation generator $A$, $\beta$, number of negatives $n$
$x \leftarrow \text{select}(\mathcal{D})$                                    ▷ Sample
$\tilde{x} \leftarrow A(x)$                                                   ▷ Augment
$z \leftarrow \text{sample}(p_\varphi(Z|\tilde{x}))$                          ▷ Encode
$\text{rate\_loss} \leftarrow -\log q_\theta(z)$                              ▷ Entropy Bottleneck
$x^+ \leftarrow A(x)$                                                         ▷ Sample Positive
$\{x_i^-\}_{i=1}^n \leftarrow \text{select}(\mathcal{D})$ $n$ times           ▷ Sample Negatives
$\mathbf{x} \leftarrow [x^+, x_1^-, \ldots, x_n^-]$                           ▷ Concatenate
$\mathbf{z} \leftarrow \text{sample}(p_\varphi(Z|\mathbf{x}))$                ▷ Encode each $x$
$\text{softmax} \leftarrow \frac{\exp f(z^+, z)}{\sum_{z' \in \mathbf{z}} \exp f(z', z)}$
$\text{distortion\_loss} \leftarrow -\log(\text{softmax})$                    ▷ Contrastive Distortion
return $\text{rate\_loss} + \beta \cdot \text{distortion\_loss}$
```

Although the variational mutual information bound does not give rise to efficient coders, they can still be of interest for other tasks such as for representation learning. For example, VIC with a variational mutual information bound becomes a standard VAE where the input is augmented, but the target reconstructions $M(X)$ are not. This is also equivalent to a VIB where the task is to predict the maximal invariant $M(X)$, which to the best of our knowledge is the first self-supervised formulation of the information bottleneck. To train VIC or BINCE using the variational mutual information bound instead of the entropy bound it suffices to use stochastic encoders (typically neural networks predicting a mean and covariance of a mutlivariate Gaussian) and replacing $\mathrm{E}_{p(X)p(Z|X)}[-\log q(Z)]$ by $\mathrm{E}_{p(X)}[\mathrm{D}_{\mathrm{KL}}[p(Z|X)]\, q(Z)]$ in the "entropy bottleneck" step of both algorithms.

E Extended Previous Work

Lossy image compression The connection between the rate-distortion objective and variational inference is well known (MacKay, 1992; **?**). Ballé et al. (2017) and Theis et al. (2017) first implemented large scale neural image compression based on Variational Autoencoders (VAEs). Later approaches improved upon the original model by adding a hierarchical layers and autoregressive components (Ballé et al., 2018; Minnen et al., 2018; Lee et al., 2019; Johnston et al., 2018). Further recent studies improve the compressor by improving the inference procedure of these models (Yang et al., 2020). Another aspect, is computational complexity (Johnston et al., 2019).

Task-Centric Compression At the core of our work is the idea that compression should be task-centric, i.e., optimized for a set of tasks of interest. To the best of our knowledge, the only theoretical framework that considers this question is Tishby et al.'s (2000) Information Bottleneck (IB), which was derived using rate distortion theory for the specific distortion $\mathrm{H}[Y \mid Z]$. IB was nevertheless never (to our knowledge) used for compression/The lack of use of IB in compression probably comes from its requirement of (1) a single task; (2) knowing the labels at compression time. The first problem could easily be overcome by using a "multi-task" distortion $\mathrm{H}[\mathcal{T} \mid Z]$. The second point is the main drawback of IB for compression. Indeed, it would be easier to directly compress the labels if we had access to them. Our paper can be seen as a self-supervised extension of IB that is useful for compression as it does not require the labels at compression time.

From a practical perspective, there have recent task-centric compression methods. Mentzer et al. (2020) shows that they can significantly decrease the bit-rates by using a generative adversarial network (Goodfellow et al., 2014). This can be seen as a task-specific compression where the task is to ensure that a discriminator cannot discriminate between a highly compressed reconstruction and the original image. More related is Singh et al.'s (2020) work on end-to-end compression of pre-trained features for transfer learning. From a very practical perspective, their method is similar to our idea of compressing pretrained self-supervised features. Their work does not provide any theory justification, and constrained to cases where downstream tasks to similar the task on which the featurizer was pre-trained. Nevertheless the intuitive idea is similar and shows that there is interest in taking advantage of recent advancements and open sourcing of pretrained models to perform task specific compression.

F Reproducibility

In this section we provide further details of the hyperparameters chosen for the various experiments in the main text. Unless stated otherwise, all the models are trained for 200 epochs, using Adam (**?**) as the optimizer, a learning rate of $1e-3$, a batch-size of 128. We checkpoint and use the model which achieves the smallest *validation* loss for evaluation. Results are averaged over 5 random seeds, and standard errors are reported. For all convolutional layers we use Kaiming normal initialization(He et al., 2015), for all linear layers we use Kaiming uniform initialization(He et al., 2015), while all biases are always initialized at 0. Activation functions are ReLUs while other unspecified parameters are PyTorch (Paszke et al., 2019) defaults. Throughout this section, instead of optimizing $\mathrm{I}[Z;X] + \beta\,\mathrm{D}_{\sim}[X,Z]$ we optimize $\lambda\,\mathrm{I}[Z;X] + \mathrm{D}_{\sim}[X,Z]$, which is a more standard formulation for VIB, VAE, VC.

F.1 Banana Experiments

For the Banana dataset most of the arguments were selected so as to replicate Fig.1.B. from (Ballé et al., 2020). [13]

The data distribution is obtained by starting from a bivariate Gaussian $X \sim \mathcal{N}(X; \mathbf{0}, \mathrm{diag}([3, 0.5]))$. It is then transformed to a banana distribution using the following transformation: $x_2 = x_2 + 0.1x_1^2 - 9$. We then rotate it and shift it: $\mathbf{X} = (\mathrm{Rot}(-40) \cdot \mathbf{X}) + [-3, -4]^T$. For every epoch we resample 1024000 new points, i.e., examples are never seen twice during training).

For all Banana experiments we use a 2 dimensional representation $Z \in \mathbb{R}^2$, and a batch size of 8192. The encoders (and decoders if there is one) is always a 2-hidden layer MLP with 1024 hidden neurons, batch norm (Ioffe & Szegedy, 2015), and softplus activation. The learning rate scheduling consists in decreasing the learning rate by a factor 10 at epochs 50,75,87, and 120. We train both a standard variational compressor (VC) and our variational invariant compressor (VIC), in both cases we use the factorized prior entropy model from (Ballé et al., 2018). To obtain RD curves we sweep over $\lambda = 0.00001, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000$.

For Fig. 2 we use $\lambda = 0.1$ for both plots. For the classical compression we use a variational upper bound on $\mathrm{H}[X \mid Z]$, namely $\min_{q' \in \mathcal{Q}'} \mathrm{E}_{p(Z,X)}[-\log q'(X \mid Z)]$. Essentially we are trying to reconstruct the input X using an MLP $\mathcal{Q}'$. For our compression we use the variational distortion, namely $\min_{q' \in \mathcal{Q}'} \mathrm{E}_{p(Z,X)}[-\log q'(M(X) \mid Z)]$. Where $M(X)$ is a representative of each equivalence class, which we select to be $M : x \mapsto \|x\|_2 \cdot [-0.7071, -0.7071]^T$, i.e., the point with the same radius but at 225 degrees. Essentially we are trying to reconstruct a representative of the equivalence clas of X using the same MLP $\mathcal{Q}'$.

[13] Their code can be found at `https://github.com/tensorflow/compression/blob/master/models/toy_sources/toy_sources.ipynb`

F.2 MNIST Experiments

For our MNIST (LeCun et al., 1998) experiments we compare again our VIC (as described in Algorithm 1) against a standard VC criterion (same as in Algorithm 1 but the direct distortion is evaluated at $\tilde{x}$ instead of x).

Each image first passes through a ResNet18 (He et al., 2016) encoder, which maps it to a 128 dimensional representation $Z \in \mathbb{R}^{128}$. We then pass Z through an entropy bottleneck with a scaled hyperprior entropy model from (Ballé et al., 2018) which gives us the reconstruction $\hat{Z}$. The reconstructed $\hat{Z}$ is finally passed through a 5-layer transposed CNN decoder, which reconstructs the augmented input (in the standard case) or the non-augmented input (in the invariant case).

Once the compressor is trained we freeze it, apply it to the dataset and train a new ResNet18 to classify the digits using the reconstructions. This thus simulated how well you could perform downstream tasks, by considering one possible task, namely, classifying the digits. This classifier is trained using SGD for 100 epochs, an initial learning rate 0.1, and a scheduler that decreases the learning rate every 20 epochs. To obtain RD curves we sweep over $\lambda = 0.001, 0.01, 0.03, 0.1, 0.3, 1, 3, 10, 100$.

We consider three settings of data augmentations.

Highly augmented test and train First we consider a highly augmented MNIST datasets, which is augmented both at train and test time. Specifically, we apply random rotations sampled from $[-45, 45]$ degrees, random translations between $[0, \frac{1}{4}]$ percentage of pixels, random shearing between $[0, 25]$ degrees, and random scaling by a factor in $[0.6, 1.4]$.

Mildly augmented test and train We consider more mild data augmentations applied both at train and test time. Specifically, we apply random rotations sampled from $[-30, 30]$ degrees, random translations between $[0, \frac{1}{10}]$ percentage of pixels, random shearing between $[0, 10]$ degrees, and random scaling by a factor in $[0.8, 1.2]$.

Mildly augmented train Finally we consider the same mildly augmented data described above but only applied at training time

F.3 ImageNet Experiments

For ImageNet (Deng et al., 2009) experiments, we downloaded a pretrained SimCLR model with a ResNet50 architecture, with a 2048 representations $\mathcal{Z} = \mathbb{R}^{2048}$. We used this SimCLR to encode a 256×256 ImagNet, and zipped the resulting features to get the average bit-rate.

For our preliminary result on BINCE, we start with the pretrained SimCLR, and add an entropy bottleneck with Ballé et al.'s (2018) factorized prior model. We then finetune both losses in an end-to-end fashion for 3 epochs with Adam optimizer and a learning rate of 1e-6. During finetuning we also train a single layer (1024 hidden neurons) MLP to classify ImageNet labels from the reconstrcuted representations.

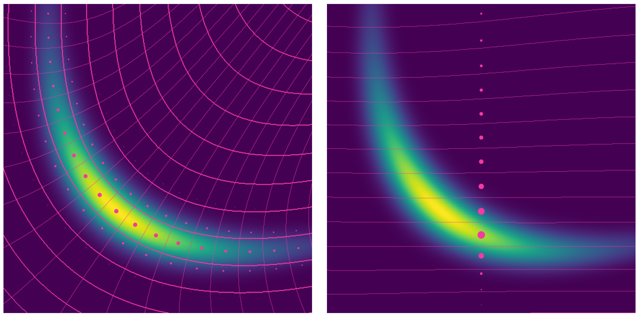
G Additional Results

G.1 Banana

G.1.1 Different Invariances

Fig. 2 compares a classical compressor to our invariant distortion in the case of rotation invariant tasks. The standard compressor achieves a rate 5.42 ± 0.00 bits for an invariance distortion of $\mathrm{D}_{\sim}[X, Z] = 7.04\text{e-}2 \pm 0.19\text{e-}2$. While our compressor achieves a rate 2.54 ± 0.00 bits for an invariance distortion of $\mathrm{D}_{\sim}[X, Z] = 5.25\text{e-}2 \pm 0.08\text{e-}2$.Here we show the same experiment for other invariances.

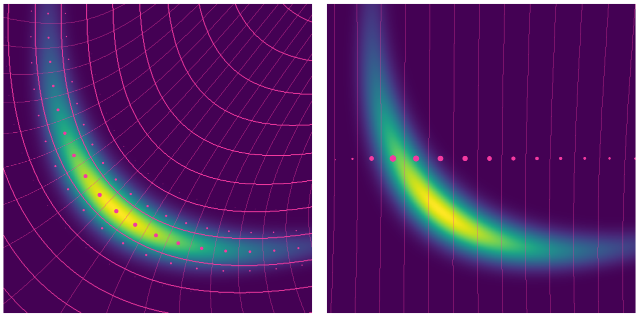
Fig. 4 considers the case where downstream tasks are invariant to translations on the x axis. The maximal invariant used during training is chosen to be $M : x \mapsto [0, x_2]^T$. We only ran a single run for visualization. We see that our model can essentially perform as well on all downstream tasks (similar invariant distortion) for only 60% of the bit-rate. Unsurprisingly we see that the codebook is



(a) Standard Compression. (b) x-Translation Compression.

Figure 4: Similar to Fig.1. for the case of downstream tasks that are invariant to translation on the x-axis. (left) standard compression with a rate of 5.4 bits and an invariant distortion of 5.59e-2 ; (right) our compression with a rate of 3.19 bits and an invariant distortion of 5.35e-2.

in shape of horizontal stripes as these can cover the entire distribution with a few codes (small bit rate) while incurring a small invariance distortion (which only depends on the y value).



(a) Standard Compression. (b) y-Translation Compression.

Figure 5: Similar to Fig.1. for the case of downstream tasks that are invariant to translation on the x-axis. (left) standard compression with a rate of 5.55 bits and an invariant distortion of 5.25e-2 ; (right) our compression with a rate of 3.53 bits and an invariant distortion of 5.48e-2.

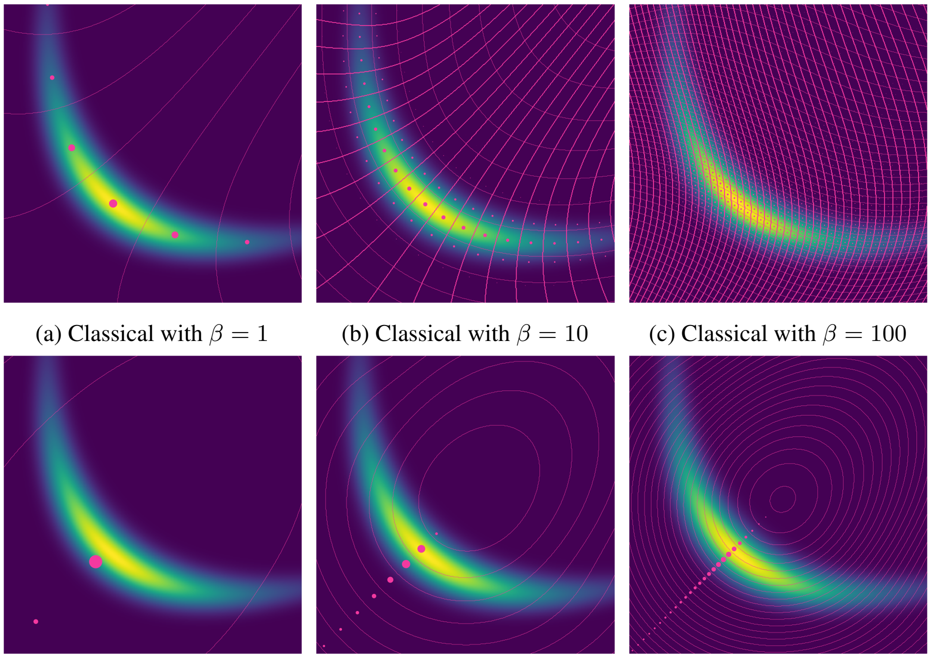
Similar results can be seen in Fig. 5 in the case of downstream tasks that are invariant to translations on the y axis.

G.1.2 Rate-Distortion Curves

By sweeping over the hyperparameter β we can can move along the rate distortion curve, thus balancing the respective rate and distortion term. Specifically, by increasing β in Eq. (5) we give more importance to the distortion term, and thus will have to increase the size of the codebook. This is exactly what can be seen in Fig. 6.

To get a better sense of the gains in rates that can be achieved we show in Fig. 7 the (average) rate-distortion curve for our invariant neural compressor and the non-invariant compressor. We see that our invariant compressor significantly outperforms the non invariant compressor. Indeed, the area-under-the RD curve for the invariant model is 35.8 ± 4.2 bits while it is 48.1 ± 0.3 bits for the the non invariant case, this means that, if you are interested in rotation invariant tasks, you can increase your compression rates by an average of 12.3 bits without hindering your downstream performance.

Note that our Thm. 1 shows that the rate and the distortion have a linear relationship $rate(\delta) = (const) - \delta$, yet the rate distortion curve in Fig. 7 shows an approximately logarithmic relationship (notice the log-scale in the x axis). This happens because we are estimating the invariant distortion using the mean squared error (MSE) or conditional variance, and the conditional entropy is proportional to the logarithm of the conditional variance $\mathrm{H}[M(X) \mid Z] \propto \log(\mathrm{Var}[M(X) \mid Z]) + (const)$.



(a) Classical with $\beta = 1$ (b) Classical with $\beta = 10$ (c) Classical with $\beta = 100$

(d) Ours with $\beta = 1$ (e) Ours with $\beta = 10$ (f) Ours with $\beta = 100$

Figure 6: Similar to Fig.1. but for different values of β.

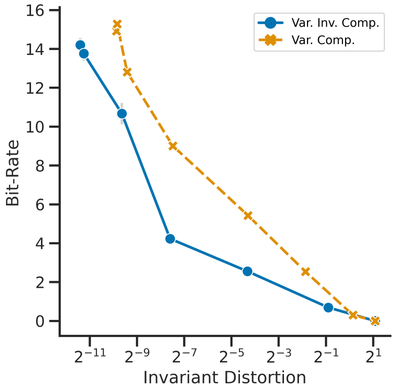
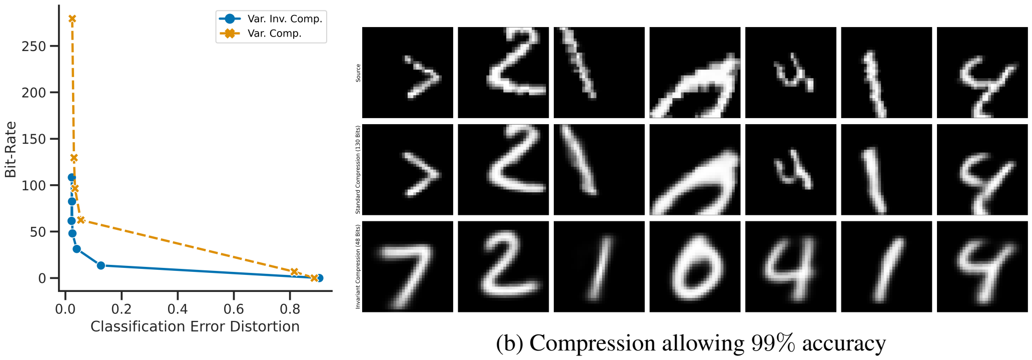


Figure 7: Rate-Distortion curve for a neural compressor and our invariant neural compressor, for compressing a banana distributed source for rotation invariant downstream tasks. We see that the invariant compressor significantly outperforms the non-invariant compressor, especially in low distortion regimes. RD curves are generated by sweeping over β, the plotted curve is the average over 5 runs, standard errors for each β are shown in gray (on the x and y axis).

G.2 MNIST

Highly Augmented Test and Train. As discussed in the main paper, we are interested in knowing the gains in bit-rate that can be achieved by our practical loss VIC (compared to standard VC) in the case where you have access to some augmentations w.r.t. which your tasks of interest are invariant to. To simulate this we apply large augmentations at training and test time. The results are summarized in Fig. 8b.

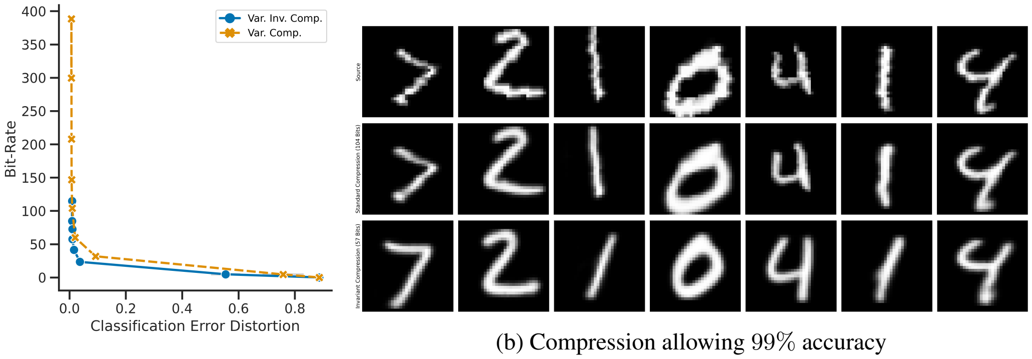
Mildly Augmented Test and Train. We then consider the same experiment but with less data augmentation. The results are summarized in Fig. 9b. Unsurprisingly, the gains in bit rate decrease compared to the highly augmented case.



(b) Compression allowing 99% accuracy

(a) Rate-Error curve

Figure 8: Compression of a highly augmented MNIST dataset by our invariant compressor (Inv. VAE) and a standard compressor (VAE). Left) The rate-error curve for MNIST classification, the area under the curve is 6.2 ± 4.0 for the invariant case and 12.0 ± 10.0 for the non invariant case. Right) Reconstructions for the a non-invariant (second row) and invariant compressor (last row) that retains enough information for a downstream ResNet18 to classify the highly augmented MNIST with 99% test accuracy. Our model only requires a a bit-rate of 48.1 ± 0.5 bits compared to 129.6 ± 0.5 bits in the non invariant case. All quantitative results are averaged over 5 runs.

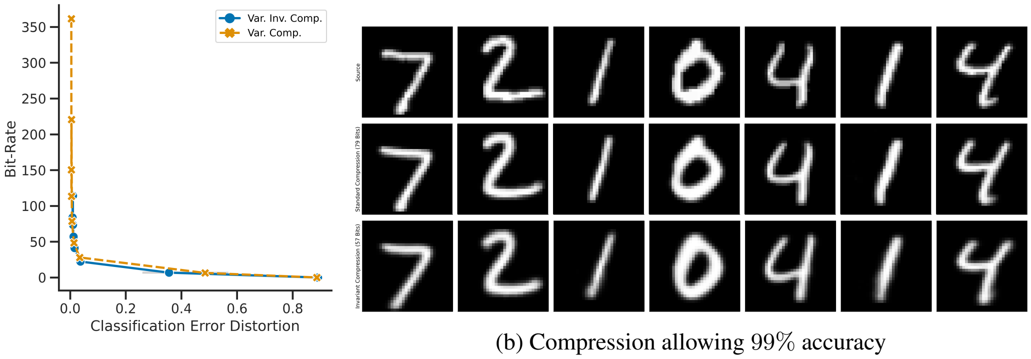


(b) Compression allowing 99% accuracy

(a) Rate-Error curve

Figure 9: Compression of an augmented MNIST dataset by our invariant compressor (Inv. VAE) and a standard compressor (VAE). Left) The rate-error curve for MNIST classification, the area under the curve is 9.5 ± 0.1 for the invariant case and 17.1 ± 0.1 for the non invariant case. Right) Reconstructions for the a non-invariant (second row) and invariant compressor (last row) that retains enough information for a downstream ResNet18 to classify the augmented MNIST with 99% test accuracy. Our model only requires a a bit-rate of 57.4 ± 0.2 bits compared to 103.9 ± 0.5 bits in the non invariant case. All quantitative results are averaged over 5 runs.

Mildly Augmented Test and Train. Finally, we consider the case where the augmentations are not known, and so we apply augmentations at training time w.r.t. which you might not be invariant to. We thus do not apply the same augmentations at test time, and can only hope that MNIST has similar inherent invariances. The results are summarized in Fig. 10b. In this case the gains in bit-rate is much smaller but still significant. Importantly we see that the rate for VC is highly dependent on whether the images are augmented at test time or not, while the it is not for VIC. This makes sense as VIC essentially learned to compress the maximal invariant $M(X)$ which is the same regardless as to whether the input is augmented or not.



(b) Compression allowing 99% accuracy

(a) Rate-Error curve

Figure 10: Compression of a (non augmented) MNIST dataset by our invariant compressor (Inv. VAE) and a standard compressor (VAE). The invariant compressor was trained with the augmented MNIST but tested on the non augmented MNIST. Left) The rate-error curve for MNIST classification, the area under the curve is 8.0 ± 0.7 for the invariant case and 10.3 ± 0.5 for the non invariant case. Right) Reconstructions for the a non-invariant (second row) and invariant compressor (last row) that retains enough information for a downstream ResNet18 to classify MNIST with 99% test accuracy. Our model only requires a a bit-rate of 57.2 ± 0.2 bits compared to 78.4 ± 0.4 bits in the non invariant case. All quantitative results are averaged over 5 runs.