Towards Empirical Sandwich Bounds on the Rate-Distortion Function

Yibo Yang and **Stephan Mandt**

University of California, Irvine

Hasso Plattner Institut
Digital Engineering · Universität Potsdam

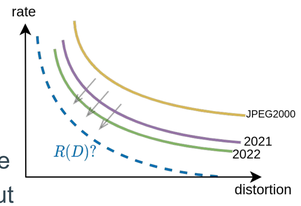
paper: https://arxiv.org/abs/2111.12166
code & data: https://github.com/mandt-lab/RD-sandwich

Summary:

- What the rate-distortion (R-D) function is to lossy data compression = what Shannon entropy is to lossless data compression.
- Establishing the R-D function has been a hard problem in info theory.
- We develop ML methods to estimate *sandwich* bounds on R-D functions:
  - Can handle general (discrete, continuous, etc.), high-dim data;
  - Work by training generative models (e.g. VAEs) on i.i.d. data samples.
- We estimate R-D sandwich bounds on a variety of real-world data (particle physics, speech, images), and assess optimality of SOTA (neural / traditional) lossy image compression algorithms.

Motivation: how far are we from info-theoretic limits?

- ML has made great strides in improving lossy data compression performance.
- However, any lossy compression algorithm must face the rate ("avg file size") and distortion ("loss of quality") tradeoff.
- The R-D function of the data determines the best R-D tradeoff we can possibly attain, but it is mostly unknown for real-world data.



Background: lossy compression and $R(D)$

X —> 010110 —> Y

enc *dec*

X is the data **source** r.v., following distribution P_X

Y is the **reproduction** r.v.

Limits of compression – lossless (H[X]) v.s. lossy (R(D)):

	Requirement	*Objective*	*Fundamental Limit*
lossless compression:	$X = Y$	minimize E[|*enc*(X)|]	H$[X]$:= E[-log $P(X)$]
lossy compression:	E[$\rho(X, Y)$] ≤ D	minimize E[|*enc*(X)|]	$R(D)$

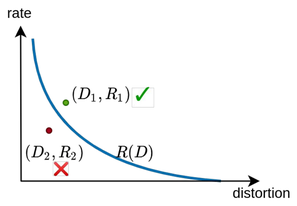
distortion function ρ: $\mathcal{X} \times \mathcal{Y} \to [0, \infty)$

$$R(D) := \inf_{Q_{Y|X}: \mathbb{E}[\rho(X,Y)] \le D} I(X;Y)$$

where

$$\mathbb{E}[\rho(X,Y)] := \int_{\mathcal{X}\times\mathcal{Y}} \rho(x,y)\ \mathrm{d}P_X Q_{Y|X}(x,y)$$

$$I(X;Y) := KL(P_X Q_{Y|X} \| P_X Q_Y)$$



$R(D)$ is "the **minimum** number of bits (per sample) needed, by **any** algorithm, to transmit data samples from P_X with an average distortion not exceeding D".

Prior state-of-the-art method for R(D) estimation

The Blahut-Arimoto (BA) algorithm [Blahut 1972; Arimoto 1972]:
Solve the unconstrained problem by coordinate-descent on the variational Lagrangian:

$$\min_{Q_{Y|X}, Q_Y} \mathcal{L}(Q_{Y|X}, Q_Y, \lambda) := \underbrace{\mathbb{E}_{x\sim P_X}[KL(Q_{Y|X=x}\|Q_Y)]}_{:= \boldsymbol{\mathcal{R}} \ge I(X;Y)} + \lambda \underbrace{\mathbb{E}[\rho(X,Y)]}_{:= \boldsymbol{\mathcal{D}}}$$

✓ Converges to the global minimum; the associated point $(\boldsymbol{\mathcal{D}}, \boldsymbol{\mathcal{R}})$ converges to a point on $R(D)$ from above.

✘ Only works when the data (and reproductions) are finite, and the source distribution is known.

✘ Otherwise, we will need to discretize and/or estimate source probabilities by a histogram (runs into the curse of dimensionality)).

Proposed: R(D) upper bound via β-VAEs

- Basic idea: keep the objective of the BA algorithm, but do *(stochastic) gradient descent* instead of coordinate descent.
- Parameterize the variational distributions with neural nets (e.g., flows).
- Can be reduced to fitting a β-VAE, with $p(x|z) \propto e^{-\rho(x,\omega(z))}$

NELBO =

$$\underbrace{\mathbb{E}_{x\sim P_X}[KL(Q_{Z|X=x}\|Q_Z)]}_{:= \boldsymbol{\mathcal{R}} \ge I(X;Z)} + \lambda \underbrace{\mathbb{E}_{P_X}[\rho(X, \omega(Z))]}_{:= \boldsymbol{\mathcal{D}}} + \text{const}$$

- Claim (*Theorem A.3*): the point $(\boldsymbol{\mathcal{D}}, \boldsymbol{\mathcal{R}})$ lies on an upper bound of the data R(D); the upper bound becomes *tight* in the infinite capacity limit, as long as the decoder is bijective (sufficient condition).

✓ Works for cont. or discrete (or neither) data; only need i.i.d. samples.

✓ Converges to a local minimum, yields only an *upper bound* on R(D).

Proposed: R(D) lower bound via Lagrange dual

R(D) is also the solution of constrained maximization over a family of functions [Csiszár 1974]:

$$R(D) = \max_{g, \lambda \ge 0} \{\mathbb{E}[-\log g(X)] - \lambda D\}$$

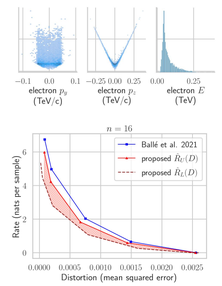
$$\text{subj to } \mathbb{E}\left[\frac{\exp(-\lambda\rho(X,y))}{g(X)}\right] = \int \frac{\exp(-\lambda\rho(x,y))}{g(x)}\,\mathrm{d}P_X(x) \le 1, \forall y \in \mathcal{Y}$$

Gist of the proposed lower bound method:

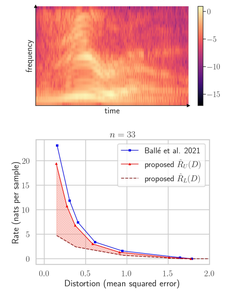
- Make the optimization problem unconstrained by reparameterizing $g(x)$.
- An IWAE-style estimator is proposed to obtain a tractable lower bound.
- *Represent $g(x)$ by a neural network*; train it by *(stochastic) grad ascent*.
- A few technical approximations were involved; see paper for details.

Results

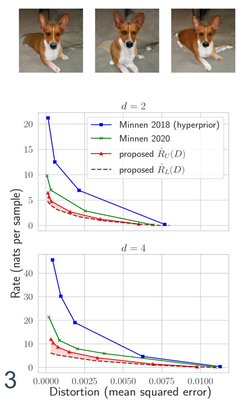
Particle physics



Speech spectrograms



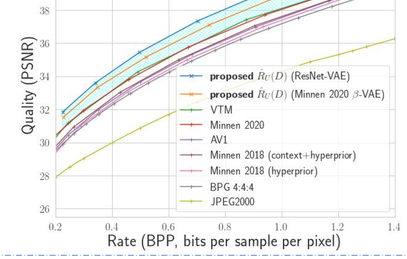
GAN generated images



- Proposed sandwiched region colored in red. Neural compression methods in blue/green.
- Blahut-Arimoto no longer feasible in more than 3 dimensions; results (when feasible) agree with ours.
- No known algorithm has been applied at the scale of these problems.

High-resolution image compression, compared to SOTA methods

- Estimate R-D upper bound of natural images by fitting hierarchical β-VAEs.
- Resulting bound (blue) suggests *theoretical room for improving SOTA traditional / neural compression performance by ~ 1 dB PSNR*.



Main takeaways

- ML is starting to revolutionize data compression, yet
- we don't have a good theoretical understanding of neural compression.
- ML can help bridge the gap between information theory and practice.

- Fitting a suitable β-VAE to your data naturally yields an **upper bound** on the data R(D), analogous to in lossless compression (bits-back coding).
- More expressive VAE ≈ tighter NELBO = tighter upper bound on R(D).

- R(D) **lower bound** has been "notoriously hard to obtain" [Riegler et al., 2018], but is more useful for assessing optimality of compression algorithms.
- The proposed LB algorithm seems to have sample complexity that is exponential in the *intrinsic* dimension of the data.
- More work is needed to understand if the difficulties are fundamental.

References

[Blahut 1972]. R. Blahut. "Computation of channel capacity and rate-distortion functions". *IEEE Transactions on Information Theory*, 18(4):460–473, 1972. doi: 10.1109/TIT.1972.1054855.
[Arimoto 1972] S. Arimoto. "An algorithm for computing the capacity of arbitrary discrete memoryless channels". *IEEE Transactions on Information Theory*, 18(1):14–20, 1972.
[Csiszár 1974] I. Csiszár, "On an extremum problem of information theory," *Studia Scientiarum Mathematicarum Hungarica*, vol. 9, no. 1, pp. 57–71, Jan. 1974.
[Riegler et al., 2018] Erwin Riegler, Gunther Koliander, and Helmut Bolcskei. "Rate-distortion theory for general sets and measures". *arXiv preprint arXiv:1804.08980*, 2018.