Rate-Distortion via Energy-Based Models

Qing Li*, Yongjune Kim†, and Cyril Guyot*

*Western Digital Research
Milpitas, CA, 95035
cyril.guyot@wdc.com

†Pohang University of Science and Technology
Pohang, Gyeongbuk, 37673, South Korea
yongjune@postech.ac.kr

Abstract

In this work, we estimate rate-distortion via energy-based models (EBMs). We begin by providing a framework for estimating rate-distortion with neural networks. We then instantiate the framework with EBMs and provide Discriminative-Blahut-Arimoto. Our empirical results show that our estimates agree with closed-form expressions and known bounds.

1 Introduction

Our main goal is to estimate rate-distortion via energy-based models (EBMs). Source coding [1] is a technique that represents a source with fewer bits and less-than-perfect fidelity. Rate-distortion presents the theoretical limits of source coding. It is important to compute rate-distortion and find rate-distortion approaching posteriors. This is because they provide insights to help design good source codes. Classical numerical algorithms such as Blahut-Arimoto (BA) [2, 3] efficiently calculate rate-distortion when sources are independent and identically distributed.

EBMs have a long history in physics, statistics, and machine learning [4]. EBMs define Boltzmann distributions, which include rate-distortions approaching posteriors, so EBMs can be used to represent rate-distortions approaching posteriors and to estimate rate-distortions.

In this paper, we show how to estimate rate-distortion with EBMs. We provide a paradigm for using neural networks to estimate rate-distortion. The framework is then instantiated using EBMs, and Discriminative-Blahut-Arimoto is provided. Our empirical estimates agree with closed-form expressions and known bounds.

2 Background

2.1 Rate-Distortion

Given a source distribution $p(\mathbf{y})$ and a distortion constraint $d \in \mathbb{R}^+$ associated with a distortion metric $\rho(\cdot)$, *rate-distortion* is defined as

$$R(d) := \min_{\{p(\mathbf{x}), p(\mathbf{x}|\mathbf{y}) : \mathbb{E}[\rho(\mathbf{x},\mathbf{y})] \leq d\}} I(\mathbf{x};\mathbf{y}), \tag{1}$$

where $I(\mathbf{x};\mathbf{y})$ denotes the mutual information and $\mathbb{E}[\cdot]$ is the expectation operator.

Let denote

$$\mathcal{L}_{RD}(p(\mathbf{x}), p(\mathbf{x}|\mathbf{y})) := I(\mathbf{x};\mathbf{y}) + \beta \mathbb{E}[\rho(\mathbf{x},\mathbf{y})], \tag{2}$$

where β controls the trade-off between rate ($I(\mathbf{x};\mathbf{y})$) and distortion ($\mathbb{E}[\rho(\mathbf{x},\mathbf{y})]$).

Let denote *optimized* $p(\mathbf{x}|\mathbf{y})$ and $p(\mathbf{x})$ achieving $R(d)$ by $p^*_{RD}(\mathbf{x}|\mathbf{y})$ and $p^*_{RD}(\mathbf{x})$. That is,

$$p^*_{RD}(\mathbf{x}|\mathbf{y}) := \underset{\{p(\mathbf{x}|\mathbf{y}):\mathbf{y},\mathbf{x}\ \sim p(\mathbf{y})p(\mathbf{x}|\mathbf{y}),\mathbb{E}[\rho(\mathbf{x},\mathbf{y})]\leq d\}}{\arg\min} I(\mathbf{x};\mathbf{y}), p^*_{RD}(\mathbf{x}) = \int p(\mathbf{y})p^*_{RD}(\mathbf{x}|\mathbf{y})d\mathbf{y}. \quad (3)$$

$p^*_{RD}(\mathbf{x}|\mathbf{y})$ and $p^*_{RD}(\mathbf{x})$ are characterized by [1, chapter 10, pp. 330]:

$$p^*_{RD}(\mathbf{x}|\mathbf{y}) = \frac{1}{Z_\beta(\mathbf{y})}p^*_{RD}(\mathbf{x})\exp[-\beta\rho(\mathbf{y},\mathbf{x})], Z_{\beta,RD}(\mathbf{y}) := \int p^*_{RD}(\mathbf{x})\exp[-\beta\rho(\mathbf{x},\mathbf{y})]d\mathbf{x}. \quad (4)$$

2.2 Energy-Based Models

Let $E_\phi(\mathbf{x}) \in \mathbb{R}^+$ be the energy function represented by a neural network ϕ given data $\mathbf{x}$. An energy-based model (EBM) [4] defines Boltzmann distributions:

$$p_\phi(\mathbf{x}) := \frac{\exp[-E_\phi(\mathbf{x})]}{Z_\phi}, \quad (5)$$

where $Z_\phi := \int E_\phi(\mathbf{x})d\mathbf{x}$ denotes the partition function.

Langevin dynamics (LD) describes a sampling approach from $p_\phi(\mathbf{x})$ using $\nabla_\mathbf{x}\log p_\phi(\mathbf{x})$. Specifically, given a step size $\lambda > 0$, a total number of iterations K, and an initial prior $\mathbf{x}_0 \sim \pi(\mathbf{x})$, it iterates the following

$$\mathbf{x}_i := \mathbf{x}_{i-1} - \lambda\nabla_{\mathbf{x}_{i-1}}E_\phi(\mathbf{x}_{i-1}) + \sqrt{2\lambda}\mathbf{z}_i, \quad \mathbf{z}_i \sim \mathcal{N}(0,\mathbf{I}). \quad (6)$$

Under some regularity criteria, the distribution of $\{\mathbf{x}_K\}$ will be close to $p_\phi(\mathbf{x})$ when λ is sufficiently small and K is sufficiently large [5, 6].

3 Minimax Game of Rate-Distortion

3.1 Rate-Distortion-Generative-Network

We first formulate (2) as a minimax game. To do so, we relax $I(\mathbf{x};\mathbf{y})$ by using the variational lower bound given by Nguyen et al. [7, Equation 8]. That is,

$$I(\mathbf{x};\mathbf{y}) \geq \sup_{D\in\mathcal{D}} \mathbb{E}_{\mathbf{x},\mathbf{y}\sim p(\mathbf{x},\mathbf{y})}[D(\mathbf{x},\mathbf{y})] - \mathbb{E}_{\mathbf{x},\mathbf{y}\sim p(\mathbf{x})p(\mathbf{y})}[\exp(D(\mathbf{x},\mathbf{y})-1)], \quad (7)$$

where $\mathcal{D}$ is a function class $D : (\mathbf{x},\mathbf{y}) \to \mathbb{R}^+$ [1].

Let us denote

$$\mathcal{L}_{MI}(D) := \mathbb{E}_{\mathbf{x},\mathbf{y}\sim p(\mathbf{x},\mathbf{y})}[D(\mathbf{x},\mathbf{y})] - \mathbb{E}_{\mathbf{x},\mathbf{y}\sim p(\mathbf{x})p(\mathbf{y})}[\exp(D(\mathbf{x},\mathbf{y})-1)], \quad (8)$$

[1] Based on [7, Equation 8], $I(\mathbf{x};\mathbf{y}) \geq \sup_{F\in\mathcal{F}}\mathbb{E}_{\mathbf{x},\mathbf{y}\sim p(\mathbf{x},\mathbf{y})}[\log F(\mathbf{x},\mathbf{y})] - \mathbb{E}_{\mathbf{x},\mathbf{y}\sim p(\mathbf{x})p(\mathbf{y})}[F(\mathbf{x},\mathbf{y})] + 1$, where $\mathcal{F}$ is a class of functions $F : (\mathbf{x},\mathbf{y}) \to \mathbb{R}^+$. (7) is derived by setting $\log F := D - 1$.

and

$$\begin{aligned}\mathcal{L}'_{RD}[p(\mathbf{x}), p(\mathbf{x}|\mathbf{y}), D] &:= \mathbb{E}_{\mathbf{x},\mathbf{y}\sim p(\mathbf{x},\mathbf{y})}[D(\mathbf{x},\mathbf{y})] - \mathbb{E}_{\mathbf{x},\mathbf{y}\sim p(\mathbf{x})p(\mathbf{y})}[\exp(D(\mathbf{x},\mathbf{y}) - 1)] \\ &+ \beta\mathbb{E}[\rho(\mathbf{x},\mathbf{y})]. \end{aligned} \tag{9}$$

As a result, (2) is formulated as the following minimax game:

$$\min_{p(\mathbf{x}),p(\mathbf{x}|\mathbf{y})} \max_{D} \mathcal{L}'_{RD}[p(\mathbf{x}), p(\mathbf{x}|\mathbf{y}), D], \tag{10}$$

where min over $p(\mathbf{x}), p(\mathbf{x}|\mathbf{y})$ is due to the definition of rate-distortion, and max over D is to maximize lower bound of $I(\mathbf{x};\mathbf{y})$, i.e., (7).

Due to the fact that deep neural networks are universal approximators [8], we introduce three neural networks to model $p(\mathbf{x}|\mathbf{y})$, $p(\mathbf{x})$, and D: an encoder neural network parameterized by θ to model $p(\mathbf{x}|\mathbf{y})$, a generator neural network parameterized by ϕ to model $p(\mathbf{x})$, and one discriminative network parameterized by ω to model D. As a result, (10) is:

$$\min_{\phi,\theta} \max_{\omega} \mathcal{L}'_{RD}(\phi,\theta,\omega). \tag{11}$$

That is, 1) ϕ, θ, and ω constitute a *minimax game* with the objective (11): ω is trained to maximize $\mathcal{L}_{MI}(\omega)$ with fixed ϕ and θ, and while ϕ and θ are trained to minimize $\mathcal{L}'_{RD}(\phi,\theta,\omega)$ with fixed ω; and 2) ϕ, θ, and ω define a generative model and a conditional generative model, and we call it Rate-Distortion-Generative-Network (RD-GEN). Fig. 1a summarizes this.

3.1.1 Special case: $\beta = \infty$

When $\beta = \infty$, $p^*_{RD}(\mathbf{x}|\mathbf{y}) = \mathbb{1}_{\mathbf{y}=\mathbf{x}}$ (i.e., $p^*_{RD}(\mathbf{x}|\mathbf{y}) = 1$ if $\mathbf{y} = \mathbf{x}$ otherwise 0), $d = 0$, $p^*_{RD}(\mathbf{x})$ and $p(\mathbf{y})$ are identical, and $Z_{\beta,RD}(\mathbf{y}) = \ln p(\mathbf{y})$ based on (4). Furthermore, as $p^*_{RD}(\mathbf{x}|\mathbf{y}) = \mathbb{1}_{\mathbf{y}=\mathbf{x}}$, the encoder is an identity function, thus skipped for optimization, and (9) degenerates to the objective of f-GAN [9]. Fig. 1b summarizes this.

3.2 Optimal discriminator

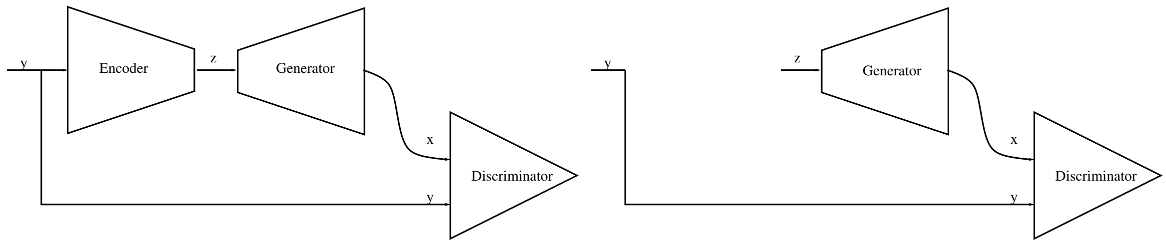
Lemma 1. For given $\mathbf{x}$ and $\mathbf{y}$, the optimal discriminator $D^*(\mathbf{x},\mathbf{y})$ according to (9), i.e., $D^*(\mathbf{x},\mathbf{y}) = \arg\max_{\omega} \mathcal{L}'_{RD}[p^*_{RD}(\mathbf{x}|\mathbf{y}), p^*_{RD}(\mathbf{x}), \omega]$, is given by

$$D^*(\mathbf{x},\mathbf{y}) = 1 - \beta\rho(\mathbf{x},\mathbf{y}) - \ln Z_{\beta,RD}(\mathbf{y}). \tag{12}$$

The proof is deferred to Appendix A. That is, for a given $\mathbf{y}$, the encoder and generator's task is to return a reconstruction of $\mathbf{y}$, i.e., $\mathbf{x}$, satisfying the average distortion constraint; for a given $\mathbf{y}$ and its reconstruction $\mathbf{x}$, the optimal discriminator returns a *soft score*, i.e., $1 - \beta\rho(\mathbf{x},\mathbf{y}) - \ln Z_{\beta,RD}(\mathbf{y})$.

Corollary 2. When $\beta = \infty$, then $p^*_{RD}(\mathbf{x}|\mathbf{y}) = \mathbb{1}_{\mathbf{y}=\mathbf{x}}$, $p^*_{RD}(\mathbf{x})$ and $p(\mathbf{y})$ are identical, and $D^*(\mathbf{x},\mathbf{y}) = 1 - \ln p(\mathbf{y})$.

That is, the generator's task is to return a sample $\mathbf{x}$; given $\mathbf{x}$ and $\mathbf{y}$, the optimal discriminator returns a *hard score*, i.e., $D^*(\mathbf{x},\mathbf{y}) = 1 - \ln p(\mathbf{y})$ if $\mathbf{x} \sim p(\mathbf{y})$, otherwise $D^*(\mathbf{x},\mathbf{y}) = \infty$.



(a) RD-GEN; (b) GAN.

Figure 1: Comparison between RD-GEN and GAN. Given $p(\mathbf{y})$, RD-GEN is to learn $p^*_{RD}(\mathbf{x})$ and $p^*_{RD}(\mathbf{x}|\mathbf{y})$ achieving $R(d)$. When $d = 0$ and the encoder is optional, RD-GEN degenerates to GAN.

4 Rate Distortion Via EBM

In this part, we implement RD-GEN with EBMs. There are two advantages with EBMs: first it is possible to represent both $p^*_{RD}(\mathbf{x})$ and $p^*_{RD}(\mathbf{x}|\mathbf{y})$ by one EBM ϕ, i.e., Section 4.1; secondly the training of EBMs with the objective $\mathcal{L}'_{RD}(\phi, \theta, \omega)$ can be accomplished in a manner similar to Blahut-Arimoto, i.e., Section 4.2.

4.1 Represent $p^*_{RD}(\mathbf{x})$ and $p^*_{RD}(\mathbf{x}|\mathbf{y})$ by one EBM

Lemma 3. Suppose $p^*_{RD}(\mathbf{x})$ is represented by one EBM, i.e., $p^*_{RD}(\mathbf{x}) = \exp\left[-E_\phi(\mathbf{x})\right]/Z_\mathbf{x}$, then $p^*_{RD}(\mathbf{x}|\mathbf{y})$ can be represented by a related EBM, i.e., $p^*_{RD}(\mathbf{x}|\mathbf{y} = \exp\{-[E_\phi(\mathbf{x}) + \beta\rho(\mathbf{y}, \mathbf{x})]\}/Z_{\mathbf{y}|\mathbf{x}}$, where $Z_{\mathbf{y}|\mathbf{x}} = \int \exp\{-[E_\phi(\mathbf{y}) + \beta\rho(\mathbf{y}, \mathbf{x})]\}d\mathbf{x}$.

The proof is in Appendix A. That is, we only need to train one EBM as the generator instead of two neural networks for the encoder and the generator separately.

4.2 Discriminative-Blahut-Arimoto Algorithm

Discriminative-Blahut-Arimoto (DBA) is presented in Algorithm 1, where ω^t, ϕ^t, and $R^t(d)$ denote the trained discriminator, the trained EBM, and the estimated rate distortion at the t^{th}-iteration, respectively.

More specifically, the algorithm first initializes ω^t, ϕ^t randomly, i.e., Line 2; After that the algorithm alternatively loops between two steps until both ω^t, ϕ^t converge: optimize ω^t based on (8) i.e., Line 7, where $\mathbf{x} \sim p_{\phi^t}(\mathbf{x}|\mathbf{y})$ ($p_{\phi^t}(\mathbf{x}|\mathbf{y})$ is the same as (3) except $p^*_{RD}(\mathbf{x})$ is replaced by $p_{\phi^k}(\mathbf{x})$) and $\mathbf{x}' \sim p_{\phi^t}(\mathbf{x})$ are obtained via LD, i.e., Line 5; optimize ϕ^t based on (9), i.e., Line 8.

That is, an EBM is trained to model $p^*_{RD}(\mathbf{x})$ and thus $p^*_{RD}(\mathbf{x}|\mathbf{y})$ due to Lemma 3 and a discriminator is trained to estimate mutual information.

Algorithm 1 DBA

procedure DBA($p(\mathbf{y}), \beta, \rho(\cdot)$)
  $t \leftarrow 0$ and initialize ω^t, ϕ^t arbitrarily
  while not converged **do**
    for $\mathbf{y} \sim p(\mathbf{y})$ **do**
      sample $\mathbf{x} \sim p_{\phi^t}(\mathbf{x}|\mathbf{y})$, $\mathbf{x}' \sim p_{\phi^t}(\mathbf{x})$ via LD
      feed $\mathbf{y}, \mathbf{x}, \mathbf{x}'$ to ω^t and approximate $R^t(d)$
      update ω^t by stochastic gradient ascent of $\mathcal{L}_{MI}$
      update ϕ^t by stochastic gradient descent of $\mathcal{L}'_{RD}$
    end for
    $t \leftarrow t + 1$
  end while
  return ω^t, ϕ^t and $R^t(d)$
end procedure

4.3 Theoretical analysis

Theorem 1. 1. Algorithm 1 converges, that is,

$$\mathcal{L}_{RD}(\phi^t) \geq \mathcal{L}_{RD}(\phi^{t+1});$$

2. Assume ϕ and ω have enough capacity to represent $p^*_{RD}(\mathbf{x}|\mathbf{y})$ and $D^*(\mathbf{x}, \mathbf{y})$, $\mathcal{L}_{RD}(\phi^t) \to \min_{\{p(\mathbf{x}), p(\mathbf{x}|\mathbf{y})\}} \mathcal{L}_{RD}(p(\mathbf{y}), p(\mathbf{y}|\mathbf{x}))$ as $t \to \infty$.

The proof is in Appendix A. Theorem 1 states that $(p_{\phi^t}(\mathbf{x}), p_{\phi^t}(\mathbf{x}|\mathbf{y}))$ learned by DBA converges to rate-distortion posterior $(p^*_{RD}(\mathbf{x}), p^*_{RD}(\mathbf{x}|\mathbf{y}))$ when $t \to \infty$.

5 Experiments

We compare our estimated rate distortion functions with theoretical predictions for rate distortion functions with closed-form expressions, i.e., Section 5.1; for rate distortion functions with known bounds, we compare with both theoretical bounds and prior best empirical results [10], i.e., Section 5.2. Appendix B contains the experiment's specifics.

5.1 Estimation of rate distortion with closed-form expressions

We first consider a binary symmetric source (BSS) with Hamming distortion. Its rate distortion is given by [1, Theorem 10.3.1], i.e.,

$$R(d) = \begin{cases} 1 - H(d) & \text{if } 0 \leq d \leq \frac{1}{2}, \\ 0 & \text{if } d > \frac{1}{2}, \end{cases}$$

where $H(\cdot)$ is the binary entropy function.

We next consider a Gaussian source $\mathcal{N}(0, \sigma^2)$ with L2 distortion. Its rate distortion is given by [1, Theorem 10.3.2], i.e.,

$$R(d) = \begin{cases} \frac{1}{2} \log \frac{\sigma^2}{d} & \text{if } 0 \leq d \leq \sigma^2, \\ 0 & \text{if } d > \sigma^2. \end{cases}$$

We finally consider a Laplacian source, i.e., $p(x, \lambda) = \frac{\lambda}{2}\exp(-\lambda\|x\|_1)$, with L1-norm distortion. Its rate distortion is given by [11], i.e.,

$$R(d) = \begin{cases} -\log(\lambda d) & \text{if } 0 \le d \le \frac{1}{\lambda}, \\ 0 & \text{if } d > \frac{1}{\lambda}. \end{cases}$$

We present approximation results of $R(d)$ in Fig. 2. Comparing with theoretical results, DBA approximates theoretical results closely.

5.2 Estimation of rate distortion with known bounds

5.2.1 Binary Symmetric Markov Source and L1-norm distortion

We now consider a long-standing problem ([12–14]) in information theory: determination of the rate distortion function for a binary symmetric Markov source (BSMS). Let $\{x_k, k = 1, 2, \cdots, n\}$ be a binary symmetric Markov source with transition parameter q. Mathematically, $\{x_k\}$ is a 2-state Markov chain with $\Pr(x_1 = 0) = \Pr(x_1 = 1) = 1/2$ and a probability transition matrix

$$\begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} 1-q & q \\ q & 1-q \end{pmatrix} \end{array}.$$

For simplicity, we assume $q \le 1/2$. Computation of $R(d)$ of BSMS has been investigated by Gray in [13], where only bounds were provided.

In Fig. 3a, we show the approximation results of $R(d)$ of BSMS ($p = 0.25$) with L1 distance. We present empirical lower bounds based on [15] and the best empirical results [10] for reference. DBA aligns with theoretical bounds and approximates better than [10] when compared to previous best empirical results.

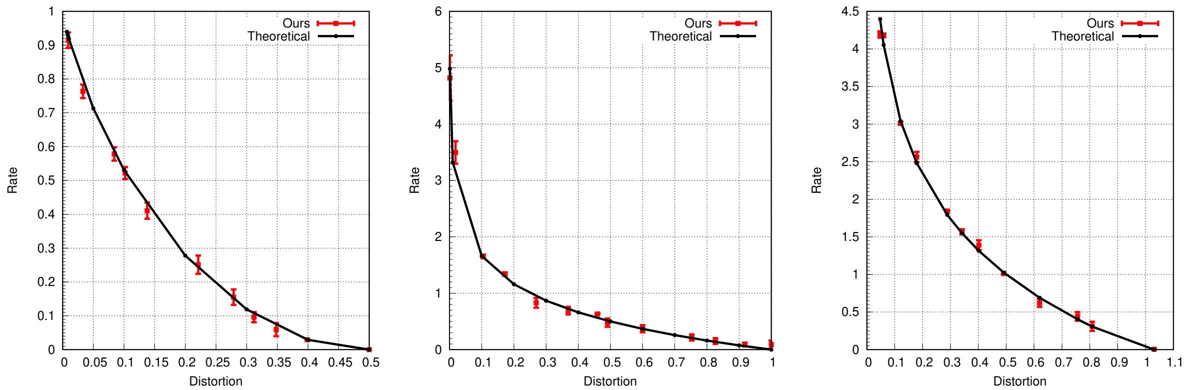
5.2.2 Binary Asymmetric Markov Source with L1-norm Distortion

Binary Asymmetric Markov Source (BAMS) $\{x_k, k = 1, 2, \cdots, n\}$ is a binary Markov source with transition probabilities between the two states p and q. Mathematically, $\{x_k\}$ is a 2-state Markov chain with $\Pr(x_1 = 0) = \Pr(x_1 = 1) = 1/2$ and a probability transition matrix

$$\begin{array}{c} \\ 0 \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} 0 & 1 \end{array} \\ \begin{pmatrix} 1-q & q \\ p & 1-p \end{pmatrix} \end{array}.$$

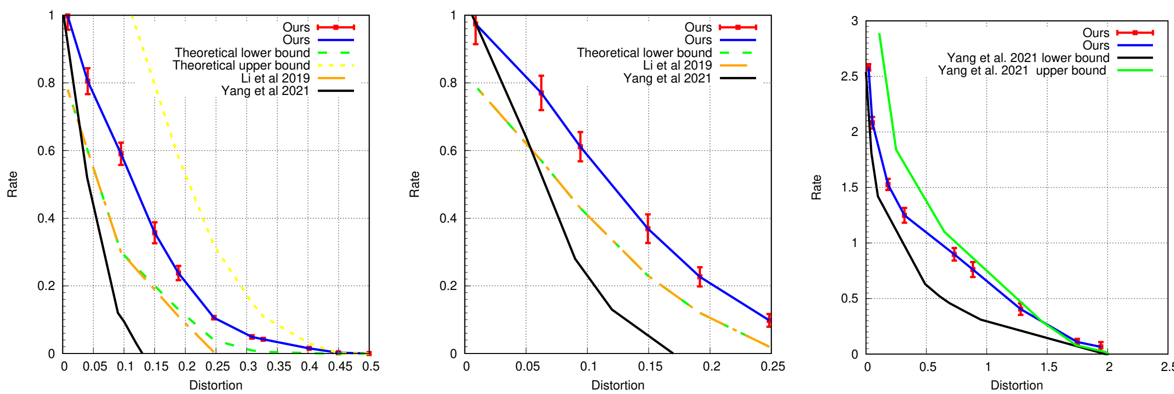
For simplicity, assume that $p < q \le 1/2$. $R(d)$ of BAMS source has not been solved yet except lower bounds [13].

In Fig. 3b, we shown the approximation results of $R(d)$ of BAMS ($p = 0.25, q = 0.3$) with L1-norm. Similarly, based on [15] and best empirical results [10], we give empirical lower bounds. In most cases, the empirical bounds from [15] are less stringent than theoretical bounds. DBA aligns with theoretical bounds and approximates better than [10] when compared to previous best empirical results.



(a) BSS with L1 distortion; (b) Gaussian with L2 distortion; (c) Laplacian with L1 distortion.

Figure 2: Estimations of rate distortion with closed-form expressions via RD-EBM.



(a) BSMS with L1 distortion; (b) BAMS with L1 distortion; (c) Laplacian with L2 distortion.

Figure 3: Estimations of rate distortion without closed-form expressions via RD-EBM.

5.3 Estimation of rate distortion with unknown bounds

This part focuses on rate distortion of a Laplacian source with L2-norm distortion, which has unknown theoretical bounds. In Fig. 3c, we approximate it with DBA and compare it to experimental upper and lower bounds based on [15]. DBA aligns with experiment bounds and is hence useful for investigating general rate distortion with unknown bounds.

6 Related Work

Recently, there has been a surge in interest in deep learning applications for rate-distortion estimation [10, 15–18]. Our study focuses on rate-distortion estimations and differs from previous work [10, 15–18] in that we theoretically investigate the connection between rate-distortion and energy-based models and effectively employ energy-based models to estimate various rate-distortion functions.

7 Conclusions

In this work, we estimate rate distortion using energy-based models (EBMs). The central idea is to model rate-distortion as a minimax game, which provides a framework for neural networks for estimating rate-distortion. After that, the framework is then instantiated with EBMs, and DBA is provided to approximate rate-distortion. Our empirical results show that our estimates agree with closed-form expressions and known bounds.

8 References

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.

[2] Suguru Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 14–20, 1972.

[3] Richard Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE transactions on Information Theory*, vol. 18, no. 4, pp. 460–473, 1972.

[4] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang, "A tutorial on energy-based learning," *Predicting structured data*, vol. 1, no. 0, 2006.

[5] Gareth O Roberts and Richard L Tweedie, "Exponential convergence of langevin distributions and their discrete approximations," *Bernoulli*, pp. 341–363, 1996.

[6] Max Welling and Yee W Teh, "Bayesian learning via stochastic gradient langevin dynamics," in *Proceedings of the 28th international conference on machine learning (ICML-11)*. Citeseer, 2011, pp. 681–688.

[7] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp. 5847–5861, 2010.

[8] Kurt Hornik, Maxwell Stinchcombe, and Halbert White, "Multilayer feedforward networks are universal approximators," *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.

[9] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka, "f-gan: Training generative neural samplers using variational divergence minimization," in *Advances in neural information processing systems*, 2016, pp. 271–279.

[10] Qing Li and Yang Chen, "Rate distortion via deep learning," *IEEE Transactions on Communications*, vol. 68, no. 1, pp. 456–465, 2019.

[11] William HR Equitz and Thomas M Cover, "Successive refinement of information," *IEEE Transactions on Information Theory*, vol. 37, no. 2, pp. 269–275, 1991.

[12] Toby Berger, "Explicit bounds to r (d) for a binary symmetric markov source," *IEEE Transactions on Information Theory*, vol. 23, no. 1, pp. 52–59, 1977.

[13] R Gray, "Rate distortion functions for finite-state finite-alphabet markov sources," *IEEE Transactions on Information Theory*, vol. 17, no. 2, pp. 127–134, 1971.

[14] Shirin Jalali and Tsachy Weissman, "New bounds on the rate-distortion function of a binary markov source," in *Information Theory, 2007. ISIT 2007. IEEE International Symposium on*. IEEE, 2007, pp. 571–575.

[15] Yibo Yang and Stephan Mandt, "Towards empirical sandwich bounds on the rate-distortion function," *arXiv preprint arXiv:2111.12166*, 2021.

[16] Qing Li, Yang Chen, and Yongjune Kim, "Compression By and For Deep Boltzmann Machines," *IEEE Transactions on Communications*, 2020.

[17] Shirin Jalali and Tsachy Weissman, "Rate-distortion via markov chain monte carlo," in *2008 IEEE International Symposium on Information Theory*. IEEE, 2008, pp. 852–856.

[18] Eric Lei, Hamed Hassani, and Shirin Saeedi Bidokhti, "Neural estimation of the rate-distortion function with applications to operational source coding," *arXiv preprint arXiv:2204.01612*, 2022.

A Appendix

A.1 Proof for Lemma 1

Lemma 4. For given ϕ, θ, $\mathbf{x}$ and $\mathbf{y}$, the optimal discriminator $D^*(\mathbf{x}, \mathbf{y}, \phi, \theta)$ according to (9), i.e., $D^*(\mathbf{x}, \mathbf{y}, \phi, \theta) = \arg\max_\omega \mathcal{L}'_{RD}(\phi, \theta, \omega)$, is given by

$$D^*(\mathbf{x}, \mathbf{y}, \phi, \theta) = \ln p_\phi(\mathbf{x}|\mathbf{y}) - \ln p_\theta(\mathbf{x}) + 1. \tag{13}$$

Proof. Define $\mu(\mathbf{x}, \mathbf{y}) := p(\mathbf{x}, \mathbf{y}) D(\mathbf{x}, \mathbf{y}) - p(\mathbf{x}) p(\mathbf{y}) \exp[D(\mathbf{x}, \mathbf{y}) - 1]$

Thus

$$\frac{d\mu}{dD} = p(\mathbf{x}, \mathbf{y}) - p(\mathbf{x}) p(\mathbf{y}) \exp[D(\mathbf{x}, \mathbf{y}) - 1].$$

By setting the above equation to zero, we obtain (13). □

Proof for Lemma 1 The Lemma 1 holds because of Lemma 4 and (4).

A.2 Proof for Lemma 3

Proof.

$$\begin{aligned}
p^*_{RD}(\mathbf{x}|\mathbf{y}) &= \frac{1}{Z_\beta(\mathbf{y})} p^*_{RD}(\mathbf{x}) \exp[-\beta\rho(\mathbf{y}, \mathbf{x})], && (14)\\
&= \frac{p^*_{RD}(\mathbf{x}) \exp[-\beta\rho(\mathbf{y}, \mathbf{x})]}{\int p^*_{RD}(\mathbf{x}) \exp[-\beta\rho(\mathbf{y}, \mathbf{x})] d\mathbf{x}}, && (15)\\
&= \frac{\frac{\exp[-E_\phi(\mathbf{x})]}{Z_\mathbf{x}} \exp[-\beta\rho(\mathbf{y}, \mathbf{x})]}{\int \frac{\exp[-E_\phi(\mathbf{x})]}{Z_\mathbf{x}} \exp[-\beta\rho(\mathbf{y}, \mathbf{x})] d\mathbf{x}}, && (16)\\
&= \frac{\exp\{-[E_\phi(\mathbf{x}) + \beta\rho(\mathbf{y}, \mathbf{x})]\}}{\int \exp\{-[E_\phi(\mathbf{x}) + \beta\rho(\mathbf{y}, \mathbf{x})]\} d\mathbf{x}}, && (17)\\
&= \frac{\exp\{-[E_\phi(\mathbf{x}) + \beta\rho(\mathbf{y}, \mathbf{x})]\}}{Z_{\mathbf{y}|\mathbf{x}}}, && (18)
\end{aligned}$$

where

(14) is based on (4);

(15) is based on (5);

(16) is based on the assumption that $p^*_{RD}(\mathbf{x})$ can be represented by one EBM of the form (5);

(16) is based on the notation $Z_{\mathbf{y}|\mathbf{x}} = \int \exp\{-[E_\phi(\mathbf{y}) + \beta\rho(\mathbf{y}, \mathbf{x})]\} d\mathbf{x}$.

□

A.3 Proof for Theorem 1

Proof. 1. The convergence part follows by:

$$\begin{aligned}
\mathcal{L}_{RD}(\phi^t) &\geq \max_{\omega} \mathcal{L}'_{RD}(\phi^t, \omega), && (19)\\
&= \mathcal{L}'_{RD}(\phi^t, \omega^{t+1}), && (20)\\
&\geq \min_{\phi} \mathcal{L}'_{RD}(\phi, \omega^{t+1}), && (21)\\
&= \mathcal{L}_{RD}(\phi^{t+1}), && (22)
\end{aligned}$$

where

(19) is due to (2) or [7, Lemma1];

(20) is due to (10) and the universal assumption on neural networks.

2. The second part follows because $\{\mathcal{L}_{RD}(\phi^t)\}$ is decreasing and bounded, thus $\{\mathcal{L}_{RD}(\phi^t)\}$ must converge.

□

B Experimental details

B.1 Architecture and hyperparameters for Section 5.1 and Section 5.3

- Energy-Based Model: $1 \to 64 \to 64 \to 64 \to 1$ with Sigmoid activation function;
- Discriminator: $2 \to 640 \to 640 \to 640 \to 1$ with LeakyReLu activation function;
- Langevin dynamic: $K = 200$ and $\lambda = 0.001$.

The training epoch is 400, data length is 1, total dataset size is 81920, which are randomly sampled from a given source distribution, learning rate is $1e^{-4}$, and batch size is 128. For each rate distortion function, we report its mean and standard deviation over five runs. The experiments are run on a single GPU.

B.2 Architecture and hyperparameters for Section 5.2

- Energy-Based Model: $100 \to 640 \to 640 \to 640 \to 1$ with Sigmoid activation function;
- Discriminator: $200 \to 640 \to 640 \to 640 \to 1$ with LeakyReLu activation function;
- Langevin dynamic: $K = 200$ and $\lambda = 0.001$.

The training epoch is 400, data length is 100, and total dataset size is 8192.