Rate-Distortion via Energy-Based Models

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

* Western Digital Research, Milpitas, CA, 95035

† Department of Electrical Engineering at Pohang University of Science and Technology, Pohang, Gyeongbuk, South Korea

Rate Distortion

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

$$p^*_{RD}(\mathbf{x}|\mathbf{y}) := \underset{\{p(\mathbf{x}|\mathbf{y}):\mathbf{y},\mathbf{x}\sim p(\mathbf{y})p(\mathbf{x}|\mathbf{y}),\mathbb{E}[\rho(\mathbf{x},\mathbf{y})]\leq d\}}{\arg\min} I(\mathbf{x};\mathbf{y}), \quad (2)$$

inducing the following distribution

$$p^*_{RD}(\mathbf{x}) = \int p(\mathbf{y})p^*_{RD}(\mathbf{x}|\mathbf{y})d\mathbf{y}. \quad (3)$$

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

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

and

$$Z_{\beta,RD}(\mathbf{y}) := \int p^*_{RD}(\mathbf{x})\exp[-\beta\rho(\mathbf{x},\mathbf{y})]d\mathbf{x}. \quad (5)$$

EBM

- Definition:

$$p_\phi(\mathbf{x}) := \frac{\exp[-E_\phi(\mathbf{x})]}{Z_\phi}, \quad (6)$$

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

$$\mathcal{L}_{ML}(\phi) := \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}[-\log p_\phi(\mathbf{x})], \quad (7)$$

  where $p(\mathbf{x})$ represents the underlying data distribution.
- Generating samples with the Langevin dynamics (LD) [2]

$$\mathbf{x}_i := \mathbf{x}_{i-1} - \lambda\nabla_{\mathbf{x}_{i-1}}E_\phi(\mathbf{x}_{i-1}) + \sqrt{2\lambda}\mathbf{z}_i, \quad \mathbf{z}_i \sim \mathcal{N}(0,\mathbf{I}), \quad (8)$$

Notations

$$\mathcal{L}_{MI}(D) := \mathbb{E}_{\mathbf{x},\mathbf{y}\sim p(\mathbf{x},\mathbf{y})}[D(\mathbf{x},\mathbf{y})] - \mathbb{E}_{\mathbf{x},\mathbf{y}\sim p(\mathbf{x})p(\mathbf{y})}[\exp(D(\mathbf{x},\mathbf{y})-1)], \quad (9)$$

$$\mathcal{L}'_{RD}[p(\mathbf{x}),p(\mathbf{x}|\mathbf{y}),D] := \mathcal{L}_{MI}(D) + \beta\mathbb{E}[\rho(\mathbf{x},\mathbf{y})]. \quad (10)$$

MAIN RESULTS

- Representing $p^*_{RD}(\mathbf{x})$ and $p^*_{RD}(\mathbf{x}|\mathbf{y})$ using one EBM;
- Learning a single EBM ϕ that represents both $p^*_{RD}(\mathbf{x})$ and $p^*_{RD}(\mathbf{x}|\mathbf{y})$ with the EBMs-Blahut-Arimoto (EBA) algorithm.

Methods

Algorithm 1 EBA

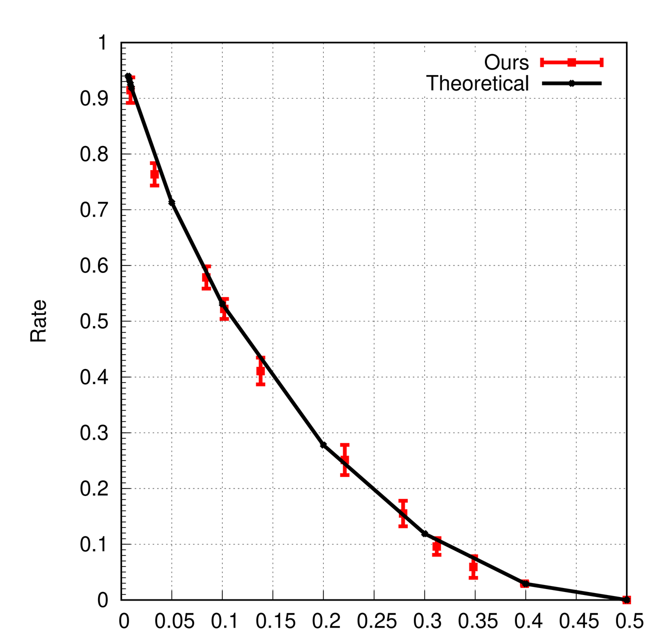
```
procedure EBA(p(y), β, ρ(·))
    t ← 0 and initialize ω^t, φ^t arbitrarily
    while not converged do
        for y ∼ p(y) do
            sample x ∼ p_{φ^t}(x|y), x' ∼ p_{φ^t}(x) via LD
            feed y, x, x' to ω^t and approximate R^t(d)
            update ω^t by stochastic gradient ascent of L_MI
            update φ^t by stochastic gradient descent of L'_RD
        end for
        t ← t + 1
    end while
    return ω^t, φ^t and R^t(d)
end procedure
```

Main Theorem

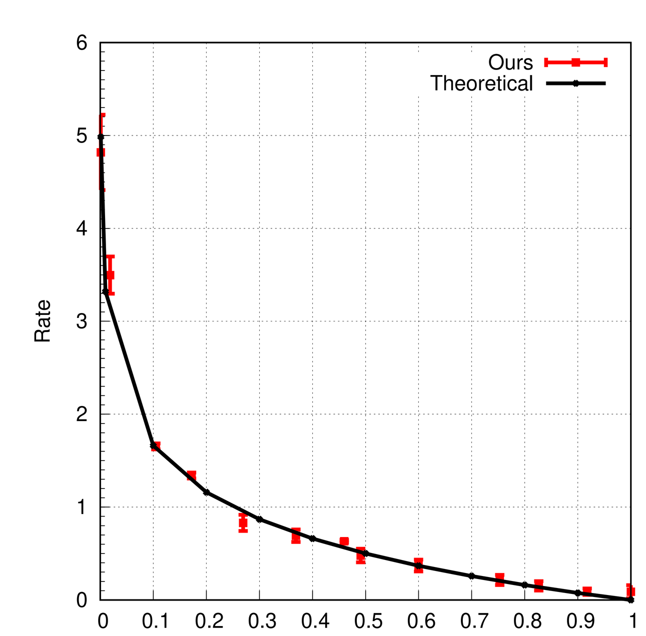
The following properties hold for Algorithm 1:

1. *Convergence: Algorithm 1 converges, meaning that $\mathcal{L}_{RD}(\phi^t) \geq \mathcal{L}_{RD}(\phi^{t+1})$.*
2. *Asymptotic rate-distortion approaching posterior: If ϕ has enough capacity to fully represent any probability distribution, then Algorithm 1 learns the rate-distortion approaching posterior asymptotically, i.e., $(p_{\theta^t}(\mathbf{x}), p_{\theta^t}(\mathbf{x}|\mathbf{y})) \to (p^*_{RD}(\mathbf{x}), p^*_{RD}(\mathbf{x}|\mathbf{y}))$ when $t \to \infty$.*

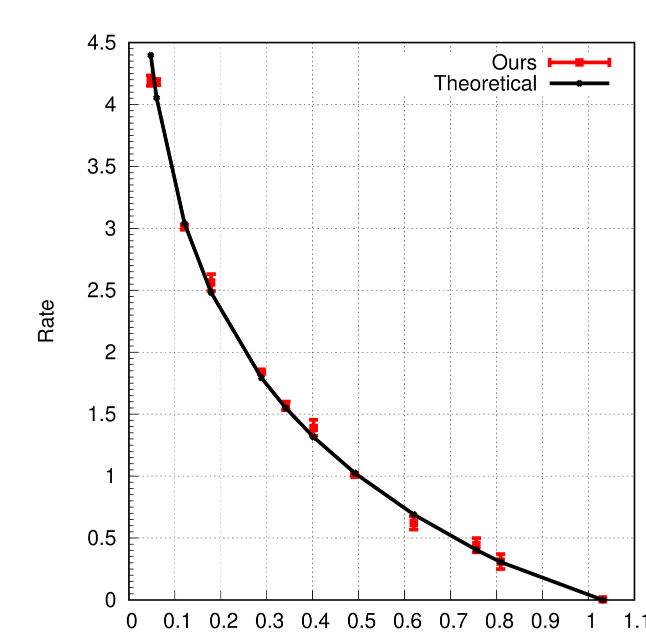
Experiment Results



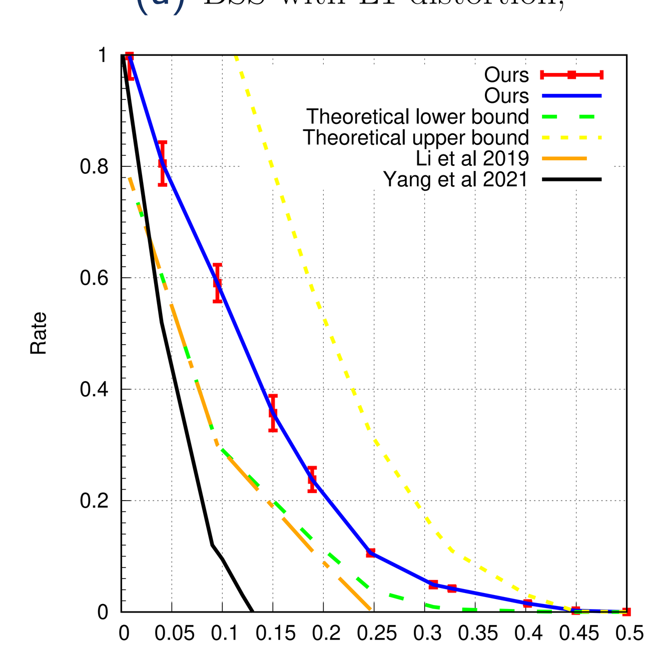
(a) BSS with L1 distortion;



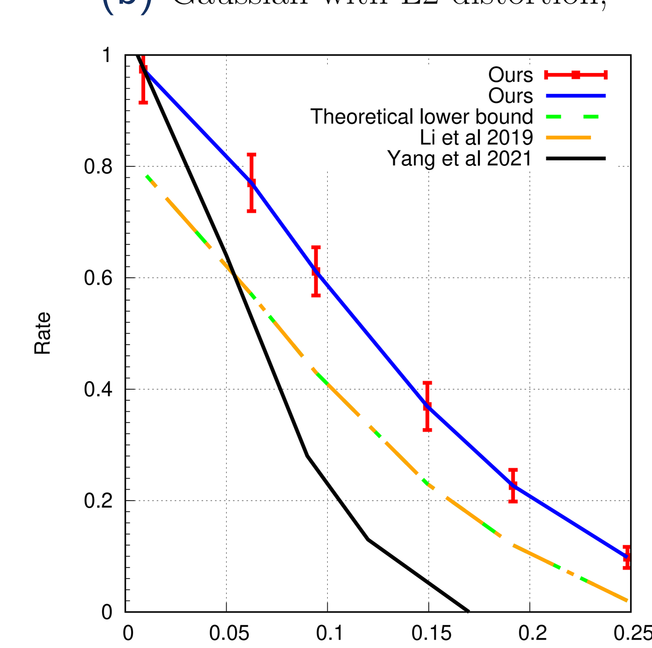
(b) Gaussian with L2 distortion;



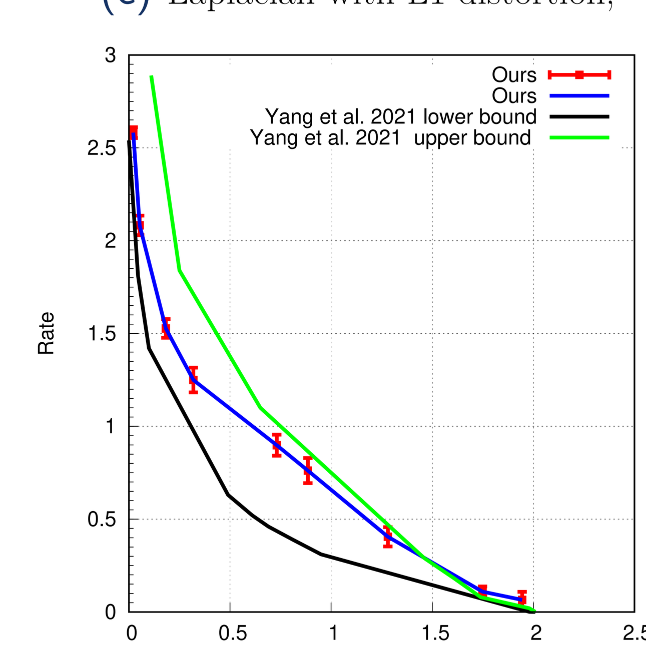
(c) Laplacian with L1 distortion;



(d) BSMS with L1 distortion;



(e) BAMS with L1 distortion;



(f) Laplacian with L2 distortion.

Conclusion

- We demonstrate that EBMs can be used to approximate the rate-distortion approaching posterior, as in the Blahut-Arimoto (BA) algorithm.
- Our empirical results show that our estimates match closed-form expressions and known bounds.

References

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

[2] 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)*, pages 681–688. Citeseer, 2011.

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

[4] Richard Turner. Cd notes. http://www.gatsby.ucl.ac.uk/~turner/Notes/ContrastiveDivergence/CDv3.pdf.

Contact Information

- Email: qingklee@gmail.com

Western Digital

Western Digital