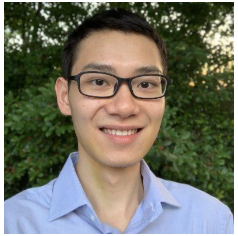
Estimating the Rate-Distortion Function by Wasserstein Gradient Descent




Yibo Yang[1], Stephan Eckstein[2], Marcel Nutz[3], and Stephan Mandt[1]

[1] University of California, Irvine

[2] ETH Zurich [3] Columbia University

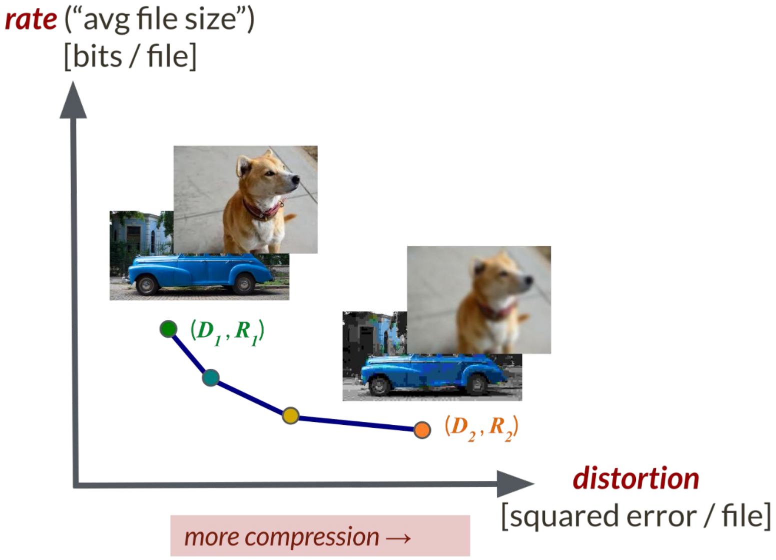
paper (NeurIPS 2023): https://arxiv.org/abs/2310.18908

code & data: https://github.com/yiboyang/wgd

Motivation: info-theoretic limit of lossy compression

Lossy compression algorithms (e.g., JPEG) are typically evaluated on

- **rate** (“average file size”)
- **distortion** (reconstruction error)



https://yiboyang.com/posts/estimating-the-rate-distortion-function-of-real-world-data-part-1/

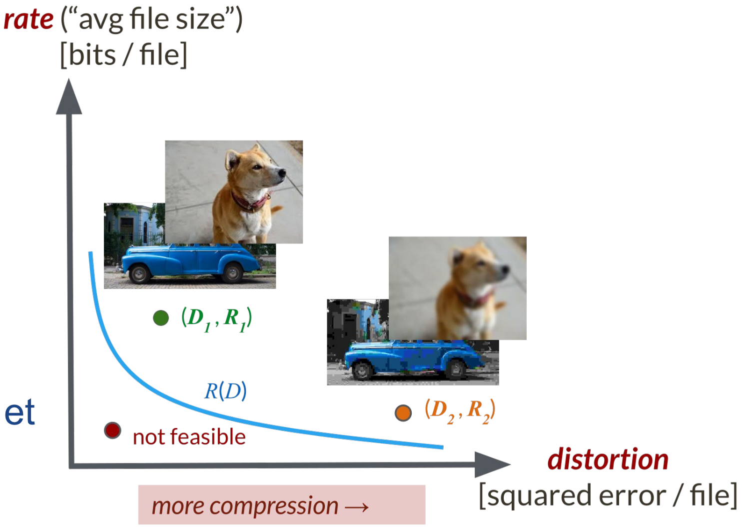
Motivation: info-theoretic limit of lossy compression

Question: Given a data source and distortion metric, what's the best possible rate-distortion (R-D) tradeoff?

Answer: the rate-distortion function [Shannon 1959]

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

No closed-form



This work (also see [Gibson 2017, Yang & Mandt 2022, Lei et al., 2023]): a new, neural network-free algorithm for estimating R(D) over continuous spaces, based on ideas/techniques from Optimal Transport.

The R-D problem, formally

Given: 1. $(\mathcal{X}, \mathcal{Y})$, a.k.a. the (data, reconstruction) alphabets (Polish spaces here).

2. Source distribution μ on $\mathcal{X}$.

3. Distortion metric $\rho: \mathcal{X} \times \mathcal{Y} \to [0, \infty)$

$$R(D) := \inf_{\pi \in \Pi(\mu, \cdot) : \int \rho d\pi \leq D} H(\pi | \pi_1 \otimes \pi_2)$$

We work with an equivalent “Lagrangian” parameterization of R(D), following [Blahut 1972, Arimoto 1972]

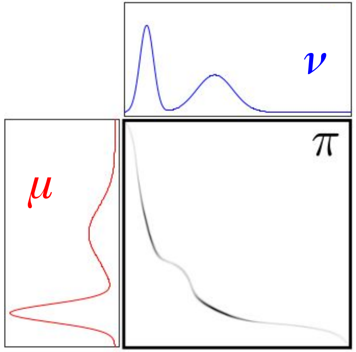
$$F(\lambda) := \inf_{\nu \in \mathcal{P}(\mathcal{Y})} \underbrace{\inf_{\pi \in \Pi(\mu, \cdot)} \lambda \int \rho d\pi + H(\pi | \mu \otimes \nu)}_{\mathcal{L}^{\lambda}_{BA}(\mu, \nu)}$$

Background: optimal transport (OT)

Given: 1. $(\mathcal{X}, \mathcal{Y})$, a.k.a. the (source, destination) spaces.

2. Source distribution μ on $\mathcal{X}$., target distribution ν on $\mathcal{Y}$.

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



The Kantorovich problem:

$$\inf_{\pi \in \Pi(\mu,\nu)} \int \rho(x,y) d\pi(x,y)$$

Defines a metric (Wasserstein distance) b/w prob. measures if the cost ρ is a metric.

Entropic regularization [Peyré and Cuturi, 2019, Chapter 4]:

$$\mathcal{L}^{\epsilon}_{EOT}(\mu,\nu) := \inf_{\pi \in \Pi(\mu,\nu)} \int \rho d\pi + \epsilon H(\pi | \mu \otimes \nu)$$

Theoretical insights – part 1

The R-D problem (1) is equivalent to

(2) Projection under an entropic OT cost;

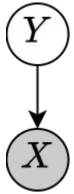
(3) Deconvolution/denoising of the source (e.g., quadratic cost = Gaussian noise)

$$(1)\quad \min_{\nu \in \mathcal{P}(\mathcal{Y})} \mathcal{L}_{BA}^{\lambda}(\mu, \nu) \qquad \longleftrightarrow \qquad (2)\quad \min_{\nu \in \mathcal{P}(\mathcal{Y})} \mathcal{L}_{EOT}^{1/\lambda}(\mu, \nu)$$

Also see [Csiszár, 1974]
and [Lei et al., 2023]

[Rigollet and Weed, 2018]

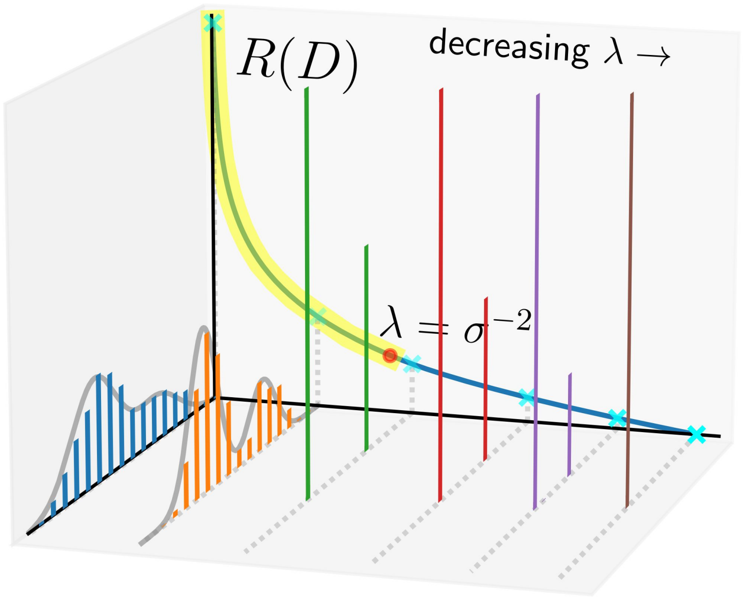
$$(3)\quad \max_{\nu \in \mathcal{P}(\mathcal{Y})} \mathbb{E}_{x \sim \mu}\left[\log\left(\int e^{-\lambda \rho(x,y)} \nu(dy)\right)\right]$$



Theoretical insights – part 1

Thus,

- The convolution between a Gaussian and any distribution (e.g., Gaussian mixture with shared covariance) has a segment of R(D) available in closed-form;
- Provides a wide class of sources that can serve as test cases for algorithms.



Wasserstein gradient descent

Suppose $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, ρ continuously differentiable. Goal:

$$\min_{\nu \in \mathcal{P}(\mathbb{R}^d)} \mathcal{L}(\nu), \qquad \mathcal{L}(\cdot) \in \{\mathcal{L}_{BA}(\mu, \cdot), \mathcal{L}_{EOT}(\mu, \cdot)\}$$

Idea: simulate the gradient flow of the $\mathscr{L}$ in the 2-Wasserstein space of probability measures [Santambrogio 2015]:

$$\nu^{(t)} = \left(\mathrm{id} - \gamma \underbrace{\nabla \frac{\delta \mathcal{L}}{\delta \nu}(\nu^{(t-1)})}_{\text{W. gradient : } \mathbb{R}^d \longrightarrow \mathbb{R}^d}\right)_{\#} \nu^{(t-1)}$$

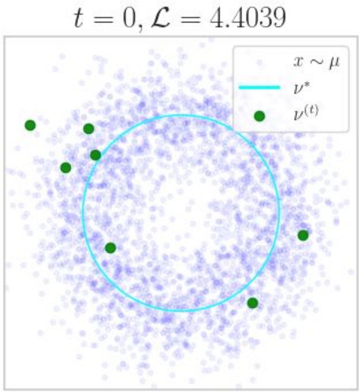
The Wasserstein gradient can be tractably computed by

- Sinkhorn's algorithm, for $\mathscr{L} = \mathscr{L}_{\mathrm{EOT}}$, or
- A **single** Sinkhorn iteration, for $\mathscr{L} = \mathscr{L}_{\mathrm{BA}}$ (orders of magnitude faster!)

Wasserstein gradient descent

$$\nu^{(t)} = \left(\mathrm{id} - \gamma \nabla \frac{\delta \mathcal{L}}{\delta \nu}(\nu^{(t-1)})\right)_{\#} \nu^{(t-1)}$$

In practice, we maintain/update particles:



$$\nu = \frac{1}{n} \sum_{i=1}^{n} \delta_{y_i}$$

$$y_i^{(t)} = y_i^{(t-1)} - \gamma \nabla \frac{\delta \mathcal{L}}{\delta \nu}(\nu^{(t-1)})[y_i^{(t-1)}], \quad i = 1, 2, ..., n$$

Theoretical insights (2)

The R-D problem is equivalent to “EOT projection”, therefore:

Finite-sample bounds on estimating $\mathscr{L}_{\mathrm{EOT}}$ [Mena and Niles-Weed, 2019, Genevay et al., 2019, Rigollet and Stromme, 2022]

$\Downarrow$ implies

Finite-sample bounds on estimating R(D) [also see Harrison and Kontoyiannis, 2008]:

Proposition 4.3. *Let μ be σ^2-subgaussian. Consider $\mathcal{L} := \mathcal{L}_{EOT}$. Then the optimal reproduction distribution ν^* is also σ^2-subgaussian. For a constant C_d only depending on d, we have*

$$\mathbb{E}\left[\left|\min_{\nu\in\mathcal{P}(\mathbb{R}^d)} \mathcal{L}(\mu,\nu) - \min_{\nu_n\in\mathcal{P}_n(\mathbb{R}^d)} \mathcal{L}(\mu^m,\nu_n)\right|\right] \le C_d\,\epsilon\left(1+\frac{\sigma^{\lceil 5d/2\rceil+6}}{\epsilon^{\lceil 5d/4\rceil+3}}\right)\left(\frac{1}{\sqrt{m}}+\frac{1}{\sqrt{n}}\right),$$

for all $n, m \in \mathbb{N}$, where $\mathcal{P}_n(\mathbb{R}^d)$ is the set of probability measures over $\mathbb{R}^d$ supported on at most n points, μ^m is the empirical measure of μ with m independent samples and the expectation $\mathbb{E}[\cdot]$ is over these samples. The same inequalities hold for $\mathcal{L} := \lambda^{-1}\mathcal{L}_{BA}$, with the identification $\epsilon = \lambda^{-1}$.

Empirical results: maximum-likelihood deconvolution

- Compared to Blahut-Arimoto and SOTA neural methods NERD [Lei et al., 2023] and RD-VAE [Yang & Mandt, 2022].
- Faster convergence.
- Better solution quality.

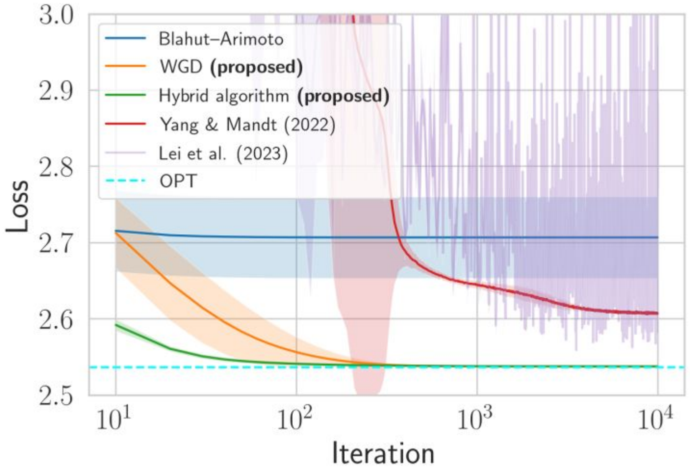


Figure 2: Losses over iterations. Shading corresponds to one standard deviation over random initializations.

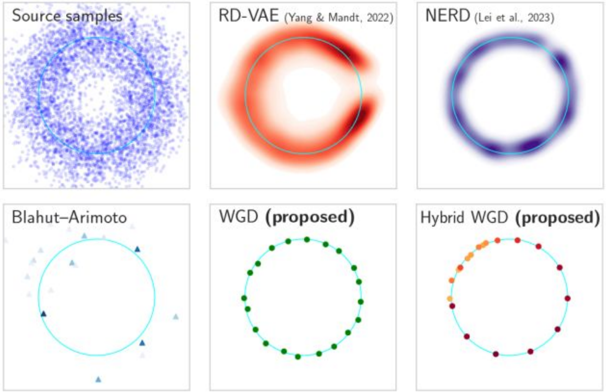
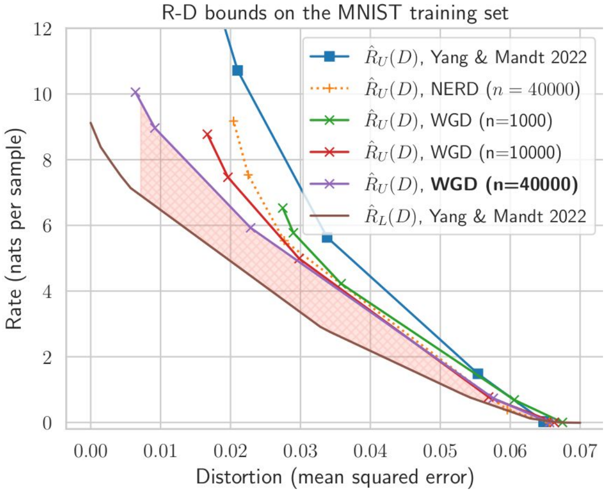
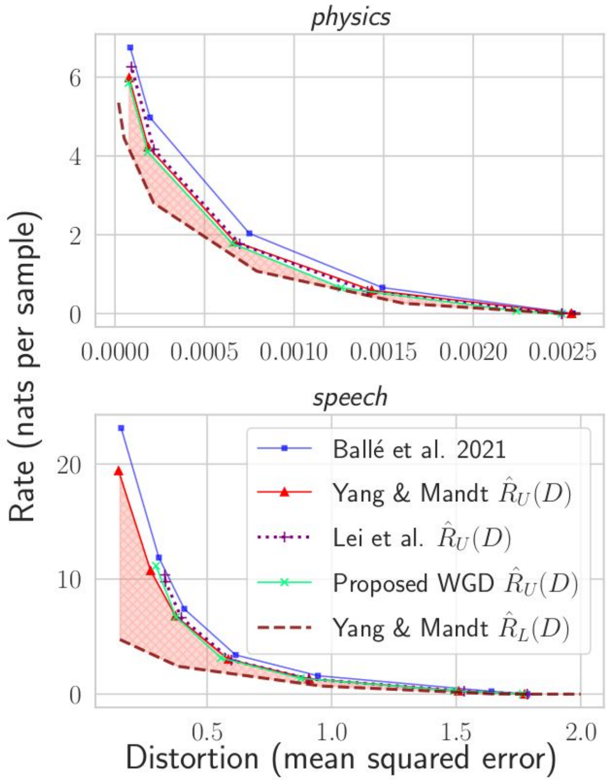


Figure 3: Visualizing μ samples (top left), as well as the ν returned by various algorithms compared to the ground truth ν^* (cyan).

Neural-network free upper bounds on R(D)

- Significantly faster convergence than neural-network-based methods.
- Bound tightness depends on the number of particles used; no neural network architecture tuning!





References

CE Shannon. Coding theorems for a discrete source with a fidelity criterion. IRE Nat. Conv. Rec.,March 1959, 4:142–163, 1959.

Jerry Gibson. Rate distortion functions and rate distortion function lower bounds for real-world sources. Entropy, 19(11):604, 2017.

Yibo Yang and Stephan Mandt. Towards empirical sandwich bounds on the rate-distortion function. In International Conference on Learning Representations, 2022.

Eric Lei, Hamed Hassani, and Shirin Saeedi Bidokhti. Neural estimation of the rate-distortion function with applications to operational source coding. IEEE Journal on Selected Areas in Information Theory, 2023.

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.

Suguru Arimoto. An algorithm for computing the capacity of arbitrary discrete memoryless channels. IEEE Transactions on Information Theory, 18(1):14–20, 1972.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. Foundations and Trends in Machine Learning, 11(5-6):355–607, 2019.

References, cont’d

Imre Csiszár. On an extremum problem of information theory. Studia Scientiarum Mathematicarum Hungarica, 1974

F. Santambrogio. Optimal Transport for Applied Mathematicians. Birkhäuser, 2015

Gonzalo Mena and Jonathan Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. Advances in Neural Information Processing Systems, 32, 2019.

Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of Sinkhorn divergences. In The 22nd International Conference on Artificial Intelligence and Statistics, pages 1574–1583. PMLR, 2019.

Thank you!

paper (NeurIPS 2023): https://arxiv.org/abs/2310.18908
code & data: https://github.com/yiboyang/wgd