Estimating the Rate-Distortion Function by Wasserstein Gradient Descent

Yibo Yang, Stephan Eckstein, Marcel Nutz, and **Stephan Mandt**

UCI University of California, Irvine · ETH Zürich · Columbia University

paper & code:

Overview:

- We apply ideas and techniques from optimal transport to make advances on a basic problem in information theory — estimating the rate-distortion (R-D) function from data.
- Our R-D estimator is based on minimizing an appropriate functional in the space of probability measures, approximated by moving particles.
- We draw close connections between R-D estimation, entropic optimal transport, and deconvolution, and leverage the connections to:
  - introduce a new class of sources with known solutions to the R-D problem as test cases for algorithms, and
  - derive sample complexity for our R-D / deconvolution estimator.
- We obtain comparable or improved R-D estimates compared to SOTA methods based on neural networks [Yang & Mandt 2022; Lei et al. 2023], while requiring significantly less computation and tuning.

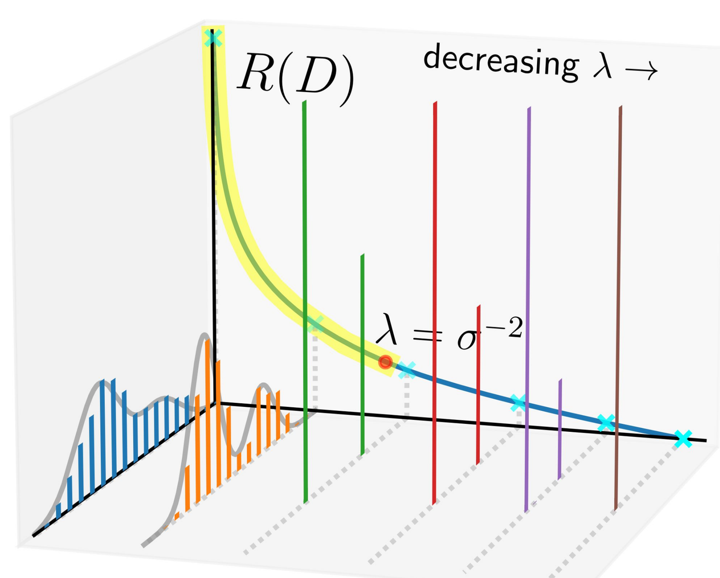
Background: lossy compression and $R(D)$

In lossy compression, we are given

1. The spaces ("alphabets") of data and reproductions, $(\mathcal{X}, \mathcal{Y})$.
2. The source (data) distribution μ (a prob. measure) on $\mathcal{X}$.
3. A distortion function ρ: $\mathcal{X} \times \mathcal{Y} \to [0, \infty)$

A lossy compression algorithm maps the *source* measure μ on $\mathcal{X}$ to a *reproduction* measure ν on $\mathcal{Y}$, incurring

- a **distortion**/transportation cost ("reconstruction error") and
- a **rate** cost ("avg. file size").



Q: what is the best possible rate-distortion trade-off?
A: the rate-distortion function $R(D)$.

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

Following [Blahut 1972, Arimoto 1972], we work with an equivalent variational "Lagrangian" representation of R(D):

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

Connections to EOT and denoising

We show that the Lagrangian R-D problem (1) is equivalent to:

(2) projecting the source measure under an entropic optimal transport (EOT) cost;

(3) denoising/deconvolving the source by maximum-likelihood.

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

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

[Rigollet and Weed, 2018]

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

$Y \to X$

$Y \sim \nu^*$

$X|_{Y=y} \sim \mathcal{N}(y, \frac{1}{\lambda})$

Thus our algorithm/results transfer across all three problems.

In particular:

- (1) $\leftrightarrow$ (2): we leverage sample complexity results for EOT [Mena and Niles-Weed, 2019] to obtain finite-sample bounds for R-D estimation / projection under EOT / maximum-likelihood deconvolution;
- (1) $\leftrightarrow$ (3): we leverage the solution to the deconvolution problem to introduce a new class of sources with closed-form R(D) segments.

Proposed: Wasserstein Gradient Descent (WGD)

Let $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, ρ continuously differentiable. We aim to solve

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

Idea: simulate the gradient flow of $\mathcal{L}$ in the 2-Wasserstein space of probability measures:

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

Particle scheme in practice:

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

The Wasserstein gradient can be tractably computed by

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

Finite-sample bounds for R-D estimation

Given m samples from a σ^2-sub-Gaussian source, the best empirical loss achievable with n particles converges to the true optimum as follows (ϵ=$1/\lambda$):

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

Empirical results

- We compare against other R-D upper bound algorithms: Blahut-Arimoto [Blahut 1972; Arimoto 1972] and SOTA deep learning methods RD-VAE [Yang & Mandt 2022] and NERD [Lei et al. 2023].
- For a given per-iteration compute budget, we obtain much faster convergence and better approximation quality (deconv example):

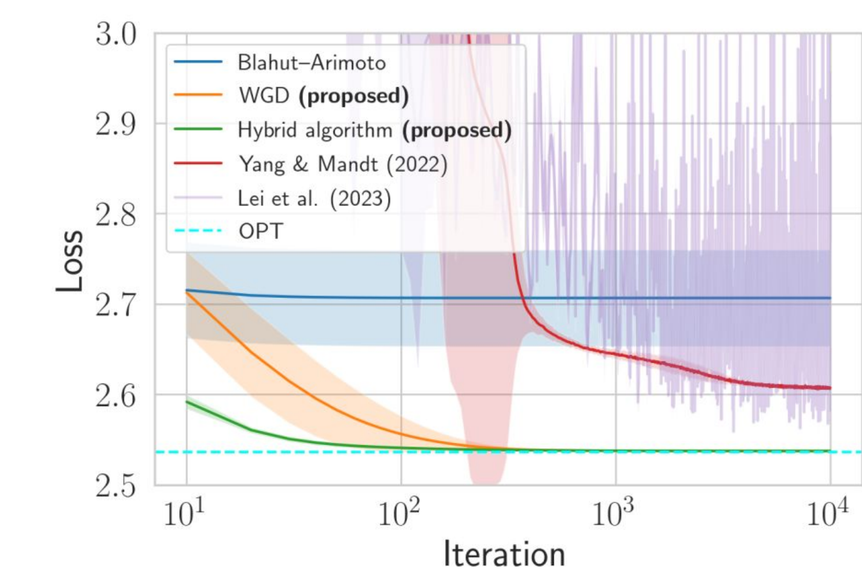


Figure 2: Losses over iterations. Shading corresponds to one standard deviation over random initializations.

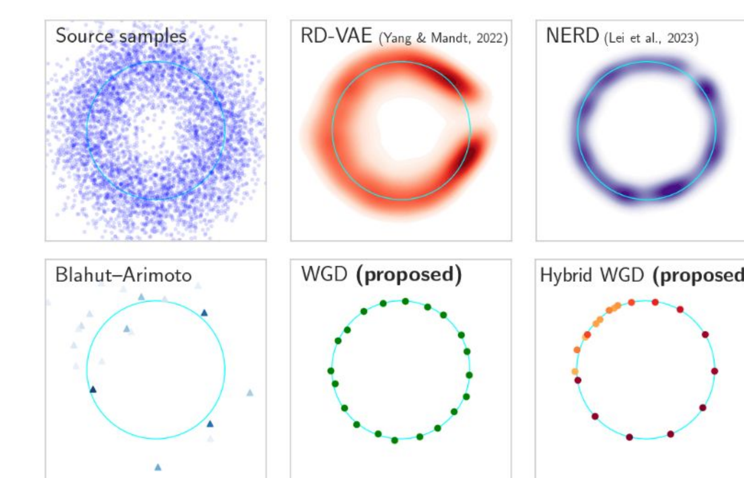
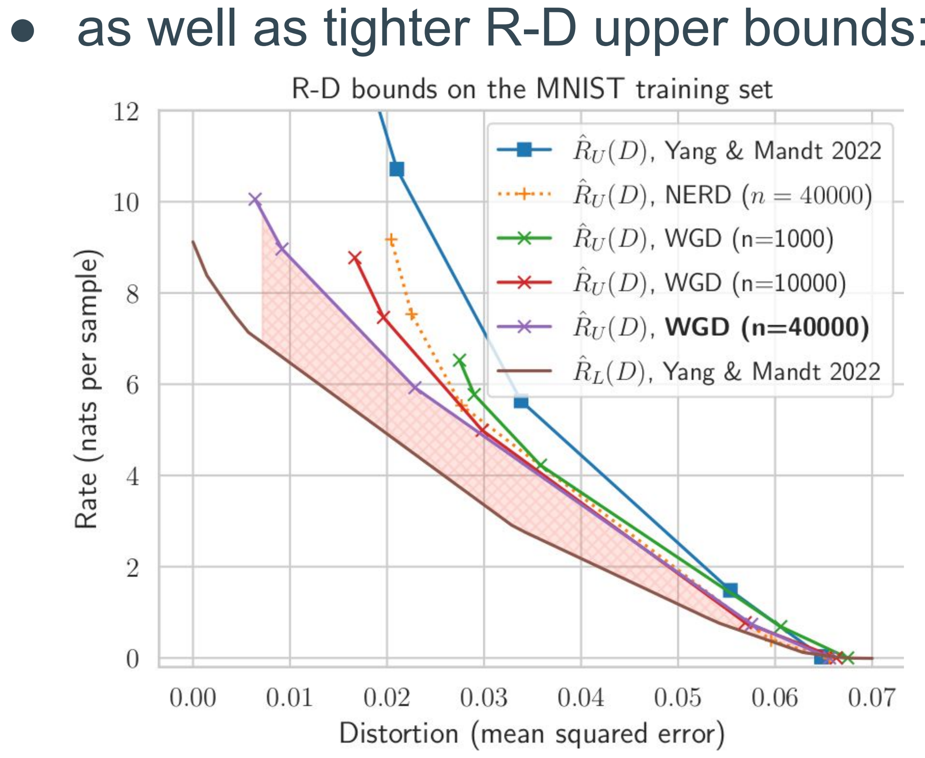
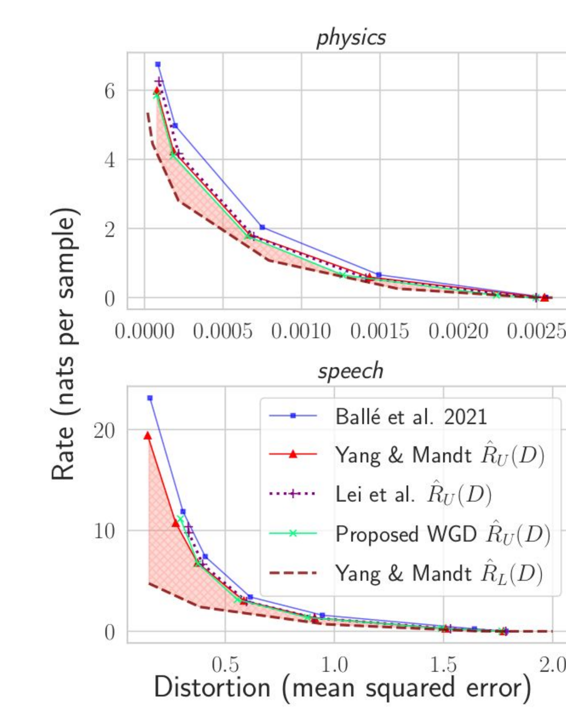


Figure 3: Visualizing μ samples (top left), as well as the ν returned by various algorithms compared to the ground truth ν^* (cyan).

- as well as tighter R-D upper bounds:





Limitations and future work

- Like NERD [Lei et al. 2023], our method can only target an R(D) point with a rate ≤ log(n) nats/sample, where n = number of particles.
- It remains to be studied how best to convert our R-D estimator into a practical data communication/compression algorithm.

References

[Blahut 1972] R. Blahut. "Computation of channel capacity and rate-distortion functions". *IEEE Transactions on Information Theory*, 18(4):460–473, 1972.
[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.
[Rigollet and Weed, 2018] Philippe Rigollet and Jonathan Weed. Entropic optimal transport is maximum-likelihood deconvolution. Comptes Rendus Mathematique, 356(11-12):1228–1235, 2018.
[Mena and Niles-Weed, 2019] 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.
[Yang & Mandt 2022] Y. Yang, S. Mandt. Towards empirical sandwich bounds on the rate-distortion function",
International Conference on Learning Representations, 2022
[Lei et al. 2023] E. Lei, H. Hassani, and S. Saeedi Bidokhti. "Neural estimation of the rate-distortion function with applications to operational source coding". *IEEE Journal on Selected Areas in Information Theory*, 2023.