AN OPTIMAL DIFFUSION APPROACH TO QUADRATIC RATE-DISTORTION PROBLEMS: NEW SOLUTION AND APPROXIMATION METHODS

Anonymous authorsPaper under double-blind review

ABSTRACT

When compressing continuous data, it is inevitable to suffer some loss of information, which occurs in some distortion at reconstruction. The Rate-Distortion function is the lowest rate possible for a code whose decoding allows a given amount of such distortion. We exploit the connection between rate-distortion (RD) and entropic optimal transport to propose a novel stochastic control formulation, and use a classic result dating back to Schrödinger to show that the tradeoff between rate and mean squared error distortion is equivalent to a tradeoff between control energy and the differential entropy of the terminal state, whose probability law yields the reconstruction distribution. For a special class of sources, we show that the optimal control law and trajectory in the space of probability measures are given by solving a backward heat equation. In the more general case, our approach gives rise to a numerical solution method, estimating the RD function using diffusion processes with a constant diffusion coefficient. We demonstrate our method in various examples.

1 Introduction

As is well known from information theory (Shannon, 1948; Cover, 1999), any probabilistic data source cannot be compressed in a lossless manner at an encoding rate lower than its entropy. Rate distortion (RD) theory (Shannon, 1959; Berger, 2003; Cover, 1999) complements this result by considering the effect of lower encoding rates on the reconstruction of the decoder. This reconstruction is inevitably lossy: Given a distortion measure between a pair of data samples, this loss is quantified by the total accumulated distortion between the data source samples and the reconstruction samples produced by the decoder from the compressed bits sent by the encoder. The *operational RD function* thus characterizes the minimal encoding rate required for a given (average) distortion level *D*, under the prescribed distortion measure, in the limit of a large number of samples.

For independent and identically distributed (i.i.d.) samples, the celebrated lossy source-coding theorem of information theory (Shannon, 1959; Berger, 2003; Cover, 1999) establishes that the operational RD function is equal to the informational RD function R(D), where the latter is expressed as a single-letter expression (though a single sample may be a vector on its own). To wit, it is given as minimization of the mutual information (MI) between a random variable X, representing a single sample from the source, and a random variable \hat{X} , representing a single reconstruction sample. This minimization is over the test channel – the conditional distribution of \hat{X} given X, under the constraint that the average distortion between X and \hat{X} is below D (see § 2).

The optimization problem involved in the computation of R(D) should thus be solved for the given distortion measure and the required distortion level. However, even though being presented for almost 70-years (Shannon, 1959), closed-form solutions of this problem are known only for a few canonical examples, such as binary sources under the Hamming distortion, or a Gaussian source under the mean squared error (MSE) Cover (1999). It is therefore necessary to develop efficient algorithms for the computation of the RD function. When the data source alphabet is discrete and finite, the classical Blahut–Arimoto (BA) algorithm (Arimoto, 1972; Blahut, 1972) is such an algorithm, especially effective for small alphabets. However, the typical setup for lossy compression is that of a *continuous* data source, and in this setting, efficient computation of the RD function is a

long-standing challenge. Recently, Lei et al. (2023); Yang et al. (2024) identified an interesting connection between the BA algorithm and entropy-regularized Optimal Transport (OT), which allows to approximate the RD function in cases where BA is intractable (Yang et al., 2024). OT (Ambrosio et al., 2008) is a widespread and useful formulation, which lies at the theoretical basis of training deep generative models, such as Generative Adversarial Networks (GANs) (Arjovsky et al., 2017) and variational auto-encoders (VAEs) (Tolstikhin et al., 2017), and defined as follows: Given a pair of distributions, the OT objective is to find an optimal *plan* between them, namely a joint distribution with the given marginals (coupling), minimizing some metric. The *entropy-regularized* optimal transport (*a.k.a. weak* OT, *entropic* OT, or EOT) was suggested by Cuturi (2013) as an approximation to OT, for which efficient solution methods exist, *e.g.*, Sinkhorn's algorithm (Altschuler et al., 2017). More recently, Gushchin et al. (2022) suggested to find the optimal plan of an EOT problem using diffusion processes. Their method is based on a reduction of the EOT problem to a stochastic control problem known as the *Schrödinger bridge* (SB) (Schrödinger, 1932; Chen et al., 2021).

As we state later, a solution to SB can be written as a *finite-energy diffusion processs*. Diffusion processes are popular for generative modeling Ho et al. (2020); Song et al. (2020), where a sample from a data distribution is gradually drifted and becomes more noisy towards a completely noisy data sample, typically Gaussian. The celebrated paper of Song et al. (2021) suggested finding the *drift* term of the model by learning the *score* function and plugging it back to the reverse stochastic differential equation (SDE) (Anderson, 1982).

In this paper we focus on the computation of the RD function for continuous data sources, under the MSE distortion. This leads to an OT problem with the quadratic cost. More specifically, we exploit the connection between RD and EOT to propose a novel stochastic control formulation to RD, where the classical result of Schrödinger (1932) implies that the tradeoff between rate and MSE distortion is equivalent to a tradeoff between a controller *energy* and the *differential entropy* of the terminal state, whose probability law yields the *reconstruction* distribution. For a special class of sources, we show that the optimal control law and trajectory in the space of probability measures are given by solving a *Backward* Heat Equation (BHE). In the more general case, our approach gives rise to a numerical solution method where the RD function is estimated using diffusion processes, with a constant diffusion coefficient.

Our contributions: (1) We present Terminal-Entropy Control (TEC), a novel stochastic control formulation that is regularized by terminal uncertainty, and show that this formulation is equivalent to the RD problem. (2) We characterize the optimal solution to TEC under some regularity conditions. (3) We find a closed-form solution for the reconstruction distribution of a Gaussian-mixture source. We also demonstrate our approach on other mixture distributions, with non-Gaussian components, via Fourier analysis. To the best of our knowledge, such results were unknown to date, which emphasizes the unique theoretical contribution of our approach. (4) Based on our new approach we propose R2D2, a novel neural method for estimating the RD function and the reconstruction distributions, using a simple diffusion model.

1.1 RELATED WORK

Neural estimation of information-theoretic quantities Following the ongoing research of neural methods in AI and ML, a plethora of methods have emerged in recent years for estimating information measures such as the RD function, as well as for the design of optimal compression methods aiming to achieve these fundamental limits (see survey in Yang et al. (2023)). Belghazi et al. (2018), for example, used the Donsker–Varadhan identity to estimate MI. Kholkin et al. (2025) used samples from Brownian bridges to estimate both the MI between datasets, as well as differential entropies. Lei et al. (2022) suggested approximating the RD function using a deep neural network (DNN) model, and proposed an operational coding scheme. Tsur et al. (2024) suggested approximating RD by modeling unknown input distributions, both continuous and discrete. Recently, Lei et al. (2023); Yang et al. (2024) pointed out an intriguing connection between the BA algorithm and EOT. This connection was further exploited in Yang et al. (2024) to estimate the RD function using a discrete approximation of reconstruction probability law. Finally, Zou et al. (2025) used the connection between EOT and the Schrödinger problem to characterize this density. In this paper, we approximate the reconstruction distribution in RD problems with MSE loss, *i.e.*, a quadratic cost, using a *continuous* diffusion model.

Diffusion models for lossy compression Lossy compression methods that hinge on diffusion processes have become popular in recent years. In Elata et al. (2025) pre-trained diffusion models were sampled for zero-shot image compression. Recently, Ohayon et al. (2025) replaced the noise at every time-step of the reverse diffusion with samples from a sequence of pre-defined codebooks, achieving high perceptual quality at reconstruction. In Theis et al. (2022) a noisy version of the source data was compressed, and then used in a reverse diffusion model for reconstruction. Here, we use a forward model to achieve the reconstruction distribution, and compute both its rate and distortion.

EOT and the Schrödinger Bridge Recently, Gushchin et al. (2022) has drawn an equivalency between SB and EOT. They offered a game-theoretic formulation for the former, which can be solved by optimizing a diffusion model. In § 3 we suggest a slightly modified stochastic control problem, in which the target probability is free, but the total objective is penalized for being uncertain. We show (Thm. 3.1) that this formulation is equivalent to the RD problem under the MSE distortion.

Entropy-regularized stochastic control In addition to presenting a novel approach to the RD problem, our results can also be considered as a contribution to control theory. As we state in § 2, SB can be formulated as a stochastic control problem. This form, as well as its KL-equivalence, are given in Chen et al. (2021). Entropy regularization is a common practice in stochastic control (Lambert et al., 2025) and reinforcement learning (Haarnoja et al., 2018; Ziebart et al., 2008). However, most studies aim to *maximize* the control policy entropy, encouraging diverse actions. Alternatively, Fridman and Shaked (2000) proposed to minimize the steady-state entropy of closed-loop linear systems in H_{∞} control problems with infinite horizons. Here, we focus on *penalizing* the terminal-state uncertainty, leading to a novel tradeoff between energy and entropy. Surprisingly, maybe, to the best of our knowledge, this is the first work to present this formalism, which may open the door for additional applications in the broader field of stochastic control.

2 Preliminaries

Rate-Distortion theory Let $X \sim p_X = \mathbb{P}_0 \in \mathcal{P}(\mathbb{R}^d)$ for $d \geq 1$ denote a single sample from the source, where $\mathcal{P}(\mathbb{R}^d)$ is the set of probability measures on \mathbb{R}^d , and where we assume that different source samples are i.i.d. Let $\hat{X} \in \mathbb{R}^d$ denote a reconstruction sample, where the pair follows the joint distribution $(X,\hat{X}) \sim p_{X\hat{X}} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$. Further let $p_{\hat{X}|X}$ denote the induced conditional distribution, which is also called the *reconstruction law*. Let $d(\cdot,\cdot) \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ denote a distortion measure between $x \in \mathbb{R}^d$ and $\hat{x} \in \mathbb{R}^d$, and let the average distortion be $D(\hat{X},X) \triangleq \mathbb{E}[d(\hat{X},X)]$, which is an implicit function of $p_{X\hat{X}}$. For a given pair of probability measures $\mathbb{P}_0, \mathbb{P}_1 \in \mathcal{P}(\mathbb{R}^d)$, let $D_{\mathrm{KL}}(\mathbb{P}_0||\mathbb{P}_1) \triangleq \mathbb{E}_{X \sim \mathbb{P}_0}[\log \frac{d\mathbb{P}_0}{d\mathbb{P}_1}(X)]$ denote the Kullback–Leibler divergence, and let the mutual information $\mathcal{I}(X;\hat{X}) = D_{\mathrm{KL}}(p_{X\hat{X}}||p_X \otimes p_{\hat{X}})$. The lossy compression theorem of information theory (Shannon, 1959; Berger, 2003; Cover, 1999) then states that the operational RD function is equivalent to the informational RD function, given by

$$R(D) \triangleq \min_{p_{\hat{X}|X}: D(\hat{X},X) \le D} \mathcal{I}(X;\hat{X}). \tag{1}$$

The BA algorithm (Cover, 1999) computes R(D) by optimizing the Lagrangian (with a Lagrange multiplier $\epsilon > 0$)

$$L_{BA}(p_{\hat{X}|X}, \mathbb{P}_0, \epsilon) = D(\hat{X}, X) + \epsilon \mathcal{I}(\hat{X}; X)$$
(2)

w.r.t $p_{\hat{X}|X}$. Alternatively, if we let $\mathbb{P}_1 = p_{\hat{X}}$ be the marginal distribution of the reconstruction (Yang et al., 2024), then an equivalent formulation is

$$\min_{p_{\hat{X}|X}} L_{BA}(p_{\hat{X}|X}, \mathbb{P}_0, \epsilon) = \min_{\mathbb{P}_1 \in \mathcal{P}(\mathbb{R}^d)} \min_{(X, \hat{X}) \sim \pi \in \Pi(\mathbb{P}_0, \mathbb{P}_1)} \left\{ D(\hat{X}, X) + \epsilon D_{\text{KL}}(\pi || \mathbb{P}_0 \otimes \mathbb{P}_1) \right\}, \quad (3)$$

where $\Pi(\mathbb{P}_0, \mathbb{P}_1)$ is the set of all couplings, that is, the set of joint distributions $p_{X\hat{X}} \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ whose X-marginal (resp. \hat{X} -marginal) is \mathbb{P}_0 (resp. \mathbb{P}_1). In this work, we focus on the quadratic cost $d(\hat{x}, x) = \frac{1}{2} \|\hat{x} - x\|^2$. The average distortion is then given by $D(\hat{X}, X) \triangleq \frac{1}{2} \mathbb{E}[\|X - \hat{X}\|^2]$, and the

minimization problem Eq. (3) reads

$$\min_{p_{\hat{X}|X}} L_{BA}(p_{\hat{X}|X}, \mathbb{P}_0, \epsilon) = \min_{\mathbb{P}_1} \min_{(X, \hat{X}) \sim \pi \in \Pi(\mathbb{P}_0, \mathbb{P}_1)} \left\{ \frac{1}{2} \mathbb{E}_{\pi} \left[\|X - \hat{X}\|^2 \right] + \epsilon D_{\text{KL}}(\pi || \mathbb{P}_0 \otimes \mathbb{P}_1) \right\}. \tag{4}$$

Entropic optimal transport For a probability measure $\mathbb{P} \in \mathcal{P}(\mathbb{R}^d)$ with density p(x), $H(\mathbb{P}) \triangleq -\mathbb{E}_{X \sim \mathbb{P}} \log(p(X))$ denotes its *differential* entropy. Now, considering probability measures $\mathbb{P}_0, \mathbb{P}_1 \in \mathcal{P}(\mathbb{R}^d)$, the EOT problem (Cuturi, 2013) is given by

$$\inf_{\pi \in \Pi(\mathbb{P}_0, \mathbb{P}_1)} \left\{ \int_{\mathbb{R}^d \times \mathbb{R}^d} \frac{||x - \hat{x}||^2}{2} d\pi(x, \hat{x}) + \epsilon D_{\mathrm{KL}}(\pi || \mathbb{P}_0 \otimes \mathbb{P}_1) \right\}. \tag{5, EOT}$$

As one may readily recognize, for values of ϵ where the optimal reconstruction has a density, the inner minimization of Eq. (4) coincides with Eq. (5, EOT) (Lei et al., 2023; Yang et al., 2024).

The Schrödinger Bridge The SB problem (Schrödinger, 1932) with parameter ϵ is formulated as

$$\inf_{u} \frac{1}{2} \mathbb{E} \left[\int_{0}^{1} \|u(X_{t}, t)\|^{2} dt \right] \text{ s.t.} \begin{cases} X_{0} \sim \mathbb{P}_{0}, \ X_{1} \sim \mathbb{P}_{1} \\ dX_{t} = u(X_{t}, t) dt + \sqrt{\epsilon} dW_{t} \end{cases} , \tag{6, SB}$$

where $\mathbb{P}_0, \mathbb{P}_1$ are absolutely continuous probability measures w.r.t. the Lebesgue measure on \mathbb{R}^d , and W_t is a standard Wiener process, independent of X_0 . The $drift\ u: \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ can be seen as a controller, designed to steer $X_0 \sim \mathbb{P}_0$ into $X_1 \sim \mathbb{P}_1$ with $minimal\ average\ energy$, acting against a 'natural' force W_t . Recent developments (Gushchin et al., 2022) has drawn an equivalency (up to an additive constant, depends on $\mathbb{P}_0, \mathbb{P}_1, \epsilon$) between SB and EOT, where the latter can be optimized via a game-theoretic formulation. The optimal joint probability π in Eq. (5, EOT) is then given by the joint probability law of (X_0, X_1) .

3 RATE-DISTORTION FUNCTIONS AND TERMINAL-ENTROPY STOCHASTIC CONTROL

3.1 PROBLEM STATEMENT

Let the source $X_0 \sim \mathbb{P}_0$ where $\mathbb{P}_0 \in \mathcal{P}_2(\mathbb{R}^d)$ is absolutely continuous *w.r.t*. Lebesgue measure, and $\mathcal{P}_2(\mathbb{R}^d)$ is the subset of $\mathcal{P}(\mathbb{R}^d)$ for which $\mathbb{E}\left[\|X\|^2\right] < \infty$. We further assume that $H(\mathbb{P}_0)$ is finite. As in the BA algorithm, we aim to compute the RD Largrangian

$$\mathcal{L}_{RD}(\mathbb{P}_{0}, \epsilon) \triangleq \min_{\mathbb{P}_{1}} \min_{\pi \in \Pi(\mathbb{P}_{0}, \mathbb{P}_{1})} \left\{ \frac{1}{2\epsilon} \int_{\mathbb{P}^{d} \times \mathbb{P}^{d}} \|\hat{x} - x\|^{2} d\pi(x, \hat{x}) - H(\pi) + H(\mathbb{P}_{1}) \right\} + H(\mathbb{P}_{0}), (7, RD)$$

where from standard decomposition of MI to entropy terms (Cover, 1999), $D_{\mathrm{KL}}(\pi||\mathbb{P}_0 \otimes \mathbb{P}_1) = H(\mathbb{P}_1) + H(\mathbb{P}_0) - H(\pi)$. Here, $\epsilon > 0$ is a tuning parameter and $H(\mathbb{P}_0)$ is not subject to optimization.

Assumption A1: Eq. (7, RD) admits a solution, where the optimal reconstruction distribution is absolutely continuous and satisfies $\mathbb{P}_1 \in \mathcal{P}_2(\mathbb{R}^d)$, with finite differential entropy $H(\mathbb{P}_1)$.

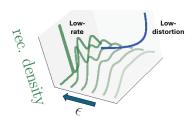
While **A1** is necessarily violated for low rates (high ϵ values), in many cases it is still expected to be in effect in the low-distortion regime, where \mathbb{P}_1 is close to \mathbb{P}_0 , as illustrated in Fig. 1. This holds true in a variety of settings, as we demonstrate in § 3.3 ,§ 5.

Let $\mathrm{d}W^\epsilon = \sqrt{\epsilon}\mathrm{d}W_t,\ W^\epsilon_0 \sim \mathbb{P}_0$ be the (scaled) Brownian motion starting at \mathbb{P}_0 , and denote $\pi^\epsilon = \mathbb{P}_{W^\epsilon_0,W^\epsilon_1}$, the joint law of its start and end points. It is known (Gushchin et al., 2022; Chen et al., 2021) that

$$D_{\mathrm{KL}}(\pi||\pi^{\epsilon}) = \frac{1}{2\epsilon} \int_{\mathbb{P}^d \times \mathbb{P}^d} \|\hat{x} - x\|^2 \mathrm{d}\pi - H(\pi) + H(\mathbb{P}_0) + \frac{d}{2} \log(2\pi\epsilon)$$
 (8)

for every π with marginal distribution \mathbb{P}_0 . Let $\mathcal{F}(\mathbb{P}_0, \mathbb{P}_1)$ be the class of random trajectories $T_t \in \mathbb{R}^d$, $t \in [0, 1]$ with a joint distribution $\pi_T = \mathbb{P}_{T_0, T_1} \in \Pi(\mathbb{P}_0, \mathbb{P}_1)$. For every $T \in \mathcal{F}(\mathbb{P}_0, \mathbb{P}_1)$,

$$D_{\mathrm{KL}}(T||W^{\epsilon}) = D_{\mathrm{KL}}(\pi_T||\pi^{\epsilon}) + \int_{\mathbb{R}^d \times \mathbb{R}^d} D_{\mathrm{KL}}(T|_{x,\hat{x}}||W^{\epsilon}|_{x,\hat{x}}) \mathrm{d}\pi_T(x,\hat{x}), \tag{9}$$



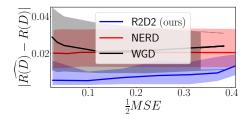


Figure 1: Evolution of reconstruction distri- Figure 2: Estimation error on a 1-D Gaussian butions w.r.t. parameter ϵ . Typically, we assource. $X_0 \sim \mathcal{N}(0,1)$ and we applied R2D2 sume the reconstruction density to be close to the (Alg. 1) to $\epsilon \in [0.05, 0.95]$. We compare our continuous source in the low distortion regime algorithm to NERD, where we observe the im-(small ϵ 's), while eventually becoming singular proved accuracy of R2D2 over existing methods for a low enough rate (large ϵ 's).

where there exists a process $T_{\mathbb{P}_1}^*$, minimizing $D_{\mathrm{KL}}(T||W^\epsilon)$ over $\mathcal{F}(\mathbb{P}_0,\mathbb{P}_1)$ with $D_{\mathrm{KL}}(T_{\mathbb{P}_1}^*|_{x,\hat{x}}||W^\epsilon|_{x,\hat{x}})=0$ for all $x,\hat{x}\in\mathbb{R}^d$ (Léonard, 2013, Prop. 4.1,2.3). The process $T_{\mathbb{P}_1}^*$ is known to take the form $T_{u_{\mathbb{P}_1}}^*$ of an Itô diffusion (Gushchin et al., 2022)

$$T_u: dX_t = u(X_t, t)dt + \sqrt{\epsilon}dW_t,$$
 (10)

with drift $u_{\mathbb{P}_1}$, where $\mathbb{E}\left[\int_0^1 \|u_{\mathbb{P}_1}(X_t,t)\|^2 dt\right] < \infty$. Furthermore, for such a finite-energy process we have (Pavon and Wakolbinger, 1991),

$$D_{\mathrm{KL}}(T_{\mathbb{P}_1}^*||W^{\epsilon}) = \frac{1}{2\epsilon} \mathbb{E}\left[\int_0^1 \|u_{\mathbb{P}_1}(X_t, t)\|^2 \mathrm{d}t\right]. \tag{11}$$

Considering Eq. (7, RD)-Eq. (11), we suggest the surrogate loss

$$\tilde{\mathcal{L}}_{RD}(\mathbb{P}_0, \epsilon) \triangleq \min_{\mathbb{P}_1} \min_{T_u \in \mathcal{F}(\mathbb{P}_0, \mathbb{P}_1)} \left\{ \frac{1}{2\epsilon} \mathbb{E} \left[\int_0^1 \|u(X_t, t)\|^2 dt \right] + H(\mathbb{P}_1) \right\} - \frac{d}{2} \log(2\pi\epsilon), \quad (12)$$

where T_u is a finite-energy diffusion Eq. (10) with drift $u \in \mathcal{U} \triangleq \{u(x,t) : H(\mathbb{P}_1) > -\infty\}$, namely \mathbb{P}_1 has a finite differential entropy. The above discussion leads to the following equivalency between Eq. (7, RD) and Eq. (12).

Theorem 3.1. Given \mathbb{P}_0 and ϵ , under AI we have $\mathcal{L}_{RD} = \tilde{\mathcal{L}}_{RD}$. Furthermore, let $u^*(x,t)$ minimize the surrogate objective

$$u^* \in \arg\min_{u \in \mathcal{U}} \left\{ \frac{1}{2\epsilon} \mathbb{E} \left[\int_0^1 \|u(X_t, t)\|^2 dt \right] + H(X_1) \right\},\tag{13}$$

under the law Eq. (10). Then, the distribution $\mathbb{P}_{X_1^*}$ of X_1^* associated with u^* through Eq. (10) is the minimizer in Eq. (12), where $\mathbb{P}_1^* = \mathbb{P}_{X_1^*}$ is the optimal reconstruction distribution in Eq. (7, RD) and $\pi^* = \mathbb{P}_{X_0^*, X_1^*}$ is the optimal plan.

The proof is given in App. C, where we also establish the opposite direction; whenever (\mathbb{P}_1, π) minimizes Eq. (7, RD), there exists a drift term u(x,t) minimizing Eq. (13) under Eq. (10), where $(X_0, X_1) \sim \pi$.

3.2 TERMINAL-ENTROPY STOCHASTIC CONTROL: THE ENERGY-ENTROPY TRADEOFF

Motivated by Thm. 3.1, we present the problem of Terminal-Entropy regularized stochastic Control

$$\inf_{u \in \mathcal{U}} \left\{ \frac{1}{2} \mathbb{E} \left[\int_0^1 \|u(X_t,t)\|^2 \mathrm{d}t \right] + \epsilon H(X_1) \right\} \text{ s.t. } \begin{cases} X_0 \sim \mathbb{P}_0, \ \mathbb{P}_{X_1} \text{ is } \textit{free} \\ \mathrm{d}X_t = u(X_t,t) \mathrm{d}t + \sqrt{\epsilon} \mathrm{d}W_t \end{cases}, \quad (14, \text{TEC})$$

where the admissible control set is again $\mathcal{U} = \{u(x,t) : H(\mathbb{P}_1) > -\infty\}$. In light of Eq. (6, SB), here u can be viewed as a control law for reducing the terminal state uncertainty, while spending

minimal energy. As a consequence of Thm. 3.1, in order to estimate the RD function for the source $X_0 \sim \mathbb{P}_0$, one should solve Eq. (14, TEC) with a dynamic range of ϵ values. Given the drift term u, X_1 can be efficiently sampled (e.g. using the Euler–Maruyama algorithm, see App. A). Taking Assumption A1 into account, Eq. (14, TEC) takes the variational form

$$\inf_{u \in \mathcal{U}} \left\{ \frac{1}{2} \int_{\mathbb{R}^d} \mathrm{d}x \int_0^1 \mathrm{d}t \|u(x,t)\|^2 p_t(x) - \epsilon \int_{\mathbb{R}^d} \mathrm{d}x p_1(x) \log p_1(x) \right\},\tag{15, var-TEC}$$

whereas for a diffusion process Eq. (10), the density state $p_t(x)$ is governed by the Fokker-Planck equation (Oksendal, 2013)

$$\frac{\partial}{\partial t} p_t(x) = -\nabla \cdot (p_t(x)u(x,t)) + \frac{1}{2} \epsilon \Delta_{xx} p_t(x), \ p_0(x) = \mathbb{P}_0(x), \tag{16, FPE}$$

with $\nabla \cdot = \sum \frac{\partial}{\partial x_i} e^{(i)}$ being the *divergence* operation, and $\Delta_{xx} = \sum \frac{\partial^2}{\partial x_i^2}$ is the *Laplace* operator. Interestingly, under suitable regularity conditions, the solution to Eq. (14, TEC) can be characterized by a simple equation, as we show next:

Theorem 3.2. Let $p_t^*(x) \in C^{1,2}([0,1] \times \mathbb{R}^d)^1$ satisfy the backward heat equation (BHE)

$$\frac{\partial}{\partial t} p_t^*(x) = -\frac{1}{2} \epsilon \Delta_{xx} p_t^*(x), \ p_0^*(x) \sim \mathbb{P}_0, \tag{17. BHE}$$

such that $\log p_t^*(x) \in \mathcal{C}^{1,2}([0,1] \times \mathbb{R}^d)$ and $\|\nabla p_t^*(x)\| \log p_t^*(x) \to 0$ as $\|x\| \to \infty$ for all $t \in [0,1]$. Let $u^* = \epsilon \nabla \log p_t^*(x)$, where $\nabla \log p_t(x)$ is the Stein score function. Then, (u^*, p_t^*) is an optimal pair in Eq. (15, var-TEC)-Eq. (16, FPE) and the optimal solution to Eq. (14, TEC) admits the SDE

$$dX_t^* = \epsilon \nabla \log p_t^*(X_t) dt + \sqrt{\epsilon} dW_t. \tag{18}$$

The proof is given in App. C.

Fourier analysis of Eq. (17, BHE) Let us assume d=1 for simplicity; similar arguments hold in higher dimensions. Let the source $X_0 \sim \mathbb{P}_0$ in \mathbb{R} with density p_0 and characteristic function $\hat{p}(\omega)$, namely $p_0(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{i\omega x} \hat{p}(\omega) d\omega$. It is easy to verify that a solution to Eq. (17, BHE) is given by

$$p_t(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{i\omega x + \frac{1}{2}\epsilon\omega^2 t} \hat{p}(\omega) d\omega,$$
 (19)

whenever the latter converges for all $t \in [0,1]$ and $\nabla \log p_t(x)$ is defined for all $x, t \in \mathbb{R} \times [0,1]$.

3.3 SPECIAL CASES

Backward heat conductance problems are generally ill-posed and unstable (Miranker, 1961; Fu et al., 2007). However, Thm. 3.2 yields an exact solution for some special cases, as we show next.

Gaussian source We begin with the canonical example of a scalar Gaussian source $\mathbb{P}_0 = \mathcal{N}\left(0,\sigma_0^2\right)$, and show how our formulation recovers its known RD function. For $\epsilon < \sigma_0^2$, a solution to Eq. (17, BHE) is given by $p_t(x) = \frac{1}{\sqrt{2\pi(\sigma_0^2 - \epsilon t)}} e^{-\frac{x^2}{2(\sigma_0^2 - \epsilon t)}}$. The optimal controller is hence given by $u(x,t) = \epsilon \nabla \log p_t(x) = -\frac{\epsilon}{\sigma_0^2 - \epsilon t} x$. Evidently, under u(x,t), X_0 and X_1 are jointly-Gaussian

where $D = \frac{1}{2}\mathbb{E}[(X_0 - X_1)^2] = \frac{1}{2}\epsilon$, and $R = \mathcal{I}(X_0; X_1) = -\frac{1}{2}\log\left(\frac{\epsilon}{\sigma_0^2}\right)$, and we recover the known closed-form result

$$R_{\text{Gauss}}(D) = \frac{1}{2} \log \left(\frac{\sigma_0^2}{2D} \right), \ \ 0 < 2D < \sigma_0^2.$$
 (20)

Note that the factor of 2 is due to our distortion definition has half of the MSE. This result can be easily generalized for Gaussian vector sources and ϵ 's smaller than the eigenvalues of the covariance matrix Σ . In this case, the solution to Eq. (17, BHE) is $p_t(x) = \mathcal{N}(0, \Sigma - \epsilon tI)$.

 $^{{}^{1}\}mathcal{C}^{1,2}([0,1]\times\mathbb{R}^{d})$ is the set of functions that are continuously differentiable w.r.t. t, and twice continuously differentiable w.r.t. x.

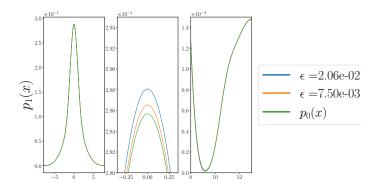


Figure 3: Reconstruction distribution $p_1(x)$ of non-Gaussian mix source. In the (left) pane, we approximate reconstruction distribution Eq. (23) by numerical integration. A closer look at different points is provided in the (middle) and (right) panes.

Mixture of Gaussians The Gaussian example can be easily extended to the case of a *Gaussian mixture*, for which no closed-form solution for the RD function is known. In this case,

$$p_0(x) = \sum_{i=1}^{N} \frac{p_i}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}}.$$
 (21)

As Eq. (17, BHE) is a linear equation, the solution is now given by the superposition

$$p_t(x) = \sum_{i=1}^{N} \frac{p_i}{\sqrt{2\pi(\sigma_i^2 - \epsilon t)}} e^{-\frac{(x - \mu_i)^2}{2(\sigma_i^2 - \epsilon t)}}, \ \epsilon \in \left(0, \min_i \sigma_i^2\right). \tag{22}$$

The optimal controller $u(x,t) = \epsilon \nabla \log p_t(x)$ is derived accordingly. Knowing $p_t(x)$ and u(x,t), it is possible to acquire rate and distortion values through a Monte-Carlo simulation of Eq. (14, TEC), or via neural estimation as we suggest in § 4.

Non-Gaussian mixture To show the wide applicability of Thm. 3.2, we now apply its result to a mixture of sinc^4 functions. Being band-limited, for settings where $p_0(x) > 0$, such a source satisfies the conditions of the theorem, and allows the desirable frequency-domain analysis of § 3.2. We emphasize that while this is a toy problem, to the best of our knowledge, no other approach is known to tackle this case.

We consider the source X_0 drawn from the mixture $p_0(x) = \sum_{i=1}^N p_i C_i^{-1} \operatorname{sinc}^4(\frac{x}{m_i})$ where $\operatorname{sinc}(x) \triangleq \frac{\sin(x)}{x} \in \mathcal{C}^{\infty}(\mathbb{R})$, and $C_i = \frac{2\pi}{3}m_i$ are appropriate normalization factors. In this case, the characteristic function is $\hat{p}_0(\omega) = \sum_{i=1}^N p_i C_i^{-1} m_i \tilde{p}_0(m_i \omega)$ where $\tilde{p}_0(\omega) = \frac{1}{2\pi} \left[\pi(1-\frac{1}{2}|\omega|)_+ *\pi(1-\frac{1}{2}|\omega|)_+\right]$, and * is the *convolution* operation. Now, for non-vanishing mixture distributions and small enough ϵ 's, the Fourier analysis of Eq. (19) implies that we can write the solution to Eq. (17, BHE) as

$$p_t(x) = \frac{1}{2\pi} \int_{-4}^4 e^{i\omega x + \frac{1}{2}\epsilon\omega^2 t} \hat{p}_0(\omega) d\omega = \frac{2}{2\pi} \sum_{i=1}^N p_i C_i^{-1} m_i \int_0^4 \tilde{p}_0(m_i \omega) \cos(\omega x) e^{\frac{1}{2}\epsilon\omega^2 t} d\omega.$$
 (23)

Fig. 3 demonstrates this result for $N=4, m_i=\left[1,\sqrt{2},\pi,e\right]$ and $p_i=\frac{1}{4}$, where we numerically integrated Eq. (23) in order to approximate the reconstruction distribution $p_1(x)$ for different values of ϵ . Parameters were chosen such that $p_t(x)>0$ everywhere on $\mathbb R$.

4 R2D2: NEURAL ESTIMATION OF RATE-DISTORTION FUNCTIONS

Although being rather elegant, as may be pointed out, the assumptions of Thm. 3.2 may be restrictive: They require $p_t(x)$ to be non-vanishing and twice-differentiable everywhere in \mathbb{R}^d ; even under these requirements, Eq. (17, BHE) might be ill-posed, depending on the initial condition \mathbb{P}_0 ; Finally, the explicit source distribution is hardly known in practice, and instead is accessible only from

samples. Therefore, in this section, we propose R2D2 (Alg. 1), a sample-based method for solving Eq. (14, TEC) and approximating the RD function in the general case (under Assumption A1).

Algorithm 1 Revealing RD functions with Diffusion (R2D2)

```
382
                       1: Input: source X_0 \sim \mathbb{P}_0 \in \mathcal{P}(\mathbb{R}^d), initial controller u_\theta, batch size M, timestep \Delta_t, \epsilon_{\min}, \epsilon_{\max} > 1
                              0. learning rate \alpha.
384
                       2: while Training do
                                                                                                                                                                                                                                 ▶ Training
                                     Choose \epsilon \sim \text{Uniform}[\epsilon_{\min}, \epsilon_{\max}].

Sample batch \{X_0^m\}_{m=1}^M \sim \mathbb{P}_0.

Sample trajectory \{u_{\theta}(X_{t_i}^m, t_i, \epsilon), X_1^m\}_{m=1}^M \leftarrow \text{EuMa}(u_{\theta}, \{X_0^m\}_{m=1}^M, \epsilon, \Delta_t).
                       3:
386
                       4:
387
                       5:
388
                                      Estimate energy L^{\epsilon}_{\theta} \leftarrow \frac{1}{2M} \sum_{m} \sum_{t_i} \|u_{\theta}(X^m_{t_i}, t_i, \epsilon)\|^2 \Delta_t.
                       6:
389
                                      Estimate terminal entropy \hat{H}(X_1) (see App. B).
                       7:
390
                                      RD loss \mathcal{L}_{\theta}^{\epsilon} \leftarrow L_{\theta}^{\epsilon} + \epsilon \hat{H}(X_1).
Step \theta \leftarrow \theta - \alpha \nabla_{\theta} \mathcal{L}_{\theta}^{\epsilon}.
                       8:
391
                       9:
392
                    10: end while
393
                    12: Sample batch \{X_0^m\}_{m=1}^M \sim \mathbb{P}_0. \triangleright Evaluate specific \epsilon \in [\epsilon_{\min}, \epsilon_{\max}] 13: Sample trajectory \{u_{\theta}(X_{t_i}^m, t_i, \epsilon), X_1^m\}_{m=1}^M \leftarrow \operatorname{EuMa}(u_{\theta}, \{X_0^m\}_{m=1}^M, \epsilon, \Delta_t). 14: Obtain RD loss \mathcal{L}_{\theta}^{\epsilon} (lines 6-8).
394
                    15: Estimate distortion: \hat{D} = \frac{1}{2M} \sum_{m=1}^{M} \|X_1^m - X_0^m\|^2.
397
                    16: Estimate rate: \hat{R} = \frac{\mathcal{L}_{\theta}^{\epsilon} - \hat{D}}{\epsilon} - \frac{d}{2} \log(2\pi\epsilon).
398
399
                    17: Output: (\hat{R}, \hat{D}).
400
```

Our method R2D2 is based on modeling the controller function $u_{\theta}(x,t,\epsilon)$ using a DNN with parameters θ . The flexibility and generalizability offered by DNNs allow us to capture multiple positions on the RD-curve (different ϵ values) using a single controller model. To train our model, we access the data source X_0 to draw a batch of M samples. Using the Euler–Maruyama method (EuMa, Alg. 2 in App. A), we sample discretized random trajectories X_{t_i} from Eq. (10). The minimization object Eq. (12) is approximated (up to an additive factor of $\frac{d}{2}\log(2\pi\epsilon)$) by $\mathcal{L}_{\theta}^{\epsilon} = L_{\theta}^{\epsilon} + \epsilon \hat{H}(X_1)$ where the estimated controller energy is $L_{\theta}^{\epsilon} \approx \frac{1}{2M} \sum_{m=1}^{M} \sum_{t_i} \|u_{\theta}(X_{t_i}^m, t_i, \epsilon)\|^2 \Delta_t$. The terminal entropy $\hat{H}(X_1)$ is estimated through the approximated negative-entropy, or through a kernel method (see App. B for details). To evaluate R(D), after training, we re-calculate $\mathcal{L}_{\theta}^{\epsilon}$. According to Eq. (12), we can compute the empirical values

$$\hat{D}(\epsilon) = \frac{1}{2M} \sum_{m=1}^{M} \|X_1^m - X_0^m\|^2, \quad \hat{R}(\epsilon) = \hat{\mathcal{I}}(X_0^m; X_1^m) = \frac{\mathcal{L}_{\theta}^{\epsilon} - \hat{D}(\epsilon)}{\epsilon} - \frac{d}{2} \log(2\pi\epsilon).$$
 (24)

The method is summarized in Algorithm 1.

5 Numerical results

In this section, we apply our results to a variety of toy and real-world problems. For full details and more numerical results, we kindly refer the reader to App. D. For full simulation details, we refer to the technical Section of App. E.

5.1 Gaussian data

In Fig. 2 we demonstrate the efficiency of Alg. 1 on the 1-D Gaussian case² of § 3.3. We compare our method with NERD (Lei et al., 2022) and WGD (Yang et al., 2024), over 64 independent experiments (seeds) and plot median absolute error with interquartile ranges. We observe that R2D2 is clearly superior to the existing methods in terms of estimation error, in both the high-rate and low-rate regimes.

²All our codes will be made publicly available upon publication.

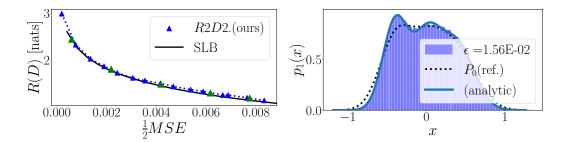


Figure 4: The R(D) function for a mixture of Gaussians. Here, $X_0 \sim \mathbb{P}_0$ is a mixture of Gaussians. (left) We apply R2D2 (Alg. 1) to $\epsilon \in [4 \times 10^{-4}, 1.64 \times 10^{-2}]$ and compare the result with SLB. *Green* markers indicate higher precision. (right) For $\epsilon = 1.56 \times 10^{-2}$, we plot the reconstruction distribution \mathbb{P}_1 . The empirical distribution matches the analytical result (bold line).

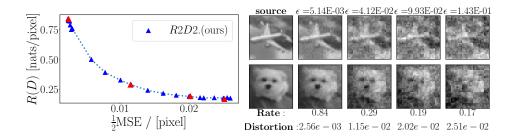


Figure 5: The R(D) function of Cifar10 images. Here, X_0 is a random 4×4 patch from a grayscale image. (left) The RD function, estimated by R2D2 (Alg. 1). (right) Patches drawn from the reconstruction distribution X_1 obtained for different ϵ 's. The corresponding locations on the RD-plane are marked in red.

Results for a Gaussian mixture Eq. (21) are shown in Fig. 4, where N=3 and $\mu_i=-.4,0,.4,\ \sigma_i^2=4\times 10^{-2}, 5\times 10^{-2}, 6\times 10^{-2},\ p_i=\frac{1}{3}.$ We apply Alg. 1 to $\epsilon\in[4\times 10^{-4},1.64\times 10^{-2}]$ and compare the estimated RD function with the approximated Shannon's lower bound (SLB) (Cover, 1999; Berger, 2003), given in this case by $H(\mathbb{P}_0)-\frac{1}{2}\log(4\pi eD)$. For $\epsilon=1.56\times 10^{-2}$, we further plot the reconstruction distribution \mathbb{P}_1 , which is the probability law of the diffusion process' outcome. We observe that empirical distribution obtained by Alg. 1 matches the closed-form Eq. (22).

5.2 AN EXPERIMENT: CIFAR10 DATASET

We now demonstrate the efficiency of Alg. 1 on a realistic high-dimensional source. More specifically, as input to R2D2, we sample 4×4 -pixel grayscale image patches from the 'Cifar10' dataset (Krizhevsky and Hinton, 2009). Fig. 5 demonstrates the efficiency of our method in solving this problem. In the left pane, we present the RD function, as estimated by R2D2 (Alg. 1). In the right pane, we present images, drawn from the reconstruction distribution \mathbb{P}_1 for different ϵ 's. The corresponding locations of these reconstructions on the RD-plane are marked in *red* on the left pane.

6 CONCLUSION

In this paper, we considered the computation of the RD function and the optimal reconstruction distribution for continuous data sources, under the MSE distortion. We exploited the connection between RD and EOT to estimate the RD function using diffusion processes, through a novel control formulation in which the RD tradeoff is equivalent to a tradeoff between energy and entropy. Under regularity conditions, the optimal control is given by a BHE. We demonstrated our results in some special cases to obtain closed-form solutions, and in a real-world setting, using a numerical method. This work paves the way for solving RD in settings beyond MSE loss and continuous distributions.

REFERENCES

- Jason Altschuler, Jonathan Niles-Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. *Advances in neural information processing systems*, 30, 2017.
- Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: In metric spaces and in the space of probability measures.* Springer Science & Business Media, 2008.
- Brian D.O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.
- Suguru Arimoto. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, 18(1):14–20, 1972.
- Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.
- Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R. Devon Hjelm. MINE: Mutual information neural estimation. *arXiv preprint* arXiv:1801.04062, 2018.
- Toby Berger. Rate-distortion theory. Wiley Encyclopedia of Telecommunications, 2003.
 - Richard Blahut. Computation of channel capacity and rate-distortion functions. *IEEE transactions on Information Theory*, 18(4):460–473, 1972.
 - Yongxin Chen, Tryphon T. Georgiou, and Michele Pavon. Stochastic control liaisons: Richard Sinkhorn meets Gaspard Monge on a Schrödinger bridge. *Siam Review*, 63(2):249–313, 2021.
 - Thomas M. Cover. Elements of information theory. John Wiley & Sons, 1999.
 - Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.
 - Noam Elata, Tomer Michaeli, and Michael Elad. Psc: Posterior sampling-based compression, 2025. URL https://arxiv.org/abs/2407.09896.
 - Giulio Franzese, Mustapha Bounoua, and Pietro Michiardi. MINDE: Mutual information neural diffusion estimation. *arXiv preprint arXiv:2310.09031*, 2023.
 - Emilia Fridman and Uri Shaked. Robust h_{∞} minimum entropy static output-feedback control of singularly perturbed systems. *Automatica*, 36(8):1181–1188, 2000.
 - Chu-Li Fu, Xiang-Tuan Xiong, and Zhi Qian. Fourier regularization for a backward heat equation. *Journal of Mathematical Analysis and Applications*, 331(1):472–480, 2007.
 - Nikita Gushchin, Alexander Kolesov, Alexander Korotin, Dmitry Vetrov, and Evgeny Burnaev. Entropic neural optimal transport via diffusion processes. *arXiv preprint arXiv:2211.01156*, 2022.
 - Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. Pmlr, 2018.
 - Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- Sergei Kholkin, Ivan Butakov, Evgeny Burnaev, Nikita Gushchin, and Alexander Korotin. Info-Bridge: Mutual information estimation via bridge matching. *arXiv preprint arXiv:2502.01383*, 2025.
 - Diederik P Kingma. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.
 - Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

- Marc Lambert, Francis Bach, and Silvère Bonnabel. Entropy regularized variational dynamic programming for stochastic optimal control. 2025.
- 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*, 3(4):674–686, 2022.
 - Eric Lei, Hamed Hassani, and Shirin Saeedi Bidokhti. On a relation between the rate-distortion function and optimal transport, 2023. URL https://arxiv.org/abs/2307.00246.
 - Christian Léonard. A survey of the Schrödinger problem and some of its connections with optimal transport. *arXiv preprint arXiv:1308.0215*, 2013.
 - Willard L. Miranker. A well posed problem for the backward heat equation. *Proceedings of the American Mathematical Society*, 12(2):243–247, 1961.
 - Guy Ohayon, Hila Manor, Tomer Michaeli, and Michael Elad. Compressed image generation with denoising diffusion codebook models. *arXiv preprint arXiv:2502.01189*, 2025.
 - Erikki Oja and A. Hyvarinen. Independent component analysis: Algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.
 - Bernt Oksendal. Stochastic differential equations: An introduction with applications. Springer Science & Business Media, 2013.
 - A Paszke. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.
 - Michele Pavon and Anton Wakolbinger. On free energy, stochastic control, and schrödinger processes. In *Modeling, Estimation and Control of Systems with Uncertainty: Proceedings of a Conference held in Sopron, Hungary, September 1990*, pages 334–348. Springer, 1991.
 - Georg Pichler, Pierre Jean A Colombo, Malik Boudiaf, Günther Koliander, and Pablo Piantanida. A differential entropy estimator for training neural networks. In *International Conference on Machine Learning*, pages 17691–17715. PMLR, 2022.
 - Erwin Schrödinger. Sur la théorie relativiste de l'électron et l'interprétation de la mécanique quantique. In *Annales de l'institut Henri Poincaré*, volume 2, pages 269–310, 1932.
 - Claude E. Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.
 - Claude E. Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec*, 4(142-163):1, 1959.
 - Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv* preprint arXiv:2010.02502, 2020.
 - Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021.
 - Lucas Theis, Tim Salimans, Matthew D. Hoffman, and Fabian Mentzer. Lossy compression with Gaussian diffusion. *arXiv* preprint arXiv:2206.08889, 2022.
 - Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein autoencoders. *arXiv preprint arXiv:1711.01558*, 2017.
 - Dor Tsur, Bashar Huleihel, and Haim H. Permuter. On rate distortion via constrained optimization of estimated mutual information. *IEEE Access*, 2024.
 - Yibo Yang, Stephan Mandt, and Lucas Theis. An introduction to neural data compression. *Foundations and Trends*® *in Computer Graphics and Vision*, 15(2):113–200, 2023.

Yibo Yang, Stephan Eckstein, Marcel Nutz, and Stephan Mandt. Estimating the rate-distortion function by Wasserstein gradient descent. Advances in Neural Information Processing Systems, 36, 2024. Brian D. Ziebart, Andrew L. Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In Aaai, volume 8, pages 1433–1438. Chicago, IL, USA, 2008. Jiayang Zou, Luyao Fan, Jiayang Gao, and Jia Wang. A revisit to rate-distortion problems via optimal weak transport theory. arXiv preprint arXiv:2501.09362, 2025.

A EULER-MARUYAMA SAMPLING

Here we review the *Euler-Maruyama* (EuMa) Algorithm, which we used in our simulations to sample from the diffusion process Eq. (10). The sampling procedure is given in Alg. 2.

Algorithm 2 Euler-Maruyama (EuMa)

- 1: Input: $\epsilon > 0$, drift $u(x, t, \epsilon)$, initial batch $\{X_0^m\}_{m=1}^M \sim \mathbb{P}_0$, timestep Δ_t .
- 2: for $t_i = 0, \Delta_t, \dots, 1 \Delta_t$ and $m = 1, \dots, M$ do
- 3: Sample $z_i^m \sim \mathcal{N}(0, I)$.
- 4: $X_{t_{i+1}}^m \leftarrow X_{t_i}^m + u(X_{t_i}^m, t_i, \epsilon)\Delta_t + \sqrt{\epsilon \Delta_t} z_i^m.$
- 5: end for

6: Return $\{u(X_{t_i}^m, t_i, \epsilon), X_1^m\}_{m=1}^M$.

B ENTROPY ESTIMATION

For the sake of completeness of Alg. 1, here we present the techniques used in the paper for estimating entropy (line 9 in the Algorithm). We emphasize, though, that we use the entropy estimator as a black-box, where any method could be plugged-in, orthogonally to our main ideas.

In our experiments on low-dimensional settings we used the approximated negative entropy method, for its simplicity and ease of compute. This could be hardly scaled to higher dimensions since it requires the computation of large covariance matrices. For real-world settings we used the scalable kernel method of Pichler et al. (2022).

B.1 NEGENTROPY

Negative entropy, or *negentropy* (Oja and Hyvarinen, 2000) of a random variable $X_1 \in \mathbb{R}^d$ is the difference in entropy from the Gaussian distribution with the same second-order statistics. Explicitly

negentropy
$$(X_1) \triangleq H(Z) - H(X_1) = -H(X_1) + \frac{1}{2} \sum_{i=1}^{d} \log \lambda_i + \frac{d}{2} \log(2\pi e),$$
 (25)

where λ_i are the eigenvalues of the covariance matrix Σ_{X_1} . We have the following connection between negentropy and KL divergence

Lemma B.1. (Kholkin et al., 2025, Corollary A.3) Let $Z \sim \mathcal{N}(\mu_1, \Sigma_1)$ where μ_1, Σ_1 are the mean and covariance of X_1 , respectively. Then,

$$H(X_1) = H(Z) - D_{\mathrm{KL}}(\mathbb{P}_{X_1}||\mathbb{P}_Z)$$
(26)

Using Eq. (26), we approximate the negentropy through the *Donsker-Varadhan* identity (Belghazi et al., 2018)

negentropy
$$(X_1) = D_{\mathrm{KL}}(\mathbb{P}_{X_1}||\mathbb{P}_Z) = \sup_{f} \left[\mathbb{E}_{z \sim \mathbb{P}_{X_1}} f(z) - \log \mathbb{E}_{z \sim \mathbb{P}_Z} e^{f(z)} \right],$$
 (27)

which can be estimated from samples (c.f. (Kholkin et al., 2025, § A.2) and (Franzese et al., 2023, § 3.2)).

In our simulations, we model the argument in Eq. (27) as a parametric model $Z_{\omega}(\cdot, \epsilon)$, where now we approximate the negentropy from M samples of X_1 as

negentropy
$$(X_1, \epsilon) \approx \frac{1}{M} \sum_{m=1}^{M} Z_{\omega}(X_1^m, \epsilon) - \log \left[\frac{1}{M} \sum_{m=1}^{M} e^{Z_{\omega}(z^m, \epsilon)} \right].$$
 (28)

 z^m are i.i.d. Gaussian samples with the empirical mean and covariance of X_1 . The entropy $H(X_1)$ can now be estimated through Eq. (25).

B.2 KNIFE

The KNIFE estimator (Pichler et al., 2022) is a Gaussian-mixture plug-in estimator for differential entropies. For $x \in \mathbb{R}^d$ the empirical distribution is approximated using

$$\hat{p}_{\text{KNIFE}}(x;\theta) = \sum_{k=0}^{K-1} u_k g_{\text{pdf}}(x;\mu_k, A_k), \tag{29}$$

where $g_{\mathrm{pdf}}(\cdot;\mu_k,A_k)$ are Gaussian kernels with mean and variance μ_k,A_k respectively, and $u_k>0$ are weights, $\Sigma_k u_k=1$.

The plug-in estimation is then given by

$$\hat{H}(X_1; \theta) = -\mathbb{E}_{x \sim \mathbb{P}_1} \log \left[\hat{p}_{\text{KNIFE}}(x; \theta) \right] \ge H(X_1). \tag{30}$$

For tight estimation, $\hat{H}(X_1; \theta)$ is minimized over $\theta := \{(u_k, \mu_k, A_k)\}_{k=0}^{K-1}$.

C PROOFS OF THINGS

C.1 PROOF OF THM. 3.1

Theorem C.1. (Thm. 3.1 in the Main text) $\mathcal{L}_{RD} = \tilde{\mathcal{L}}_{RD}$. Furthermore, let $u^*(x,t)$ minimize the surrogate objective

$$u^* \in \arg\min_{u} \left\{ \mathbb{E} \left[\frac{1}{2\epsilon} \int_0^1 \|u(X_t, t)\|^2 dt \right] + H(X_1) \right\}, \tag{31}$$

under

$$dX_t = u(X_t, t)dt + \sqrt{\epsilon}dW_t, X_0 \sim \mathbb{P}_0.$$
(32)

Then, the distribution $\mathbb{P}_{X_1^*}$ of X_1^* associated with u^* through Eq. (32) is the minimizer in Eq. (12), where $\mathbb{P}_1^* = \mathbb{P}_{X_1^*}$ is the optimal reconstruction distribution in Eq. (7, RD) and $\pi^* = \mathbb{P}_{X_0^*, X_1^*}$ is the optimal plan.

Here, we also prove the opposite direction; whenever (\mathbb{P}_1, π) minimizes Eq. (7, RD), there exists a drift term u(x,t) minimizing Eq. (31) under Eq. (32), where $X_0, X_1 \sim \pi$.

Proof. Let (\mathbb{P}_1, π) be a solution to Eq. (7, RD), where $\pi \in \Pi(\mathbb{P}_0, \mathbb{P}_1)$. Then, we choose

$$T(\omega) = \int_{x,\hat{x}} W^{\epsilon}|_{x,\hat{x}}(\omega) d\pi(x,\hat{x})$$
(33)

as a probability law in $\mathcal{F}(\mathbb{P}_0, \mathbb{P}_1)$. It can be easily deduced that

$$D_{\mathrm{KL}}(T|_{x,\hat{x}}||W^{\epsilon}|_{x,\hat{x}}) = 0, \tag{34}$$

hence, from Eq. (8)-Eq. (9) we have

$$D_{\mathrm{KL}}(T||W^{\epsilon}) = D_{\mathrm{KL}}(\pi||\pi^{W^{\epsilon}}) = \mathbb{E}_{\pi} \left[\frac{\|X - \hat{X}\|^2}{2\epsilon} \right] - H(\pi) + \frac{d}{2}\log(2\pi\epsilon). \tag{35}$$

which is equal to Eq. (7, RD) up to an additive constant, depends only on \mathbb{P}_0 , \mathbb{P}_1 and ϵ . Since π minimizes Eq. (7, RD) for \mathbb{P}_1 , it minimizes $D_{\mathrm{KL}}(\pi||\pi^{W^\epsilon})$ over $\Pi(\mathbb{P}_0,\mathbb{P}_1)$, thus T minimizes $D_{\mathrm{KL}}(T||W^\epsilon)$ over $\mathcal{F}(\mathbb{P}_0,\mathbb{P}_1)$ (see Chen et al. (2021, Problem 4.2, Eqn. (4.8)); mutatis mutandis). As a solution to $\min_{T\in\mathcal{F}(\mathbb{P}_0,\mathbb{P}_1)}D_{\mathrm{KL}}(T||W^\epsilon)$, T takes the form Eq. (10) with some drift function u (Léonard, 2013, Prop. 4.1).

Using Eq. (11) we get

$$\tilde{\mathcal{L}}_{RD} \le \frac{1}{2\epsilon} \mathbb{E} \left[\int_0^1 \|u(X_t, t)\|^2 dt \right] + H(\mathbb{P}_1) - \frac{d}{2} \log(2\pi\epsilon)$$
(36)

$$= D_{\mathrm{KL}}(T||W^{\epsilon}) + H(\mathbb{P}_1) - \frac{d}{2}\log(2\pi\epsilon)$$
(37)

$$= D_{\mathrm{KL}}(\pi || \pi^{\epsilon}) + H(\mathbb{P}_1) - \frac{d}{2} \log(2\pi\epsilon)$$
(38)

$$= \frac{1}{2\epsilon} \mathbb{E}_{\pi} \left[\|X - \hat{X}\|^2 \right] - H(\pi) + H(\mathbb{P}_0) + H(\mathbb{P}_1)$$
 (39)

$$=\mathcal{L}_{RD},\tag{40}$$

implying that $\mathcal{L}_{RD} \geq \tilde{\mathcal{L}}_{RD}$.

On the other hand, let u^* as in Eq. (31), and let $\mathbb{P}_1^* = \mathbb{P}_{X_1^*}$. Clearly, $\mathbb{P}_{X_1^*}, u^*$ minimize Eq. (12). Furthermore, the law T^* induced by u^* ,

$$T^*: \mathrm{d}X_t^* = u^*(X_t^*, t)\mathrm{d}t + \sqrt{\epsilon}\mathrm{d}W_T, \ X_0^* \sim \mathbb{P}_0, \tag{41}$$

minimizes $D_{\mathrm{KL}}(T^*||W^{\epsilon}) = \frac{1}{2\epsilon} \mathbb{E}\left[\int_0^1 \|u^*(X_t^*,t)\|^2 \mathrm{d}t\right]$ over $\mathcal{F}(\mathbb{P}_0,\mathbb{P}_1^*)$, hence it is a solution to $\min_{T \in \mathcal{F}(\mathbb{P}_0,\mathbb{P}_1^*)} D_{\mathrm{KL}}(T||W^{\epsilon})$. Thus, according to Léonard (2013, Prop. 2.3),

 $D_{\mathrm{KL}}(T^*|_{x,\hat{x}}||W^{\epsilon}|_{x,\hat{x}})=0, \forall x,\hat{x}. \ \mathrm{Let} \ \pi^*=\mathbb{P}_{X_0^*,X_1^*}\in \Pi(\mathbb{P}_0,\mathbb{P}_1^*). \ \mathrm{We \ have \ from \ Eq. \ (9)}$

$$D_{\text{KL}}(\pi^* || \pi^{\epsilon}) = \mathbb{E}_{\pi^*} \left[\frac{\|X - \hat{X}\|^2}{2\epsilon} \right] - H(\pi^*) + H(\mathbb{P}_0) + \frac{d}{2} \log(2\pi\epsilon)$$
 (42)

$$= D_{\mathrm{KL}}(T^*||W^{\epsilon}) \tag{43}$$

$$= \frac{1}{2\epsilon} \mathbb{E} \left[\int_0^1 \|u^*(X_t^*, t)\|^2 dt \right]. \tag{44}$$

Thus,

$$\mathcal{L}_{RD} \le \mathbb{E}_{\pi^*} \left[\frac{\|X - \hat{X}\|^2}{2\epsilon} \right] - H(\pi^*) + H(\mathbb{P}_1^*) + H(\mathbb{P}_0)$$
 (45)

$$= \frac{1}{2\epsilon} \mathbb{E} \left[\int_0^1 \|u^*(X_t^*, t)\|^2 dt \right] + H(\mathbb{P}_1^*) - \frac{d}{2} \log(2\pi\epsilon)$$
 (46)

$$=\tilde{\mathcal{L}}_{RD},\tag{47}$$

yielding $\mathcal{L}_{RD} = \tilde{\mathcal{L}}_{RD}$, which completes the proof. Note that arguments similar to Eq. (45)–Eq. (47) yield that under $\mathbf{A1}$, $\tilde{\mathcal{L}}_{RD} > -\infty$.

C.2 PROOF OF THM. 3.2

Theorem C.2. (Thm. 3.2 in the Main text) Let $p_t^*(x)$ such that $\log p_t^*(x) \in \mathcal{C}^{1,2}([0,1] \times \mathbb{R}^d)$ and $\|\nabla p_t^*(x)\| \log p_t^*(x) \to 0$ as $\|x\| \to \infty$ for all $t \in [0,1]$, satisfying the BHE

$$\frac{\partial}{\partial t} p_t(x) = -\frac{1}{2} \epsilon \Delta_{xx} p_t(x), \ p_0(x) \sim \mathbb{P}_0, \tag{48, BHE}$$

and let $u^* = \epsilon \nabla \log p_t^*(x)$ where $\nabla \log p_t(x)$ is the Stein score function. Then, (u^*, p_t^*) is an optimal pair in Eq. (15, var-TEC)-Eq. (16, FPE) and the solution to Eq. (14, TEC) admits the SDE

$$dX_t = \epsilon \nabla \log p_t^*(X_t) dt + \sqrt{\epsilon} dW_t. \tag{49}$$

Proof. Let us define the Lagrangian

$$L(u, p, \mu, \lambda) = \frac{1}{2} \int_{\mathbb{R}^d} dx \int_0^1 dt \|u(x, t)\|^2 p_t(x)$$

$$+ \int_{\mathbb{R}^d} dx \int_0^1 dt \mu(x, t) \left(\dot{p}_t(x) + \nabla \cdot (up(x, t)) - \frac{1}{2} \epsilon \Delta_{xx} p_t(x)\right)$$

$$- \epsilon \int_{\mathbb{R}^d} dx p_1(x) \log p_1(x) + \int_{\mathbb{R}^d} dx \lambda(x) (p_0(x) - d\mathbb{P}_0(x)), \tag{50}$$

where $d\mathbb{P}_0$ denote the *density* function of \mathbb{P}_0 .

We now apply the following integration by parts property of the divergence: For $g: \mathbb{R}^d \to \mathbb{R}$ and $f: \mathbb{R}^d \to \mathbb{R}^d$, and a bounded domain $D \subseteq \mathbb{R}^d$ with boundary ∂D ,

$$\int_{D} g(x)\nabla \cdot f(x)dx = -\int_{D} \nabla g(x) \cdot f(x)dx + \oint_{\partial D} g(x)f(x) \cdot \vec{n}da$$
 (51)

where $\nabla, \nabla \cdot$ are the gradient and divergence operators, respectively. If $\|g(x)f(x)\|$ decays as $\|x\| \to \infty$, we can integrate over domains with large enough diameters, thus ignoring the boundary term and practically integrate over \mathbb{R}^d . In our setting, p, μ are scalar functions, and u is a field.

Integrating by parts, we have

$$\int_{0}^{1} dt \mu(x,t) \dot{p}_{t}(x) = \mu(x,1) p_{1}(x) - \mu(x,0) p_{0}(x) - \int_{0}^{1} dt \dot{\mu}(x,t) p_{t}(x)$$
 (52)

and

$$\int_{\mathbb{R}^d} \mathrm{d}x \mu(x,t) \nabla \cdot (up(x,t)) = -\int_{\mathbb{R}^d} \mathrm{d}x p u \cdot \nabla \mu(x,t) + \oint \mu p u \cdot \vec{n} \mathrm{d}a$$
 (53)

as well as

$$\int_{\mathbb{R}^d} dx \mu(x,t) \Delta_{xx} p(x,t)$$

$$= \int_{\mathbb{R}^d} dx \mu(x,t) \nabla \cdot \nabla p(x,t)$$
(54)

$$= -\int_{\mathbb{R}^d} dx \nabla \mu(x, t) \cdot \nabla p(x, t) + \oint \mu(\nabla p(x, t)) \cdot \vec{n} da$$
 (55)

$$= -\int_{\mathbb{R}^d} dx \nabla \mu(x,t) \cdot \nabla p(x,t) + \oint p \nabla \mu \cdot \vec{n} da + \oint \left[\mu(\nabla p(x,t)) - p \nabla \mu \right] \cdot \vec{n} da \qquad (56)$$

$$= \int_{\mathbb{R}^d} dx \nabla_{xx} \mu(x, t) p_t(x) - \oint p \nabla \mu \cdot \vec{n} da + \oint \mu(\nabla p(x, t)) \cdot \vec{n} da.$$
 (57)

Provided that all boundary terms vanish as $||x|| \to \infty$, and putting everything back together, we obtain

$$L(u, p, \mu, \lambda) = \frac{1}{2} \int_{\mathbb{R}^d} dx \int_0^1 dt \|u(x, t)\|^2 p_t(x)$$

$$- \int_0^1 dt \int_{\mathbb{R}^d} dx \left[\dot{\mu}(x, t) + u(x, t) \cdot \nabla \mu(x, t) + \frac{1}{2} \epsilon \Delta_{xx} \mu(x, t) \right] p_t(x)$$

$$+ \int_{\mathbb{R}^d} dx (\mu(x, 1) - \epsilon \log p_1(x)) p_1(x)$$

$$- \int_{\mathbb{R}^d} dx (\mu(x, 0) - \lambda(x)) p_0(x) - \int_{\mathbb{R}^d} dx \lambda(x) d\mathbb{P}_0(x).$$
 (58)

Taking the first variation to zero, $\frac{\delta L}{\delta u} = 0$ yields

$$u^*(x,t) = \nabla \mu(x,t). \tag{59}$$

From $\frac{\delta L}{\delta p} = 0$ we obtain (*Hamilton–Jacobi equation*),

$$\dot{\mu}(x,t) + \frac{1}{2} \|\nabla \mu(x,t)\|^2 + \frac{1}{2} \epsilon \Delta_{xx} \mu(x,t) = 0, \ \mu(x,1) = \epsilon (1 + \log p_1(x)). \tag{60}$$

We also know that (Fokker–Planck equation)

$$\dot{p}_t(x) + \nabla \cdot (p_t(x)\nabla \mu(x,t)) - \frac{1}{2}\epsilon \Delta_{xx} p_t(x) = 0, p_0(x) = d\mathbb{P}_0.$$
(61)

We now substitute a solution of the form $\mu(x,t) = \epsilon(1 + \log p_t(x))$ into Eq. (60), and verify it satisfies our equations:

$$\dot{p}_{t}/p_{t} + \frac{1}{2}\epsilon \|\nabla p_{t}/p_{t}\|^{2} + \frac{1}{2}\epsilon \nabla \cdot (\nabla p_{t}/p_{t})$$

$$= \dot{p}_{t}/p_{t} + \frac{1}{2}\epsilon \|\nabla p_{t}/p_{t}\|^{2} + \frac{1}{2}\epsilon \sum_{i} (p_{x_{i}x_{i}}/p_{t} - (p_{x_{i}}/p_{t})^{2})$$
(62)

$$= \dot{p}_t/p_t + \frac{1}{2}(\epsilon - \epsilon) \|\nabla p_t/p_t\|^2 + \frac{1}{2}\epsilon \Delta_{xx} p_t/p_t = 0$$
 (63)

where the last equality stems from Eq. (48, BHE).

D EXTENDED RESULTS AND TECHNICAL DETAILS

For the sake of completeness and in-depth reading, in this section we extend § 5 with full technical details and additional results. We also present detailed and full-sized figures for improved accessibility.

D.1 GAUSSIAN SOURCES

Let $\mathbb{P}_0 = \mathcal{N}\left(0, \sigma_0^2\right)$. A solution to Eq. (17, BHE) is given by

$$p_t(x) = \frac{1}{\sqrt{2\pi(\sigma_0^2 - \epsilon t)}} e^{-\frac{x^2}{2(\sigma_0^2 - \epsilon t)}}.$$
 (64)

The optimal controller is hence given by

$$u(x,t) = \epsilon \nabla \log p_t(x) = -\frac{\epsilon}{\sigma_0^2 - \epsilon t} x. \tag{65}$$

Let us denote $a_t = \frac{\epsilon}{\sigma_0^2 - \epsilon t}$ and $r = \frac{\epsilon}{\sigma_0^2}$. It is easy to see that under u(x,t), X_0 and X_1 are jointly-Gaussian where

$$dX_t = -a_t X_t dt + \sqrt{\epsilon} dW_t \tag{66}$$

$$U_t^{-1} X_t = X_0 + \sqrt{\epsilon} \int_0^t U_s^{-1} dW_s,$$
 (67)

with

$$U_t = e^{-\int_0^t a_s ds} = 1 - rt. ag{68}$$

That is,

$$X_1 = (1 - r)X_0 + N_1 (69)$$

$$\sigma_{N_1}^2 = \epsilon (1 - r)^2 \int_0^1 U_s^{-2} ds = \epsilon (1 - r)^2 \frac{\sigma_0^2}{\sigma_0^2 - \epsilon} = \epsilon (1 - r). \tag{70}$$

where N_1 is a Gaussian noise, independent of X_0 . We can now compute the distortion

$$\mathbb{E}\left[(X_1 - X_0)^2\right] = r^2 \sigma_0^2 + \epsilon (1 - r) = \epsilon - \epsilon^2 \sigma_0^{-2} + \epsilon^2 \sigma_0^{-2} = \epsilon,\tag{71}$$

and the correlation

$$\rho \triangleq \mathbb{E}\left[X_0 X_1\right] = (1 - r)\sigma_0^2 = \sigma_0^2 - \epsilon \tag{72}$$

and then also compute the MI by plugging

$$\sigma_1^2 = (1-r)^2 \sigma_0^2 + \epsilon (1-r) = \sigma_0^2 - 2\epsilon + \epsilon^2 \sigma_0^{-2} + \epsilon - \epsilon^2 \sigma_0^{-2} = \sigma_0^2 - \epsilon$$
 (73)

into

$$\mathcal{I}(X_0; X_1) = -\frac{1}{2} \log \left(1 - \frac{\rho^2}{\sigma_1^2 \sigma_0^2} \right) \tag{74}$$

$$= -\frac{1}{2}\log\left(1 - \frac{(1-r)^2\sigma_0^4}{(\sigma_0^2 - \epsilon)\sigma_0^2}\right)$$
 (75)

$$= -\frac{1}{2}\log\left(1 - \frac{(1-r)^2}{1-r}\right) \tag{76}$$

$$= -\frac{1}{2} \log \left(\frac{\epsilon}{\sigma_0^2} \right). \tag{77}$$

To summarize, we obtained

$$D = \frac{1}{2}\mathbb{E}\left[\left(X_0 - X_1\right)^2\right] = \frac{1}{2}\epsilon, \quad R = \mathcal{I}(X_0; X_1) = -\frac{1}{2}\log\left(\frac{\epsilon}{\sigma_0^2}\right),\tag{78}$$

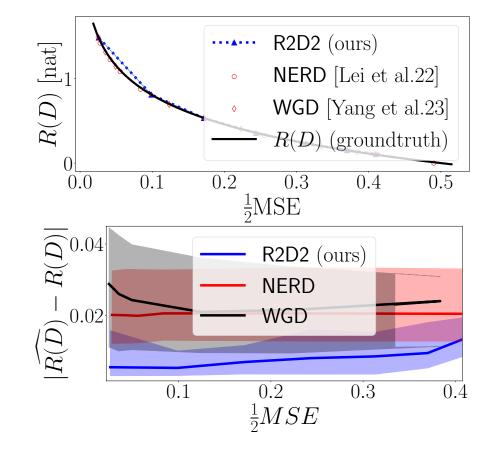


Figure 6: The R(D) function of a 1-D Gaussian source. (top) Here, $X_0 \sim \mathcal{N}(0,1)$ and we applied Alg. 1 to $\epsilon \in [0.05, 0.95]$. Green markers indicate higher precision. We also plot the analytical result (black line). (bottom) We compare our algorithm to NERD and WGD, where we observe that our method is more accurate.

and recovered the well-known result

$$R_{\text{Gauss}}(D) = \frac{1}{2} \log \left(\frac{\sigma_0^2}{2D} \right), \quad 0 \le 2D \le \sigma_0^2.$$
 (79)

Note that the factor of 2 is due to our convention $D = \frac{1}{2}MSE$. This result can be easily generalized for Gaussian vector sources and ϵ 's smaller than the eigenvalues of the covariance matrix Σ . In this case, the solution to Eq. (17, BHE) is

$$p_t(x) = \mathcal{N}(0, \Sigma - \epsilon t I). \tag{80}$$

In Fig. 6 we demonstrate the efficiency of Alg. 1 on the 1-D case. We compare our method with NERD (Lei et al., 2022) WGD (Yang et al., 2024), where we observe that R2D2 is clearly superior to the existing methods in terms of estimation error, in both the high-rate and low-rate regimes.

D.2 MIXTURE OF GAUSSIANS

The latter Example can be easily extended to the case of a Gaussian mixture, where

$$p_0(x) = \sum_{i=1}^{N} \frac{p_i}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}}.$$
 (81)

As Eq. (17, BHE) is a linear equation, the solution is now given by the superposition

$$p_t(x) = \sum_{i=1}^{N} \frac{p_i}{\sqrt{2\pi(\sigma_i^2 - \epsilon t)}} e^{-\frac{(x - \mu_i)^2}{2(\sigma_i^2 - \epsilon t)}}, \ \epsilon \in \left[0, \min_i \sigma_i^2\right). \tag{82}$$

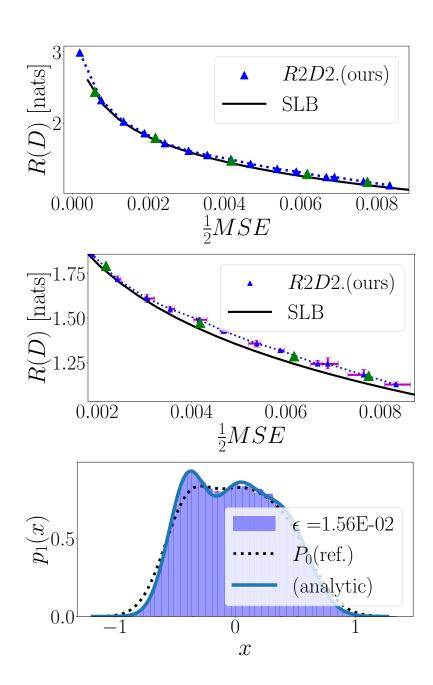


Figure 7: The R(D) function for a mixture of Gaussians. Here, $X_0 \sim \mathbb{P}_0$ is a mixture of Gaussians where $\mu_i = -.4, 0, .4, \ \sigma_i^2 = 4 \times 10^{-2}, 5 \times 10^{-2}, 6 \times 10^{-2}, \ p_i = \frac{1}{3}$. (top) We apply Alg. 1 to $\epsilon \in [4 \times 10^{-4}, 1.64 \times 10^{-2}]$ and compare the result with SLB. *Green* markers indicate higher precision. (middle) A closer look on error bars (inter-quartile range) over 8 evaluations. (bottom) For $\epsilon = 1.56 \times 10^{-2}$, we plot the reconstruction distribution \mathbb{P}_1 which is the distribution of the diffusion process' outcome, X_1 . Observe that the empirical distribution matches the analytical result (bold line).

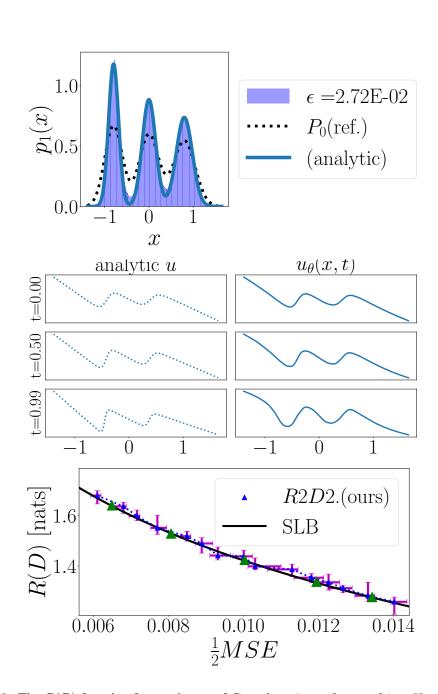


Figure 8: The R(D) function for a mixture of Gaussians (second example). Here, X_0 is a mixture of Gaussians where $\mu_i=-.8,0,.8,~\sigma_i^2=4\times10^{-2},5\times10^{-2},6\times10^{-2},~p_i=\frac{1}{3}.$ (top) The empirical reconstruction distribution for $\epsilon=2.72\times10^{-2}$ matches the analytical result. (middle) Trained controller model u_{θ} (for $\epsilon=2.8\times10^{-2}$), compared to analytical result at times t=0,0.5,0.99. (bottom) RD function estimated over 16 evaluation steps (medians and interquartile ranges).

The optimal controller $u(x,t)=\epsilon\nabla\log p_t(x)$ is derived accordingly. We illustrate this result in Fig. 7, where N=3 and $\mu_i=-.4,0,.4,\ \sigma_i^2=4\times 10^{-2},5\times 10^{-2},6\times 10^{-2},\ p_i=\frac{1}{3}$. We apply Alg. 1 to $\epsilon\in[4\times 10^{-4},1.64\times 10^{-2}]$ and compare the estimated RD function with Shannon's lower bound (SLB)(Cover, 1999; Berger, 2003)

$$H(\mathbb{P}_0) - \frac{1}{2}\log(4\pi eD),\tag{83}$$

approximated here from $M=2^{11}$ i.i.d. samples $X^m \sim \mathbb{P}_0$ by

$$H(\mathbb{P}_0) \approx -\frac{1}{M} \sum_{m=1}^{M} \log p_0(X^m). \tag{84}$$

In practice, we estimated Eq. (84) for 8 independent trials, and used the median value for our approximation. For $\epsilon = 1.56 \times 10^{-2}$. We further plot the reconstruction distribution \mathbb{P}_1 , which is the probability law of the diffusion process' outcome. We observe that empirical distribution obtained by Alg. 1 matches the closed-form result Eq. (82).

In the setting of Gaussian mixtures, we conducted an additional experiment in which $\mu_i = -.8, 0, .8$. The results are given in Fig. 8, where we also plot the outcome of the deep controller model we train, compared to the desired product.

D.3 Non-Gaussian mixture

Consider the source X_0 drawn from the mixture $p_0(x) = \sum_{i=1}^N p_i C_i^{-1} sinc^4(\frac{x}{m_i})$ where $sinc(x) \triangleq \frac{\sin(x)}{x} \in \mathcal{C}^{\infty}(\mathbb{R})$, and $C_i = \frac{2\pi}{3} m_i$ are appropriate normalization factors. Recall that the characteristic function in this case is

$$\hat{p}_0(\omega) = \int_{-\infty}^{\infty} p_0(t)e^{-iwt} dt = \sum_{i=1}^{N} p_i C_i^{-1} m_i \tilde{p}_0(m_i \omega),$$
 (85)

where (with * being the *convolution* operation)

$$\tilde{p}_0(\omega) =$$
 (86)

$$= \frac{1}{2\pi} \left[\pi (1 - \frac{1}{2}|\omega|)_{+} * \pi (1 - \frac{1}{2}|\omega|)_{+} \right]$$
 (87)

$$= \frac{1}{96}\pi \left((w-4)^3 \operatorname{sign} w - 4 - 4(w-2)^3 \operatorname{sign} w - 2 \right)$$
 (88)

$$+6w^3 \operatorname{sign} w - 4(2+w)^3 \operatorname{sign} 2 + w + (4+w)^3 \operatorname{sign} 4 + w$$
, (89)

which vanishes outside $\{|w| \le 4\}$. Now, for non-vanishing mix distributions and small enough ϵ 's, by Eq. (19) we can write the solution to Eq. (17, BHE) as

$$p_t(x) = \frac{1}{2\pi} \int_{-4}^4 e^{i\omega x + \frac{1}{2}\epsilon\omega^2 t} \hat{p}_0(\omega) d\omega = \frac{2}{2\pi} \sum_{i=1}^N p_i C_i^{-1} m_i \int_0^4 \tilde{p}_0(m_i \omega) \cos(\omega x) e^{\frac{1}{2}\epsilon\omega^2 t} d\omega.$$
 (90)

We numerically approximate

$$p_1(x) \approx \frac{1}{\pi} \sum_{i=1}^{N} p_i C_i^{-1} m_i \sum_{k=0}^{K-1} \frac{4}{K} \tilde{p}_0(m_i \frac{4k}{K}) \cos(\frac{4k}{K} x) e^{\frac{1}{2}\epsilon(\frac{4k}{K})^2}.$$
(91)

Figure 9 demonstrates this result for $N=4, m_i=\left[1,\sqrt{2},\pi,e\right]$ (m_i 's were chosen such that $p_0(x)>0$ everywhere on $\mathbb R$) and $p_i=\frac{1}{4}$, where we numerically integrated Eq. (91) with $K=4\times32\times10^4$ in order to approximate the reconstruction distribution $p_1(x)$ for different values of ϵ . We emphasize that although this is a toy problem, to the best of our knowledge, no other technique is known to tackle this case.

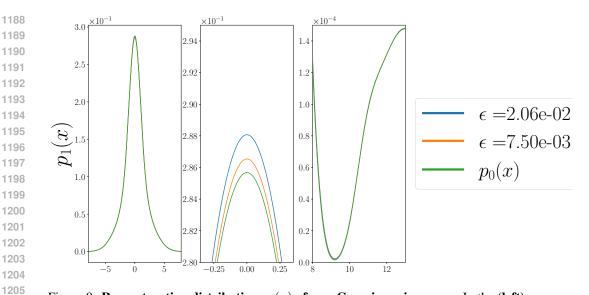


Figure 9: **Reconstruction distribution** $p_1(x)$ **of non-Gaussian mix source.** In the (**left**) pane, we approximate reconstruction distribution Eq. (90) by numerical integration. A closer look at different points is provided in the (**middle**) and (**right**) panes.

D.4 EXAMPLE: CIFAR10 DATASET

We now demonstrate the efficiency of Alg. 1 on a realistic high-dimensional source. More specifically, as input to the Algorithm, we sample 4×4 grayscale image patches from the 'Cifar10' dataset (Krizhevsky and Hinton, 2009). Pixel values are normalized to [0,1]. Fig. 5 demonstrates the efficiency of our method in solving this problem. In the left pane, we present the RD function, as estimated by R2D2 (Alg. 1). In the right pane, we present images, drawn from the reconstruction distribution \mathbb{P}_1 for different ϵ 's. The corresponding locations of these reconstructions on the RD-plane are marked in *red* on the left pane.

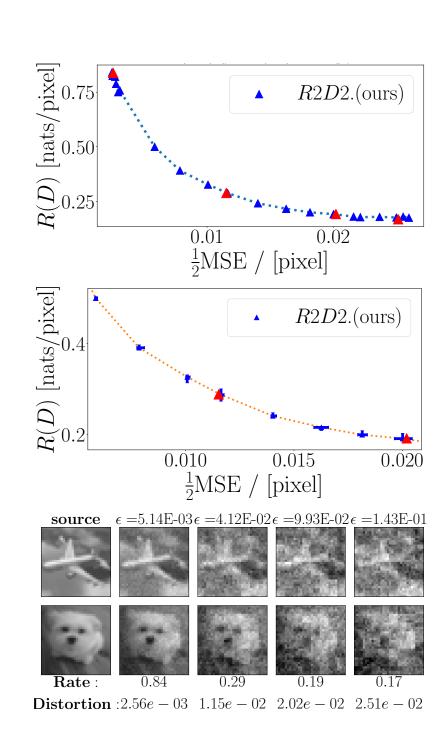


Figure 10: **The** R(D) **function of Cifar10 images.** Here, X_0 is a random 4×4 patch from a grayscale image. **(top)** The RD function, estimated by R2D2 (Alg. 1). **(middle)** A closer look on error bars (inter-quartile range) over 8 evaluations. **(bottom)** Patches drawn from the reconstruction distribution X_1 obtained for different ϵ 's. The corresponding locations on the RD-plane are marked in red.

E IMPLEMENTATION NOTES

E.1 GENERAL DETAILS

1296

1297 1298

1299 1300

1301

1302 1303 1304

1305

1306

1307

1308

1309

1310

1311

1312

1313 1314

1315

1316

1317

1318

1319

1320 1321

1322 1323

1324

1325 1326

1328

1332

1333 1334

1336 1337

1338 1339

1340

1341

1348

1349

For all experiments, we used 2 fully-connected DNN models:

- The controller u_{θ} taking $(X_t, t, \epsilon) \in \mathbb{R}^d \times [0, 1] \times [\epsilon_{\min}, \epsilon_{\max}]$ as input and whose output is in \mathbb{R}^d .
- The Z_{ω} network for estimating the negentropy, taking $(z, \epsilon) \in \mathbb{R}^d \times [\epsilon_{\min}, \epsilon_{\max}]$ and returning a scalar value (see App. B for details).

Despite having different input and output layers, both models are of the same depth and hidden-layer sizes. We used LeakyReLU activation following each hidden layer. The two models were trained in a 1 : 3-ratio of update steps, using the ADAM optimizer (Kingma, 2014) with parameters $\beta = (.9, .999)$ and learning rate α .

At each evaluation step, we draw a batch of samples and evaluate R(D) according to Alg. 1. Whenever there are more than one evaluation step or independent seeds, the presented (R,D)-values are the medians over all steps, while error bars indicate the inter-quartile (25%-75%) range.

Codes for the NERD baseline (Lei et al., 2022) are provided by the authors at https://github.com/leieric/NERD-RCC. Codes for the WGD baseline (Yang et al., 2024) are provided by the authors at https://github.com/yiboyang/wgd. Our codes will be publicly available upon publication.

All experiments were implemented in PyTorch (Paszke, 2019) environment, and performed using a NVIDIA RTX A6000 GPU.

E.2 SIMULATION PARAMETERS

Gaussian sources (Fig. 2)

- $\epsilon_{\min}, \epsilon_{\max} = 0.025, 0.975$
- DNN hidden layers: 1
- Hidden layer size: 128
- Step size: $\Delta_t = \frac{1}{100}$
- Train steps: 25,000
 - Batch size (train): M = 512
 - Experiments: 64
 - Batch size (evaluation): 1024
- Learning rate: $\alpha = 1\text{e-}3$

Mixture of Gaussians (Fig. 4)

- ϵ_{\min} , ϵ_{\max} =4e-4,1.64e-2
- DNN hidden layers: 4
- Hidden layer size: 128
- Step size: $\Delta_t = \frac{1}{100}$
- Train steps: 3.7M
 - Batch size (train): M = 256
- Evaluation steps: 8 (32 at high-precision)
 - Batch size (evaluation): 1024
 - Learning rate: $\alpha = 5e-4$

Mixture of Gaussians #2 (Fig. 8) • ϵ_{\min} , $\epsilon_{\max} = 1.2e-2,2.8e-2$ • DNN hidden layers: 4 • Hidden layer size: 100 • Step size: $\Delta_t = \frac{1}{100}$ • Train steps: 855,000 • Batch size (train): M = 256• Evaluation steps: 16 (64 at high-precision) • Batch size (evaluation): 1024 • Learning rate: $\alpha = 1e-3$ Cifar10 dataset (Fig. 5) • ϵ_{\min} , $\epsilon_{\max} = 0.05, 0.15$ • DNN hidden layers: 3 • Hidden layer size: 100 • Step size: $\Delta_t = \frac{1}{256}$ • Train steps: 23,300 • Batch size (train): M = 1024• Evaluation steps: 8 (32 at high-precision) • Batch size (evaluation): 512 • Learning rate: $\alpha = 5e-4$