Generating DDPM-based Samples from Exponentially Tilted Distributions

Anonymous Author
Anonymous Institution

Abstract

Given independent samples from a multidimensional probability distribution, our aim is to generate samples from a distribution obtained by exponentially tilting the original distribution where the exponent in the tilt may be a linear or nonlinear function of the random vectors. Application areas of such samples include finance, weather and climate modeling, and many other domains, where the aim may be to generate samples from a tilted distribution that satisfies practically motivated moment constraints. We rigorously show that under certain conditions running DDPM on the reweighed samples is capable of faithfully reproducing the tilted samples. Our theoretical results are supported by simulations.

1 INTRODUCTION

Diffusion based generative-AI typically considers observing many samples from a high-dimensional distribution and using them to generate many independent samples from that distribution. Traditionally, the samples corresponded to images or text, although recently there has been research on the underlying samples corresponding to high-dimensional data, for example, in weather or climate modeling(Li et al. (2024), Ling et al. (2024)) or in finance (see Cont et al. (2025), for GAN based application)).

However, in many practically important settings, samples may be available under one distribution $\mu(x)$, and our interest may be in generating samples from a related tilted distribution $\nu(x) \propto \exp(\theta^T g(x))\mu(x)$ (that is, a distribution obtained by appropriately multiplicatively biasing the underlying distribution). This is relevant in portfolio optimization (see Meucci (2010)) and in option pricing in finance (see Buchen & Kelly (1996), Stutzer (1996), Avellaneda (1998)) where exponentially tilted versions of pricing distributions are sought to better model asset prices, where the exponent in the tilt is a linear function of stock returns. Goll & Rueschendorf (2002) consider exponentially tilted distributions in mathematical finance where the exponent may be a nonlinear function of financial securities.

Exponential tilting of random variables or vectors is a common technique in the field of rare event sampling and Monte Carlo simulation, that effects an exponential change on an underlying probability measure in order to bias its rare events into becoming more common. It also has applications Gerber & Shiu (1994) to phase transitions in physical phenomena Lee et al. (2025), rare event sampling from discrete Markov processes Aguilar & Gatto (2024), and finding robust best response strategies in security games Kong et al. (2025). We refer the reader to Alvo (2022), Chapter 1 for an excellent review on applications of exponential tilting.

These tilted distributions arise as a solution to a well-motivated optimization problem. For example, given a probability measure μ on $\mathbb{R}^d$, if we look for a probability measure ν that is closest to μ in relative entropy, and satisfies certain moment constraints, the solution is a distribution obtained by appropriately exponentially tilting μ with a linear term in the exponent. When more general f-divergences are used instead of relative entropy, the solution can be an exponentially tilted distribution with a non-linear term in the exponent (see, Csiszár (1967), Csiszár (2008)).

Previous works have largely approached tilted or biased sampling through importance sampling or MCMC biasing schemes, but these typically require some knowledge of the target density. More recently, diffusion-based methods have been proposed to enable sampling from tilted distributions via score guidance

Preliminary work. Under review by AISTATS 2026. Do not distribute.

methods (see Wang et al. (2024)).

In this article, we consider samples from an unknown high-dimensional probability distribution μ. Our aim is to establish conditions under which a diffusion sampling-based algorithm is capable of producing high-quality samples from μ exponentially tilted by any desired vector θ i.e. $\nu(x) \propto \exp(\theta^T g(x))\mu(x)$, where $g(x)$ is some desirable twisting function.

The proposed algorithm proceeds in two steps. First, we re-weight the given samples $\{X_i\}$ using the weights $w_i = \exp(\theta^T g(x_i))$ and then normalize with $nw_n^* = \sum_i w_i$. Then, we perform diffusion sampling by training a denoiser using the reweighted samples and running the reverse diffusion to generate samples. Note that we must provide two theoretical guarantees:

1. The accuracy of the reweighted samples in comparison to empirical samples from the true twisted distribution as a function of the tilt amount, the number of generated samples and the underlying distribution μ.

2. The accuracy of diffusion output if input samples do not come from the actual distribution but instead from a nearby one.

In this article, we positively address both questions.

Outline: Recall that exponential tilts are a solution to an optimization problem where the new probability measure is selected that minimizes an entropic distance with respect to a given one, subject to moment constraints on the new measure. In Section 2, we concretize this for entropic distances corresponding to Kullback-Leibler, Renyi and Tsallis entropy.

Our key goal is to find conditions under which DDPM-sampling leads to samples that match the appropriate exponentially tilted distribution of the input samples. To do this, we first recall (Chen et al., 2023) that the total variation (TV) error between the DDPM-samples and the true distribution can be controlled by controlling the relevant denoiser error. Our approach to proving that the denoiser error under the tilted distribution is small is by first characterizing the expected Wasserstein distance between the weighted empirical distribution and the true twisted distribution in Section 3, and, then, under a Lipschitz assumption, using that to establish bounds on the difference in the relevant denoiser errors in Section 4. We end the discussion with extensive experimentation in Section 5.

2 KULLBACK-LEIBLER (KL), RENYI, TSALLIS ENTROPIES

In this section, we justify exponentially tilting the distribution via KL, Renyi and Tsallis entropies. Note the maximizer densities follow the $\exp(\theta^T g(x))$ form for some function $g(x)$.

Theorem 1 (KL ⇒ exponential tilt)**.** *Let Q be a reference probability measure and let Φ be measurable with the integrability required. Consider*

$$\mathcal{J}_{\mathrm{KL}}(P) = \mathbb{E}_P[\Phi] - D_{\mathrm{KL}}(P\|Q)$$

$$D_{\mathrm{KL}}(P\|Q) = \int \log\frac{dP}{dQ}\, dP$$

Maximizing $\mathcal{J}_{\mathrm{KL}}$ over probability measures $P \ll Q$ yields the unique optimizer $P^\star$ whose density $h^\star := \dfrac{dP^\star}{dQ}$ satisfies

$$\log h^\star(x) = \Phi(x) + \mathrm{const} \quad \textit{(a.e.)},$$

hence

$$\frac{dP^\star}{dQ}(x) = \frac{\exp(\Phi(x))}{\displaystyle\int \exp(\Phi)\, dQ}.$$

Theorem 2 (Rényi ⇒ power/escort tilt)**.** *Fix $\alpha > 0$, $\alpha \neq 1$. For $h = dP/dQ$ define*

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1}\log\Big(\int h^\alpha\, dQ\Big).$$

Consider the functional $\mathcal{J}_\alpha(P) = \mathbb{E}_P[\Phi] - D_\alpha(P\|Q)$. A maximizer $P^\star$ (if it exists) has density $h^\star$ satisfying a.e.

$$\frac{\alpha}{\alpha-1}\frac{(h^\star(x))^{\alpha-1}}{\int (h^\star)^\alpha dQ} = \Phi(x) - \mu$$

for some constant μ, equivalently (after absorbing constants)

$$h^\star(x) \;\propto\; \big(a + b\,\Phi(x)\big)^{1/(\alpha-1)},$$

i.e. a power/escort–type tilt.

Theorem 3 (Tsallis ⇒ q–exponential tilt)**.** *Fix $q > 0$, $q \neq 1$. Using one convenient form of Tsallis relative entropy*

$$D_q^{\mathrm{Ts}}(P\|Q) = \frac{1}{q-1}\Big(1 - \int h^q\, dQ\Big), \qquad h = \frac{dP}{dQ},$$

consider $\mathcal{J}_{\mathrm{Ts}}(P) = \mathbb{E}_P[\Phi] - D_q^{\mathrm{Ts}}(P\|Q)$. A maximizer (when it exists) satisfies a.e.

$$\frac{q}{q-1}h^\star(x)^{\,q-1} = \mu - \Phi(x)$$

for some constant μ, and hence after reparameterization

$$h^\star(x) \;\propto\; \big[1 + (1-q)\,c\,\Phi(x)\big]^{1/(1-q)} = \exp_q(c\,\Phi(x)),$$

the q–exponential (with the usual normalization).

3 ON THE ACCURACY OF REWEIGHED SAMPLES

3.1 Setup

Throughout this article, let x be a random vector in $\mathbb{R}^d$ with its multivariate CDF given by F. Given a vector $\theta \in \mathbb{R}^d$, an *exponential tilting* of x by the vector θ produces a random vector x_θ whose distribution is given by

$$\mathbb{P}(\mathrm{x}_\theta \in A) = \frac{\mathbb{E}[e^{\theta^T g(\mathrm{x})}\mathbf{1}_{\mathrm{x}\in A}]}{\mathbb{E}[e^{\theta^T g(\mathrm{x})}]} \tag{1}$$

for some suitable function g which can be considered a site-specific tilt. For now we assume that $\|g(x)\| \leq g_{\max}$ for all x in the support of x.

Let F_θ be the CDF of x_θ and μ_θ be the probability measure corresponding to x_θ.

Now, given independent, identically distributed samples $\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_n$ from F, we define an empirical reweighed distribution by

$$\mu_{n,\theta} = \frac{1}{n}\sum_{i=1}^{n} \frac{w_i}{w_n^*}\delta_{\mathrm{x}_i}, \qquad \text{where } w_n^* = \frac{1}{n}\sum_{j=1}^{n} w_j, \tag{2}$$

and $w_i = e^{\theta^T g(\mathrm{x}_i)}, w^* = \mathbb{E}[\exp(\theta^T g(\mathrm{x}))]$. Denote the corresponding CDF by $F_{n,\theta}$. Note that $\mu_{n,\theta}$ is a random point measure, or more precisely a point process in $\mathbb{R}^d$. Define

$$M_q(\mu) = \mathbb{E}_{\mathrm{x}\sim\mu}\|\mathrm{x}\|^q$$

$$W_k = \mathbb{E}[w^k/(\mathbb{E}w^{k/2})^2]^{1/k}$$

$$C_w = \mathbb{E}w^{-2}\mathbb{E}w^2$$

Before presenting our first result, we define the p-Wasserstein distances W_p for $p \geq 1$ as follows. For any measures $\nu_i, i = 1, 2$ on possibly different probability spaces $(\Omega_i, \mathcal{F}_i), i = 1, 2$, we call a probability measure π on $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2)$ a *coupling* of ν_1 and ν_2, if ν_1 and ν_2 are the marginals of π(i.e. $\pi(\Omega_1 \times \cdot) = \nu_2(\cdot)$ and $\pi(\cdot \times \Omega_2) = \nu_1(\cdot)$. Let $\Pi(\nu_1, \nu_2)$ be the set of all couplings of ν_1 and ν_2. Define

$$\mathcal{W}_p(\nu_1, \nu_2) = \inf_{\pi\in\Pi(\nu_1,\nu_2)} \left(\mathbb{E}_{\mathrm{x}\sim\nu_1,\mathrm{y}\sim\nu_2}\|\mathrm{x}-\mathrm{y}\|^p\right)^{\frac{1}{p}}. \tag{3}$$

Before presenting our results on the rates of expected Wasserstein distance between a weighted empirical sampler and the true tilted distribution, we will briefly review the results obtained by (Fournier & Guillin, 2013) on the rates of the iid empirical sampler, instead.

Theorem. *Let $\mu \in \mathcal{P}(\mathrm{d})$ and let $p > 0$. Assume that $M_q(\mu) < \infty$ for some $q > p$. There exists a constant C depending only on p, d, q such that, for all $N \geq 1$, $p < d/2$ and $d > \frac{qp}{p-q}$ we have*

$$\mathbb{E}\left(\mathcal{W}_p(\mu_N, \mu)\right) \leq CM_q(\mu)^{\frac{p}{q}}[N^{-p/d} + N^{-1/2}]$$

We now present our results. We believe that these results are of independent interest. We adapt the proof techniques of Fournier & Guillin (2013) to prove these. The first result assumes moment conditions on μ; whereas the second assumes that $\|g(x)\| \leq g^m$. This clearly leads to a better rate, although, for high enough dimensions the rates remain the same.

Theorem 4. *For measures $\mu_\theta, \mu_{N,\theta}$, as defined in equation 1 and equation 2 respectively, with $M_q(\mu) = \mathbb{E}[\|\mathrm{x}\|^q] < \infty$ where $\mathrm{x} \sim \mu$ and $d > \frac{qp}{q-2p}$ with $q > 2p$, we have*

$$\mathbb{E}[W_p(\mu_{N,\theta}, \mu_\theta)] \leq CC_w[W_4 M_q^{1/4} N^{-p/d} + 4W_2 M_q^{1/2} N^{-1/2}]$$

where $W_k = \mathbb{E}[w^k/(\mathbb{E}w^{k/2})^2]^{1/k}$, $C_w = \mathbb{E}w^{-2}\mathbb{E}w^2$ and C is a constant independent of μ and N.

Theorem 5. *For measures $\mu_\theta, \mu_{N,\theta}$ defined in equation 1 and equation 2 respectively, with $\|g(\mathrm{x})\| \leq g^m$ where $\mathrm{x} \sim \mu$, $d > \frac{qp}{q-p}$ with $q > p$, we have*

$$\mathbb{E}[W_p(\mu_{N,\theta}, \mu_\theta)] \leq CVS(N^{-\frac{p}{d}} + N^{-1/2})$$

where $V = \exp(\|\theta\| g^m)/w^*$ and $S = \max(\sqrt{M_q}, M_q)$.

The result demonstrates that even with weighted empirical samplers you obtain similar rates for high enough dimensions, albeit the constants vary. For d not too big, however, the rates do get slightly worse. However, if the function $g(X)$ has bounded norm, then, you end up getting the good rates (with different constants), yet again.

Here, we were not able obtain to $M_q(\mu)^{\frac{p}{q}}$ in the constants because the scaling argument used in (Fournier & Guillin, 2013) requires homogeneous behavior in the function $g(X)$, which is a strong assumption. We don't make that assumption in this paper.

The following lemma will be used to prove the theorems. The complete proof is in Appendix A.

Lemma 1. *For a Borel set A,*

$$\mathbb{E}[|\mu_\theta - \mu_{n,\theta}|(A)] \leq \frac{1}{\sqrt{n}}\sqrt{\mathbb{E}\left[w^{-2}\right]\mathbb{E}w^2}\left[\sqrt{\mu_{2\theta}(A)} + \mu_\theta(A)\right].$$

The following corollary of the theorem immediately implies asymptotic accuracy of reweighed sampling as N, θ converge to ∞ under suitable conditions.

Corollary 1. *Under the assumptions of 4, if for a sequence* $\{(N, \theta_N)\}$,

$$C_w[W_4 M_q^{1/4} N^{-p/d} + 4W_2 M_q^{1/2} N^{-1/2}] \to 0$$

with the data dimension $d > \frac{qp}{(q-2p)}$ *and* $q > 2p$*, then,*

$$\mathbb{E}W_p(\mu_{N,\theta_N}, \mu_{\theta_N}) \to 0.$$

We remark that under boundedness and appropriate tail assumptions on μ*, one can prove that* W_k *grows at most polynomially in* θ*, thus demonstrating the ease of sampling from bounded distributions. This is achieved using regular variation techniques and Tauberian theorems (see Bingham et al. (1987) for further details). We defer this discussion to Appendix A.*

4 ON THE ACCURACY OF DIFFUSION

In this section, we prove that if two bounded distributions are close in either the W_1 or W_2 distance, then diffusion accompanied by good score-matching generates samples which are highly accurate with respect to the TV distance. Hasty readers can quickly go to 8 to look at our main result.

We begin by stating our assumptions, which are inspired from Chen et al. (2023). Let μ be a measure whose samples are provided as an initial input to a diffusion sampler. For completeness, we shall now explain briefly the process of diffusion sampling.

Diffusion begins with the forward process, which is an Ornstein-Uhlenbeck process started at time 0 from the measure μ. More specifically, let $\{\mathrm{x}_t\}_{t\geq 0}$ be the solution of the stochastic differential equation (SDE)

$$d\mathrm{x}_t = -\eta \mathrm{x}_t dt + \sqrt{2}\sigma d\mathrm{b}_t \tag{4}$$

on $[0, T]$, where $\{\mathrm{b}_t\}_{t\geq 0}$ is a Brownian motion, $\eta > 0$ is some drift parameter, σ is a noise parameter and $T > 0$ is some fixed time endpoint which must equal $+\infty$ for diffusion to be theoretically accurate. It is well known that x_t converges to a centered Gaussian in distribution as $t \to \infty$. In particular, the solution to this equation is given by

$$\mathrm{x}_t = e^{-\eta t}\mathrm{x}_0 + \sigma \int_0^t e^{-\eta(t-s)} db_s. \tag{5}$$

Therefore, assuming that x_T is close to the limiting normal, one wishes to reverse the process conducted above starting from normal samples, to obtain new samples of x_t. This can, in fact, be done. We can show that $\mathrm{x}_t^{\leftarrow} = \{\mathrm{x}_{T-t}\}_{t\geq 0}$ satisfies the SDE

$$d\mathrm{x}_t^{\leftarrow} = \{\eta \mathrm{x}_t^{\leftarrow} + 2\nabla \ln q_{T-t}(\mathrm{x}_t^{\leftarrow})\}dt + \sqrt{2}d\mathrm{b}_t^{\leftarrow}, \tag{6}$$

where $\mathrm{b}_t^{\leftarrow}$ is the reversed Brownian motion. Here, $\nabla \ln q_t$ is referred to as the *score* function, and note that this must be ascertained before the reverse process begins. That is done using a score-matching algorithm using a neural network; we refer the reader to (Chen et al., 2023, Section A.1) for the details. Note that diffusion, as performed above, suffers from three inaccuracies, namely inefficient score estimation, insufficiency of T, and any discretization that may be used to simulate the solutions of the SDEs.

Our focus is on the inefficiency of score matching, which we capture using a loss function. For each $t \in [0, T]$, let $s(\mathrm{x}_t, t)$ be the estimator of the score $\nabla \ln q_t(\mathrm{x}_t)$ at time t and space point x_t. Let $f_t(\mathrm{x}_t) = s(\mathrm{x}_t, t) - \ln q_t(\mathrm{x}_t)$, and define

$$l(\mu) = \frac{1}{T}\int_0^T \mathbb{E}\|f_t(\mathrm{x}_t)\|^2 dt. \tag{7}$$

We shall now make a Lipschitz assumption on $f_t(x)$.

Assumption 1. *There is a real-valued function* $L_t, t \in [0, T]$ *such that* $f_t(x)$ *is* L_t *Lipschitz for each* $t \in [0, T]$*, and*

$$C_\eta = \frac{1}{T}\int_0^T L_{t_0}^2 e^{-2\eta t_0} dt_0 < \infty.$$

A brief discussion of this assumption follows. A Lipschitz assumption on the score function is made quite often. However, one can also show that by assuming the score network to have the same Lipschitz constant as that of the true score function, the denoiser error can only decrease. This means by choosing NN architecture that behaves appropriately, one could justify this assumption. Also, note, in particular, we allow L_{t_0} to be a constant in t_0, something which is seen in Chen et al. (2023) as well, whose result we will need to use. However, L_{t_0} may very well be unbounded near 0 or T, something that is rather unusual as a condition in diffusion-based sampling. This adds to the generality of our second main result, which we shall now state and believe is one that is of independent interest. Define

$$\Delta(\mu, \nu, f) = \left|\mathbb{E}_\mu\|f_t(\mathrm{x}_t)\|^2 - \mathbb{E}_\nu\|f_t(\mathrm{y}_t)\|^2\right|$$

where $t \sim \mathrm{Unif}[0, T]$ is random as well.

Theorem 6. *Let* μ, ν *be probability distributions and* $\varepsilon > 0$ *be such that* $l(\nu) \leq \varepsilon^2$*. Let* $\{\mathrm{x}_t\}_{t\in[0,T]}$ *and* $\{\mathrm{y}_t\}_{t\in[0,T]}$ *be the forward processes specified by equation 4 with* $\mathrm{x}_0 = \mu$ *and* $\mathrm{y}_0 = \nu$ *respectively. Then,*

1. *we have*

$$\Delta(\mu, \nu, f) \leq C_\eta W_2^2(\mu, \nu) + 2\sqrt{C_\eta}W_2(\mu, \nu)\varepsilon. \tag{8}$$

where C_η *is as in Assumption 1.*

2. *If, in addition, μ and ν are concentrated on the set $\{\|x\| \leq M\}$, then*

$$\Delta(\mu, \nu, f) \leq (2C_\eta M + 2\sqrt{C_\eta}\varepsilon)W_2(\mu, \nu). \quad (9)$$

3. *Furthermore, under the boundedness assumption of the previous point,*

$$\Delta(\mu, \nu, f) \leq 2MC_\eta W_1(\mu, \nu) + 2\sqrt{2MC_\eta \varepsilon}\sqrt{W_1(\mu, \nu)}. \quad (10)$$

Informally, probability measures close in the Wasserstein distances can be diffusion sampled with high accuracy provided at least one of them can be, as the denoiser error completely determines the performance of sampling via diffusion. We state this a bit later. Observe that the stated bounds cover both the bounded and unbounded regimes, and proximity in both W_2 and W_1, thereby adding to its versatility.

Theorem 7. *Following the setup of 6, take $\mu = \mu_\theta$ and $\nu = \mu_{N,\theta}$. Then,*

1. *If μ is concentrated on the set $\{\|x\| \leq M\}$, then*

$$\mathbb{E}[\Delta(\mu_\theta, \mu_{N,\theta}, f)] \leq (2C_\eta M + 2\sqrt{C_\eta}\varepsilon) \cdot \mathbb{E}W_2(\mu_\theta, \mu_{N,\theta}). \quad (11)$$

2. *Furthermore, under the boundedness assumption of the previous point,*

$$\mathbb{E}[\Delta(\mu_\theta, \mu_{N,\theta}, f)] \leq 2MC_\eta \mathbb{E}W_1(\mu_\theta, \mu_{N,\theta}) + 2\sqrt{2MC_\eta\varepsilon}\sqrt{\mathbb{E}W_1(\mu_\theta, \mu_{N,\theta})}. \quad (12)$$

where the first expectation is on the randomness from the random measure $\mu_{N,\theta}$.

The following is trivial from the previous theorem and $\mathbb{E}\sqrt{\mathrm{x}} \leq \sqrt{\mathbb{E}\mathrm{x}}$.

Corollary 2. *Set $f_t(x)$ to be the map $x \to s(x, t) - \nabla \log q_t(x)$. Under Assumption 1, and $\ell(\mu_{N,\theta}) \leq \varepsilon^2$. Then,*

1. *If μ is concentrated on the set $\{\|x\| \leq M\}$, then*

$$\mathbb{E}_{\mu_\theta}\|f_t(\mathrm{x}_t)\|^2 \leq \varepsilon^2 + (2C_\eta M + 2\sqrt{C_\eta}\varepsilon) \cdot \mathbb{E}W_2(\mu_\theta, \mu_{N,\theta}) \quad (13)$$

2. *Furthermore, under the boundedness assumption of the previous point,*

$$\mathbb{E}_{\mu_\theta}\|f_t(\mathrm{x}_t)\|^2 \leq \varepsilon^2 + 2MC_\eta \mathbb{E}W_1(\mu, \nu) + 2\sqrt{2MC_\eta\varepsilon}\sqrt{\mathbb{E}W_1(\mu_\theta, \mu_{N,\theta})} \quad (14)$$

We remark that it is possible to obtain a similar result with an assumption of the form $\mathbb{E}\ell(\mu_{N,\theta}) \leq \varepsilon^2$ as well.

It is difficult to obtain a result like this without a tail condition like μ being supported on $\{\|x\| \leq M\}$. This is because it is difficult to obtain concentration inequalities on $W_2(\mu_\theta, \mu_{N,\theta})$ (and hence difficult to bound $\mathbb{E}W_2^2(\mu_\theta, \mu_{N,\theta})$). Recently, (Lei, 2020) obtained concentration inequalities on $W_2(\mu, \mu_N)$ where μ_N is the iid empirical sampler under a Log Sobolev assumption, and using the Lipschitz-ness of $W_2(\mu, \mu_N)$ as a function from $\mathcal{X}$ to $\mathbb{R}$.

Let μ_N and μ'_N be two iid samplers through the points $(x_i)_i$ and $(x'_i)_i$ respectively. The Lipschitz condition is proved using $|W_2(\mu_N, \mu) - W_2(\mu, \mu'_N)| \leq W_2(\mu_N, \mu'_N)$, and then noticing that the optimal coupling is the trivial coupling that assigns mass $\frac{1}{N}$ to the pair (x_i, x'_i).

However, even under Log Sobolev, it isn't possible to obtain a Lipschitz condition for our weighted sampler. This is because we don't obtain a trivial coupling of $W_2(\mu_{N,\theta}, \mu'_{N,\theta})$ anymore as the weights could behave arbitrarily badly.

Our final theorem asserts the accuracy of DDPM in this setting. Note we use a different target than in Corollary 1. We do this as it is often easier to optimize on this target as opposed to the earlier one. This is justified via noticing that

$$W_k \leq \left[\mathbb{E}\frac{w^k}{(w^*)^k}\right]^{1/k}.$$

Theorem 8. *Under Assumption 1, $\|\mathrm{x}\| \leq M$, $d > qp/(q-2p)$, and with twist with weight $w = \exp(\theta^T g(\mathrm{x}))$, if*

$$C_w S\left[\left[\mathbb{E}\frac{w^4}{(w^*)^4}\right]^{1/4} + 4\left[\mathbb{E}\frac{w^2}{(w^*)^2}\right]^{1/2}\right] N^{-p/d} \leq \delta$$

where $S = \max(\sqrt{M_q}, M_q)$ for $p = 1$ or 2 and diffusion is run on the twisted empirical measure $\mu_{N,\theta}$, then the output of the diffusion process has $\mathcal{O}(C_\eta M\delta + \sqrt{MC_\eta}\sqrt{\delta})$ or $\mathcal{O}(C_\eta M\delta)$ TV-error with respect to μ_θ for $p = 1$ or $p = 2$ respectively, assuming the other sources of error (discretization and the time the reverse process is run) is small.

5 EXPERIMENTS

We now demonstrate our approach on both unbounded and bounded high-dimensional distributions. In each case, we compare three methods: i) reweighted sampling, ii) reweighted sampling combined with diffusion, and iii) the force-guided diffusion strategy proposed in Wang et al. (2024).

5.1 Multivariate Gaussian

As a first experiment, we consider an unbounded setting: a 50-dimensional multivariate Gaussian distribution which is exponentially twisted with different values of θ. Figure 1 plots the sum of the entries of the sampled vectors under the three methods above. This illustrates the difficulty of twisting unbounded variables: beyond a small threshold, the number of samples required to achieve a stable twist grows exponentially.

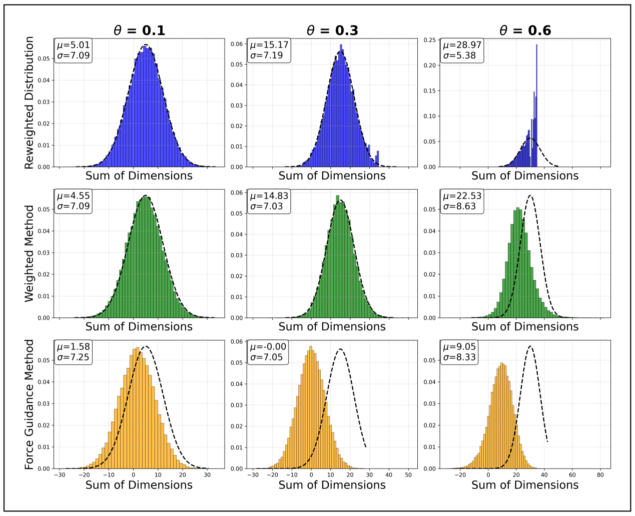

Figure 1: Samples generated by twisting a multivariate gaussian in 50 dimensions by $\boldsymbol{\theta} = \theta \cdot (1, \ldots, 1)$, for $\theta = 0.1,\ 0.2,\ 0.3$, and 0.6 using reweighed sampling, weighted diffusion, and the force guided method.

5.2 Bounded, Correlated Target

In the second experiment we consider a distribution with a bounded support.

Starting from independent bounded marginals X, we form $Y = AX$ with an arbitrary matrix A, whose column sums to 1, so each coordinate of Y is a convex combination of the X_i's. This yields a correlated, non-Gaussian target with a bounded support. We report performance across tilt parameters θ; full construction details are in Appendix B.

Figure 2 plots the sum of the entries of the vectors under the three methods above.

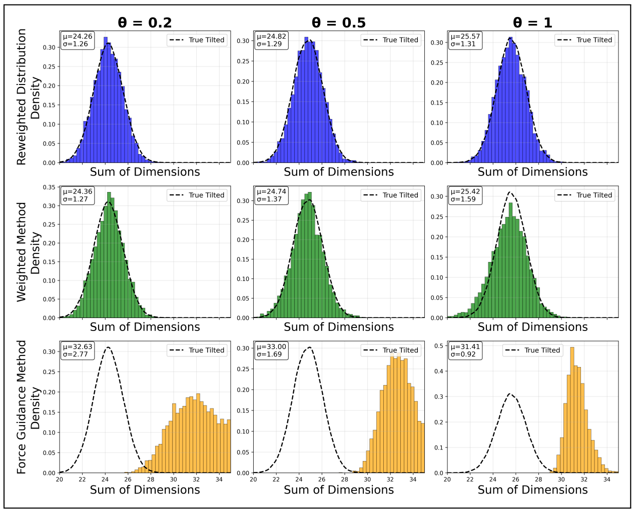

Figure 2: Samples generated by twisting a bounded in 50 dimensions by $\boldsymbol{\theta} = \theta \cdot (1, \ldots, 1)$, for $\theta = 0.2,\ 0.5,\ 1$ using reweighed sampling, weighted diffusion, and the force guided method.

Figure 3 shows that as $\left[\left[\mathbb{E}w^4\right]^{1/4} + \left[\mathbb{E}w^2\right]^{1/2}\right] \frac{N^{-p/d}}{w^*}$ reduces, so does the W_1 distance of the distributions generated by diffusion.

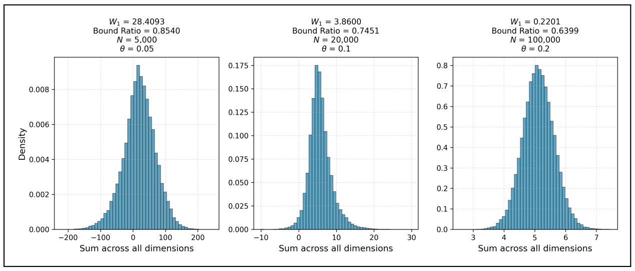

Figure 3: Empirical validation of the proposed bound: as the ratio reduces, the Wasserstein-1 distance correspondingly reduces.

6 CONCLUSION

In this work, our goal was to prove that diffusion run on a weighted empirical sampler would, under certain conditions, also lead to accurate samples! Our approach to this was by first deriving upper bounds on the expected Wasserstein distance between the weighted empirical sampler and the true twisted distribution. And, then, we use that to obtain bounds on the TV error between the twisted samples and the true twisted distribution.

The main limitation of this work is the bounded support assumption required for the final result. We have discussed the reasons of the difficulty in 4. Future work could try to use different styles of arguments to fix this difficulty, as we believe the Wasserstein style of arguments probably wouldn't work for this.

References

Javier Aguilar and Riccardo Gatto. Unified perspective on exponential tilt and bridge algorithms for rare trajectories of discrete markov processes. *Phys. Rev. E*, 109:034113, Mar 2024. doi: 10.1103/PhysRevE.109.034113. URL https://link.aps.org/doi/10.1103/PhysRevE.109.034113.

Mayer Alvo. *Exponential Tilting and Its Applications*, pp. 171–193. Springer International Publishing, Cham, 2022. ISBN 978-3-031-06784-6. doi: 10.1007/978-3-031-06784-6_6. URL https://doi.org/10.1007/978-3-031-06784-6_6.

M Avellaneda. Minimum entropy calibration of asset pricing models, internat. *J. Theoret. Appl. Finance*, 1:447472, 1998.

N. H. Bingham, C. M. Goldie, and J. L. Teugels. *Regular Variation*, volume 27 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 1987. ISBN 0521307872. doi: 10.1017/CBO9780511721434.

Peter W Buchen and Michael Kelly. The maximum entropy distribution of an asset inferred from option prices. *Journal of Financial and Quantitative Analysis*, 31(1):143–159, 1996.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R. Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions, 2023. URL https://arxiv.org/abs/2209.11215.

Rama Cont, Mihai Cucuringu, Renyuan Xu, and Chao Zhang. Tail-gan: Learning to simulate tail risk scenarios. *Management Science*, 2025.

Imre Csiszár. On topology properties of f-divergences. *Studia Scientifica Mathematica Hungerica*, 2:329–339, 1967.

Imre Csiszár. Axiomatic characterizations of information measures. *Entropy*, 10(3):261–273, 2008.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure, 2013. URL https://arxiv.org/abs/1312.2128.

H.U. Gerber and E.S.W. Shiu. Option pricing by esscher transforms, 1994.

T. Goll and Ludger Rueschendorf. *Minimal distance martingale measures and optimal portfolios consistent with observed market prices*. Taylor and Francis, 01 2002.

Lingkai Kong, Haichuan Wang, Yuqi Pan, Cheol Woo Kim, Mingxiao Song, Alayna Nguyen, Tonghan Wang, Haifeng Xu, and Milind Tambe. Robust optimization with diffusion models for green security. *arXiv preprint arXiv:2503.05730*, 2025.

Suemin Lee, Ruiyu Wang, Lukas Herron, and Pratyush Tiwary. Exponentially tilted thermodynamic maps (exptm): Predicting phase transitions across temperature, pressure, and chemical potential, 2025. URL https://arxiv.org/abs/2503.15080.

Jing Lei. Convergence and concentration of empirical measures under wasserstein distance in unbounded functional spaces. *Bernoulli*, 26(1), 2020. ISSN 1350-7265. doi: 10.3150/19-bej1151. URL http://dx.doi.org/10.3150/19-BEJ1151.

Lizao Li, Robert Carver, Ignacio Lopez-Gomez, Fei Sha, and John R. Anderson. Generative emulation of weather forecast ensembles with diffusion models. *Science Advances*, 10(13):eadk4489, 2024. doi: 10.1126/sciadv.adk4489. URL https://www.science.org/doi/10.1126/sciadv.adk4489.

Fenghua Ling, Zeyu Lu, Jing-Jia Luo, Lei Bai, Swadhin K. Behera, Dachao Jin, Baoxiang Pan, Huidong Jiang, Toshio Yamagata, et al. Diffusion model-based probabilistic downscaling for 180-year east asian climate reconstruction. *npj Climate and Atmospheric Science*, 7:131, 2024. doi: 10.1038/s41612-024-00679-1. URL https://www.nature.com/articles/s41612-024-00679-1.

Attilio Meucci. Fully flexible views: Theory and practice, 2010. URL https://arxiv.org/abs/1012.2848.

Michael Stutzer. A simple nonparametric approach to derivative security valuation. *The Journal of Finance*, 51(5):1633–1652, 1996. doi: https://doi.org/10.1111/j.1540-6261.1996.tb05220.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1996.tb05220.x.

Yan Wang, Lihao Wang, Yuning Shen, Yiqun Wang, Huizhuo Yuan, Yue Wu, and Quanquan Gu. Protein conformation generation via force-guided se(3) diffusion models, 2024. URL https://arxiv.org/abs/2403.14088.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Not Applicable. We use the usual DDPM sampling algorithm.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Not Applicable

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes.
   (b) Complete proofs of all theoretical results. Yes. In the Appendix. A
   (c) Clear explanations of any assumptions. Yes.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes (in the supplementary material).
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes (in the supplementary material).
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes (in the supplementary material).
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes (in the supplementary material).

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Not Applicable
   (b) The license information of the assets, if applicable. Not Applicable
   (c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable
   (d) Information about consent from data providers/curators. [Yes/No/Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable

Supplementary Materials: Proofs and Experiments

A PROOFS

A.1 Theorems 1, 2 and 3

Theorem (KL ⇒ exponential tilt). *Let Q be a reference probability measure and let Φ be measurable with the integrability required. Consider*

$$\mathcal{J}_{\mathrm{KL}}(P) = \mathbb{E}_P[\Phi] - D_{\mathrm{KL}}(P\|Q)$$

$$D_{\mathrm{KL}}(P\|Q) = \int \log\frac{dP}{dQ}\, dP$$

Maximizing $\mathcal{J}_{\mathrm{KL}}$ over probability measures $P \ll Q$ yields the unique optimizer $P^\star$ whose density $h^\star := \dfrac{dP^\star}{dQ}$ satisfies

$$\log h^\star(x) = \Phi(x) + \mathrm{const} \quad \textit{(a.e.)},$$

hence

$$\frac{dP^\star}{dQ}(x) = \frac{\exp(\Phi(x))}{\displaystyle\int \exp(\Phi)\, dQ}.$$

Theorem (Rényi ⇒ power/escort tilt). *Fix $\alpha > 0$, $\alpha \neq 1$. For $h = dP/dQ$ define*

$$D_\alpha(P\|Q) = \frac{1}{\alpha - 1}\log\Big(\int h^\alpha\, dQ\Big).$$

Consider the functional $\mathcal{J}_\alpha(P) = \mathbb{E}_P[\Phi] - D_\alpha(P\|Q)$. A maximizer $P^\star$ (if it exists) has density $h^\star$ satisfying a.e.

$$\frac{\alpha}{\alpha-1}\frac{(h^\star(x))^{\alpha-1}}{\int (h^\star)^\alpha dQ} = \Phi(x) - \mu$$

for some constant μ, equivalently (after absorbing constants)

$$h^\star(x) \;\propto\; \big(a + b\,\Phi(x)\big)^{1/(\alpha-1)},$$

i.e. a power/escort–type tilt (which approaches the exponential tilt as $\alpha \to 1$).

Theorem (Tsallis ⇒ q–exponential tilt). *Fix $q > 0$, $q \neq 1$. Using one convenient form of Tsallis relative entropy*

$$D_q^{\mathrm{Ts}}(P\|Q) = \frac{1}{q-1}\Big(1 - \int h^q\, dQ\Big), \qquad h = \frac{dP}{dQ},$$

consider $\mathcal{J}_{\mathrm{Ts}}(P) = \mathbb{E}_P[\Phi] - D_q^{\mathrm{Ts}}(P\|Q)$. A maximizer (when it exists) satisfies a.e.

$$\frac{q}{q-1} h^\star(x)^{\,q-1} = \mu - \Phi(x)$$

for some constant μ, and hence after reparameterization

$$h^\star(x) \;\propto\; \big[1 + (1-q)\,c\,\Phi(x)\big]^{1/(1-q)} = \exp_q(c\,\Phi(x)),$$

the q–exponential (with the usual normalization).

Proof to Theorem 1. Write $h = dP/dQ$. Form the Lagrangian with multiplier μ for $\int h\,dQ = 1$:

$$\mathcal{L}(h,\mu) = \int \Phi\,h\,dQ - \int h\log h\,dQ - \mu\Big(\int h\,dQ - 1\Big).$$

For any perturbation φ with $\int \varphi\,dQ = 0$ the first variation vanishes:

$$0 = \frac{d}{d\varepsilon}\Big|_{\varepsilon=0}\mathcal{L}(h+\varepsilon\varphi,\mu) = \int \varphi(x)\big(\Phi(x) - (\log h(x) + 1) - \mu\big)\,dQ(x).$$

Since this holds for all admissible φ the bracket is a.e. constant, so $\log h(x) = \Phi(x) - C$. Exponentiating and normalizing yields the displayed exponential form. □

Proof to Theorem 2. Put $A(h) := \int h^\alpha\,dQ$. The Lagrangian with multiplier μ is

$$\mathcal{L}(h,\mu) = \int \Phi\,h\,dQ - \frac{1}{\alpha-1}\log A(h) - \mu\Big(\int h\,dQ - 1\Big).$$

For perturbation φ with $\int \varphi\,dQ = 0$,

$$0 = \frac{d}{d\varepsilon}\Big|_{\varepsilon=0}\mathcal{L}(h+\varepsilon\varphi,\mu) = \int \varphi(x)\Big(\Phi(x) - \frac{\alpha}{\alpha-1}\frac{h(x)^{\alpha-1}}{A(h)} - \mu\Big)\,dQ(x).$$

Thus the bracket is a.e. constant. Rearranging gives the algebraic relation displayed above, and solving for h (absorbing multiplicative/additive constants into a, b and the normalizer) yields the power–law tilt. □

Proof to Theorem 3. Up to an additive constant one may write the objective as $\int \Phi\,h\,dQ + \frac{1}{q-1}\int h^q\,dQ$. Form the Lagrangian

$$\mathcal{L}(h,\mu) = \int \Phi\,h\,dQ + \frac{1}{q-1}\int h^q\,dQ - \mu\Big(\int h\,dQ - 1\Big).$$

The first variation for perturbations φ with $\int \varphi\,dQ = 0$ yields

$$0 = \int \varphi(x)\Big(\Phi(x) + \frac{q}{q-1}h(x)^{q-1} - \mu\Big)\,dQ(x).$$

Thus the bracket is a.e. constant, giving the algebraic relation above; solving and absorbing constants gives the q–exponential representation. □

A.2 Theorems 4 and 5

Theorem 9. *For measures $\mu_\theta, \mu_{N,\theta}$, as defined in equation 1 and equation 2 respectively, with $M_q(\mu) = \mathbb{E}[\|\mathrm{x}\|^q] < \infty$ where $\mathrm{x} \sim \mu$ and $d > \frac{qp}{q-2p}$ with $q > 2p$, we have*

$$\mathbb{E}[W_p(\mu_{N,\theta},\mu_\theta)] \le CC_w[W_4 M_q^{1/4} N^{-p/d} + 4W_2 M_q^{1/2} N^{-1/2}]$$

where $W_k = \mathbb{E}[w^k/(\mathbb{E}w^{k/2})^2]^{1/k}$, $C_w = \mathbb{E}w^{-2}\mathbb{E}w^2$ and C is a constant independent of μ and N.

Theorem 10. *For measures $\mu_\theta, \mu_{N,\theta}$ defined in equation 1 and equation 2 respectively, with $\|g(\mathrm{x})\| \le g^m$ where $\mathrm{x} \sim \mu$, $d > \frac{qp}{q-p}$ with $q > p$, we have*

$$\mathbb{E}[W_p(\mu_{N,\theta},\mu_\theta)] \le CVS(N^{-\frac{p}{d}} + N^{-1/2})$$

where $V = \exp(\|\theta\| g^m)/w^*$ and $S = \max(\sqrt{M_q}, M_q)$.

Lemma 2. *For a Borel set A,*

$$\mathbb{E}[|\mu_\theta - \mu_{n,\theta}|(A)] \le \frac{1}{\sqrt{n}}\sqrt{\mathbb{E}\left[w^{-2}\right]\mathbb{E}w^2}\left[\sqrt{\mu_{2\theta}(A)} + \mu_\theta(A)\right].$$

Proof. Redefine the probability measure as a fraction of non-normalized measures and their mass on $\mathbb{R}^d$. That is, let $\mu_\theta = \tilde{\mu}_\theta / w^*, \mu_{n,\theta} = \tilde{\mu}_{n,\theta}/w_n^*$ where $w^* = \mathbb{E}[w], w_n^* = \frac{1}{n}\sum_i w_i$. Now,

$$\begin{aligned}
|\mu_\theta - \mu_{n,\theta}|(A) = \left|\frac{\tilde{\mu}_\theta}{w^*} - \frac{\tilde{\mu}_{n,\theta}}{w_n}\right|(A) \le & \frac{1}{w_n}\left|\tilde{\mu}_\theta - \tilde{\mu}_{n,\theta}\right|(A) + \tilde{\mu}_\theta \left|\frac{w_n - w^*}{w^* w_n}\right|(A) \\
\le & \frac{1}{w_n}\left[|\tilde{\mu}_\theta - \tilde{\mu}_{n,\theta}|(A) + \tilde{\mu}_\theta \left|\frac{w_n - w^*}{w^*}\right|(A)\right].
\end{aligned}$$

Taking the expectation of both sides and applying the Cauchy-Schwarz inequality to the right hand side,

$$\begin{aligned}
\mathbb{E}|\mu_\theta - \mu_{n,\theta}|(A) \le & \sqrt{\mathbb{E}\left[\frac{1}{w_n^2}\right]}\left[\sqrt{\mathbb{E}|\tilde{\mu}_\theta - \tilde{\mu}_{n,\theta}|^2(A)} + \frac{\tilde{\mu}_\theta}{w^*}\sqrt{\mathbb{E}|w_n - w^*|^2}\right] \\
\le & \sqrt{\mathbb{E}\left[\frac{1}{w_n^2}\right]}\left[\sqrt{V(\tilde{\mu}_{n,\theta}(A))} + \frac{\tilde{\mu}_\theta}{w^*}\sqrt{V(w_n)}\right] \\
\le & \sqrt{\mathbb{E}\left[\frac{1}{w_n^2}\right]}\left[\frac{1}{\sqrt{n}}\sqrt{\mathbb{E}(w^2 I(\mathrm{x} \in A))} + \frac{\tilde{\mu}_\theta}{w^*}\frac{1}{\sqrt{n}}\sqrt{\mathbb{E}w^2}\right].
\end{aligned} \tag{15}$$

Here, in the second line, we note that for any centered random variable x we have $(\mathbb{E}|\mathrm{x}|)^2 \le \mathbb{V}\mathrm{ar}\,(\mathrm{x})$, and used this observation with $\mathrm{x} = \tilde{\mu}_{n,\theta}(A)$ and $\mathrm{x} = w_n$ respectively. In the third line, we used the definitions of $\tilde{\mu}_{n,\theta}$ and w_n.

In order to finish the proof, we make the following two observations : $\mu_{2\theta}(A) = \frac{\mathbb{E}w^2 I(\mathrm{x}\in A)}{\mathbb{E}w^2}$ by definition, and

$$\mathbb{E}\frac{1}{w_n^2} = \mathbb{E}\left[\frac{1}{(\frac{1}{n}\sum_{i=1}^n \exp(\theta^T f(\mathrm{x}_i))^2}\right] \le \mathbb{E}\left[\exp\left(-2\theta^T\left[\frac{1}{n}\sum_{i=1}^n f(\mathrm{x}_i)\right]\right)\right] = \mathbb{E}[w^{-2}].$$

Substituting these into equation 15 completes the proof of theorem. □

Before proceeding to the next lemma, we will set some notation.

Notation 1. *We have the following as in (Fournier & Guillin, 2013).*

(a) For $\ell \ge 0$, denote by $\mathcal{P}_\ell$ the natural partition of $(-1,1]^d$ into $2^{d\ell}$ translations of $(-2^{-\ell}, 2^{-\ell}]^d$. For two probability measures μ, ν on $(-1,1]^d$ and for $p > 0$, define

$$\mathcal{D}_p(\mu,\nu) := \frac{2^p - 1}{2}\sum_{\ell \ge 1} 2^{-p\ell} \sum_{F \in \mathcal{P}_\ell} |\mu(F) - \nu(F)|,$$

which is a distance on $\mathcal{P}((-1,1]^d)$ that is bounded by 1.

(b) Define $B_0 := (-1,1]^d$ and, for $n \ge 1$, $B_n := (-2^n, 2^n]^d \setminus (-2^{n-1}, 2^{n-1}]^d$. For $\mu \in \mathcal{P}(\mathbb{R}^d)$ and $n \ge 0$, denote by $\mathcal{R}_{B_n}\mu$ the probability measure on $(-1,1]^d$ defined as the image of $\mu|_{B_n}/\mu(B_n)$ by the map $x \mapsto x/2^n$. For two probability measures μ, ν on $\mathbb{R}^d$ and for $p > 0$, define

$$\mathcal{D}_p(\mu,\nu) := \sum_{n \ge 0} 2^{pn}\big(|\mu(B_n) - \nu(B_n)| + (\mu(B_n) \wedge \nu(B_n))\mathcal{D}_p(\mathcal{R}_{B_n}\mu, \mathcal{R}_{B_n}\nu)\big).$$

We will now state the following two lemmas from Fournier & Guillin (2013). The proofs may be found in that paper.

Lemma 3. *Let $d \ge 1$ and $p > 0$. For all pairs of probability measures μ, ν on $\mathbb{R}^d$, $\mathcal{T}_p(\mu,\nu) \le \kappa_{p,d}\mathcal{D}_p(\mu,\nu)$, with $\kappa_{p,d} := 2^{p(1+d/2)}(2^p+1)/(2^p-1)$.*

Lemma 4. *Let $p > 0$ and $d \geq 1$. There is a constant C, depending only on p, d, such that for all $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$,*

$$\mathcal{D}_p(\mu, \nu) \leq C \sum_{n \geq 0} 2^{pn} \sum_{\ell \geq 0} 2^{-p\ell} \sum_{F \in \mathcal{P}_\ell} |\mu(2^n F \cap B_n) - \nu(2^n F \cap B_n)|$$

with the notation $2^n F = \{2^n x \ : \ x \in F\}$.

Given these lemmas, we can now prove t

Proof of Theorem 4. We have from Lemma 2,

$$\mathbb{E}|\mu_\theta - \mu_{N,\theta}|(A) \leq \frac{1}{\sqrt{N}} \sqrt{\mathbb{E}\left[w^{-2}\right] \mathbb{E} w^2} \left[\sqrt{\mu_{2\theta}(A)} + \mu_\theta(A)\right].$$

Taking $A = 2^n F \cap B_n$ and summing over $F \in \mathcal{P}_\ell$, using Cauchy-Schwarz and $\#(\mathcal{P}_\ell) = 2^{dl}$, we have for all $n \geq 0$ that

$$\sum_{F \in \mathcal{P}_\ell} \mathbb{E}(|\mu_{N,\theta}(2^n F \cap B_n) - \mu_\theta(2^n F \cap B_n)|) \leq \frac{\sqrt{\mathbb{E}\left[w^{-2}\right] \mathbb{E} w^2}}{\sqrt{N}} \left[2^{d\ell/2} \sqrt{\mu_{2\theta}(B_n)} + \mu_\theta(B_n)\right].$$

Using $\mu_{2\theta}(B_n) = \mathbb{E}[(w^2/(w^{**}))I(\mathrm{x} \in B_n)] \leq \sqrt{\mathbb{E}[\frac{w^4}{(w^{**})^2}]}\sqrt{\mu(B_n)}$, where $w^{**} = \mathbb{E}w^2$. Similarly bound $\mu_\theta(B_n)$ and define $W_k = \mathbb{E}[\frac{w^k}{(\mathbb{E}w^{k/2})^2}]^{\frac{1}{k}}$.

Also bound $\mu(B_n) \leq P(2^{n-1} \leq ||\mathrm{x}||) \leq \mathbb{E}||\mathrm{x}||^q / 2^{q(n-1)} = M_q / 2^{q(n-1)}$.

Then, using 4,

$$\mathbb{E}(\mathcal{D}_p(\mu_{N,\theta}, \mu_\theta)) \leq C \frac{\sqrt{\mathbb{E}[w{-}2]\mathbb{E}w^2}}{\sqrt{N}} \sum_{n \geq 0} 2^{pn} \sum_{\ell \geq 0} 2^{-p\ell} [2^{d\ell/2} W_4 M_q^{1/4} / 2^{qn/4} + W_2 M_q^{1/2} / 2^{qn/2}]$$

Collect constants to get

$$\mathbb{E}(\mathcal{D}_p(\mu_{N,\theta}, \mu_\theta)) \leq C \frac{C_w}{\sqrt{N}} \sum_{n \geq 0} 2^{pn} \sum_{\ell \geq 0} 2^{-p\ell} [D_4 \cdot 2^{d\ell/2} \cdot 2^{-qn/4} + D_2 \cdot 2^{-qn/2}]$$

where $D_k = W_k M_q^{1/k}$.

We have, from Fournier & Guillin (2013),

$$\sum_{\ell \geq 0} 2^{-p\ell} 2^{d\ell/2} (\varepsilon/N)^{1/2} \leq C \begin{cases} (\varepsilon/N)^{1/2} & \text{if} \quad p > d/2, \\ (\varepsilon/N)^{1/2} \log(2 + \varepsilon N) & \text{if} \quad p = d/2, \\ (\varepsilon N)^{-p/d} & \text{if} \quad p \in (0, d/2). \end{cases}$$

Final step:

$$\mathbb{E}(\mathcal{D}_p(\mu_{N,\theta}, \mu_\theta)) \leq C C_w [\sum_{n \geq 0} 2^{pn} [D_4 2^{-\frac{q}{2} n(1 - p/d)} N^{-p/d}] + \frac{4 D_2}{\sqrt{N}}]$$

Note $pn - \frac{q}{2} n(1 - p/d) = n[\frac{d(2p-q)+pq}{2d}]$ i.e. $q > 2p$ and $d > \frac{pq}{q-2p}$, and if that condition matches, then, you get rates like

$$C C_w [D_4 N^{-p/d} + 4 D_2 N^{-1/2}]$$

which finishes the result. □

Proof of Theorem 5. Note if $||\mathrm{x}|| \leq B$ then $w = \exp(\theta^T g(\mathrm{x})) \leq \exp(||\theta|| ||g_{\max}||) = K$.

Using the previous analysis we get,

$$\sum_{F\in\mathcal{P}_\ell} \mathbb{E}(|\mu_{N,\theta}(2^nF\cap B_n)-\mu_\theta(2^nF\cap B_n)|) \le \frac{\sqrt{\mathbb{E}\left[w^{-2}\right]\mathbb{E}w^2}}{\sqrt{N}}\left[2^{d\ell/2}\sqrt{\mu_{2\theta}(B_n)}+\mu_\theta(B_n)\right]$$

But this time use $\mu_{2\theta}(B_n)=\mathbb{E}[(w^2/(w^{**}))I(\mathrm{x}\in B_n)] \le \frac{K^2}{(w^{**})}\mu(B_n)$ and $\mu_\theta(B_n)\le\frac{K}{w^*}\mu(B_n)$. Let $V=\frac{K}{w^*}$. Therefore,

$$\mathbb{E}(\mathcal{D}_p(\mu_{N,\theta},\mu_\theta)) \le C\frac{\sqrt{\mathbb{E}[w-2]\mathbb{E}w^2}}{\sqrt{N}}\sum_{n\ge0}2^{pn}\sum_{\ell\ge0}2^{-p\ell}V[2^{d\ell/2}M_q^{1/2}/2^{qn/2}+M_q/2^{qn}]$$

Continuing the analysis similarly you get the result. □

A.3 Theorem 6

Theorem. *Let μ,ν be probability distributions and $\varepsilon>0$ be such that $l(\nu)\le\varepsilon^2$. Let $\{\mathrm{x}_t\}_{t\in[0,T]}$ and $\{\mathrm{y}_t\}_{t\in[0,T]}$ be the forward processes specified by equation 4 with $\mathrm{x}_0=\mu$ and $\mathrm{y}_0=\nu$ respectively. Assume that*

$$W_2(\mu,\nu)\le\delta$$

for some $\delta>0$. Then,

1. *we have*
$$\left|\mathbb{E}_\mu\|f_t(\mathrm{x}_t)\|^2-\mathbb{E}_\nu\|f_t(\mathrm{y}_t)\|^2\right| \le C_\eta W_2^2(\mu,\nu)+2\sqrt{C_\eta}W_2(\mu,\nu)\varepsilon \tag{16}$$
where C_η is as in Assumption 1.

2. *If, in addition, μ and ν are concentrated on the set $\{\|x\|\le M\}$, then*
$$\left|\mathbb{E}_\mu\|f_t(\mathrm{x}_t)\|^2-\mathbb{E}_\nu\|f_t(\mathrm{y}_t)\|^2\right| \le (2C_\eta M+2\sqrt{C_\eta}\varepsilon)W_2(\mu,\nu). \tag{17}$$

3. *Furthermore, under the boundedness assumption of the previous point,*
$$\left|\mathbb{E}_\mu\|f_t(\mathrm{x}_t)\|^2-\mathbb{E}_\nu\|f_t(\mathrm{y}_t)\|^2\right| \le 2MC_\eta W_1(\mu,\nu)+2\sqrt{2MC_\eta\varepsilon}\sqrt{W_1(\mu,\nu)}. \tag{18}$$

Proof to Theorem 6. In order to proceed with the proof, we ensure that μ,ν are probability distributions on a common sample space $(\Omega,\mathcal{F})$, and $\{\mathrm{b}_t\}_{t\ge0}$ is a Brownian motion on the same space.

We begin with the proof of equation 16. Let π denote the joint distribution of μ and ν. Then,

$$\begin{aligned}
&\left|\mathbb{E}_\mu\|f_t(\mathrm{x}_t)\|^2-\mathbb{E}_\nu\|f_t(\mathrm{y}_t)\|^2\right|\\
\le&\left|\mathbb{E}_\pi\left[\|f_t(\mathrm{x}_t)\|^2-\|f_t(\mathrm{y}_t)\|^2\right]\right|\\
\le&\mathbb{E}_\pi\left|\left[\|f_t(\mathrm{x}_t)\|^2-\|f_t(\mathrm{y}_t)\|^2\right]\right|\\
=&\frac{1}{T}\int_0^T\mathbb{E}_\pi\left[\left|\left[\|f_t(\mathrm{x}_t)\|^2-\|f_t(\mathrm{y}_t)\|^2\right]\right| \;\middle|\; t=t_0\right]dt_0
\end{aligned} \tag{19}$$

where the equality in the last line, we used the tower rule of conditional expectation.

Observe that for any two arbitrary vectors $u,v\in\mathbb{R}^d$,

$$|\|u\|^2-\|v\|^2|=|(u+v)^T(u-v)|=|(u-v)^T(u-v)+2(u-v)^Tv|\le\|u-v\|^2+2\|u-v\|\|v\|,$$

where we used the triangle inequality and the Cauchy-Schwarz inequality in the last step. From this point on, for $t_0\in[0,T]$ let $E_{t_0,\pi}$ denote the measure E_π conditioned on $t=t_0$. Applying the previous inequality to the last step of equation 19,

$$
\begin{aligned}
&\frac{1}{T}\int_0^T \mathbb{E}_\pi\left[\left|\left[\|f_t(\mathrm{x}_t)\|^2 - \|f_t(\mathrm{y}_t)\|^2\right]\right| \;\;\middle|\;\; t = t_0\right] dt_0 \\
\leq &\frac{1}{T}\int_0^T \mathbb{E}_{\pi,t_0}\left[\|f_t(\mathrm{x}_t) - f_t(\mathrm{y}_t)\|^2\right] + 2\mathbb{E}_{\pi,t_0}\|f_t(\mathrm{x}_t) - f_t(\mathrm{y}_t)\|\|f_t(\mathrm{y}_t)\| dt_0 \\
\leq &\frac{1}{T}\int_0^T \mathbb{E}_{\pi,t_0}\left[L_t^2\|\mathrm{x}_t - \mathrm{y}_t\|^2\right] + 2L_t\sqrt{\mathbb{E}_{\pi,t_0}\|\mathrm{x}_t - \mathrm{y}_t\|^2}\sqrt{\mathbb{E}_{\pi,t_0}\|f_t(\mathrm{y}_t)\|^2} dt_0,
\end{aligned}
\tag{20}
$$

where in the last line we used the L_t Lipschitz assumption on both terms and the Cauchy-Schwarz inequality on the final term.

So far, we have not made an assumption on π. Let π be a minimizer of equation 3 characterizing $W_2(\mu,\nu)$. If there is no minimizer, we can repeat the above argument with a sequence of approximations Π_n to the minimizing value, hence without loss of generality we assume that there is a minimizer.

Since x_t and y_t satisfy equation 5, it follows that $\mathrm{x}_t - \mathrm{y}_t = e^{-\eta t}(\mathrm{x}_0 - \mathrm{y}_0)$, and therefore

$$
\mathbb{E}_{\pi,t_0}\|\mathrm{x}_t - \mathrm{y}_t\|^2 = \mathbb{E}_{\pi,t_0} e^{-2\eta t}\|\mathrm{x}_0 - \mathrm{y}_0\|^2 = e^{-2\eta t_0} W_2^2(\mu,\nu). \tag{21}
$$

Applying the above equality in equation 20 we see that

$$
\begin{aligned}
&\frac{1}{T}\int_0^T \mathbb{E}_{\pi,t_0}\left[L_t^2\|\mathrm{x}_t - \mathrm{y}_t\|^2\right] + 2L_t\sqrt{\mathbb{E}_{\pi,t_0}\|\mathrm{x}_t - \mathrm{y}_t\|^2}\sqrt{\mathbb{E}_{\pi,t_0}\|f_t(\mathrm{y}_t)\|^2} dt_0 \\
&\leq \frac{1}{T}\int_0^T \left[L_{t_0}^2 e^{-2\eta t_0} dt_0\right] W_2^2(\mu,\nu) + \sqrt{\mathbb{E}_{\pi,t_0}\|f_t(\mathrm{y}_t)\|^2}\frac{1}{T}\int_0^T \left[2\varepsilon L_{t_0} e^{-\eta t_0} W_2(\mu,\nu) dt_0\right] \\
&\leq W_2^2(\mu,\nu) C_\eta + 2W_2(\mu,\nu)\sqrt{\frac{1}{T}\int_0^T \left[\mathbb{E}_{\pi,t_0}\|f_t(\mathrm{y}_t)\|^2\right]}\sqrt{\frac{1}{T}\int_0^T \left[L_{t_0}^2 e^{-2\eta t_0}\right] dt_0} \\
&\leq W_2^2(\mu,\nu) C_\eta + 2\varepsilon\sqrt{C_\eta} W_2(\mu,\nu)
\end{aligned}
$$

where C_η and ε are is as in Assumption 1 and we used the Cauchy-Schwarz inequality in the last line. This completes the proof of equation 16.

The proof of equation 17 follows by noting that if x_0 and y_0 have distribution μ and ν respectively, then $\|\mathrm{x}_0 - \mathrm{y}_0\| \leq 2M$ by the triangle inequality. Therefore, for any coupling Π of μ and ν,

$$
W_2^2(\mu,\nu) \leq \mathbb{E}_\pi[\|\mathrm{x}_0 - \mathrm{y}_0\|^2] \leq 4M^2.
$$

The inequality $0 \leq W_2(\mu,\nu) \leq 2M$ follows. For any $0 \leq x \leq 2M$, it is clear that $x^2 \leq 2Mx$. Applying this to $x = W_2(\mu,\nu)$, we have $W_2(\mu,\nu)^2 \leq 2MW_2(\mu,\nu)$. Plugging this inequality into equation 16 directly yields equation 17.

In order to prove equation 18, we return to the step equation 20 used in the proof of equation 16. Having made a different choice of π for that proof, in this proof we choose π to be the minimizer of equation 3 used to define $W_1(\mu,\nu)$ (Note : as seen before, without loss of generality one can assume the existence of a minimizer). Then, we have

$$
\begin{aligned}
&\frac{1}{T}\int_0^T \mathbb{E}_{\pi,t_0}\left[L_t^2\|\mathrm{x}_t - \mathrm{y}_t\|^2\right] + 2L_t\sqrt{\mathbb{E}_{\pi,t_0}\|\mathrm{x}_t - \mathrm{y}_t\|^2}\sqrt{\mathbb{E}_{\pi,t_0}\|f_t(\mathrm{y}_t)\|^2} dt_0 \\
= &\frac{1}{T}\int_0^T \mathbb{E}_{\pi,t_0}\left[L_t^2 e^{-2\eta t_0}\|\mathrm{x}_0 - \mathrm{y}_0\|^2\right] + 2L_t\sqrt{\mathbb{E}_{\pi,t_0} e^{-2\eta t_0}\|\mathrm{x}_0 - \mathrm{y}_0\|^2}\sqrt{\mathbb{E}_{\imath,t_0}\|f_t(\mathrm{y}_t)\|^2} dt_0.
\end{aligned}
\tag{22}
$$

We will now introduce $W_1(\mu,\nu)$ as follows : observe that

$$
\mathbb{E}_\pi[\|\mathrm{x}_0 - \mathrm{y}_0\|^2] \leq \mathbb{E}_\pi[(\|\mathrm{x}_0\| + \|\mathrm{y}_0\|)\|\mathrm{x}_0 - \mathrm{y}_0\|] \leq 2M\mathbb{E}_\pi[\|\mathrm{x}_0 - \mathrm{y}_0\|] \leq 2MW_1(\mu,\nu),
$$

where we used the boundedness of the supports of μ, ν by M. Applying this to equation 22,

$$
\begin{aligned}
&\frac{1}{T}\int_0^T \mathbb{E}_{\pi,t_0}\left[L_t^2 e^{-2\eta t_0}\|\mathrm{x}_0-\mathrm{y}_0\|^2\right] + 2L_t\sqrt{\mathbb{E}_{\pi,t_0} e^{-2\eta t_0}\|\mathrm{x}_0-\mathrm{y}_0\|^2}\sqrt{\mathbb{E}_{\pi,t_0}\|f_t(\mathrm{y}_t)\|^2}dt_0\\
\leq&\frac{1}{T}\int_0^T 2MW_1(\mu,\nu)L_{t_0}^2 e^{-2\eta t_0} + 2L_{t_0}e^{-\eta t_0}\sqrt{2MW_1(\mu,\nu)}\sqrt{\mathbb{E}_{\pi,t_0}\|f_t(\mathrm{y}_t)\|^2}dt_0\\
\leq&2MC_\eta W_1(\mu,\nu) + 2\sqrt{2MW_1(\mu,\nu)}\frac{1}{T}\int_0^T L_{t_0}e^{-\eta t_0}\sqrt{\mathbb{E}_{\pi,t_0}\|f_t(\mathrm{y}_t)\|^2}dt_0. \qquad (23)
\end{aligned}
$$

The final term is easily bounded by Cauchy Schwarz :

$$
\begin{aligned}
&\frac{1}{T}\int_0^T L_{t_0}e^{-\eta t_0}\sqrt{\mathbb{E}_{\pi,t_0}\|f_t(\mathrm{y}_t)\|^2}dt_0\\
\leq&\sqrt{\frac{1}{T}\int_0^T L_{t_0}^2 e^{-2\eta t_0}dt_0}\sqrt{\frac{1}{T}\int_0^T \sqrt{\mathbb{E}_{\pi,t_0}\|f_t(\mathrm{y}_t)\|^2}dt_0}\\
\leq&\sqrt{C_\eta}\sqrt{\varepsilon},
\end{aligned}
$$

where we used Assumption 1 and a final Cauchy-Schwarz application in the final line. Combining the above with equation 23 completes the proof of equation 18 and hence the entire theorem. □

A.4 Theorem 7

Theorem. *Under Assumption 1, $\|\mathrm{x}\| \leq M$ and $d > qp/(q-2p)$, and with twist with weight $w = \exp(\theta^T g(\mathrm{x}))$, if*

$$\left[\left[\mathbb{E}\frac{w^4}{(w^*)^4}\right]^{1/4} + \left[\mathbb{E}\frac{w^2}{(w^*)^2}\right]^{1/2}\right] N^{-p/d} \leq \delta$$

for $p = 1$ or 2 and diffusion is run on the twisted empirical measure $\mu_{N,\theta}$, then the output of the diffusion process has small TV-error with respect to μ_θ.

Proof to Theorem 8. Using 1 and assuming the empirical loss minimization process gives low error, we know $\mathbb{E}W_p(\mu_\theta, \mu_{N,\theta}) \leq \mathcal{O}(\delta)$. Therefore, then, by 2 we obtain

$$\mathbb{E}_{\mu_\theta}\|f_t(\mathrm{x}_t)\|^2 \lesssim \varepsilon^2 + (2C_\eta M + 2\sqrt{C_\eta}\varepsilon)\delta,$$

and, similarly,

$$\mathbb{E}_{\mu_\theta}\|f_t(\mathrm{x}_t)\|^2 \lesssim \varepsilon^2 + 2MC_\eta\delta + 2\sqrt{2MC_\eta\varepsilon}\sqrt{\delta},$$

depending on whether you choose to use the W_1 or W_2. Both of this essentially means that the denoiser error under the true tilted distribution is small.

Let's recall the following result from Chen et al. (2023)

Theorem (DDPM)**.** *Suppose that Assumptions 1, 2, and 3 hold. Let $p_{\theta,T}$ be the output of the DDPM algorithm (Section 2.1) at time T, and suppose that the step size $h := T/N$ satisfies $h \lesssim 1/L$, where $L \geq 1$. Then, it holds that*

$$\mathrm{TV}(p_T, q) \lesssim \underbrace{\sqrt{D_{\mathrm{KL}}(q \parallel \gamma^d)}\exp(-T)}_{\textit{convergence of forward process}} + \underbrace{(L\sqrt{dh} + L\mathfrak{m}_2 h)\sqrt{T}}_{\textit{discretization error}} + \underbrace{\varepsilon_{score}\sqrt{T}}_{\textit{score estimation error}}.$$

Using this, we can conclude:

$$\mathrm{TV}(p_{\theta,T}, \mu_\theta) \leq \mathcal{O}(\exp(-T)) + \mathcal{O}(\sqrt{h}\sqrt{T}) + \mathcal{O}(\varepsilon^2 + \sqrt{\delta}).$$

Therefore, the proof is complete. □

A.5 On the polynomial growth of W_k

Recall the quantity

$$W_k = \mathbb{E}[w^k/(\mathbb{E}w^{k/2})^2]^{1/k}.$$

For the sake of demonstration, we will be in the one-dimensional setting. It is possible, under suitable assumptions, to show that W_k grows only polynomially in k. We have the following result, whose proof we only sketch.

Lemma 5. *Suppose $g(X)$ is a random variable with CDF $F_{g(X)}$ and maximum value B, and suppose $G_{g(X)} = 1 - F_{g(X)}(B - x)$ is regularly varying at 0 with index $\alpha > 0$. Then, W_k grows at most polynomially with θ_n as $\theta_n \to \infty$.*

Proof. Let θ_n be a sequence converging to infinity. By the definition of expectation, and following a simple rewrite,

$$\mathbb{E}[e^{\theta_n g(X)}] = e^{\theta_n B} \int_0^\infty e^{-\theta_n u} dG_{g(X)}(u).$$

Now, $G_{g(X)}(u)$ is regularly varying at 0 of order $\alpha > 0$. By Karamata Tauberian theorem (Bingham et al., 1987, Proposition 1.5.8) it follows that

$$\frac{\mathbb{E}[e^{\theta_n g(X)}]}{e^{\theta_n B}\Gamma(\rho+1)G_{g(X)}\left(\frac{1}{\theta_n}\right)} \to 1.$$

By the above, and simple manipulations, we have asymptotically that

$$W_k = \frac{\mathbb{E}[e^{k\theta_n g(X)}]^{\frac{1}{k}}}{\mathbb{E}[e^{\frac{k}{2}\theta_n g(X)}]^{2/k}} \sim C_1 G_{g(X)}\left(\frac{1}{\theta_n}\right)^{-C_2}$$

for some $C_1, C_2 > 0$. Since $G_{g(X)}$ decays at most polynomially in θ_n due to its regular variation at 0, the result follows. □

B EXPERIMENTS

All code for the experiments is available at: https://github.com/aistats20262406/codebase.

B.1 Reweighted DDPM Objective and Training Algorithm

We train a standard ϵ–prediction DDPM on the *tilted* target

$$\nu(dx) \;\propto\; w(x)\,\mu(dx), \qquad w(x) \;=\; \exp\big(\boldsymbol{\theta}^\top g(x)\big),$$

by weighting the per–sample loss with $w(x)$. Writing the usual forward noising as

$$x_t \;=\; \alpha_t x + \sigma_t \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I), \quad t \sim \text{Unif}[0, T],$$

and denoting the denoiser by $\varepsilon_\phi(\,\cdot\,, t)$, the *tilted* ϵ–prediction objective is

$$\mathcal{L}_{\text{tilt}}(\phi) \;=\; \mathbb{E}_{t\sim\text{Unif},\, x\sim\nu,\, \varepsilon\sim\mathcal{N}}\Big[\, \big\|\varepsilon - \varepsilon_\phi(x_t, t)\big\|^2 \Big]. \tag{24}$$

Using $\nu(dx) = \frac{w(x)}{\mathbb{E}_\mu[w]}\mu(dx)$, equation 24 can be written as the importance–weighted risk under μ:

$$\mathcal{L}_{\text{tilt}}(\phi) \;=\; \frac{\mathbb{E}_{t,x\sim\mu,\varepsilon}[w(x)\,\ell_\phi(x,t,\varepsilon)]}{\mathbb{E}_\mu[w(x)]}, \qquad \ell_\phi(x,t,\varepsilon) = \big\|\varepsilon - \varepsilon_\phi(x_t,t)\big\|^2. \tag{25}$$

Thus, multiplying the per–sample MSE by $w(x)$ optimizes the DDPM objective for the tilted law ν, matching the theoretical setup of Sections 3–4.

Global-weight objective (single normalization). In our implementation we *precompute* weights once for the entire dataset and use a *fixed global normalization*, rather than re-normalizing within each mini-batch. Let

$$\hat{Z} \;=\; \frac{1}{N}\sum_{i=1}^{N} w(x_i), \qquad \tilde{w}(x_i) \;=\; \frac{w(x_i)}{\hat{Z}}.$$

The mini-batch loss used for training is then

$$\hat{\mathcal{L}}_{\text{global}} \;=\; \frac{1}{B}\sum_{i\in\mathcal{B}} \tilde{w}(x_i)\,\big\|\varepsilon^{(i)} - \varepsilon_\phi(x_t^{(i)}, t^{(i)})\big\|^2, \tag{26}$$

where $x_t^{(i)} = \alpha_{t^{(i)}} x_i + \sigma_{t^{(i)}} \varepsilon^{(i)}$, with $t^{(i)} \sim \text{Unif}[0,T]$ and $\varepsilon^{(i)} \sim \mathcal{N}(0,I)$. Note that equation 26 is an unbiased stochastic estimator of equation 25 up to a constant factor, and does *not* require any per-batch reweighting.

Practical stability (dataset-level). To control variance while preserving the target ν:

- **Log–space weights:** store $\log w_i = \boldsymbol{\theta}^\top g(x_i)$ and set $w_i = \exp(\log w_i - c)$ with a fixed dataset-wide offset $c = \max_j \log w_j$ (or $c = \text{median}$) before computing $\hat{Z}$.
- **Weighted sampling (optional):** sample indices with probabilities $p_i \propto w_i$ and use the *unweighted* per-sample MSE; this is equivalent in expectation to equation 26.

Training procedure. The training procedure is described in Algorithm 1.

Algorithm 1 Reweighted–DDPM training with global weights

```
Require: Dataset $\{x_i\}_{i=1}^N \sim \mu$, twist $g(\cdot)$, parameter $\boldsymbol{\theta}$, schedule $\{\alpha_t, \sigma_t\}$, batch size $B$
Precompute $\ell_i^{(w)} \leftarrow \boldsymbol{\theta}^\top g(x_i)$ for all $i$
(stabilize) $c \leftarrow \max_i \ell_i^{(w)}$;   $w_i \leftarrow \exp(\ell_i^{(w)} - c)$
$\hat{Z} \leftarrow \frac{1}{N}\sum_{i=1}^N w_i$;   $\tilde{w}_i \leftarrow \frac{w_i}{\hat{Z}}$          (fixed for entire training)
while not converged do
    Sample indices $\mathcal{B} = \{i_1, \ldots, i_B\}$ (uniform or $p_i \propto w_i$)
    Sample $t^{(k)} \sim \text{Unif}[0,T]$, $\varepsilon^{(k)} \sim \mathcal{N}(0,I)$
    $x_t^{(k)} \leftarrow \alpha_{t^{(k)}} x_{i_k} + \sigma_{t^{(k)}} \varepsilon^{(k)}$
    $\ell_\phi^{(k)} \leftarrow \big\|\varepsilon^{(k)} - \varepsilon_\phi(x_t^{(k)}, t^{(k)})\big\|^2$
    Loss: $\hat{\mathcal{L}} \leftarrow \frac{1}{B}\sum_{k=1}^B \tilde{w}_{i_k}\, \ell_\phi^{(k)}$
    Minimize $\hat{\mathcal{L}}$
end while
```

Remarks

(i) We normalize *once* at the dataset level.

(ii) The forward process and schedule remain unchanged; only the training loss is reweighted.

B.2 Bounded, Correlated Target Construction

For the bounded setting described in Section 5, we construct a high-dimensional non-Gaussian target distribution with bounded support as follows.

Step 1: Independent bounded marginals. We first sample independent random variables

$$X = (X_1, \ldots, X_d),$$

where each X_i follows a Beta distribution

$$X_i \sim \text{Beta}(\alpha_i, \beta_i),$$

with parameters $\alpha_i \sim \text{Unif}[1,5]$ and $\beta_i \sim \text{Unif}[1,5]$. This ensures that all coordinates of X are supported on $[0,1]$ while exhibiting varying degrees of skewness and concentration.

Step 2: Inducing correlations. To introduce correlations among coordinates, we define

$$Y = AX,$$

where $A \in \mathbb{R}^{d\times d}$ is a dense matrix with entries

$$A_{ij} \sim \mathrm{Unif}[0,1].$$

Each column of A is normalized so that

$$\sum_{i=1}^{d} A_{ij} = 1,$$

ensuring that every coordinate Y_i is a convex combination of the independent bounded variables $\{X_j\}$. This normalization preserves the overall scaling of the variables so that the magnitude of the exponential tilt induced by θ remains unaffected. Consequently, Y remains supported on a bounded support but exhibits non-trivial correlations and non-Gaussian behavior.

Step 3: Twisting and evaluation. We perform exponential tilting on the resulting distribution using

$$\boldsymbol{\theta} = \theta \cdot (1,\ldots,1),$$

for several values of θ, and compare three methods: i) reweighted sampling, ii) reweighted sampling combined with diffusion, and iii) the force-guided diffusion method from Wang et al. (2024).

All experiments are conducted in $d = 50$ dimensions with $N = 5 \times 10^5$ samples per setting. Performance metrics and Wasserstein-1 distances are computed over repeated trials to assess stability and sample efficiency.

B.3 Force-Guided Diffusion

We briefly outline the force-guided diffusion framework proposed in Wang et al. (2024), which we employ as a baseline in our experiments.

The key idea is to couple diffusion-based sampling with an auxiliary energy model that captures intermediate energy changes during the diffusion process. Say we want to sample from the distribution

$$p_0(\boldsymbol{x}_0) = q_0(\boldsymbol{x}_0)\frac{e^{-\theta\mathcal{E}_0(\boldsymbol{x}_0)}}{Z}$$

where $z := \int q_0(x_0)e^{-\theta\mathcal{E}_0(x_0)}$. Then, at intermediate diffusion times $t \in (0,1)$, the corresponding marginal distribution satisfies

$$p_t(\boldsymbol{x}_t) \propto q_t(\boldsymbol{x}_t)\, e^{-\theta\, \mathcal{E}_t(\boldsymbol{x}_t)},$$

where $q_t(\boldsymbol{x}_t)$ is the data-based marginal and the intermediate energy $\mathcal{E}_t(\boldsymbol{x}_t)$ is defined as

$$\mathcal{E}_t(\boldsymbol{x}_t) = -\frac{1}{\theta}\log \mathbb{E}_{q_t(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[e^{-\theta\,\mathcal{E}_0(\boldsymbol{x}_0)}\right].$$

This formulation interprets $\mathcal{E}_t$ as a dynamic energy landscape that evolves along the diffusion trajectory.

To learn $\mathcal{E}_t$, the force-guided approach introduces a contrastive energy prediction objective

$$\mathcal{L}_{\mathrm{CEP}} = \mathbb{E}_{p(t)}\mathbb{E}_{q_0(\boldsymbol{x}_0),q_t(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[-e^{\theta\mathcal{E}_0(\boldsymbol{x}_0)}\log\frac{e^{-f_\phi(\boldsymbol{x}_t,t)}}{\sum e^{-f_\phi(\boldsymbol{x}_t,t)}}\right],$$

where $f_\phi(\boldsymbol{x}_t,t)$ is an energy-predicting network. It has been shown that the optimal predictor satisfies

$$\nabla_{\boldsymbol{x}_t} f_\phi^*(\boldsymbol{x}_t,t) = \theta\nabla_{\boldsymbol{x}_t}\mathcal{E}_t(\boldsymbol{x}_t),$$

linking the learned energy to the physical force field driving the reverse diffusion.

The training procedure minimizes the following:

$$\mathcal{L} = \frac{1}{K}\sum\left\|h_\psi(\boldsymbol{x}_t,t) - e^{-\theta\mathcal{E}_0(\boldsymbol{x}_0)}\,\zeta(\boldsymbol{x}_0,\boldsymbol{x}_t)/Y\right\|_2^2.$$

where K is the batch size.

Algorithm 2 Training procedure for Force-Guided Diffusion

```
Require: Generated data $\boldsymbol{x}_0$ in a batch of size $K$, score model $s_\theta(\boldsymbol{x}_t, t)$, energy $\mathcal{E}_0(\boldsymbol{x}_0)$, force $\nabla_{\boldsymbol{x}_0}\mathcal{E}_0(\boldsymbol{x}_0)$, and intermediate force network $h_\psi(\boldsymbol{x}_t, t)$.
for each training iteration do
    Sample $t \sim \mathcal{U}(0, 1)$.
    Compute $\boldsymbol{x}_t$ = forward diffusion of $\boldsymbol{x}_0$ according to Eq. (1).
    Evaluate $q_t(\boldsymbol{x}_t|\boldsymbol{x}_0) \sim \mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_0, (1-\alpha_t)I)$.
    Compute the guided score:
        $\zeta(\boldsymbol{x}_0, \boldsymbol{x}_t) = s_\theta(\boldsymbol{x}_t, t) - \nabla_{\boldsymbol{x}_t} \log q_t(\boldsymbol{x}_t|\boldsymbol{x}_0).$
    Estimate normalization factor:
        $Y = \sum_{\boldsymbol{x}_0} q_t(\boldsymbol{x}_t|\boldsymbol{x}_0) e^{-\theta\mathcal{E}_0(\boldsymbol{x}_0)}.$
    Compute the training loss:
        $\mathcal{L} = \frac{1}{K}\sum \left\| h_\psi(\boldsymbol{x}_t, t) - e^{-\theta\mathcal{E}_0(\boldsymbol{x}_0)} \frac{\zeta(\boldsymbol{x}_0, \boldsymbol{x}_t)}{Y} \right\|_2^2.$
    Update $\psi \leftarrow \psi - \eta\nabla_\psi\mathcal{L}$.
end for
```

Modeling the tilt as a form of energy guidance provides a principled mechanism to sample from the tilted distribution, where the target density is modified by an exponential weighting factor of the form $e^{-\theta\mathcal{E}_0(\boldsymbol{x}_0)}$.

Performance and Limitations. The force-guided diffusion baseline comprises two sequential training stages. The first stage learns a score-based denoiser $\hat{s}_\theta(x, t)$, and the second trains a force network $f_\phi(x, t)$ using the learned scores. Since each update of f_ϕ requires multiple evaluations of $\hat{s}_\theta$, the overall computation grows considerably relative to single-stage diffusion methods. In addition, the coupled optimization of two networks tends to be less stable and more sensitive to hyperparameter choices, making the training process harder to tune. Empirically, in moderate to high dimensions (e.g., $d = 50$), we find that this approach yields lower sample quality and efficiency than our method, which remains easier to train and computationally more efficient.

B.4 Climate Experiment: India Temperature

We test our methodology via a small climate experiment, by tilting the temperature distribution of India.

Data. Daily temperature over India (ERA5 + CMIP6), 5° × 5° grid, months **May–June**, years **1950–2024**. Each day is one sample x; $g(x)$ is the spatial mean.

Objective. Generate from P (baseline) and from the *moment-constrained* twist

$$P_\theta \;=\; \arg\min_{Q} \mathrm{KL}(Q\|P) \;\; \text{s.t.} \;\; \mathbb{E}_Q[g(x)] = \mathbb{E}_P[g(x)] + 1,$$

which yields the exponential tilt $dP_\theta/dP \propto w_\theta(x) = \exp(\theta g(x))$ (Sec. 2).

Training. Train an ϵ–prediction DDPM on P and, separately, on P_θ using the *global* importance weights (Eq. equation 26) with log-space stabilization from Sec. B.1. Architecture and noise schedule are identical across runs.

Evaluation. Report (i) the sample mean of $g(x)$ to verify the +1 shift and (ii) distributional fidelity via W_1 on summary marginals.

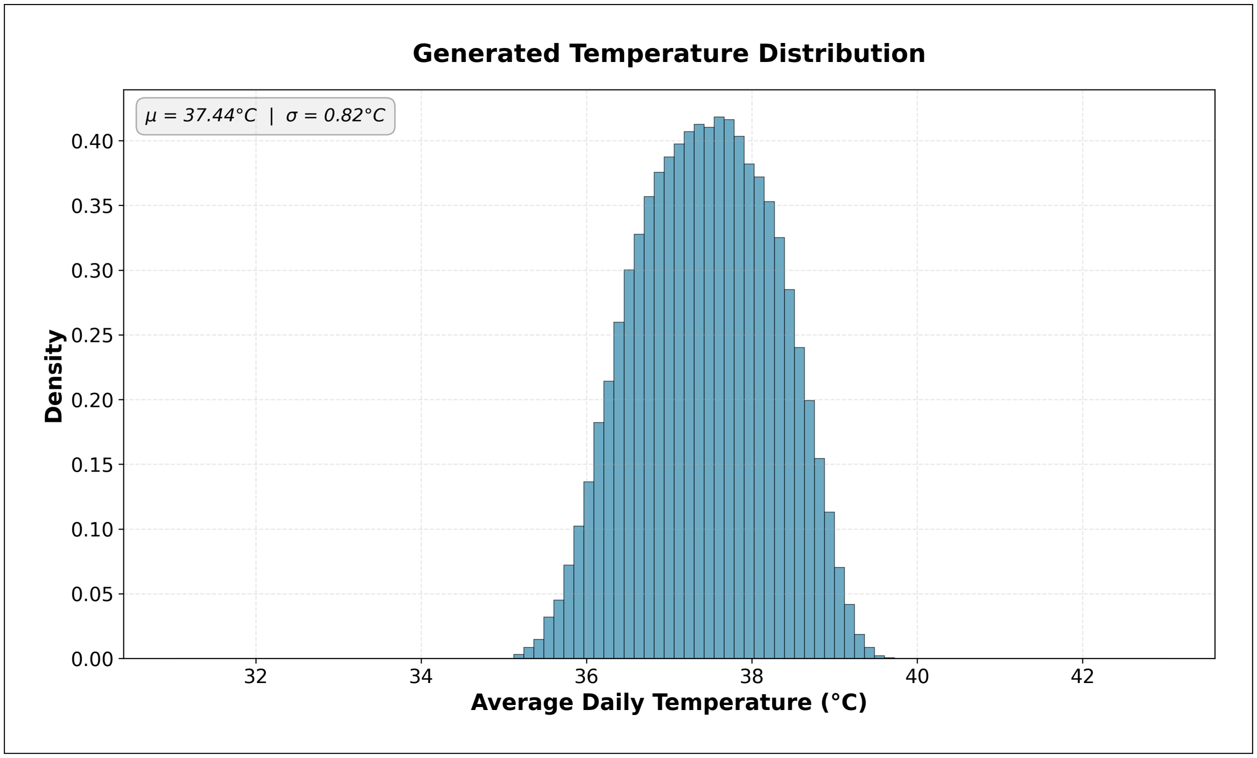

Figure 4: **DDPM samples from P.** Daily temperature fields (May–June, India, $5^\circ \times 5^\circ$).

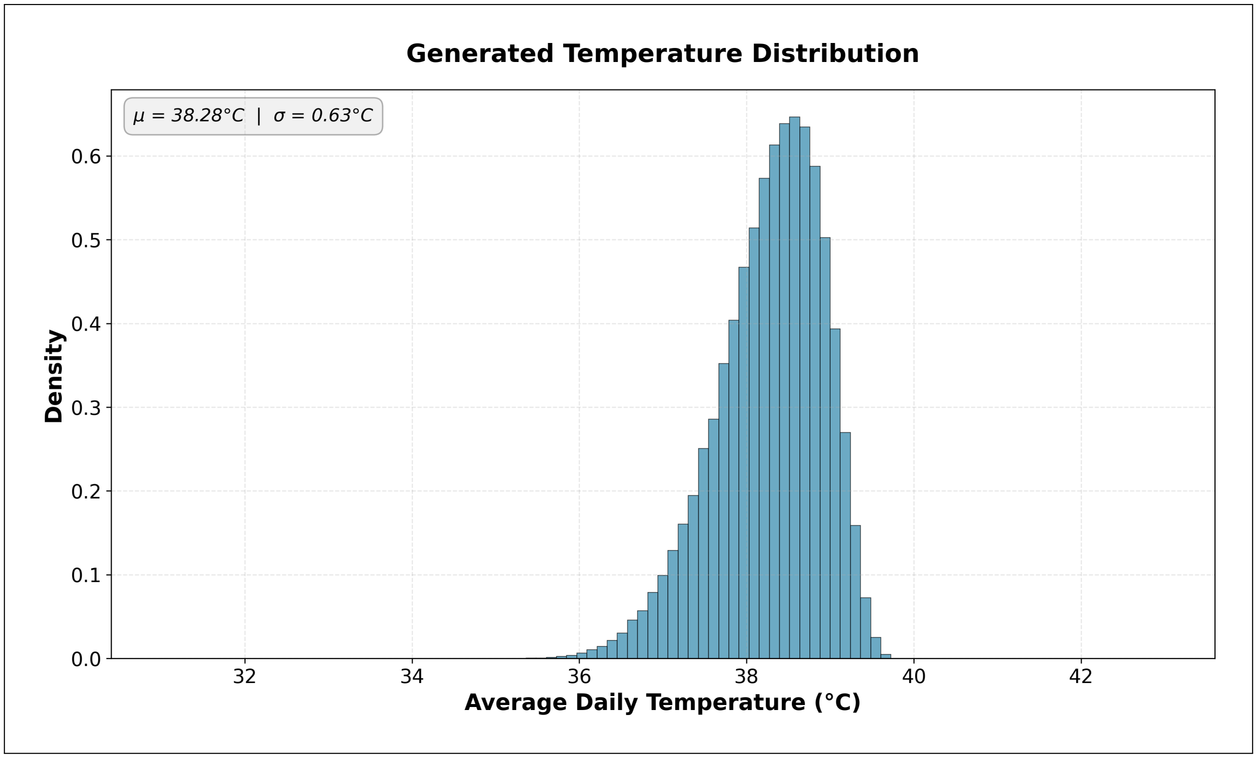

Figure 5: **DDPM samples from P_θ.** Reweighted training targets the hotter, rarer slice with $\mathbb{E}_{P_\theta}[g] = \mathbb{E}_P[g]+1$.

Discussion. Reweighting by $w_\theta(x)$ shifts the *mean* while preserving realistic spatial structure learned from P, enabling efficient sampling of rarer hot events.

C Errors and Typos

Note the restatement of Theorems 4 and 5 in the original Appendix have a different statement than what they do in the Main Paper. However, the proofs are perfectly correct. In this version of the Supplementary Material this is fixed. Kindly refer the main paper for the correct statements instead OR refer this Supplementary Material.

The definitions of W_k are different in the Main Paper and the original Appendix. This has been fixed in this Supplementary Material. One can obtain the bounds in the Appendix of the Main Paper but they are weaker than the bounds in the Main Paper/this Supplementary Material.

Different notations have been used for the Wasserstein distance in different places W_p and $\mathcal{W}_p$. This is a typographical mistake.

The mathematical expression in **Figure 3** has incorrect rendering. The correct expression is

$$\left[[\mathbb{E}w^4]^{\frac{1}{4}} + 4[\mathbb{E}w^2]^{1/2}\right] \frac{N^{-p/d}}{w^*}$$