Constrained Diffusion Models via Dual Training

Shervin Khalafi **Dongsheng Ding*** **Alejandro Ribeiro**
{shervink,dongshed,aribeiro}@seas.upenn.edu
University of Pennsylvania

Abstract

Diffusion models have attained prominence for their ability to synthesize a probability distribution for a given dataset via a diffusion process, enabling the generation of new data points with high fidelity. However, diffusion processes are prone to generating samples that reflect biases in a training dataset. To address this issue, we develop constrained diffusion models by imposing diffusion constraints based on desired distributions that are informed by requirements. Specifically, we cast the training of diffusion models under requirements as a constrained distribution optimization problem that aims to reduce the distribution difference between original and generated data while obeying constraints on the distribution of generated data. We show that our constrained diffusion models generate new data from a mixture data distribution that achieves the optimal trade-off among objective and constraints. To train constrained diffusion models, we develop a dual training algorithm and characterize the optimality of the trained constrained diffusion model. We empirically demonstrate the effectiveness of our constrained models in two constrained generation tasks: (i) we consider a dataset with one or more underrepresented classes where we train the model with constraints to ensure fairly sampling from all classes during inference; (ii) we fine-tune a pre-trained diffusion model to sample from a new dataset while avoiding overfitting.

1 Introduction

Diffusion models have become a driving force of modern generative modeling, achieving groundbreaking performance in tasks ranging from image/video/audio generation [64, 6, 43] to molecular design for drug discovery [75, 74]. Diffusion models learn diffusion processes that produce a probability distribution for a given dataset (e.g., images) from which we can generate new data (e.g., classic models [66, 34, 68, 78]). As diffusion models are used to generate data with societal impacts, e.g., art generation and content creation for media, they must comply with requirements from specific domains, e.g., social fairness in image generation [54, 57], aesthetic properties of images [12, 13], bioactivity in molecule generation [39], and more [40, 81, 8, 19, 79, 14].

Classic diffusion models [66, 34, 68] have been extended to generate data under different requirements through either first-principle methods or fine-tuning. In image generation with fairness, for instance, first principle methods directly mitigate biases towards social/ethical identities by revising the training loss functions per biased/unbiased data [49, 41]; while fine-tuning methods align biased diffusion models with desired data distributions by optimizing the associated metrics [24, 72, 71, 70]. Although these methods allow us to equip diffusion models with specific requirements, they are often designed for particular generation tasks and do not provide transparency on how these requirements are satisfied. Since diffusion models are trained by minimizing the loss functions of diffusion processes, it is natural to incorporate requirements into diffusion models by imposing constraints on these optimization

*Corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-0836

problems. Therefore, it is imperative to develop diffusion models under constraints by generalizing constrained learning methods and theory (e.g., [9, 10, 22, 36]) for diffusion models.

In this work, we formulate the training of diffusion models under requirements as a constrained distribution optimization problem in which the objective function measures a training loss induced by the original data distribution, and the constraints require other training losses induced by some desirable data distributions to be small. This constrained formulation can be instantiated for several critical requirements. For instance, to promote fairness for unrepresented groups, the constraints can encode the closeness of the model to the distributions of underrepresented data. Compared with the typical importance re-weighting method [41], our constrained formulation provides an optimal trade-off between matching given data distribution and following reference distribution. Not limited to constraints with other desirable data distributions, our constrained formulation captures more general requirements. For instance, when adapting a pretrained diffusion model to new data, the constraints require the model to be close to the pretrained model, not degrading the original generation ability. Specifically, our main contribution is three-fold.

(i) We propose and analyze constrained diffusion models from the perspective of constrained optimization in an infinite-dimensional distribution space, where the constraints require KL divergences between the model and our desired data distributions to be under some thresholds. We exploit the strong duality in convex optimization to show that our constrained diffusion models generate new data from a mixture data distribution that achieves the optimal trade-off among objective and constraints.

(ii) To train constrained diffusion models, we introduce parametrized constrained diffusion models and develop a Lagrangian-based dual training algorithm. We exploit the relation between un/parametrized problems to show that constrained diffusion models generate new data from the optimal mixture distribution, up to some optimization/parametrization errors.

(iii) We empirically demonstrate the merit of our constrained diffusion models in two aforementioned requirements. In the fair generation task, we show that our constrained model promotes sampling more from the minority classes compared to unconstrained models, leading to fair sampling across all classes. In the adaptation task, we show that our fine-tuned constrained model learns to generate new data without significantly degrading the original generation ability, compared to the unconstrained model which tends to overfit to new data.

Related work. As a first-principle method, our constrained diffusion approach is more relevant to diffusion models that incorporate requirements in distribution space [20, 35, 49, 21, 60, 28], rather than those applied in sample space [37, 53, 29, 27, 26, 17, 48, 25]. In comparison with conditional diffusion models that restrict generation through conditional information [20, 35, 1], our constrained diffusion models impose distribution constraints within the constrained optimization framework. Compared to compositional generation [49, 21, 60] and fair diffusion [28], our work provides a constrained learning approach to balance different distribution models using Lagrangian multipliers, which is different from equal weights [49, 21], hyperparameter [60] or fair guidance [28]. Our constrained approach is also relevant to the importance re-weighting method for diffusion models [41] and GANs [16], which reduces bias in a dataset by re-weighting it with a pre-trained debiased density ratio. In contrast, we design our diffusion models to mitigate the bias in a dataset by imposing distribution constraints without pre-training. In addition to being distinct from existing methods, we provide a systematic study of our constrained diffusion models, covering duality analysis, dual-based algorithm design, and convergence analysis, which is absent in the diffusion model literature.

Our work is also pertinent to recently surging fine-tuning and alignment methods that aim to improve pre-trained diffusion models by optimizing their downstream performance, e.g., aesthetic scores of images. In reward-guided fine-tuning methods, such as supervised learning [45, 80, 76], control-based feedback learning [77, 65, 71, 70, 18], and reinforcement learning [23, 32, 24, 72, 5, 82], reward functions have to be pre-trained from an authentic evaluation dataset, and the trade-off for reward functions in pre-trained diffusion models is often regulated heuristically. In contrast, our constrained approach directly minimizes the gap between a fine-tuning model and a high-quality dataset with the desired properties, while ensuring the generated outputs being close to that of pre-trained models.

Compared to other generative models under requirements (e.g., VAEs [61], GANs [16]), and classical sampling methods (e.g., Langevin dynamics and Stein variational gradient descent [50]), our work is different because we focus on diffusion-based generative models.

2 Preliminaries

We overview diffusion models from the perspective of variational inference [55] by presenting forward/backward processes in Section 2.1, and the evidence lower bound in Section 2.2.

2.1 Forward and backward processes

The forward process begins with a true data sample $x_0 \in \mathbb{R}^d$, and generates latent variables $\{x_t\}_{t=1}^T$ from $t = 1$ to T by adding white noise recursively. The joint distribution of latent variables results from conditional probabilities $q(x_{1:T} \,|\, x_0) = \prod_{t=1}^T q(x_t \,|\, x_{t-1})$, where the distribution of latent variable x_t conditioned on the previous latent x_{t-1} is given by a Gaussian distribution $q(x_t \,|\, x_{t-1}) = \mathcal{N}(x_t; \mu_t, \sigma_q^2(t) I)$, where $\mu_t = \sqrt{\alpha_t}\, x_{t-1}$ and $\sigma_q^2(t) = 1 - \alpha_t$. Since the forward process is a linear Gaussian model with pre-selected mean and variance, it is solely determined by the data distribution $q(x_0)$. We often refer to $q(x_0)$ as a forward process.

The backward process begins with the latent x_T sampled from the standard Gaussian $p(x_T) = \mathcal{N}(x_T; 0, I)$, and decodes latent variables from $t = T$ to $t = 0$ with a joint distribution

$$p(x_{0:T}) \;=\; p(x_T) \prod_{t=1}^{T} p(x_{t-1} \,|\, x_t). \tag{1}$$

Here, p is our distribution model that can be used to generate new samples. We denote by $\mathcal{P}$ the set of all joint distributions over $x_{0:T}$ in form of (1), where $p(x_{t-1} \,|\, x_t)$ is a conditional Gaussian with a fixed variance (see Appendix A). Throughout the paper, we work in the convergent regimes of the backward process (e.g., [15, 11, 2]). Without loss of generality, we adopt the convergent regime in [47] by taking the scheduling parameter $\alpha_1 = 1 - 1/T^{c_0}$ and $\alpha_t = 1 - c_T \min\left((1-\alpha_1)(1+c_T)^t, 1\right)$ for $t > 1$, and the variance as $\sigma_p^2(t) = 1/\alpha_t - 1$, where $c_T := c_1 \log(T)/T$ and c_0, c_1 are some constants. Hence, $\bar{\alpha}_T \approx 0$ implies $q(x_T) \simeq \mathcal{N}(x_T; 0, I)$. Thus, $q(x_{0:T}) \in \mathcal{P}$. Also, $\bar{\alpha}_1 \approx 1$ implies $q(x_1) \simeq q(x_0)$. It is ready to generalize our results to other diffusion processes (e.g., [46, 38, 3]).

2.2 The evidence lower bound (ELBO)

Denote the KL divergence of distribution q from distribution p by $D_{\mathrm{KL}}(q \,\|\, p) := \mathbb{E}_{x \sim q(x)} \log(\frac{q(x)}{p(x)})$. Generative diffusion modeling aims to generate samples whose distribution is close to that of an observed dataset of samples. Formally, we express this objective as maximizing the log-likelihood of an observation generated by the diffusion model: $\operatorname{maximize}_{p \in \mathcal{P}} \mathbb{E}_{q(x_0)}[\log p(x_0)]$, where

$$\mathbb{E}_{q(x_0)}\left[\log p(x_0)\right] \;=\; E(p;q) \;+\; \mathbb{E}_{q(x_0)}\left[D_{\mathrm{KL}}\left(q(x_{1:T} \,|\, x_0) \,\|\, p(x_{1:T} \,|\, x_0)\right)\right] \tag{2}$$

and $E(p;q) := \mathbb{E}_{q(x_0)} \mathbb{E}_{q(x_{1:T} \,|\, x_0)} \log \frac{p(x_{0:T})}{q(x_{1:T} \,|\, x_0)}$ is known as ELBO in variational inference [7, 55]. Alternatively, we aim to minimize the KL divergence between the forward/backward processes,

$$D_{\mathrm{KL}}\left(q(x_{0:T}) \,\|\, p(x_{0:T})\right) \;=\; -\,E(p;q) \;+\; \mathbb{E}_{q(x_0)}\left[\log q(x_0)\right]. \tag{3}$$

Thus, we connect the log-likelihood maximization to the KL divergence minimization via ELBO.

Lemma 1 (Equivalent formulations)**.** *The ELBO maximization and the KL divergence minimization are equivalent over the distribution space $\mathcal{P}$, and the unique solution of these two problems is a solution for the log-likelihood maximization problem, i.e.,*

$$\operatorname*{maximize}_{p \in \mathcal{P}} \; E(p;q) \;\;\Leftrightarrow\;\; \operatorname*{minimize}_{p \in \mathcal{P}} \; D_{KL}\left(q(x_{0:T}) \,\|\, p(x_{0:T})\right) \;\;\Rightarrow\;\; \operatorname*{maximize}_{p \in \mathcal{P}} \; \mathbb{E}_{q(x_0)}[\log p(x_0)]$$

See Appendix B.1 for proof. Lemma 1 states that improving the ELBO score increases the likelihood of a backward process that generates the data, together with the fit of a backward process to the forward process. Hence, finding the best backward process becomes optimizing one of three equivalent objectives. In practice, ELBO serves as a loss function approximated by

$$E(p;q) \;\approx\; -\sum_{t=2}^{T} \mathbb{E}_{q(x_0)} \mathbb{E}_{q(x_t \,|\, x_0)}\left[D_{\mathrm{KL}}\left(q(x_{t-1} \,|\, x_t, x_0) \,\|\, p(x_{t-1} \,|\, x_t)\right)\right]. \tag{4}$$

https://doi.org/10.52202/079017-0836

With the variance schedule described in Section 2.1, it is known that this approximation is almost exact (see Appendix A and also [42]), which is our focal setting. Using standard diffusion derivations [55], the ELBO maximization can be shown to equal to a quadratic loss minimization,

$$\underset{\widehat{s}\in\mathcal{S}}{\text{minimize}} \quad \mathbb{E}_{x_0,\,t,\,x_t}\left[\,\|\widehat{s}(x_t,t)-\nabla\log q(x_t)\|^2\,\right] \tag{5}$$

where $\mathbb{E}_{x_0,t,x_t}$ is an expectation over the data distribution $q(x_0)$, a discrete distribution $p_\omega(t)$ from 2 to T, and the forward process $q(x_t\,|\,x_0)$ at time t given the data sample x_0; see Appendix A for details. The minimization is done to find a function $\widehat{s}\in\mathcal{S}$ that can best predict the gradient of the forward process over data $\nabla\log q(x_t)$, commonly called the (Stein) score function, where $\mathcal{S}$ is a set of valid score functions mapping from $\mathbb{R}^d\times\mathbb{N}$ to $\mathbb{R}^d$. In practice, however, we parametrize the estimator $\widehat{s}(x_t,t)$ as $\widehat{s}_\theta(x_t,t)$ with parameter θ, which gives our focal objective of generative modeling: $\text{minimize}_{\theta\in\Theta}\,\mathbb{E}_{x_0,\,t,\,x_t}\left[\,\|\widehat{s}_\theta(x_t,t)-\nabla\log q(x_t)\|^2\,\right]$. A parametrized form of $p(x_{t-1}\,|\,x_t)$ associated with $\widehat{s}_\theta(x_t,t)$ is denoted by $p_\theta(x_{t-1}\,|\,x_t)$ and the backward process has a parametrized joint distribution $p_\theta(x_{0:T})$. We remark that the prediction problem (5) can be also be formulated as data or noise prediction instead [55], with our results directly transferable to these formulations.

3 Variational constrained diffusion models

We introduce constrained diffusion models by considering the unparametrized set of joint distributions in Section 3.1, and illustrating constraints via two examples in Section 3.2.

3.1 KL divergence-constrained diffusion model: unparametrized case

The standard diffusion model specifies a single data distribution, denoted by q in Lemma 1. To account for other generation requirements, we introduce m additional data distributions $\{q^i\}_{i=1}^m$ that represent m desired properties on generated data. To incorporate new properties into the diffusion model, we formulate an unparametrized KL divergence-constrained optimization problem,

$$\begin{array}{ll}\underset{p\in\mathcal{P}}{\text{minimize}} & D_{\text{KL}}\left(q(x_{0:T})\,\|\,p(x_{0:T})\right)\\ \text{subject to} & D_{\text{KL}}\left(q^i(x_{0:T})\,\|\,p(x_{0:T})\right)\;\le\;b_i\quad\text{for } i=1,\dots,m.\end{array} \tag{U-KL}$$

Let an optimal solution to Problem (U-KL) be $p^\star$. Then the optimal value of the objective function is $F^\star:=D_{\text{KL}}(q\,\|\,p^\star)$. Problem (U-KL) aims to find a model $p^\star$ that generates data from the original distribution q while staying close to m distributions $\{q^i\}_{i=1}^m$ that encode our desired properties, e.g., unbiasedness towards minorities. Let the Lagrangian for Problem (U-KL) be

$$\mathcal{L}(p,\lambda)\;=\;D_{\text{KL}}\left(q(x_{0:T})\,\|\,p(x_{0:T})\right)\;+\;\sum_{i=1}^m\lambda_i\left(D_{\text{KL}}\left(q^i(x_{0:T})\,\|\,p(x_{0:T})\right)-b_i\right) \tag{6}$$

for $\lambda\ge 0$. The dual function $g(\lambda)$ is given by $g(\lambda):=\min_{p\in\mathcal{P}}\mathcal{L}(p,\lambda)$, which is always concave.

To make Problem (U-KL) meaningful, we assume the constraints are strictly satisfied by some model.

Assumption 1 (Strict feasibility)**.** *There exists a model $p\in\mathcal{P}$ and $\zeta>0$ such that $D_{\text{KL}}(q^i(x_{0:T})\,\|\,p(x_{0:T}))\le b_i-\zeta$ for all $i=1,\dots,m$.*

Let an optimal dual variable of Problem (U-KL) be $\lambda^\star\in\text{argmax}_{\lambda\ge 0}\,g(\lambda)$, and the optimal value of the dual function be $D^\star:=g(\lambda^\star)$. From weak duality, the duality gap is non-negative, i.e., $F^\star-D^\star\ge 0$. Moreover, due to the convexity of KL divergence, Problem (U-KL) is a convex optimization problem, and thus it satisfies strong duality; see Appendix B.2 for proof.

Lemma 2 (Strong duality)**.** *Let Assumption 1 hold. Then, Problem* (U-KL) *has zero duality gap, i.e., $F^\star=D^\star$. Moreover, $(p^\star,\lambda^\star)$ is an optimal primal-dual pair of Problem* (U-KL)*.*

Let a mixture data distribution be $q_{\text{mix}}^{(\lambda)}:=\left(q+\sum_{i=1}^m\lambda^i q^i\right)/(1+\lambda^\top 1)$ for $\lambda\ge 0$. We denote by $q_{\text{mix}}^{(\lambda)}(x_{0:T})$ a joint distribution of the forward process with data distribution $q_{\text{mix}}^{(\lambda)}$. Leveraging strong duality, we show that an optimal model can be obtained by solving an equivalent unconstrained problem in Theorem 1 and its proof is deferred to Appendix B.3.

Theorem 1 (Optimal constrained model)**.** *Let Assumption 1 hold. Then, Problem* (U-KL) *equals*

$$\underset{p \in \mathcal{P}}{\text{minimize}} \quad D_{\text{KL}}\left(q_{\text{mix}}^{(\lambda^\star)}(x_{0:T}) \,\|\, p(x_{0:T})\right) \tag{U-MIX}$$

where $q_{\text{mix}}^{(\lambda^\star)}(x_{0:T})$ *is the joint distribution of the forward process at an optimal dual variable* $\lambda^\star$*.*

Theorem 1 states that the KL divergence-constrained problem reduces to an unconstrained KL divergence minimization problem. We notice that the KL divergence is zero if and only if two probability distributions match each other. Hence, $q_{\text{mix}}^{(\lambda^\star)}(x_{0:T})$ is the optimal solution to Problem (U-MIX).

Corollary 1. *Let Assumption 1 hold. Then, the solution of Problem* (U-MIX)*, i.e.,* $p^\star(x_{0:T}) = q_{\text{mix}}^{(\lambda^\star)}(x_{0:T})$*, is the solution of Problem* (U-KL)*.*

Let $\bar{b}^i := b^i - \mathbb{E}_{q^i(x_0)}[\log q^i(x_0)]$. Application of Equality (3) to Problem (U-KL) yields an ELBO-based constrained optimization problem,

$$\begin{array}{ll} \underset{p \in \mathcal{P}}{\text{minimize}} & -E(p;q) \\ \text{subject to} & -E(p;q^i) \;\le\; \bar{b}^i \quad \text{for } i = 1, \dots, m. \end{array} \tag{U-ELBO}$$

Recall the model representation in Section 2.2, we can characterize each joint distribution $p \in \mathcal{P}$ with a function $\widehat{s} \in \mathcal{S}$. Moreover, ELBO reduces to the denoising matching term that has a simplified quadratic form given in Section 2.2. With this reformulation in mind, we cast Problem (U-ELBO) into a convex optimization problem over the function space $\mathcal{S}$,

$$\begin{array}{ll} \underset{\widehat{s} \in \mathcal{S}}{\text{minimize}} & \mathbb{E}_{q(x_0),\, t,\, x_t}\left[\left\|\widehat{s}(x_t,t) - \nabla \log q(x_t)\right\|^2\right] \\ \text{subject to} & \mathbb{E}_{q^i(x_0),\, t,\, x_t}\left[\left\|\widehat{s}(x_t,t) - \nabla \log q^i(x_t)\right\|^2\right] \;\le\; \widetilde{b}^i \quad \text{for } i = 1, \dots, m \end{array} \tag{U-LOSS}$$

where $\widetilde{b}^i := (\bar{b}^i - v)/\bar{\omega}$. Here, the notation v is a constant shift due to the variance mismatch term; see it in Appendix A. We note that scaling or shifting objective and constraints from both sides with some constants doesn't alter the solution to a constrained optimization problem. Thus, the key difference between Problems (U-KL) and (U-LOSS) is the optimization variable (respectively, p and $\widehat{s}$). Let the Lagrangian $\mathcal{L}_s(\widehat{s}, \lambda)$ for Problem (U-LOSS) be

$$\mathbb{E}_{q(x_0),\, t,\, x_t}\left[\left\|\widehat{s}(x_t,t) - \nabla \log q(x_t)\right\|^2\right] + \sum_{i=1}^{m} \lambda_i \left(\mathbb{E}_{q^i(x_0),\, t,\, x_t}\left[\left\|\widehat{s}(x_t,t) - \nabla \log q^i(x_t)\right\|^2\right] - \widetilde{b}^i\right).$$

Let the associated dual function be $g_s(\lambda) := \min_{\widehat{s} \in \mathcal{S}} \mathcal{L}_s(\widehat{s}, \lambda)$ for $\lambda \ge 0$. Hence, $g(\lambda)$ and $g_s(\lambda)$ have the same maximizer $\lambda^\star$, and the partial minimizer $\widehat{s}^\star = \operatorname{argmin}_{\widehat{s} \in \mathcal{S}} \mathcal{L}_s(\widehat{s}, \lambda^\star)$ is the solution to Problem (U-LOSS). Hence, an optimal primal-dual pair $(\widehat{s}^\star, \lambda^\star)$ to Problem (U-LOSS) gives an optimal primal-dual pair $(p^\star, \lambda^\star)$ for Problem (U-KL), where $p^\star$ is a joint distribution of the backward process induced by $\widehat{s}^\star$. By this dual property, we take a dual perspective to train constrained diffusion models: we maximize the dual function $g_s(\lambda)$ to obtain the optimal dual variable $\lambda^\star$, and then recover the solution $\widehat{s}^\star$ by minimizing the Lagrangian $\mathcal{L}_s(\widehat{s}, \lambda^\star)$.

3.2 Examples of KL divergence constraints

To illustrate our KL-divergence constraints, we provide two generation tasks of exemplary.

(i) **Fairness to underrepresented classes.** We consider a fair image generation task in which some classes are underrepresented in the available training dataset. An example of this would be the Celeb-A dataset [51] which contains pictures of celebrity faces with those labeled as male being underrepresented (42% Male vs 58% Female). To promote representation of the under-represented classes, we can pose it as an instance of Problem (U-KL), where each q^i denotes the distribution of an under-represented subset or minority class of q.

(ii) **Adapting pretrained model to new data.** Given a pretrained diffusion model over some original dataset that is no longer accessible, we aim to fine-tune the pretrained model for generating data from a new data distribution. Similarly, we can pose this as an instance of Problem (U-KL), where q^1 denotes the distribution of the new data and q is the distribution of samples generated by the pre-trained model.

https://doi.org/10.52202/079017-0836

Let $h_i := -\mathbb{E}_{q^i(x_0)}\left[\log q^i(x_0)\right]$ be the differential entropy of data distribution q^i. We relate the KL divergence constraints with the optimal dual variable through entropy in Theorem 2.

Theorem 2. *Let Assumption 1 hold, and the supports of data distributions q and $\{q^i\}_{i=1}^m$ be disjoint. Then, the optimal dual variables $\lambda^\star$ to Problem (U-ELBO) are given by*

$$\frac{\lambda_i^\star}{1 + (\lambda^\star)^T \mathbf{1}} = \mathrm{e}^{h_i - \bar{b}_i} \quad \text{for } i = 1, \ldots, m.$$

See Appendix B.4 for proof. Theorem 2 characterizes the mixture weights in the target distribution $q_{\text{mix}}^{(\lambda^\star)}$: (i) the tighter a constraint is (i.e., smaller threshold $\bar{b}_i$), the more the model will sample from the associated distribution; (ii) for the same constraint thresholds, the model will sample more often from the associated distributions that have higher entropy h_i. Assumption 1 can be relaxed to a feasibility condition for Problem (U-KL); see Lemma 6 in Appendix B.4.

4 Parametrization and dual training algorithm

Having introduced unparametrized models, we move to parametrization for constrained diffusion models in Section 4.1, provide optimality analysis of a Lagrangian-based dual method in Section 4.2, and present a practical dual training algorithm in Section 4.3.

4.1 KL divergence-constrained diffusion model: parametrized case

With the parametrized model p_θ for $\theta \in \Theta$, we present a parameterized constrained problem,

$$\begin{array}{ll} \underset{\theta \in \Theta}{\text{minimize}} & D_{\text{KL}}\left(q(x_{0:T}) \,\|\, p_\theta(x_{0:T})\right) \\ \text{subject to} & D_{\text{KL}}\left(q^i(x_{0:T}) \,\|\, p_\theta(x_{0:T})\right) \;\leq\; b^i \quad \text{for } i = 1, \ldots, m. \end{array} \tag{P-KL}$$

Let the Lagrangian for Problem (P-KL) be $\bar{\mathcal{L}}(\theta, \lambda) := \mathcal{L}(p_\theta, \lambda)$. The associated dual function $\bar{g}(\lambda)$ is given by $\bar{g}(\lambda) := \min_{\theta \in \Theta} \bar{\mathcal{L}}(\theta, \lambda)$. Let an optimal solution to Problem (P-KL) be $\theta^\star$. We denote $\bar{p} := p_\theta$ and $\bar{p}^\star := p_{\theta^\star}$, and the optimal objective by $\bar{F}^\star := D_{\text{KL}}(q \,\|\, \bar{p}^\star)$. Let an optimal dual variable be $\bar{\lambda}^\star \in \operatorname{argmax}_{\lambda \geq 0} \bar{g}(\lambda)$ and the optimal value of the dual function be $\bar{D}^\star := \bar{g}(\bar{\lambda}^\star)$.

Problem (P-KL) is non-convex in parameter space, and strong duality does not hold any more. Thus, unparametrized results in Section 3.1 don't directly apply to Problem (P-KL). For instance, it's invalid to find an optimal solution via an unconstrained problem as in Theorem 1, i.e., $\bar{p}^\star(\bar{\lambda}^\star) \in \operatorname{argmin}_{\theta \in \Theta} \bar{\mathcal{L}}(\theta, \bar{\lambda}^\star)$ doesn't equal $\bar{p}^\star$. The effect of parametrization has to be characterized. However, regardless of parametrization, weak duality always holds, i.e., $\bar{F}^\star - \bar{D}^\star \geq 0$.

To quantify the optimality of $\bar{p}^\star(\bar{\lambda}^\star)$ (closeness of it to $q_{\text{mix}}^\star$), we study a practical representation of model p_θ as a parametrized function $\widehat{s}_\theta \in \mathcal{S}_\theta$, where $\mathcal{S}_\theta$ is the set of all parametrized score functions. Problem (U-LOSS) is in a parametrized form of

$$\begin{array}{ll} \underset{\theta}{\text{minimize}} & \mathbb{E}_{q(x_0),\, t,\, x_t}\left[\, \|\widehat{s}_\theta(x_t, t) - \nabla \log q(x_t)\|^2 \,\right] \\ \text{subject to} & \mathbb{E}_{q^i(x_0),\, t,\, x_t}\left[\, \left\|\widehat{s}_\theta(x_t, t) - \nabla \log q^i(x_t)\right\|^2 \,\right] \;\leq\; \widetilde{b}^i \quad \text{for } i = 1, \ldots, m. \end{array} \tag{P-LOSS}$$

where $\widetilde{b}^i := (\bar{b}^i - v)/\bar{\omega}$. We note that Problem (P-LOSS) is equivalent to Problem (P-KL). Thus, we let the Lagrangian of Problem (P-LOSS) be $\bar{\mathcal{L}}_s(\theta, \lambda) := \mathcal{L}_s(\widehat{s}_\theta, \lambda)$, and the dual function $\bar{g}_s(\lambda) := \operatorname{minimize}_{\theta \in \Theta} \bar{\mathcal{L}}(\theta, \lambda)$. Since $\bar{g}(\lambda)$ and $\bar{g}_s(\lambda)$ have the same maximizer $\bar{\lambda}^\star$, $\bar{\theta}^\star \in \operatorname{argmin}_{\theta \in \Theta} \bar{\mathcal{L}}_s(\theta, \bar{\lambda}^\star)$, which naturally gives a dual training algorithm in Algorithm 1.

Denote $\bar{s}^\star := \widehat{s}_{\bar{\theta}^\star}$. Thus, $\bar{s}^\star$-induced diffusion model is given by $\bar{p}^\star(\bar{\lambda}^\star)$. Algorithm 1 works as a dual ascent method with two natural steps: (i) find a diffusion model with fixed dual variable $\lambda(h)$; and (ii) update the dual variable using the (sub)gradient of the Lagrangian $\mathcal{L}_s(\widehat{s}_\theta(h), \lambda)$. It is known that Algorithm 1 converges to $\bar{\lambda}^\star$ since the dual function $\bar{g}_s(\lambda)$ is concave. However, such convergence in the dual domain doesn't provide optimality guarantee on the primal solution $\bar{p}^\star(\bar{\lambda}^\star)$ due to the non-convexity in parameter space. We next exploit the optimization properties of unparametrized diffusion models in Section 3.1 to characterize the optimality of the dual training algorithm.

Algorithm 1 Constrained Diffusion Models via Dual Training

1: **Input**: total diffusion steps T, diffusion parameter α_t, total iterations H, stepsize η.
2: **Initialize**: $\lambda(1) = 0$.
3: **for** $h = 1, \cdots, H$ **do**
4: Compute model $\widehat{s}_\theta(h) \in \underset{\theta \in \Theta}{\operatorname{argmin}} \mathcal{L}_s(\widehat{s}_\theta, \lambda(h))$.
5: Update the dual variable

$$\lambda_i(h+1) \;=\; \left[\lambda_i(h) \;+\; \eta \left(\mathbb{E}_{x_0 \sim q^i, t, x_t} \left[\left\| \widehat{s}_\theta(h)(x_t, t) - \nabla \log q^i(x_t) \right\|^2 \right] \;-\; \widetilde{b}^i \right)\right]_+ \quad \text{for all } i.$$

6: **end for**

4.2 Optimality analysis of dual training algorithm

We analyze the optimality of $\bar{p}^\star(\bar{\lambda}^\star)$ as measured by its distance to $q^\star_{\text{mix}}$, i.e., $\text{TV}(q^\star_{\text{mix}},\ \bar{p}^\star(\bar{\lambda}^\star))$, where we denote the total variation distance between two probability distributions p and q by $\text{TV}(q, p) := \frac{1}{2}\int |p(x) - q(x)| dx$. We first exploit the convergence analysis of diffusion models, and then characterize the additional error induced by parametrization.

Let us begin with the difference between $\bar{p}^\star(\bar{\lambda}^\star)$ and $q^{(\bar{\lambda}^\star)}_{\text{mix}}$ at $\bar{\lambda}^\star$. Denote a partial minimizer of the Lagrangian by $\bar{p}^\star(\lambda) \in \operatorname{argmin}_{\theta \in \Theta} \bar{\mathcal{L}}_s(\theta, \lambda)$ for $\lambda \geq 0$. Noting that $\operatorname{minimize}_{\theta \in \Theta} \bar{\mathcal{L}}_s(\theta, \lambda)$ is an unconstrained diffusion problem, we are ready to quantify the difference between $\bar{p}^\star(\lambda)$ and $q^{(\lambda)}_{\text{mix}}$ for any $\lambda \geq 0$ using the convergence theory of diffusion models. To do so, we assume the boundedness of samples from a mixed data distribution $q^{(\lambda)}_{\text{mix}}$ for $\lambda \geq 0$, and a small score estimation error.

Assumption 2 (Boundedness of data)**.** *The data samples generated from $q^{(\lambda)}_{\text{mix}}$ are bounded, i.e., $\mathbb{P}\left(\|x_0\| \leq T^c \,|\, x_0 \sim q^{(\lambda)}_{\text{mix}}\right) = 1$ for any $\lambda \geq 0$ and some large constant $c > 0$.*

Assumption 3 (Boundedness of score estimation error)**.** *The score estimator $\widehat{s}_\theta(x_t, t)$ estimates the data samples from $q^{(\lambda)}_{\text{mix}}$ with bounded score matching error $\varepsilon_{\text{score}}$,*

$$\mathbb{E}_{q^{(\lambda)}_{\text{mix}}(x_0),\, t,\, x_t} \left[\left\| \widehat{s}_\theta(x_t, t) - \nabla \log q(x_t) \right\|^2 \right] \;\leq\; \varepsilon^2_{\text{score}}$$

for any $\lambda \geq 0$, where $\mathbb{E}_{q^{(\lambda)}_{\text{mix}}(x_0),\, t,\, x_t}$ is an expectation over the mixed data distribution $q^{(\lambda)}_{\text{mix}}(x_0)$, a uniform distribution over t from 2 to T, and a forward process $q(x_t \,|\, x_0)$ given the data sample x_0.

Since data samples are bounded, Assumption 2 is mild in practice. Assumption 3 is the typical score matching error that is near zero if the function class $\mathcal{S}_\theta$ is sufficiently rich.

With Assumptions 2 and 3, below we bound the TV distance between $q^{(\lambda)}_{\text{mix}}$ and $\bar{p}^\star(\lambda)$ using the convergence theory of diffusion models from [47]; see Appendix B.5 for proof.

Lemma 3 (Convergence of diffusion model)**.** *Let Assumptions 2 and 3 hold. Then, the TV distance from $\bar{p}^\star(\lambda)$ to $q^{(\lambda)}_{\text{mix}}$ is bounded by*

$$\text{TV}\left(q^{(\lambda)}_{\text{mix}},\ \bar{p}^\star(\lambda)\right) \;\leq\; \sqrt{\frac{1}{2} D_{\text{KL}}\left(q^{(\lambda)}_{\text{mix}} \,\|\, \bar{p}^\star(\lambda)\right)} \;\lesssim\; \frac{d^2 \log^3 T}{\sqrt{T}} \;+\; \sqrt{d}\, \left(\log^2 T\right) \varepsilon_{\text{score}}. \tag{7}$$

Lemma 3 states that the TV distance between $q^{(\lambda)}_{\text{mix}}$ and $\bar{p}^\star(\lambda)$ decays to zero with a sublinear rate $O(\frac{1}{\sqrt{T}})$, up to a score matching error $O(\varepsilon_{\text{score}})$. When the diffusion time T is large, the TV distance between $q^{(\lambda)}_{\text{mix}}$ and $\bar{p}^\star(\lambda)$ is dominated by the score matching error. Substitution of $\lambda = \bar{\lambda}^\star$ into (7) yields an upper bound on $\text{TV}(q^{(\bar{\lambda}^\star)}_{\text{mix}},\ \bar{p}^\star(\bar{\lambda}^\star))$, which is the second term of the inequality

$$\text{TV}\left(q^\star_{\text{mix}},\ \bar{p}^\star(\bar{\lambda}^\star)\right) \;\leq\; \text{TV}\left(q^\star_{\text{mix}},\ q^{(\bar{\lambda}^\star)}_{\text{mix}}\right) \;+\; \text{TV}\left(q^{(\bar{\lambda}^\star)}_{\text{mix}},\ \bar{p}^\star(\bar{\lambda}^\star)\right). \tag{8}$$

Next, we quantify the gap between $\bar{\lambda}^\star$ and $\lambda^\star$, which lets us bound the first term on the RHS of (8) and complete the optimality analysis. To analyze the parametrized optimal dual variable $\bar{\lambda}^\star$, we introduce the richness of the parametrized class $\mathcal{S}_\theta$ and redundancy of constraints at $\widehat{s}^\star$ below.

https://doi.org/10.52202/079017-0836

Assumption 4 (Richness of parametrization)**.** *For any function $\widehat{s} \in \mathcal{S}$, there exists parameter $\theta \in \Theta$ such that $\|\widehat{s}_\theta - \widehat{s}\|_{L_2} \leq \nu$, where $\|\cdot\|_{L_2}$ is with respect to the forward process.*

Assumption 5 (Redundancy of constraints)**.** *There exists $\sigma > 0$ such that*

$$\inf_{\|\lambda\|=1} \left\| \sum_{i=1}^{m} \lambda_i \nabla_{\widehat{s}} \mathbb{E}_{q^i(x_0),\, t,\, x_t} \left[\widehat{s}^\star(x_t, t) - \nabla \log q(x_t) \right] \right\|_{L_2} \geq \sigma \tag{9}$$

where $\nabla_{\widehat{s}}$ is the Fréchet derivative over the function $\widehat{s}$ and $\widehat{s}^\star$ is a solution to Problem (U-LOSS)*.*

Assumption 4 is mild since the gap is small for expressive neural networks [56, 31]. Assumption 5 captures the linear independence of constraints, which is similarly used in optimization [4].

Due to Assumption 2, we set the function class $\mathcal{S}$ to be bounded $\|\widehat{s}\|_{L_2} \leq T^c := R$. Using Problem (U-LOSS), we prove that the unparametrized dual function $g_s(\lambda)$ is differentiable, and strongly-concave over $\mathcal{H}$ with parameter μ, where $\mathcal{H} := \{\gamma\lambda^\star + (1-\gamma)\bar{\lambda}^\star, \gamma \in [0,1]\}$ and $\mu := \left(\sigma / \left(1 + \max\left(\|\lambda^\star\|_1, \|\bar{\lambda}^\star\|_1\right)\right)\right)^2$,which leads to Lemma 4; see Appendix B.6 for their proofs.

Lemma 4. *Let Assumptions 4 and 5 hold. Then, $\|\bar{\lambda}^\star - \lambda^\star\|^2 \leq \frac{8}{\mu} R \left(1 + \|\bar{\lambda}^\star\|_1\right) \nu$.*

Since TV $\left(q^\star_{\text{mix}},\ q^{(\bar{\lambda}^\star)}_{\text{mix}}\right)$ is bounded by $\|\bar{\lambda}^\star - \lambda^\star\|_1$ (see Appendix B.6), application of Lemma 3 and Lemma 4 to (8) leads to Theorem 3; see Appendix B.7 for proof.

Theorem 3 (Optimality of constrained diffusion model)**.** *Let Assumptions 1–5 hold. Then, the total variation distance between $\bar{p}^\star(\bar{\lambda}^\star)$ and $q^\star_{\text{mix}}$ is upper bounded by*

$$\text{TV}\left(q^\star_{\text{mix}},\ \bar{p}^\star(\bar{\lambda}^\star)\right) \lesssim \frac{d^2 \log^3 T}{\sqrt{T}} + \sqrt{\frac{8}{\mu} m R \left(1 + \|\bar{\lambda}^\star\|_1\right) \nu} + \sqrt{d} \left(\log^2 T\right) \varepsilon_{\text{score}}.$$

Theorem 3 states that the TV distance between $\bar{p}^\star(\bar{\lambda}^\star)$ and $q^\star_{\text{mix}}$ decays to zero with a sublinear rate $O(\frac{1}{\sqrt{T}})$, up to a parametrization gap $O(\sqrt{\nu})$ and a prediction error $O(\varepsilon_{\text{score}})$. When the parametrization is rich enough, the parametrization gap ν and the prediction error $O(\varepsilon_{\text{score}})$ are nearly zero. In this case, if the diffusion time T is large, then $\bar{p}^\star(\bar{\lambda}^\star)$ is close to $q^\star_{\text{mix}}$ in TV distance, which recovers the ideal optimal constrained model in the unparametrized case in Section 3.1.

4.3 Practical dual training algorithm

Having established the optimality of our dual training method, we futher turn Algorithm 1 into a practical algorithm. First, we relax the computation of a diffusion model $\widehat{s}_\theta(h)$ in line 4 of Algorithm 1 to be approximate: $\mathcal{L}_s(\widehat{s}_\theta(h), \lambda(h)) \leq \min_{\theta \in \Theta} \mathcal{L}_s(\widehat{s}_\theta, \lambda(h)) + \varepsilon^2_{\text{approx}}$, where $\varepsilon^2_{\text{approx}}$ is an approximation error of training a diffusion model given $\lambda(h)$. Second, we replace the gradient in line 5 of Algorithm 1 by a stochastic gradient $\widehat{\mathbb{E}}_{x_0 \sim q^i,\, t,\, x_t}\left[\ \|\widehat{s}_\theta(x_t, t) - \nabla \log q(x_t)\|^2\ \right]$, which enables Algorithm 1 to be a stochastic algorithm, where $\widehat{\mathbb{E}}_{x_0 \sim q^i,\, t,\, x_t}$ is an unbiased estimate of $\mathbb{E}_{x_0 \sim q^i,\, t,\, x_t}$. To analyze this approximate and stochastic variant of Algorithm 1, it is useful to introduce the maximum parametrized dual function in history up to step h by $\bar{g}_{\text{best}}(h) := \max_{h' \leq h} \bar{g}_s(\lambda(h'))$, and an upper bound of the second-order moment of stochastic gradient $S^2 := \sum_{i=1}^m \mathbb{E}\left[\ \left(\widehat{\mathbb{E}}_{x_0 \sim q^i,\, t,\, x_t}\left[\ \|\widehat{s}_\theta(h)(x_t, t) - \nabla q(x_t)\|^2\ \right] - \widetilde{b}^i\right)^2 \mid \lambda(h)\ \right]$.

Denote the dual variable that achieves $\bar{g}_{\text{best}}(h)$ by λ_{best}. To bound the TV distance between $\bar{p}^\star(\bar{\lambda}_{\text{best}})$ and $q^\star_{\text{mix}}$, we check the TV distance between $q^{(\bar{\lambda}_{\text{best}})}_{\text{mix}}$ and $\bar{p}^\star(\bar{\lambda}_{\text{best}})$ using Lemma 3. The rest is to analyze the convergence of $\bar{\lambda}_{\text{best}}$ to $\lambda^\star$ via application of martingale convergence. We defer their proofs to Appendix B.8 and present the optimality of $\bar{p}^\star(\bar{\lambda}_{\text{best}})$ in Theorem 4.

Theorem 4 (Optimality of approximate constrained diffusion model)**.** *Let Assumptions 1–5 hold. Then, the total variation distance between $\bar{p}^\star(\bar{\lambda}_{\text{best}})$ and $q^\star_{\text{mix}}$ is upper bounded by*

$$\text{TV}\left(q^\star_{\text{mix}},\ \bar{p}^\star(\bar{\lambda}_{\text{best}})\right) \lesssim \frac{d^2 \log^3 T}{\sqrt{T}} + \frac{8R\left(1 + \|\bar{\lambda}_{\text{best}}\|_1\right)}{\mu}\nu + \sqrt{d}\left(\log^2 T\right)\varepsilon_{\text{score}} + \frac{2}{\mu}\varepsilon^2_{\text{approx}} + \frac{\eta\, S^2}{\mu}.$$

Theorem 4 states that the TV distance between $\bar{p}^\star(\bar{\lambda}_{\text{best}})$ and $q^\star_{\text{mix}}$ decays to zero with a sublinear rate $O(\frac{1}{\sqrt{T}})$, up to a parametrization gap $O(\nu)$, a score matching error $O(\varepsilon_{\text{score}})$, an approximation error $O(\varepsilon_{\text{approx}})$, and stepsize $O(\eta)$. When the parametrization is rich enough, the parametrization gap ν and the score matching error $O(\varepsilon_{\text{score}})$ are near zero. Thus, if the diffusion time T is large, then the closeness of $\bar{p}^\star(\bar{\lambda}_{\text{best}})$ to $q^\star_{\text{mix}}$ in TV distance is governed by the appproximation error and stepsize.

5 Computational experiments

We demonstrate the effectiveness of constrained diffusion models trained by our dual training algorithm in two constrained settings in Section 3.2; see Appendix C for experimental details.

Fairness to underrepresented classes. We train constrained diffusion models over three datasets: MNIST digits [44], Celeb-A faces [51], and Image-Net[1] [63]. For MNIST and Image-Net, we create a dataset for the distribution q in (P-LOSS) by taking a subset of the dataset with equal number of samples from each class. Then we make some classes under-represented by removing their samples. For each distribution q^i, we use samples from the associated underrepresented class. For Celeb-A, our approach is similar to MNIST except we don't remove any samples due to the existence of class imbalance in the dataset (58% female vs 42% male). For Image-Net, since the images are of high resolution, we employ the latent diffusion scheme [62] by imposing distribution constraints in latent space. Figures 1–3 show that our constrained model samples more often from the underrepresented classes (MNIST: 4, 5, 7; Celeb-A: male; Image-Net: 'Cassette player', 'French horn', and 'Golf ball'), leading to a more uniform sampling over all classes. This reflects our theoretical insights on promoting fairness for minority classes (see Section 3.2). Quantitatively, we observe *fairly lower FID scores* when training over *the same dataset but with constraints* (see Appendix C for further discussion on FID scores). Furthermore, our Image-Net experiment shows that our approach extends to the state-of-the-art diffusion models in latent space.

Figure 1: Generation performance comparison of constrained and unconstrained models that are trained on MNIST with three minorities: 4, 5, 7. (Left) Frequencies of ten digits that are generated by an unconstrained model ( ) and our constrained model ( ); (Middle) Generated digits from unconstrained model (FID 15.9); (Right) Generated digits from our constrained model (FID 13.4).

Adapting pretrained model to new data. Given a pretrained diffusion model over some original dataset $\mathcal{D}_{\text{pretrain}}$, we fine-tune the pretrained model for generating data that resemble $\mathcal{D}_{\text{new}}$. To cast this problem into (P-KL), we let the data distribution be $\mathcal{D}_{\text{new}}$, i.e., $q(x_{0:T}) = q_{\text{new}}(x_{0:T})$ and the constrained distribution be the pre-trained model, i.e., $q^i(x_{0:T}) = p_{\theta_{\text{pre}}}(x_{0:T})$. In our experiments, we pretrain a diffusion model on a subset of MNIST digits excluding a class of digits (MNIST: 9), and fine-tune this model using samples of the excluded digit. Figure 4 shows that our constrained fine-tuned model samples from the new class as well as all previous classes, whereas the model fine-tuned without the constraint quickly overfits to the new dataset (see Appendix C for details). Our constrained model generates much better high-quality samples than the unconstrained model.

6 Conclusion

We have presented a constrained optimization framework for training diffusion models under distribution constraints. We have developed a Lagrangian-based dual algorithm to train such constrained

[1]We use a subset of ten classes from Image-Net: 'Tench Fish', 'English Springer Dog', 'Cassette Player', 'Chain saw', 'Church', 'French Horn', 'Garbage Truck', 'Gas Pump', 'Golf Ball', 'Parachute.'

https://doi.org/10.52202/079017-0836

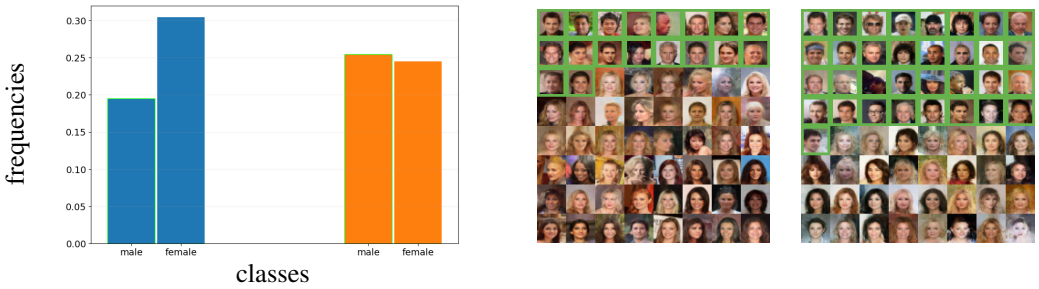

Figure 2: Generation performance comparison of constrained and unconstrained models that are trained on Celeb-A with male minority. (Left) Frequencies of two genders that are generated by an unconstrained model ( ) and our constrained model ( ); (Middle) Generated faces from unconstrained model (FID 19.6); (Right) Generated faces from our constrained model (FID 11.6).

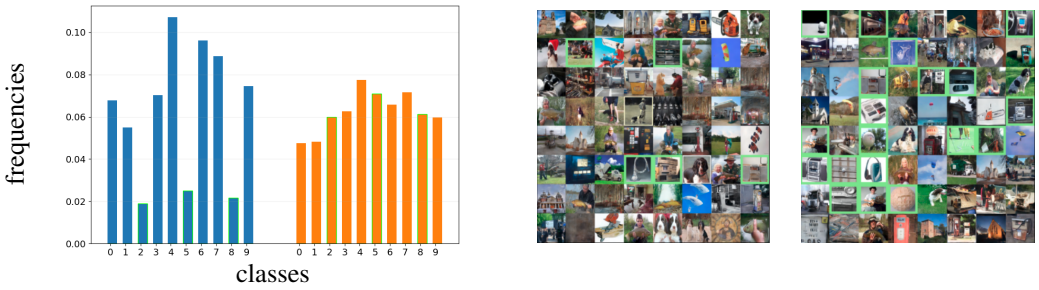

Figure 3: Generation performance comparison of constrained and unconstrained models that are trained on Image-Net with minority classes: 'Cassette player' (2), 'French horn' (5), and 'Golf ball' (8). (Left) Frequencies of ten classes that are generated by an unconstrained model ( ) and our constrained model ( ); (Middle) Generated images from unconstrained model (FID 36.0); (Right) Generated images from our constrained model (FID 27.3).

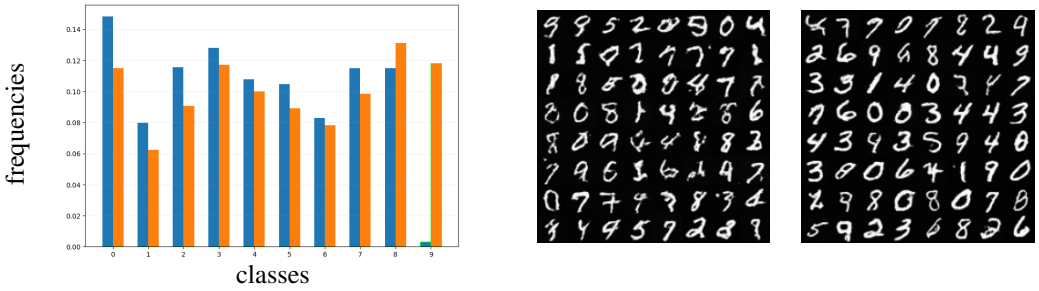

Figure 4: Fine-tuning performance comparison of constrained and unconstrained models that are trained on MNIST. (Left) Frequencies of ten digits that are generated by a pre-trained model without digit 9 ( ) and our fine-tuned constrained model ( ); (Middle) Generated digits from unconstrained model (FID 45.9); (Right) Generated digits from our constrained model (FID 25.2).

diffusion models. Our theoretical analysis shows that our constrained diffusion model generates new data from an optimal mixture data distribution that satisfies the constraints, and we have demonstrated the effectiveness of our distribution constraints in reducing bias across three widely-used datasets.

This work directly stimulates several research directions: (i) extend our distribution constraints to other domain constraints, e.g., mirror diffusion [48]; (ii) incorporate conditional generations, e.g., text-to-image generation [28, 65], into our constrained diffusion models; (iii) conduct experiments with text-to-image datasets to identify and address biases; (iv) improve the convergence theory using more advanced diffusion processes.

https://doi.org/10.52202/079017-0836

Acknowledgments

We thank reviewers and program chairs for providing helpful comments.

References

[1] A. Bansal, H.-M. Chu, A. Schwarzschild, S. Sengupta, M. Goldblum, J. Geiping, and T. Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 843–852, 2023.

[2] J. Benton, V. De Bortoli, A. Doucet, and G. Deligiannidis. Linear convergence bounds for diffusion models via stochastic localization. In *Proceedings of the International Conference on Learning Representations*, 2024.

[3] J. Benton, V. De Bortoli, A. Doucet, and G. Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. In *Proceedings of the International Conference on Learning Representations*, 2024.

[4] D. P. Bertsekas. *Nonlinear programming*. Athena Scientific, 2016.

[5] K. Black, M. Janner, Y. Du, I. Kostrikov, and S. Levine. Training diffusion models with reinforcement learning. In *Proceedings of the International Conference on Learning Representations*, 2024.

[6] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023.

[7] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[8] H. Cao, C. Tan, Z. Gao, Y. Xu, G. Chen, P.-A. Heng, and S. Z. Li. A survey on generative diffusion models. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[9] L. Chamon and A. Ribeiro. Probably approximately correct constrained learning. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 33, pages 16722–16735, 2020.

[10] L. F. Chamon, S. Paternain, M. Calvo-Fullana, and A. Ribeiro. Constrained learning with non-convex losses. *IEEE Transactions on Information Theory*, 69(3):1739–1760, 2022.

[11] H. Chen, H. Lee, and J. Lu. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *Proceedings of the International Conference on Machine Learning*, pages 4735–4763, 2023.

[12] J. Chen, Z. Shao, X. Zheng, K. Zhang, and J. Yin. Integrating aesthetics and efficiency: AI-driven diffusion models for visually pleasing interior design generation. *Scientific Reports*, 14(1):3496, 2024.

[13] J. Chen, R. Zhang, Y. Zhou, and C. Chen. Towards aligned layout generation via diffusion model with aesthetic constraints. *arXiv preprint arXiv:2402.04754*, 2024.

[14] M. Chen, S. Mei, J. Fan, and M. Wang. An overview of diffusion models: Applications, guided generation, statistical rates and optimization. *arXiv preprint arXiv:2404.07771*, 2024.

[15] S. Chen, S. Chewi, J. Li, Y. Li, A. Salim, and A. R. Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *Proceedings of the International Conference on Learning Representations*, 2023.

[16] K. Choi, A. Grover, T. Singh, R. Shu, and S. Ermon. Fair generative modeling via weak supervision. In *Proceedings of the International Conference on Machine Learning*, pages 1887–1898, 2020.

 https://doi.org/10.52202/079017-0836

[17] J. K. Christopher, S. Baek, and F. Fioretto. Projected generative diffusion models for constraint satisfaction. *arXiv preprint arXiv:2402.03559*, 2024.

[18] K. Clark, P. Vicol, K. Swersky, and D. J. Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *Proceedings of the International Conference on Learning Representations*, 2024.

[19] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[20] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794, 2021.

[21] Y. Du, C. Durkan, R. Strudel, J. B. Tenenbaum, S. Dieleman, R. Fergus, J. Sohl-Dickstein, A. Doucet, and W. S. Grathwohl. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and MCMC. In *Proceedings of the International conference on machine learning*, pages 8489–8510, 2023.

[22] J. Elenter, L. F. Chamon, and A. Ribeiro. Near-optimal solutions of constrained learning problems. In *Proceedings of the International Conference on Learning Representations*, 2024.

[23] Y. Fan and K. Lee. Optimizing DDPM sampling with shortcut fine-tuning. In *Proceedings of the International Conference on Machine Learning*, pages 9623–9639, 2023.

[24] Y. Fan, O. Watkins, Y. Du, H. Liu, M. Ryu, C. Boutilier, P. Abbeel, M. Ghavamzadeh, K. Lee, and K. Lee. Reinforcement learning for fine-tuning text-to-image diffusion models. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 36, 2023.

[25] B. T. Feng, R. Baptista, and K. L. Bouman. Neural approximate mirror maps for constrained diffusion models. *arXiv preprint arXiv:2406.12816*, 2024.

[26] N. Fishman, L. Klarner, V. De Bortoli, E. Mathieu, and M. J. Hutchinson. Diffusion models for constrained domains. *Transactions on Machine Learning Research*, 2024.

[27] N. Fishman, L. Klarner, E. Mathieu, M. Hutchinson, and V. De Bortoli. Metropolis sampling for constrained diffusion models. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 36, 2024.

[28] F. Friedrich, M. Brack, L. Struppek, D. Hintersdorf, P. Schramowski, S. Luccioni, and K. Kersting. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893*, 2023.

[29] G. Giannone, A. Srivastava, O. Winther, and F. Ahmed. Aligning optimization trajectories with diffusion models for constrained design generation. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 36, 2023.

[30] S. Gugger, L. Debut, T. Wolf, P. Schmid, Z. Mueller, S. Mangrulkar, M. Sun, and B. Bossan. Accelerate: Training and inference at scale made simple, efficient and adaptable. https://github.com/huggingface/accelerate, 2022.

[31] Y. Han, M. Razaviyayn, and R. Xu. Neural network-based score estimation in diffusion models: Optimization and generalization. *arXiv preprint arXiv:2401.15604*, 2024.

[32] Y. Hao, Z. Chi, L. Dong, and F. Wei. Optimizing prompts for text-to-image generation. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 36, 2023.

[33] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 30, 2017.

[34] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

[35] J. Ho and T. Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[36] I. Hounie, A. Ribeiro, and L. F. Chamon. Resilient constrained learning. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 36, 2023.

[37] C.-W. Huang, M. Aghajohari, J. Bose, P. Panangaden, and A. C. Courville. Riemannian diffusion models. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 35, pages 2750–2761, 2022.

[38] D. Z. Huang, J. Huang, and Z. Lin. Convergence analysis of probability flow ODE for score-based generative models. *arXiv preprint arXiv:2404.09730*, 2024.

[39] L. Huang, T. Xu, Y. Yu, P. Zhao, K.-C. Wong, and H. Zhang. A dual diffusion model enables 3D binding bioactive molecule generation and lead optimization given target pockets. *bioRxiv*, pages 2023–01, 2023.

[40] A. Kazerouni, E. K. Aghdam, M. Heidari, R. Azad, M. Fayyaz, I. Hacihaliloglu, and D. Merhof. Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis*, page 102846, 2023.

[41] Y. Kim, B. Na, M. Park, J. Jang, D. Kim, W. Kang, and I.-C. Moon. Training unbiased diffusion models from biased dataset. In *Proceedings of the International Conference on Learning Representations*, 2024.

[42] D. Kingma and R. Gao. Understanding diffusion objectives as the ELBO with simple data augmentation. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 36, 2023.

[43] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro. DiffWave: A versatile diffusion model for audio synthesis. In *Proceedings of the International Conference on Learning Representations*, 2020.

[44] Y. LeCun, C. Cortes, and C. J. Burges. The MNIST database of handwritten digits. http://yann.lecun.com/exdb/mnist/.

[45] K. Lee, H. Liu, M. Ryu, O. Watkins, Y. Du, C. Boutilier, P. Abbeel, M. Ghavamzadeh, and S. S. Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023.

[46] G. Li, Y. Huang, T. Efimov, Y. Wei, Y. Chi, and Y. Chen. Accelerating convergence of score-based diffusion models, provably. *arXiv preprint arXiv:2403.03852*, 2024.

[47] G. Li, Y. Wei, Y. Chen, and Y. Chi. Towards faster non-asymptotic convergence for diffusion-based generative models. In *Proceedings of the International Conference on Learning Representations*, 2024.

[48] G.-H. Liu, T. Chen, E. Theodorou, and M. Tao. Mirror diffusion models for constrained and watermarked generation. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 36, 2023.

[49] N. Liu, S. Li, Y. Du, A. Torralba, and J. B. Tenenbaum. Compositional visual generation with composable diffusion models. In *Proceedings of the European Conference on Computer Vision*, pages 423–439, 2022.

[50] X. Liu, X. Tong, and Q. Liu. Sampling with trusthworthy constraints: A variational gradient framework. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 23557–23568, 2021.

[51] Z. Liu, P. Luo, X. Wang, and X. Tang. Large-scale celebfaces attributes (celeba) dataset. https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html.

[52] I. Loshchilov and F. Hutter. Decoupled weight decay regularization, 2019.

[53] A. Lou and S. Ermon. Reflected diffusion models. In *Proceedings of the International Conference on Machine Learning*, pages 22675–22701, 2023.

 https://doi.org/10.52202/079017-0836

[54] A. S. Luccioni, C. Akiki, M. Mitchell, and Y. Jernite. Stable bias: Analyzing societal representations in diffusion models. *arXiv preprint arXiv:2303.11408*, 2023.

[55] C. Luo. Understanding diffusion models: A unified perspective. *arXiv preprint arXiv:2208.11970*, 2022.

[56] S. Mei and Y. Wu. Deep networks as denoising algorithms: Sample-efficient learning of diffusion models in high-dimensional graphical models. *arXiv preprint arXiv:2309.11420*, 2023.

[57] R. Naik and B. Nushi. Social biases through the text-to-image generation lens. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 786–808, 2023.

[58] G. Parmar, R. Zhang, and J.-Y. Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11410–11420, 2022.

[59] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An imperative style, high-performance deep learning library, 2019.

[60] T. Power, R. Soltani-Zarrin, S. Iba, and D. Berenson. Sampling constrained trajectories using composable diffusion models. In *Proceedings of the IROS 2023 Workshop on Differentiable Probabilistic Robotics: Emerging Perspectives on Robot Learning*, 2023.

[61] D. J. Rezende and F. Viola. Generalized ELBO with constrained optimization, GECO. In *Workshop on Bayesian Deep Learning, NeurIPS*, 2018.

[62] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[63] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[64] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 35, pages 36479–36494, 2022.

[65] X. Shen, C. Du, T. Pang, M. Lin, Y. Wong, and M. Kankanhalli. Finetuning text-to-image diffusion models for fairness. In *Proceedings of the International Conference on Learning Representations*, 2024.

[66] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the International Conference on Machine Learning*, pages 2256–2265, 2015.

[67] V. Solo and X. Kong. *Adaptive signal processing algorithms: stability and performance*. Prentice-Hall, Inc., 1994.

[68] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. In *Proceedings of the International Conference on Learning Representations*, 2021.

[69] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[70] W. Tang. Fine-tuning of diffusion models via stochastic control: entropy regularization and beyond. *arXiv preprint arXiv:2403.06279*, 2024.

[71] M. Uehara, Y. Zhao, K. Black, E. Hajiramezanali, G. Scalia, N. L. Diamant, A. M. Tseng, T. Biancalani, and S. Levine. Fine-tuning of continuous-time diffusion models as entropy-regularized control. *arXiv preprint arXiv:2402.15194*, 2024.

[72] M. Uehara, Y. Zhao, E. Hajiramezanali, G. Scalia, G. Eraslan, A. Lal, S. Levine, and T. Biancalani. Bridging model-based optimization and generative modeling via conservative fine-tuning of diffusion models. *arXiv preprint arXiv:2405.19673*, 2024.

[73] P. von Platen, S. Patil, A. Lozhkov, P. Cuenca, N. Lambert, K. Rasul, M. Davaadorj, D. Nair, S. Paul, W. Berman, Y. Xu, S. Liu, and T. Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.

[74] J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, et al. De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976):1089–1100, 2023.

[75] T. Weiss, E. Mayo Yanes, S. Chakraborty, L. Cosmo, A. M. Bronstein, and R. Gershoni-Poranne. Guided diffusion for inverse molecular design. *Nature Computational Science*, 3(10):873–882, 2023.

[76] X. Wu, K. Sun, F. Zhu, R. Zhao, and H. Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2096–2105, 2023.

[77] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 36, 2023.

[78] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Comput. Surv.*, 56(4), nov 2023.

[79] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39, 2023.

[80] H. Yuan, K. Huang, C. Ni, M. Chen, and M. Wang. Reward-directed conditional diffusion: Provable distribution estimation and reward improvement. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 36, 2023.

[81] C. Zhang, C. Zhang, S. Zheng, M. Zhang, M. Qamar, S.-H. Bae, and I. S. Kweon. Audio diffusion model for speech synthesis: A survey on text to speech and speech enhancement in generative AI. *arXiv preprint arXiv:2303.13336*, 2023.

[82] Z. Zhang, S. Zhang, Y. Zhan, Y. Luo, Y. Wen, and D. Tao. Confronting reward overoptimization for diffusion models: A perspective of inductive and primacy biases. In *Proceedings of the International Conference on Machine Learning*, 2024.

 https://doi.org/10.52202/079017-0836

Supplementary Materials for "Constrained Diffusion Models via Dual Training"

A Details on ELBO

Recall the evidence lower bound (ELBO),

$$E(p;q) \;:=\; \mathbb{E}_{q(x_0)}\mathbb{E}_{q(x_{1:T}\,|\,x_0)} \log \frac{p(x_{0:T})}{q(x_{1:T}\,|\,x_0)},$$

we can utilize conditionals to expand it into

$$\begin{aligned} E(p;q) \;=\; & \underbrace{\mathbb{E}_{q(x_0)}\mathbb{E}_{q(x_1\,|\,x_0)}\left[\log p(x_0\,|\,x_1)\right]}_{\text{reconstruction likelihood}} - \underbrace{\mathbb{E}_{q(x_0)}\left[D_{\mathrm{KL}}(q(x_T\,|\,x_0)\,\|\,p(x_T))\right]}_{\text{final latent mismatch}} \\ & \underbrace{- \sum_{t=2}^{T} \mathbb{E}_{q(x_0)}\mathbb{E}_{q(x_t\,|\,x_0)}\left[D_{\mathrm{KL}}\left(q(x_{t-1}\,|\,x_t,x_0)\,\|\,p(x_{t-1}\,|\,x_t)\right)\right]}_{\text{denoising matching term}} \end{aligned}$$

where the first term is the reconstruction likelihood of the original data given the first latent x_1, the second term is the mismatch between the final latent distribution and the Guassian prior, and the last summation measures the mismatch between the denoising transitions from forward/backward processes. With the variance schedule described in Section 2.1, it is known that the reconstruction likelihood and final latent mismatch are negligible, and thus the approximation in (4) is almost exact, which is our focal setting of this paper.

We next focus on one summand of the denoising matching term,

$$\mathbb{E}_{q(x_0)}\mathbb{E}_{q(x_t\,|\,x_0)}\left[D_{\mathrm{KL}}\left(q(x_{t-1}\,|\,x_t,x_0)\,\|\,p(x_{t-1}\,|\,x_t)\right)\right].$$

Application of the reparametrization trick leads to $x_t = \sqrt{\alpha_t}x_{t-1} + \sqrt{1-\alpha_t}\epsilon_{t-1}$, where $\epsilon_{t-1} \sim \mathcal{N}(0,I)$ is a white noise sample. Using Bayes rule, we can express $q(x_{t-1}\,|\,x_t,x_0)$ as a Guassian distribution

$$\mathcal{N}(x_{t-1};\mu_q(x_t),\sigma_q(t)I)$$

where $\mu_q(x_t) = \frac{1}{\sqrt{\alpha_t}}x_t + \frac{1-\alpha_t}{\sqrt{\alpha_t}}\nabla \log q(x_t)$ is the mean and $\sigma_q^2(t) = \frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}$ is the variance.

To stay close to the ground-truth backward conditional $q(x_{t-1}\,|\,x_t,x_0)$ as much as possible, we take $p(x_{t-1}\,|\,x_t)$ to be the same as $q(x_{t-1}\,|\,x_t,x_0)$ except replacing $\nabla \log p(x_t)$ by $\widehat{s}(x_t,t)$ and $\sigma_q^2(t)$ by $\sigma_p^2(t)$,

$$\mathcal{N}(x_{t-1};\widehat{\mu}(x_t),\sigma_p(t)I)$$

where $\widehat{\mu}(x_t) = \frac{1}{\sqrt{\alpha_t}}x_t + \frac{1-\alpha_t}{\sqrt{\alpha_t}}\widehat{s}(x_t,t)$. Thus,

$$\begin{aligned} & D_{\mathrm{KL}}\left(q(x_{t-1}\,|\,x_t,x_0)\,\|\,p(x_{t-1}\,|\,x_t)\right) \\ = \;& D_{\mathrm{KL}}(\mathcal{N}(x_{t-1};\mu_q(x_t),\sigma_q(t)I)\,\|\,\mathcal{N}(x_{t-1};\widehat{\mu}(x_t),\sigma_p(t)I)) \\ = \;& \frac{1}{2}\left(d\log\frac{\sigma_p^2(t)}{\sigma_q^2(t)} - d + d\frac{\sigma_p^2(t)}{\sigma_q^2(t)} + \frac{1}{\sigma_q^2(t)}\left\|\mu_q(x_t,x_0) - \widehat{\mu}(x_t,x_0)\right\|^2\right) \\ = \;& \underbrace{\frac{1}{2}\left(d\log\frac{\sigma_p^2(t)}{\sigma_q^2(t)} - d + d\frac{\sigma_p^2(t)}{\sigma_q^2(t)}\right)}_{\text{variance mismatch}} + \underbrace{\frac{1}{2\sigma_q^2(t)}\frac{(1-\alpha_t)^2}{\alpha_t}\left\|\widehat{s}(x_t,t) - \nabla\log q(x_t)\right\|^2}_{\text{prediction loss}} \end{aligned}$$

where the second equality is due to the KL Divergence between two Gaussians. Since $\sigma_p^2(t)$ and $\sigma_q^2(t)$ are constants, the variance mismatch term is irrelevant to optimization. Denote $\omega_t := \frac{(1-\alpha_t)^2}{2\sigma_q^2(t)\alpha_t}$ and $\bar{\omega} := \sum_{t=2}^{T}\omega_t$. We can define a discrete distribution over the set $\{2,\ldots,T\}$ as

$p_\omega(t) := \frac{\omega_t}{\bar{\omega}}$. Also denote $v := \sum_{t=2}^{T} \frac{1}{2} \left(d \log \frac{\sigma_p^2(t)}{\sigma_q^2(t)} - d + d \frac{\sigma_p^2(t)}{\sigma_q^2(t)} \right)$. Hence, the ELBO maximization: $\operatorname{maximize}_p E(p; q)$, is equivalent to the quadratic loss minimization,

$$\underset{\widehat{s}}{\text{minimize}} \quad v \, + \, \bar{\omega} \, \mathbb{E}_{x_0,\, t,\, x_t} \left[\, \| \widehat{s}(x_t, t) - \nabla \log q(x_t) \|^2 \, \right]$$

where $\mathbb{E}_{x_0,\, t,\, x_t}$ is an expectation over the data distribution $q(x_0)$, the discrete distribution $p_\omega(t)$ from 2 to T, and a forward process $q(x_t \,|\, x_0)$ given the data sample x_0. Since shifting an objective function by a constant and scaling an objective function by a constant don't change the solution of an optimization problem, we omit constants v and $\bar{\omega}$ for brevity, and only emphasize them whenever it is necessary. Hence, the ELBO maximization equals the quadratic loss minimization,

$$\underset{\widehat{s}}{\text{minimize}} \quad \mathbb{E}_{x_0,\, t,\, x_t} \left[\, \| \widehat{s}(x_t, t) - \nabla \log q(x_t) \|^2 \, \right]$$

up to some scaling and shifting constants, where $\mathbb{E}_{x_0,t,x_t}$ is an expectation over the data distribution $q(x_0)$, the discrete distribution $p_\omega(t)$ from 2 to T, and a forward process $q(x_t \,|\, x_0)$ given the data sample x_0. In practice, however, we have to parametrize the estimator $\widehat{s}(x_t, t)$ as $\widehat{s}_\theta(x_t, t)$ with parameter $\theta \in \Theta$,

$$\underset{\theta \,\in\, \Theta}{\text{minimize}} \quad \mathbb{E}_{x_0,\, t,\, x_t} \left[\, \| \widehat{s}_\theta(x_t, t) - \nabla \log q(x_t) \|^2 \, \right]$$

which is our focal objective of generative modeling. A parametrized representation of $p(x_{t-1} \,|\, x_t)$ associated with $\widehat{s}_\theta(x_t, t)$ is denoted by $p_\theta(x_{t-1} \,|\, x_t)$ and the backward process has a parametrized joint distribution $p_\theta(x_{0:T})$. We remark that the above prediction problem can be reformulated as data or noise predictions [55], with our results directly transferrable to these formulations.

B Proofs

We provide proofs of all lemmas and theorems in the main paper.

B.1 Proof of Lemma 1

Proof. The ELBO maximization has the same optimal solution with the KL divergence minimization because of the equality (3). This directly proves the second equivalence. Next, we relate these two problems to the likelihood maximization problem.

We note that the KL divergence is non-negative and is zero if and only if two distributions are the same. Since $q(x_{0:T}) \in \mathcal{P}$ for large T, the solution of the ELBO maximization and KL divergence minimization is given by $p^\star = q$. For the KL divergence minimization, the optimal value is zero. For the optimal value of the ELBO maximization, from (3) it follows that:

$$E(p^\star; q) \;=\; \mathbb{E}_{q(x_0)}[\, \log q(x_0) \,] - D_{\mathrm{KL}}(q(x_{0:T}) \,\|\, p^\star(x_{0:T})) \;=\; \mathbb{E}_{q(x_0)}[\, \log q(x_0) \,]. \tag{10}$$

It is clear that the likelihood maximization problem $\operatorname{maximize}_p \mathbb{E}_{q(x_0)}[\, \log p(x_0) \,]$ is equivalent to $\operatorname{minimize}_p D_{\mathrm{KL}}(q(x_0) \,\|\, p(x_0))$. Therefore, any distribution $p^\star(x_{0:T})$ whose marginal satisfies $p^\star(x_0) = q(x_0)$, will be a solution of the likelihood maximization problem. This includes the solution of the KL divergence minimization and ELBO maximization which is $p^\star = q$. Therefore,

$$\underset{p \,\in\, \mathcal{P}}{\text{maximize}} \;\; E(p; q) \;\Rightarrow\; \underset{p \,\in\, \mathcal{P}}{\text{maximize}} \;\; \mathbb{E}_{q(x_0)}[\, \log p(x_0) \,] \tag{11}$$

which concludes the proof.

□

B.2 Proof of Lemma 2

Proof. It is straightforward to check the zero duality gap in convex optimization; see e.g., [4, Proposition 5.3.2]. Furthermore, for a convex optimization problem, an optimal dual variable $\lambda^\star$ that maximizes the dual function is a geometric multiplier. Hence, $(p^\star, \lambda^\star)$ is an optimal primal-dual pair of the convex optimization problem. □

https://doi.org/10.52202/079017-0836

B.3 Proof of Theorem 1

Proof. From the strong duality in Lemma 2, it is known from [4, Proposition 5.1.4] that $\lambda^\star$ is also a geometric multiplier. Thus, Problem (U-KL) reduces to an unconstrained problem,

$$\underset{p \in \mathcal{P}}{\text{minimize}} \quad \mathcal{L}(p, \lambda^\star) \tag{12}$$

where the objective function results from plugging an optimal dual variable $\lambda^\star$ into Lagrangian $\mathcal{L}(p, \lambda)$.

By the definition of Lagrangian,

$$\begin{aligned}
\mathcal{L}(p, \lambda) &= D_{\mathrm{KL}}(q(x_{0:T}) \,\|\, p(x_{0:T})) + \sum_{i=1}^{m} \lambda_i \left(D_{\mathrm{KL}} \left(q^i(x_{0:T}) \,\|\, p(x_{0:T}) \right) - b_i \right) \\
&= -E(p; q) - \sum_{i=1}^{m} \lambda_i \, E(p; q^i) \\
&\quad + \mathbb{E}_{q(x_0)} \left[\log q(x_0) \right] + \sum_{i=1}^{m} \lambda_i \left(\mathbb{E}_{q(x_0)} \left[\log q^i(x_0) \right] - b_i \right)
\end{aligned}$$

By taking $\lambda = \lambda^\star$, Problem (12) is equivalent to

$$\underset{p \in \mathcal{P}}{\text{maximize}} \quad E(p; q) \, + \, \sum_{i=1}^{m} \lambda_i^\star \, E(p; q^i). \tag{13}$$

From the definition of ELBO, we know that

$$E(p; q) + \sum_{i=1}^{m} \lambda_i^\star E(p; q^i) \;=\; \left(\mathbb{E}_{q(x_0)} + \sum_{i=1}^{m} \lambda_i^\star \mathbb{E}_{q^i(x_0)} \right) \mathbb{E}_{q(x_{1:T} \,|\, x_0)} \log \frac{p(x_{0:T})}{q(x_{1:T} \,|\, x_0)}$$

where we use the fact that the forward processes have the same marginal distribution given any initial data samples. Normalization of initial data distributions leads to $q_{\mathrm{mix}}^{(\lambda^\star)}$. Thus, Problem (13) is equivalent to

$$\underset{p \in \mathcal{P}}{\text{maximize}} \;\; E(p; q_{\mathrm{mix}}^{(\lambda^\star)})$$

which, together with Lemma 1, completes the proof. □

B.4 Proof of Theorem 2 and Feasibility Criterion

We start with the proof of Theorem 2.

Proof. Similar to the proof of Theorem 1, we begin with the Lagrangian of Problem (U-ELBO),

$$\begin{aligned}
\mathcal{L}(p, \lambda) &= -E(p; q) \, - \, \sum_{i=1}^{m} \lambda_i \left(E(p; q^i) + \bar{b}_i \right) \\
&= \lambda^T \mathbf{1} \left(- \sum_{i=1}^{m} \frac{\lambda_i}{\lambda^T \mathbf{1}} E(p; q^i) \right) \, - \, \lambda^T \bar{b} \qquad (14) \\
&= \lambda^T \mathbf{1} \left(- \sum_{i=1}^{m} \frac{\lambda_i}{\lambda^T \mathbf{1}} \mathbb{E}_{q^i(x_0)} \mathbb{E}_{q(x_{1:T} \,|\, x_0)} \log \frac{p(x_{0:T})}{q(x_{1:T} \,|\, x_0)} \right) \, - \, \lambda^T \bar{b} \\
&= \lambda^T \mathbf{1} \left(- \mathbb{E}_{q^{(\lambda)}(x_0)} \mathbb{E}_{q(x_{1:T} \,|\, x_0)} \log \frac{p(x_{0:T})}{q(x_{1:T} \,|\, x_0)} \right) \, - \, \lambda^T \bar{b} \\
&= -(\lambda^T \mathbf{1}) E(p; q^{(\lambda)}) \, - \, \lambda^T \bar{b}
\end{aligned}$$

where from (14) onwards we use notation: $\lambda_0 = 1$, $\bar{b}_0 = 0$, $\lambda = [\lambda_0, \ldots, \lambda_m]^T$, $\bar{b} = [\bar{b}_0, \ldots, \bar{b}_m]^T$, and use q^0 to represent q, which will be used in the rest of proof. To formulate the dual problem, we check the minimum of the Lagrangian,

$$\begin{aligned} g(\lambda) &:= \underset{p \in \mathcal{P}}{\text{minimize}}\ \mathcal{L}(p, \lambda) \\ &= \underset{p \in \mathcal{P}}{\text{minimize}}\ -(\lambda^T \mathbf{1}) E(p; q^{(\lambda)}) - \lambda^T \bar{b} \\ &= -\lambda^T \bar{b} + (\lambda^T \mathbf{1}) \underset{p \in \mathcal{P}}{\text{minimize}}\ -E(p; q^{(\lambda)}) \end{aligned} \tag{15}$$

where the only term that depends on p is the ELBO. Recall that:

$$D_{\text{KL}}(q(x_{0:T}) \,\|\, p(x_{0:T})) = -E(p; q) + \mathbb{E}_{q(x_0)}[\log q(x_0)].$$

Since the minimum value of the KL divergence is zero (attained when $p = q$), the minimum of $-E(p; q)$ is likewise attained when $p = q$. Thus, it is straightforward that the minimum is equal to the entropy of the distribution q, denoted by $h(q) := -\mathbb{E}_{q(x_0)}[\log q(x_0)]$. With this in mind, from (15) we have

$$g(\lambda) = -\lambda^T \bar{b} + (\lambda^T \mathbf{1})\, h(q^{(\lambda)}).$$

Thus, the dual problem reads

$$\underset{\lambda \geq 0}{\text{maximize}} \quad g(\lambda) := -\lambda^T \bar{b} + (\lambda^T \mathbf{1})\, h(q^{(\lambda)}).$$

We first reformulate the entropy of the mixture distribution $q^{(\lambda)}$,

$$\begin{aligned} h(q^{(\lambda)}) &= -\mathbb{E}_{q^{(\lambda)}(x_0)}\left[\log q^{(\lambda)}(x_0)\right] \\ &= -\int \sum_{i=0}^{m} \frac{\lambda_i}{\lambda^T \mathbf{1}} q^i(x_0) \log\left(\sum_{i=1}^{m} \frac{\lambda_i}{\lambda^T \mathbf{1}} q^i(x_0)\right) dx_0 \qquad (16) \\ &= -\int \sum_{i=0}^{m} \frac{\lambda_i}{\lambda^T \mathbf{1}} q^i(x_0) \log\left(\frac{\lambda_i}{\lambda^T \mathbf{1}} q^i(x_0)\right) dx_0 \qquad (17) \\ &= -\sum_{i=0}^{m} \frac{\lambda_i}{\lambda^T \mathbf{1}} \underbrace{\int q^i(x_0) \log\left(q^i(x_0)\right) dx_0}_{:= -h_i} - \sum_{i=0}^{m} \frac{\lambda_i}{\lambda^T \mathbf{1}} \log\left(\frac{\lambda_i}{\lambda^T \mathbf{1}}\right) \\ &= \sum_{i=0}^{m} \frac{\lambda_i}{\lambda^T \mathbf{1}} h_i - \sum_{i=0}^{m} \frac{\lambda_i}{\lambda^T \mathbf{1}} \log\left(\frac{\lambda_i}{\lambda^T \mathbf{1}}\right) \\ &= \sum_{i=0}^{m} \frac{\lambda_i}{\lambda^T \mathbf{1}} h_i - \sum_{i=0}^{m} \frac{\lambda_i}{\lambda^T \mathbf{1}} \log(\lambda_i) + \log(\lambda^T \mathbf{1}) \end{aligned}$$

where going from (16) to (17) we utilize the assumption that the distributions $\{q^i\}_{i=0}^m$ have disjoint supports; see Remark 1 on when this is the case.

Now, we can compute the gradient of the dual function over λ_i, $i = 1, \ldots, m$,

$$\begin{aligned} \frac{\partial}{\partial \lambda_i}\left(-\lambda^T \bar{b} + (\lambda^T \mathbf{1})\, h(q^{(\lambda)})\right) &= \frac{\partial}{\partial \lambda_i}\left(-\lambda^T \bar{b} + \sum_{j=0}^{m} \lambda_j h_j - \sum_{j=0}^{m} \lambda_j \log \lambda_j + (\lambda^T \mathbf{1}) \log(\lambda^T \mathbf{1})\right) \\ &= h_i - \bar{b}_i - \log\left(\frac{\lambda_i}{\lambda^T \mathbf{1}}\right). \end{aligned}$$

Setting the gradient be zeros allows us to find the optimal dual variables $\lambda^\star$,

$$h_i - \bar{b}_i - \log\left(\frac{\lambda_i^\star}{(\lambda^\star)^T \mathbf{1}}\right) = 0 \quad \text{for } i = 1, \ldots, m.$$

https://doi.org/10.52202/079017-0836

Hence,

$$\frac{\lambda_i^\star}{(\lambda^\star)^T\mathbf{1}} \;=\; \mathrm{e}^{h_i - \bar{b}_i} \quad \text{for } i = 1, \ldots, m. \tag{18}$$

We clarify that in (18), $\lambda^\star = [\lambda_0^\star, \ldots, \lambda_m^\star]^T$ with its first element being $\lambda_0^\star = 1$. Finally, if we return back to notation $\lambda^\star = [\lambda_1^\star, \ldots, \lambda_m^\star]^T$, then,

$$\frac{\lambda_i^\star}{1 + (\lambda^\star)^T\mathbf{1}} \;=\; \mathrm{e}^{h_i - \bar{b}_i} \quad \text{for } i = 1, \ldots, m$$

which completes the proof. □

Remark 1. *We remark on the assumption of the distributions $\{q^i\}_{i=0}^m$ having disjoint supports. In the setting of adapting model to new data in Section 3.2, this is a reasonable assumption. Since we often finetune a pre-trained diffusion model on new data not seen in the original pre-training dataset, the new data distribution and the pre-training data distribution have mostly disjoint supports. In the minority class setting in Section 3.2, the constrained distributions $\{q^i\}_{i=1}^m$ and the objective distribution q^0 usually are not disjoint. However, since the distributions $\{q^i\}_{i=1}^m$ often are often restrictions of q^0 to subsets of the support of q^0, i.e., the minority classes, the derivation of optimal dual variables is similar to what we have provided in this section, so we omit the repeated details. Extending these results to cases where the distributions are neither disjoint nor restrictions of the objective distributions, is challenging and has been left to future work.*

To prove a feasibility criterion, we first show that the dual function is finite in Lemma 5.

Lemma 5 (Boundedness of the optimal dual function)**.** *Let the differential entropy h_i of each distribution q^i be finite. Then, the optimal value of the dual function $D^\star := \max_{\lambda \geq 0} g(\lambda)$ is finite if and only if*

$$\sum_{i=1}^m \mathrm{e}^{h_i - \bar{b}_i} \;<\; 1.$$

Proof. ($\Leftarrow$) From $\sum_{i=1}^m \mathrm{e}^{h_i - \bar{b}_i} < 1$, the otpimal dual variable $\lambda^\star$ given by (18) is finite. Thus, the optimal value of the dual function $g(\lambda^\star)$ becomes

$$g(\lambda^\star) \;=\; -(\lambda^\star)^T\bar{b} + \sum_{i=0}^m \lambda_i^\star h_i - \sum_{i=0}^m \lambda_i^\star \log \lambda_i^\star + ((\lambda^\star)^T\mathbf{1}) \log((\lambda^\star)^T\mathbf{1})$$

and $\lambda_i^\star = \mathrm{e}^{h_i - \bar{b}_i} > 0$, and also $\{h_i\}_{i=1}^m$ are all finite. Therefore, $D^\star$ is finite.

($\Rightarrow$) We prove it by contradiction. Assume $\sum_{i=1}^m \mathrm{e}^{h_i - \bar{b}_i} = \mathrm{e}^\delta \geq 1$ for some $\delta \geq 0$. For any $\lambda \geq 0$, there exists a direction in which $g(\lambda)$ increases, i.e.,

$$\frac{\partial g}{\partial \lambda_i} \;=\; h_i - \bar{b}_i - \log\left(\frac{\lambda_i}{\lambda^T\mathbf{1}}\right) > \delta \;\; \exists\, i. \tag{19}$$

To see (19) by contradiction, we check that

$$\begin{aligned}
h_i - \bar{b}_i - \log\left(\frac{\lambda_i}{\lambda^T\mathbf{1}}\right) \;&\leq\; \delta \quad \text{for } i = 1, \ldots, m \\
\implies \mathrm{e}^{h_i - \bar{b}_i - \delta} \;&\leq\; \frac{\lambda_i}{\lambda^T\mathbf{1}} \quad \text{for } i = 1, \ldots, m \\
\implies \left(\sum_{i=1}^m \mathrm{e}^{h_i - \bar{b}_i}\right)\mathrm{e}^{-\delta} \;\leq\; \frac{\sum_{i=1}^m \lambda_i}{\lambda^T\mathbf{1}} \;&=\; \frac{\sum_{i=1}^m \lambda_i}{1 + \sum_{i=1}^m \lambda_i} \;<\; 1 \\
\implies \sum_{i=1}^m \mathrm{e}^{h_i - \bar{b}_i} \;&<\; \mathrm{e}^\delta
\end{aligned}$$

which contradicts our assumption that $\sum_{i=1}^m \mathrm{e}^{h_i - \bar{b}_i} = \mathrm{e}^\delta$. By contradiction, we have (19). Furthermore, (19) implies that $g(\lambda)$ is unbounded above, which contradicts the finiteness of $D^\star$. Therefore, we must have that $\sum_{i=1}^m \mathrm{e}^{h_i - \bar{b}_i} < 1$. □

Lemma 6 (Feasibility criterion)**.** *Let the differential entropy h_i of each distribution q^i be finite. Suppose that there exists a feasible solution to Problem* (U-ELBO) *such that its objective function is bounded from below. Then, Problem* (U-ELBO) *is feasible if and only if*

$$\sum_{i=1}^{m} \mathrm{e}^{h_i - \bar{b}_i} \;<\; 1.$$

Proof. ($\Rightarrow$) Since the primal problem is feasible, the optimal objective function $F^\star$ is bounded from below and it is attained at a feasible point. For the sake of contradiction, we assume $\sum_{i=1}^{m} \mathrm{e}^{h_i - \bar{b}_i} \geq 1$, which is equivalent to $D^\star = \infty$ according to Lemma 5. However, this violates weak duality, i.e., $D^\star \leq F^\star$. By contradiction, we must have $\sum_{i=1}^{m} \mathrm{e}^{h_i - \bar{b}_i} < 1$.

($\Leftarrow$) We consider a set $\mathcal{A}$,

$$\mathcal{A} \;:=\; \left\{(u_1, \ldots, u_m, t) \,|\, -E(p, q^i) - \bar{b}_i \leq u_i \text{ for } i = 1, \ldots, m \;\text{ and }\; -E(p, q^0) \leq t \text{ for } p \in \mathcal{P}\right\}.$$

The set $\mathcal{A}$ is convex since it is the intersection of $m+1$ epigraphs of convex functions. We also introduce another convex set $\mathcal{B}$,

$$\mathcal{B} \;:=\; \{(0, \ldots, 0, t) \,|\, t \in \mathbb{R}\}.$$

We utilise proof by contradiction. Assume that the primal problem is infeasible. Then there doesn't exist any $p \in \mathcal{P}$ such that $-E(p, q^i) - \bar{b}_i \leq 0$ for all $i = 1, \ldots, m$. Hence, $\mathcal{A}$ and $\mathcal{B}$ are two disjoint convex sets. From the separating hyperplane theorem, there exists a hyperplane that separates them, i.e., $\exists v \in \mathbb{R}^{m+1}$ and $c \in \mathbb{R}$,

$$x^T v \geq c \;\text{ for all } x \in \mathcal{A} \tag{20}$$

$$y^T v \leq c \;\text{ for all } y \in \mathcal{B}. \tag{21}$$

Let $v = [\,\lambda_1, \ldots \lambda_m, \gamma\,]^T$. Then, (21) reduces to

$$y^T v \;=\; \lambda^T u + \gamma t \;=\; \gamma t \;\leq\; c \;\text{ for all } (u,t) \in \mathcal{B} \;\;\Rightarrow\;\; \gamma \;=\; 0.$$

This is because $\gamma t \leq c$ for any $t \in \mathbb{R}$. Note that $\gamma = 0$ means the separating hyperplane is vertical, i.e., being parallel to the t-axis. Furthermore, from (20) we can write

$$x^T v \;=\; \lambda^T u + \gamma t \;=\; \lambda^T u \;\geq\; c \;\text{ for all } (u,t) \in \mathcal{A} \;\;\Rightarrow\;\; \lambda \;\geq\; 0.$$

The above is true because the set of values that each u_i can take in $\mathcal{A}$ is unbounded above. Thus, since $\lambda^T u \geq c$, necessarily every λ_i has to be non-negative. Now, we consider

$$\begin{aligned}
g(\lambda) &= \inf_{(u,t)\,\in\,\mathcal{A}} \; t + \lambda^T u \\
\Longrightarrow\quad g(\alpha\lambda) &= \inf_{(u,t)\,\in\,\mathcal{A}} \; t + \alpha\lambda^T u \;\text{ for } \alpha \in \mathbb{R}_+ \\
\Longrightarrow\quad \lim_{\alpha\,\to\,\infty} g(\alpha\lambda) &= \lim_{\alpha\,\to\,\infty} \inf_{(u,t)\,\in\,\mathcal{A}} \; t + \alpha\lambda^T u \\
&= \lim_{\alpha\,\to\,\infty} \; \alpha \left(\inf_{(u,t)\,\in\,\mathcal{A}} \lambda^T u \right) \\
&\geq \lim_{\alpha\,\to\,\infty} \alpha c \\
&= \infty
\end{aligned}$$

which shows that $D^* = \infty$. This contradicts $D^\star$ being finite ($D^\star$ is finite due to $\sum_{i=1}^{m} \mathrm{e}^{h_i - \bar{b}_i} < 1$ and Lemma 5). Because of the contradiction, the primal problem has to be feasible. □

B.5 Proof of Lemma 3

Proof. The proof is an application of the convergence theory of DDPM [47, Theorem 3]. We next check all assumptions of [47, Theorem 3]. It is easy to see that we can cast $q_{\text{mix}}^{(\lambda)}$ as a target distribution

 https://doi.org/10.52202/079017-0836

of a diffusion model. By the definition,

$$\begin{aligned}\bar{\mathcal{L}}(\theta,\lambda) &= D_{\mathrm{KL}}\left(q(x_{0:T}) \,\|\, p_\theta(x_{0:T})\right) + \sum_{i=1}^m \lambda_i \left(D_{\mathrm{KL}}\left(q^i(x_{0:T}) \,\|\, p_\theta(x_{0:T})\right) - b_i\right) \\ &= -E(p_\theta; q) + \mathbb{E}_{q(x_0)}[\log q(x_0)] + \sum_{i=1}^m \lambda_i \left(-E(p_\theta; q^i) - \bar{b}_i\right).\end{aligned}$$

Thus, the partial minimization of $\bar{\mathcal{L}}(\theta,\lambda)$ over θ is equivalent to a weighted EBLO minimization,

$$\underset{\theta\in\Theta}{\text{minimize}} \quad -E(p_\theta; q) - \sum_{i=1}^m \lambda_i E(p_\theta; q^i)$$

or, equivalently,

$$\underset{\theta\in\Theta}{\text{minimize}} \quad -E(p_\theta; q^{(\lambda)}_{\text{mix}}) \tag{22}$$

where we normalize the weighted ELBO objective by introducing a mixed data distribution $q^{(\lambda)}_{\text{mix}}$. We note that $\bar{p}^\star(\lambda)$ is also a minimizer of Problem (22).

On the other hand, using Problem (P-LOSS), we can rewrite Problem (22) as

$$\underset{\theta\in\Theta}{\text{minimize}} \quad \mathbb{E}_{q^{(\lambda)}_{\text{mix}}(x_0),\, t,\, x_t}\left[\, \|\widehat{s}_\theta(x_t,t) - \nabla \log q(x_t)\|^2 \,\right]$$

which is equivalent to the score matching objective in DDPM. Therefore, the score matching assumption in [47, Assumption 1] is satisfied with the error bound $\varepsilon^2_{\text{score}}$ from Assumption 3. Viewing Assumption 2, and using appropriate stepsize and variance, all assumptions in [47, Theorem 3] are satisfied. Therefore, application of [47, Theorem 3] completes the proof. □

B.6 Proof of Lemma 4

Lemma 7. *The* TV *distance between two mixture data distributions* $q^{(\lambda)}_{\text{mix}}$, $q^{(\lambda^\star)}_{\text{mix}}$ *is bounded by*

$$\text{TV}\left(q^{(\lambda)}_{\text{mix}}, q^{(\lambda^\star)}_{\text{mix}}\right) \le \|\lambda - \lambda^\star\|_1.$$

Proof. By the definition,

$$\begin{aligned}\text{TV}\left(q^{(\lambda)}_{\text{mix}}, q^{(\lambda^\star)}_{\text{mix}}\right) &= \frac{1}{2}\int_{x_0}\left|\frac{q+\sum_{i=1}^m \lambda^i q^i}{1+\lambda^\top 1} - \frac{q+\sum_{i=1}^m \lambda^{i,\star} q^i}{1+(\lambda^\star)^\top 1}\right| \\ &= \frac{1}{2}\int_{x_0}\left|\frac{(1+(\lambda^\star)^\top 1)(q+\sum_{i=1}^m \lambda^i q^i) - (1+\lambda^\top 1)(q+\sum_{i=1}^m \lambda^{i,\star} q^i)}{(1+\lambda^\top 1)(1+(\lambda^\star)^\top 1)}\right| \\ &= \frac{1}{2}\int_{x_0}\left|\frac{\sum_{i=1}^m \lambda^i q^i + (\lambda^\star)^\top 1 q - \sum_{i=1}^m \lambda^{i,\star} q^i - \lambda^\top 1 q}{(1+\lambda^\top 1)(1+(\lambda^\star)^\top 1)}\right| \\ &\le \frac{1}{2}\int_{x_0}\left|\sum_{i=1}^m (\lambda^i - \lambda^{i,\star}) q^i + (\lambda^\star - \lambda)^\top 1 q\right| \\ &\le \sum_{i=1}^m |\lambda^i - \lambda^{i,\star}| \\ &= \|\lambda - \lambda^\star\|_1\end{aligned}$$

where the first inequality is due to $(1+\lambda^\top 1)(1+(\lambda^\star)^\top 1) \ge 1$, and we use triangle inequality in the second inequality. □

We recall the Lagrangians for Problems (U-LOSS) and (P-LOSS),

$$\begin{aligned}\mathcal{L}_s(\widehat{s},\lambda) &= \mathbb{E}_{q(x_0),\, t,\, x_t}\left[\, \|\widehat{s}(x_t,t) - \nabla \log q(x_t)\|^2 \,\right] \\ &\quad + \sum_{i=1}^m \lambda_i \left(\mathbb{E}_{q^i(x_0),\, t,\, x_t}\left[\, \|\widehat{s}(x_t,t) - \nabla \log q(x_t)\|^2 \,\right] - \widetilde{b}^i\right)\end{aligned}$$

$$\bar{\mathcal{L}}_s(\widehat{s}_\theta, \lambda) \;=\; \mathcal{L}_s(\widehat{s}_\theta, \lambda)$$

and their associated dual functions,

$$g_s(\lambda) \;=\; \underset{\widehat{s}\,\in\,\mathcal{S}}{\text{minimize}}\; \mathcal{L}_s(\widehat{s}, \lambda) \;\text{ and }\; \bar{g}_s(\lambda) \;=\; \underset{\theta\,\in\,\Theta}{\text{minimize}}\; \bar{\mathcal{L}}_s(\widehat{s}_\theta, \lambda).$$

For brevity, we use shorthand $\mathbb{E}_q$ and $\mathbb{E}_{q^i}$ for $\mathbb{E}_{q(x_0),\,t,\,x_t}$ and $\mathbb{E}_{q^i(x_0),\,t,\,x_t}$, respectively.

Lemma 8 (Parametrization gap)**.** *Let Assumption 4 hold. Then,* $0 \le \bar{g}_s(\lambda) - g_s(\lambda) \le 4R(1+\|\lambda\|_1)\nu$ *for any* $\lambda \ge 0$.

Proof. Let the partial minimizer of $\mathcal{L}_s(\widehat{s}, \lambda)$ over $\widehat{s}$ be $\widehat{s}^\star(\lambda) := \operatorname{argmin}_{\widehat{s}} \mathcal{L}_s(\widehat{s}, \lambda)$ for any $\lambda \ge 0$. For any $\lambda \ge 0$, there exists $\widetilde{\theta} \in \Theta$ such that $\left\|\widehat{s}^\star(\lambda) - \widehat{s}_{\widetilde{\theta}}\right\|_{L_2} \le \nu$ for any $\lambda \ge 0$, according to Assumption 4. Thus,

$$\begin{aligned}
\bar{\mathcal{L}}_s(\widehat{s}_{\widetilde{\theta}}, \lambda) - \mathcal{L}_s(\widehat{s}^\star(\lambda), \lambda) \;&=\; \mathbb{E}_q\left[\left\|\widehat{s}_{\widetilde{\theta}}(x_t, t) - x_0\right\|^2\right] - \mathbb{E}_q\left[\left\|\widehat{s}^\star(\lambda)(x_t, t) - x_0\right\|^2\right] \\
&\quad + \sum_{i=1}^m \lambda_i \left(\mathbb{E}_{q^i}\left[\left\|\widehat{s}_{\widetilde{\theta}}(x_t, t) - x_0\right\|^2\right] - \mathbb{E}_{q^i}\left[\left\|\widehat{s}^\star(\lambda)(x_t, t) - x_0\right\|^2\right]\right) \\
&\le\; 4R\,\mathbb{E}_q\left[\left\|\widehat{s}_{\widetilde{\theta}}(x_t, t) - \widehat{s}^\star(\lambda)(x_t, t)\right\|\right] \\
&\quad + 4R\sum_{i=1}^m \lambda_i\,\mathbb{E}_{q^i}\left[\left\|\widehat{s}_{\widetilde{\theta}}(x_t, t) - \widehat{s}^\star(\lambda)(x_t, t)\right\|\right] \\
&\le\; 4R\nu \;+\; 4R\,\|\lambda\|_1\,\nu
\end{aligned}$$

where the first inequality is due to that the quadratic function is locally Lipschitz continuous with parameter $4R$, and the second inequality is because that there exists $\widetilde{\theta} \in \Theta$ such that $\left\|\widehat{x}^\star(\lambda) - \widehat{x}_{\widetilde{\theta}}\right\|_{L_2} \le \nu$ for any $\lambda \ge 0$, according to Assumption 4.

By the definition $\widehat{s}^\star_\theta(\lambda) \in \operatorname{argmin}_{\theta\,\in\,\Theta} \bar{\mathcal{L}}_s(\widehat{s}_\theta, \lambda)$,

$$\bar{\mathcal{L}}_s(\widehat{s}^\star_\theta(\lambda), \lambda) \;\le\; \bar{\mathcal{L}}_s(\widehat{s}_{\widetilde{\theta}}, \lambda).$$

Therefore,

$$0 \;\le\; \mathcal{L}_s(\widehat{s}^\star_\theta(\lambda), \lambda) - \mathcal{L}_s(\widehat{s}^\star(\lambda), \lambda) \;\le\; \bar{\mathcal{L}}_s(\widehat{s}_{\widetilde{\theta}}, \lambda) - \mathcal{L}_s(\widehat{s}^\star(\lambda), \lambda) \;\le\; 4R(1 + \|\lambda\|_1)\nu$$

which gives our desired result by the definition of dual functions. □

Lemma 9 (Differentiability)**.** *The dual function* $g_s(\lambda)$ *is differentiable with gradient* $\nabla_\lambda \mathcal{L}_s(\widehat{s}^\star(\lambda), \lambda)$.

Proof. For any $\lambda \ge 0$, the Lagrangian $\mathcal{L}_s(\widehat{s}, \lambda)$ is strongly convex in function $\widehat{s} \in \mathcal{S}$. Since $\mathcal{S}$ is convex and compact, any partial minimizer $\widehat{s}^\star(\lambda)$ is unique. By Danskin's theorem [4], $g_s(\lambda)$ is differentiable and its gradient is the gradient of $\mathcal{L}_s(\widehat{s}, \lambda)$ over λ at $\widehat{s} = \widehat{s}^\star(\lambda)$. □

Lemma 10 (Convexity)**.** *The dual function* $g_s(\lambda)$ *is* μ*-strongly concave in* $\lambda \in \mathcal{H}$*, where*

$$\mu \;=\; \left(\frac{\sigma}{1 + \max\left(\|\lambda^\star\|_1\,,\,\left\|\bar{\lambda}^\star\right\|_1\right)}\right)^2.$$

Proof. For any $\lambda_1,\ \lambda_2 \in \mathcal{H}$, we denote $\widehat{s}^\star_1 := \widehat{s}^\star(\lambda_1)$ and $\widehat{s}^\star_2 := \widehat{s}^\star(\lambda_2)$, which are unique partial minimizers of the Lagrangians $\mathcal{L}_s(\widehat{s}, \lambda_1)$ and $\mathcal{L}_s(\widehat{s}, \lambda_2)$. Denote $\ell_0(\widehat{s}) := \mathbb{E}_q\left[\ \|\widehat{s}(x_t, t) - \nabla \log q(x_t)\|^2\ \right]$, $\ell_i(\widehat{s}) := \mathbb{E}_{q^i}\left[\ \|\widehat{s}(x_t, t) - \nabla \log q(x_t)\|^2\ \right]$ for $i = 1, \ldots, m$, and $\ell(\widehat{s}) := \left[\,\ell_1(\widehat{s}), \ldots, \ell_m(\widehat{s})\,\right]^\top$. By the convexity of ℓ_i,

$$\begin{aligned}
\ell_i(\widehat{s}^\star_1) \;&\ge\; \ell_i(\widehat{s}^\star_2) \;+\; 2\left\langle \nabla_{\widehat{s}}\ell_i(\widehat{s}^\star_2), \widehat{s}^\star_1 - \widehat{s}^\star_2\right\rangle \\
\ell_i(\widehat{s}^\star_2) \;&\ge\; \ell_i(\widehat{s}^\star_1) \;+\; 2\left\langle \nabla_{\widehat{s}}\ell_i(\widehat{s}^\star_1), \widehat{s}^\star_2 - \widehat{s}^\star_1\right\rangle.
\end{aligned}$$

If we multiply the first inequality above by $\lambda_{2,i} \ge 0$ and the second inequality above by $\lambda_{1,i} \ge 0$, and add them up from both sides, then,

$$-\left\langle \ell(\widehat{s}_2) - \ell(\widehat{s}_1), \lambda_2 - \lambda_1\right\rangle \;\ge\; 2\left\langle \lambda_1^\top \nabla_{\widehat{s}}\ell(\widehat{s}^\star_1) - \lambda_2^\top \nabla_{\widehat{s}}\ell(\widehat{s}^\star_2), \widehat{s}^\star_2 - \widehat{s}^\star_1\right\rangle.$$

https://doi.org/10.52202/079017-0836

We apply Lemma 9 to the LHS of the inequality above,

$$-\left(\nabla g_s(\lambda_2)-\nabla g_s(\lambda_1)\right)^\top(\lambda_2-\lambda_1)\;\geq\;2\left\langle\lambda_1^\top\nabla_{\widehat{s}}\ell(\widehat{s}_1^\star)-\lambda_2^\top\nabla_{\widehat{s}}\ell(\widehat{s}_2^\star),\widehat{s}_2^\star-\widehat{s}_1^\star\right\rangle. \tag{23}$$

On the other hand, by the optimality of $\widehat{s}_1^\star$ and $\widehat{s}_2^\star$,

$$\nabla_{\widehat{s}}\ell_0(\widehat{s}_1^\star)\;+\;\lambda_1^\top\nabla_{\widehat{s}}\ell(\widehat{s}_1^\star)\;=\;0 \tag{24a}$$

$$\nabla_{\widehat{s}}\ell_0(\widehat{s}_2^\star)\;+\;\lambda_2^\top\nabla_{\widehat{s}}\ell(\widehat{s}_2^\star)\;=\;0 \tag{24b}$$

which allows us to simplify the right-hand side of (23) and obtain

$$\begin{array}{rcl}-\left(\nabla g_s(\lambda_2)-\nabla g_s(\lambda_1)\right)^\top(\lambda_2-\lambda_1) & \geq & 2\left\langle\nabla_{\widehat{s}}\ell_0(\widehat{s}_2^\star)-\nabla_{\widehat{s}}\ell_0(\widehat{s}_1^\star),\widehat{s}_2^\star-\widehat{s}_1^\star\right\rangle\\ & \geq & 2\left\|\widehat{s}_1^\star-\widehat{s}_2^\star\right\|_{L_2}^2\end{array} \tag{25}$$

where the last inequality results from the strong convexity of quadratic functionals.

By the smoothness of quadratic functionals with parameter 1,

$$\begin{array}{rcl}\left\|\widehat{s}_1^\star-\widehat{s}_2^\star\right\|_{L_2} & \geq & \left\|\nabla_{\widehat{s}}\ell_0(\widehat{s}_1^\star)-\nabla_{\widehat{s}}\ell_0(\widehat{s}_2^\star)\right\|_{L_2}\\ & = & \left\|\lambda_1^\top\nabla_{\widehat{s}}\ell(\widehat{s}_1^\star)-\lambda_2^\top\nabla_{\widehat{s}}\ell(\widehat{s}_2^\star)\right\|_{L_2}\\ & = & \left\|(\lambda_2-\lambda_1)^\top\nabla_{\widehat{s}}\ell(\widehat{s}_2^\star)-\lambda_1^\top(\nabla_{\widehat{s}}\ell(\widehat{s}_1^\star)-\nabla_{\widehat{s}}\ell(\widehat{s}_2^\star))\right\|_{L_2}\\ & \geq & \left\|(\lambda_2-\lambda_1)^\top\nabla_{\widehat{s}}\ell(\widehat{s}_2^\star)\right\|_{L_2}-\left\|\lambda_1^\top(\nabla_{\widehat{s}}\ell(\widehat{s}_1^\star)\;-\;\nabla_{\widehat{s}}\ell(\widehat{s}_2^\star))\right\|_{L_2}\end{array}$$

where the equality is due to the optimality condition (24) and the last inequality is due to triangle inequality. By Assumption 5,

$$\left\|(\lambda_2-\lambda_1)^\top\nabla_{\widehat{s}}\ell(\widehat{s}_2^\star)\right\|_{L_2}\;\geq\;\sigma\left\|\lambda_2-\lambda_1\right\|.$$

We also notice that

$$\begin{array}{rcl}\left\|\lambda_1^\top(\nabla_{\widehat{s}}\ell(\widehat{s}_1^\star)-\nabla_{\widehat{s}}\ell(\widehat{s}_2^\star))\right\|_{L_2} & \leq & \displaystyle\sum_{i=1}^m\lambda_{1,i}\left\|\nabla_{\widehat{s}}\ell_i(\widehat{s}_1^\star)-\nabla_{\widehat{s}}\ell_i(\widehat{s}_2^\star)\right\|_{L_2}\\ & \leq & \displaystyle\sum_{i=1}^m\lambda_{1,i}\left\|\widehat{s}_1^\star-\widehat{s}_2^\star\right\|_{L_2}\end{array}$$

where the first inequality is due to triangle inequality and the second inequality is due to the smoothness of quadratic functionals. Hence,

$$\left\|\widehat{s}_1^\star-\widehat{s}_2^\star\right\|_{L_2}\;\geq\;\sigma\left\|\lambda_2-\lambda_1\right\|\;-\;\left\|\lambda_1\right\|_1\left\|\widehat{s}_1^\star-\widehat{s}_2^\star\right\|_{L_2}$$

or, equivalently,

$$\left\|\widehat{s}_1^\star-\widehat{s}_2^\star\right\|_{L_2}\;\geq\;\frac{\sigma}{1+\left\|\lambda_1\right\|_1}\left\|\lambda_2-\lambda_1\right\|$$

Therefore, (25) becomes

$$-\left(\nabla g_s(\lambda_2)-\nabla g_s(\lambda_1)\right)^\top(\lambda_2-\lambda_1)\;\geq\;\left(\frac{\sigma}{1+\left\|\lambda_1\right\|_1}\right)^2\left\|\lambda_2-\lambda_1\right\|^2$$

which completes the proof by choosing the smallest modulus over $\lambda_1\in\mathcal{H}$. □

Proof. By Lemmas 9 and 10, for any $\lambda\in\mathcal{H}$,

$$g_s(\lambda)\;\leq\;g_s(\lambda^\star)\;+\;\nabla g_s(\lambda^\star)^\top(\lambda-\lambda^\star)\;-\;\frac{\mu}{2}\left\|\lambda-\lambda^\star\right\|^2.$$

Thus, if we choose $\lambda=\bar{\lambda}^\star$, then

$$g_s(\bar{\lambda}^\star)\;\leq\;g_s(\lambda^\star)+\sum_{i=1}^m(\bar{\lambda}_i^\star-\lambda_i^\star)\left(\mathbb{E}_{q^i}\left[\left\|\widehat{s}^\star(\lambda^\star)(x_t,t)-\nabla\log q(x_t)\right\|^2\right]-\widetilde{b}^i\right)-\frac{\mu}{2}\left\|\bar{\lambda}^\star-\lambda^\star\right\|^2.$$

Optimality of $(\widehat{s}^\star(\lambda^\star), \lambda^\star)$ leads to the complementary slackness,

$$\sum_{i=1}^{m} \lambda_i^\star \left(\mathbb{E}_{q^i} \left[\left\| \widehat{s}^\star(\lambda^\star)(x_t, t) - \nabla \log q(x_t) \right\|^2 \right] - \widetilde{b}^i \right) \;=\; 0$$

and the feasibility,

$$\mathbb{E}_{q^i} \left[\left\| \widehat{s}^\star(\lambda^\star)(x_t, t) - \nabla \log q(x_t) \right\|^2 \right] \;\leq\; \widetilde{b}^i.$$

Therefore,

$$g_s(\bar{\lambda}^\star) \;\leq\; g_s(\lambda^\star) \;-\; \frac{\mu}{2} \left\| \bar{\lambda}^\star - \lambda^\star \right\|^2.$$

According to Lemma 8, $\bar{g}_s(\bar{\lambda}^\star) - 4R\left(1 + \left\|\bar{\lambda}^\star\right\|_1\right)\nu \leq g_s(\bar{\lambda}^\star)$. Hence,

$$\bar{g}_s(\bar{\lambda}^\star) \;-\; 4R\left(1 + \left\|\bar{\lambda}^\star\right\|_1\right)\nu \;\leq\; g_s(\lambda^\star) \;-\; \frac{\mu}{2} \left\| \bar{\lambda}^\star - \lambda^\star \right\|^2.$$

Thus,

$$\begin{aligned} \left\| \bar{\lambda}^\star - \lambda^\star \right\|^2 \;&\leq\; \frac{2}{\mu} \left(g_s(\lambda^\star) - \bar{g}_s(\bar{\lambda}^\star) \right) \;+\; \frac{8}{\mu} R \left(1 + \left\|\bar{\lambda}^\star\right\|_1\right) \nu \\ &\leq\; \frac{8}{\mu} R \left(1 + \left\|\bar{\lambda}^\star\right\|_1\right) \nu \end{aligned}$$

where the last inequality is due to that $g_s(\lambda) \leq \bar{g}_s(\lambda)$ for any $\lambda \geq 0$, and the optimality of $\bar{\lambda}^\star$,

$$g_s(\lambda^\star) \;\leq\; \bar{g}_s(\lambda^\star) \;\leq\; \bar{g}_s(\bar{\lambda}^\star).$$

□

B.7 Proof of Theorem 3

Proof. By the triangle inequality for TV distance,

$$\begin{aligned} \mathrm{TV}\left(q^\star_{\mathrm{mix}},\, \bar{p}^\star(\bar{\lambda}^\star)\right) \;&\leq\; \mathrm{TV}\left(q^\star_{\mathrm{mix}},\, q^{(\bar{\lambda}^\star)}_{\mathrm{mix}}\right) \;+\; \mathrm{TV}\left(q^{(\bar{\lambda}^\star)}_{\mathrm{mix}},\, \bar{p}^\star(\bar{\lambda}^\star)\right) \\ &\leq\; \left\|\lambda^\star - \bar{\lambda}^\star\right\|_1 \;+\; \frac{d^2 \log^3 T}{\sqrt{T}} \;+\; \sqrt{d}\left(\log^2 T\right) \varepsilon_{\mathrm{score}} \\ &\leq\; \sqrt{\frac{8}{\mu} m R \left(1 + \left\|\bar{\lambda}^\star\right\|_1\right) \nu} \;+\; \frac{d^2 \log^3 T}{\sqrt{T}} \;+\; \sqrt{d}\left(\log^2 T\right) \varepsilon_{\mathrm{score}} \end{aligned}$$

where the second inequality is due to Lemma 7 and Lemma 3, and the last inequality is due to $\|\lambda\|_1 \leq \sqrt{m}\,\|\lambda\|$ and Lemma 4. □

B.8 Proof of Theorem 4

Lemma 11. *For a stochastic variant of Algorithm 1 in Section 4.3, we have*

$$\mathbb{E}\left[\left\| \lambda(h+1) - \bar{\lambda}^\star \right\|^2 \mid \lambda(h) \right] \;\leq\; \left\| \lambda(h) - \bar{\lambda}^\star \right\|^2 \;+\; \eta^2 S^2 \;-\; 2\eta \left(\bar{D}^\star - \bar{g}(\lambda(h)) - \varepsilon^2_{\mathrm{approx}} \right).$$

Proof. For brevity, we let the stochastic gradient be $\widehat{f}(h) = [\widehat{f}_1(h), \ldots, \widehat{f}_m(h)]$ with

$$\widehat{f}_i(h) \;:=\; \widehat{\mathbb{E}}_{x_0 \sim q^i,\, t,\, x_t} \left[\left\| \widehat{s}_\theta(h)(x_t, t) - \nabla q(x_t) \right\|^2 \right] \;-\; \widetilde{b}^i.$$

By the definition of $\lambda(h+1)$,

$$\begin{aligned} \left\| \lambda(h+1) - \bar{\lambda}^\star \right\|^2 \;&=\; \left\| \left[\lambda(h) + \eta \widehat{f}(h) \right]_+ - \bar{\lambda}^\star \right\|^2 \\ &\leq\; \left\| \lambda(h) - \bar{\lambda}^\star + \eta \widehat{f}(h) \right\|^2 \\ &=\; \left\| \lambda(h) - \bar{\lambda}^\star \right\|^2 \;+\; \eta^2 \left\| \widehat{f}(h) \right\|^2 \;+\; 2\eta \widehat{f}(h)^\top \left(\lambda(h) - \bar{\lambda}^\star \right) \end{aligned}$$

https://doi.org/10.52202/079017-0836

where the inequality is due to the non-expansiveness of projection. Application of the conditional expectation over both sides of the inequality above yields,

$$\begin{aligned}\mathbb{E}\left[\left\|\lambda(h+1)-\bar{\lambda}^{\star}\right\|^{2} \mid \lambda(h)\right] \;\leq\; & \left\|\lambda(h)-\bar{\lambda}^{\star}\right\|^{2} + \eta^{2}\,\mathbb{E}\left[\left\|\widehat{f}(h)\right\|^{2} \mid \lambda(h)\right] \\ & + 2\eta\,\mathbb{E}\left[\widehat{f}(h) \mid \lambda(h)\right]^{\top}\left(\lambda(h)-\bar{\lambda}^{\star}\right)\end{aligned}$$

which gives our desired result when we use the fact that $\mathbb{E}[\widehat{f}(h) \mid \lambda(h)]$ is an approximate descent direction of the dual function $\bar{g}$,

$$\mathbb{E}\left[\widehat{f}(h) \mid \lambda(h)\right]^{\top}\left(\lambda(h)-\bar{\lambda}^{\star}\right) - \varepsilon_{\text{approx}}^{2} \;\leq\; \bar{g}(\lambda(h)) - \bar{g}(\bar{\lambda}^{\star}).$$

□

Lemma 12. *In the stochastic variant of Algorithm 1 in Section 4.3, the maximum prarametrized dual function in history up to step h satisfies*

$$\lim_{h\to\infty} \bar{g}_{\text{best}}(h) \;\geq\; \bar{D}^{\star} - \left(\frac{\eta S^{2}}{2} + \varepsilon_{\text{approx}}^{2}\right).$$

Proof. The proof is based on the supermartingale convergence theorem [67, Theorem E7.4]. We introduce two processes,

$$\alpha(h) \;:=\; \left\|\lambda(h)-\bar{\lambda}^{\star}\right\|^{2} \mathbb{1}\left(\bar{D}^{\star} - \bar{g}_{\text{best}}(h) > \frac{\eta S^{2}}{2} + \varepsilon_{\text{approx}}^{2}\right)$$

$$\beta(h) \;:=\; \left(2\eta\left(\bar{D}^{\star} - \bar{g}(\lambda(h)) - \varepsilon_{\text{approx}}^{2}\right) - \eta^{2}S^{2}\right) \mathbb{1}\left(\bar{D}^{\star} - \bar{g}_{\text{best}}(h) > \frac{\eta S^{2}}{2} + \varepsilon_{\text{approx}}^{2}\right)$$

where $\alpha(h)$ measures the gap between $\lambda(h)$ and $\bar{\lambda}^{\star}$ when the optimality gap $\bar{D}^{\star} - \bar{g}_{\text{best}}(h)$ is below the threshold, and $\beta(h)$ measures the gap bewteen $\bar{D}^{\star}$ and $\bar{g}(\lambda(h))$ (up to some optimization errors) when the optimality gap $\bar{D}^{\star} - \bar{g}_{\text{best}}(h)$ is below the threshold. Clearly, $\alpha(h)$ is non-negative. Also, $\beta(h)$ is non-negative due to

$$\bar{D}^{\star} - \bar{g}_{\text{best}}(h) - \frac{\eta S^{2}}{2} - \varepsilon_{\text{approx}}^{2} \;\leq\; \bar{D}^{\star} - \bar{g}(\lambda(h)) - \frac{\eta S^{2}}{2} - \varepsilon_{\text{approx}}^{2}.$$

Let $\mathcal{F}_h$ be the σ-algebra generated by sequences: $\alpha(h')$, $\beta(h')$, and $\lambda(h')$ for $h' \leq h$. Thus, $\{\mathcal{F}_h\}_{h\geq 1}$ is a natural filtration. We notice that $\alpha(h+1)$ and $\beta(h+1)$ are determined by $\lambda(h)$ in each step. Hence,

$$\begin{aligned}\mathbb{E}\left[\,\alpha(h+1) \,|\, \mathcal{F}_h\,\right] \;&=\; \mathbb{E}\left[\,\alpha(h+1) \,|\, \lambda(h)\,\right] \\ &=\; \mathbb{E}\left[\,\alpha(h+1) \,|\, \lambda(h), \alpha(h)=0\,\right] \Pr\left(\alpha(h)=0\right) \\ &\quad + \mathbb{E}\left[\,\alpha(h+1) \,|\, \lambda(h), \alpha(h)>0\,\right] \Pr\left(\alpha(h)>0\right).\end{aligned}$$

We first show that

$$\mathbb{E}\left[\,\alpha(h+1) \,|\, \mathcal{F}_h\,\right] \;\leq\; \alpha(h) - \beta(h). \tag{26}$$

A simple case is when $\alpha(h)=0$,

$$\mathbb{E}\left[\alpha(h+1) \,|\, \mathcal{F}_h\,\right] \;=\; \mathbb{E}\left[\,\alpha(h+1) \,|\, \lambda(h), \alpha(h)=0\,\right].$$

There are two situations for $\alpha(h)=0$. First, if $\bar{D}^{\star} - \bar{g}_{\text{best}}(h) \leq \frac{\eta S^2}{2} + \varepsilon_{\text{approx}}^2$, then $\alpha(h)=\beta(h)=0$. Due to $\bar{g}_{\text{best}}(h+1) \geq \bar{g}_{\text{best}}(h)$, we have $\beta(h)=0$ and $\bar{D}^{\star} - \bar{g}_{\text{best}}(h+1) \leq \frac{\eta S^2}{2} + \varepsilon_{\text{approx}}^2$. Thus, $\alpha(h+1)=0$, and (26) holds. Second, if $\lambda(h)=\bar{\lambda}^{\star}$, but $\bar{D}^{\star} - \bar{g}_{\text{best}}(h) > \frac{\eta S^2}{2} + \varepsilon_{\text{approx}}^2$, then $\bar{D}^{\star} = \bar{g}(\lambda(h))$. Hence, $\beta(h) < 0$, which contradicts the non-negativeness of $\beta(h)$. Therefore, $\bar{D}^{\star} - \bar{g}_{\text{best}}(h) \leq \frac{\eta S^2}{2} + \varepsilon_{\text{approx}}^2$ has to hold, which is the first situation.

We next show (26) when $\alpha(h) > 0$.

$$\begin{aligned}
\mathbb{E}\left[\,\alpha(h+1)\,|\,\mathcal{F}_h\,\right] &= \mathbb{E}\left[\,\alpha(h+1)\,|\,\lambda(h),\alpha(h)>0\,\right] \\
&= \mathbb{E}\left[\left\|\lambda(h)-\bar{\lambda}^\star\right\|^2 \mathbb{1}\left(\bar{D}^\star - \bar{g}_{\text{best}}(h) > \frac{\eta S^2}{2} + \varepsilon_{\text{approx}}^2\right) \,\middle|\, \lambda(h),\alpha(h)>0\right] \\
&\le \mathbb{E}\left[\left\|\lambda(h)-\bar{\lambda}^\star\right\|^2 \,\middle|\, \lambda(h),\alpha(h)>0\right] \\
&\le \left\|\lambda(h)-\bar{\lambda}^\star\right\|^2 + \eta^2 S^2 - 2\eta\left(\bar{D}^\star - \bar{g}(\lambda(h)) - \varepsilon_{\text{approx}}^2\right) \\
&\le \alpha(h) - \beta(h)
\end{aligned}$$

where the last inequality is due to Lemma 11, and the last equality is due to $\bar{D}^\star - \bar{g}_{\text{best}}(h) > \frac{\eta S^2}{2} + \varepsilon_{\text{approx}}^2$. Therefore, (26) holds.

Finally, application of the supermartingale convergence theorem [67, Theorem E7.4] to the processes $\alpha(h)$ and $\beta(h)$ for $h \ge 1$ concludes that $\beta(t)$ is almost surely summable,

$$\liminf_{h\to\infty}\ \beta(h) \;=\; 0 \ \text{ almost surely.}$$

This means that either

$$\liminf_{t\to\infty}\ 2\eta\left(\bar{D}^\star - \bar{g}(\lambda(h)) - \varepsilon_{\text{approx}}^2\right) - \eta^2 S^2 \;=\; 0$$

or $\bar{D}^\star - \bar{g}_{\text{best}}(h) \le \frac{\eta S^2}{2} + \varepsilon_{\text{approx}}^2$, which concludes our desired result. □

Denote the step when $\bar{g}_{\text{best}}(h)$ achieves $\bar{D}^\star - \left(\frac{\eta S^2}{2} + \varepsilon_{\text{approx}}^2\right)$ by h_{best}, and the associated dual variable be $\bar{\lambda}_{\text{best}} = \lambda(h_{\text{best}})$.

Lemma 13. *For a stochastic variant of Algorithm 1 in Section 4.3, we have*

$$\left\|\bar{\lambda}_{\text{best}} - \lambda^\star\right\|^2 \;\le\; \frac{2}{\mu}\left(\frac{\eta S^2}{2} + \varepsilon_{\text{approx}}^2 + 4R(1+\left\|\bar{\lambda}_{\text{best}}\right\|_1)\nu\right).$$

Proof. We denote the segment between $\bar{\lambda}_{\text{best}}$ and $\lambda^\star$ by $\mathcal{B}$. By Lemma 10, the dual function $g(\lambda)$ is strongly concave on $\mathcal{B}$ with parameter μ. Thus,

$$\begin{aligned}
\left\|\bar{\lambda}_{\text{best}} - \lambda^\star\right\|^2 &\le \frac{2}{\mu}\left(g(\lambda^\star) - g(\bar{\lambda}_{\text{best}})\right) \\
&\le \frac{2}{\mu}\left(\bar{g}(\lambda^\star) - \bar{g}(\bar{\lambda}_{\text{best}}) + 4R(1+\left\|\bar{\lambda}_{\text{best}}\right\|_1)\nu\right) \\
&\le \frac{2}{\mu}\left(\frac{\eta S^2}{2} + \varepsilon_{\text{approx}}^2 + 4R(1+\left\|\bar{\lambda}_{\text{best}}\right\|_1)\nu\right)
\end{aligned}$$

where the second inequality is due to Lemma 8. We note that $\bar{g}(\bar{\lambda}_{\text{best}}) = \bar{g}(\lambda(h_{\text{best}}))$ and $\bar{D}^\star \ge \bar{g}(\bar{\lambda}_{\text{best}})$, and the third inequality is due to Lemma 12. □

Proof. By the triangle inequality for TV distance,

$$\begin{aligned}
\text{TV}\left(q_{\text{mix}}^\star,\ \bar{p}^\star(\bar{\lambda}_{\text{best}})\right) &\le \text{TV}\left(q_{\text{mix}}^\star,\ q_{\text{mix}}^{(\bar{\lambda}_{\text{best}})}\right) + \text{TV}\left(q_{\text{mix}}^{(\bar{\lambda}_{\text{best}})},\ \bar{p}^\star(\bar{\lambda}_{\text{best}})\right) \\
&\le \left\|\lambda^\star - \bar{\lambda}_{\text{best}}\right\|_1 + \frac{d^2\log^3 T}{\sqrt{T}} + \sqrt{d}\left(\log^2 T\right)\varepsilon_{\text{score}} \\
&\le \frac{2}{\mu}\left(\frac{\eta S^2}{2} + \varepsilon_{\text{approx}}^2 + 4R\left(1+\left\|\bar{\lambda}_{\text{best}}\right\|_1\right)\nu\right) + \frac{d^2\log^3 T}{\sqrt{T}} + \sqrt{d}\left(\log^2 T\right)\varepsilon_{\text{score}}
\end{aligned}$$

where the second inequality is due to Lemma 7 and Lemma 3, and the last inequality is due to $\|\lambda\|_1 \le \sqrt{m}\,\|\lambda\|$ and Lemma 13. □

https://doi.org/10.52202/079017-0836

C Experimental details

We provide implementation details of our computational experiments in Section 5. The source code is available here.[2]

Algorithm details: We train our constrained diffusion models by replacing the exact primal minimization step in Algorithm 1 with N steps of gradient descent with the Lagrangian as a loss function. Without loss of generality, we take the noise prediction formulation of diffusion rather than the score-matching formulation used in our theory. Since these two formulations are equivalent, this has no bearing on our main results. Algorithm 2 depicts our practical implementation of Algorithm 1 .

Algorithm 2 Practical Implementation of Algorithm 1

Input: total diffusion steps T, diffusion parameter α_t, total dual iterations H, number of primal descent steps per dual update N, dual step size η_d, primal step size η_p, initial model parameters $\theta(0)$.
Initialize: $\lambda(1) = 0$.
for $h = 1, \cdots, H$ **do**
  for $n = 1, \cdots, N$ **do**
    $\theta_1 \;=\; \theta(h-1)$.
    $$\theta_{n+1} \;=\; \theta_n - \eta_p \nabla_\theta \left(\widehat{\mathbb{E}}_{x_0 \sim q,\, t,\, x_t} \left[\left\| \widehat{\epsilon}_{\theta_n}(x_t, t) - \epsilon_0 \right\|^2 \right] + \sum_{i=1}^{m} \lambda_i \, \widehat{\mathbb{E}}_{x_0 \sim q^i,\, t,\, x_t} \left[\left\| \widehat{\epsilon}_{\theta_n}(x_t, t) - \epsilon_0 \right\|^2 \right] \right).$$
    $\theta(h) \;=\; \theta_{N+1}$.
  end for
  Update the dual variable
  $$\lambda_i(h+1) \;=\; \left[\lambda_i(h) \;+\; \eta_d \left(\widehat{\mathbb{E}}_{x_0 \sim q^i,\, t,\, x_t} \left[\left\| \widehat{\epsilon}_{\theta(h)}(x_t, t) - \epsilon_0 \right\|^2 \right] \;-\; \widetilde{b}^i \right) \right]_+ \quad \text{for all } i.$$
end for

In Algorithm 2, the unbiased estimate of the noise prediction loss is evaluated via

$$\widehat{\mathbb{E}}_{x_0 \sim q,\, t,\, x_t} \left[\left\| \widehat{\epsilon}_\theta(x_t, t) - \epsilon_0 \right\|^2 \right] \;=\; \sum_{i=1}^{B} \left\| \widehat{\epsilon}_\theta(x_{t^{(i)}}, t^{(i)}) - \epsilon_0^{(i)} \right\|^2$$

where $\{x_0^{(i)}\}_{i=1}^B$ is a randomly chosen batch of samples from q, $\{t^{(i)}\}_{i=1}^B$ are time steps randomly sampled from the interval $[2, T]$, and $x_{t^{(i)}}$ is the noisy version of $x_0^{(i)}$ at time step $t^{(i)}$. Then, the noise ϵ_0 is sampled from the standard Gaussian, and x_t is derived from $x_t = \sqrt{\bar{\alpha}_t} x_t + \sqrt{1 - \bar{\alpha}_t} \epsilon_0$.

We remark an important implementation detail in the fine-tuning experiment. In the fine-tuning constraints, we have to evaluate the KL divergence

$$D_{\mathrm{KL}}(p_{\theta_{\mathrm{pre}}}(x_{0:T}) \,\|\, p_\theta(x_{0:T})) \;\leq\; b_i \tag{27}$$

where $p_{\theta_{\mathrm{pre}}}(x_{0:T})$ is the joint distribution of the samples and latents generated by the backward process of the pre-trained model. The KL divergence constraint in (27) further reduces to

$$D_{\mathrm{KL}}(p_{\theta_{\mathrm{pre}}}(x_{0:T}) \,\|\, p_\theta(x_{0:T})) \;=\; \mathbb{E}_{x_t \sim p_{\theta_{\mathrm{pre}}},\, t} \left[\left\| \widehat{\epsilon}_{\theta_{\mathrm{pre}}}(x_t, t) - \widehat{\epsilon}_\theta(x_t, t) \right\|^2 \right] + \text{constant}. \tag{28}$$

To estimate the expectation in (28), we need to sample latents x_t from the backward distribution $p_{\theta_{\mathrm{pre}}}(x_{0:T})$. In practice, this is computationally inefficient, since it requires running inference each time one wants to sample a latent. This is why we implement this with sampling x_t as random Gaussian noise instead. This still ensures that the predictions of the new model p_θ don't differ too much from the pre-trained distribution $p_{\theta_{\mathrm{pre}}}$ while making sampling batches much faster.

Resilient constrained learning. The choice of the constraint thresholds $\{b_i\}_{i=1}^m$ has noticeable effect on the training of constrained diffusion models. To avoid an exhaustive hyperparameter tuning process, in the minority class experiment, we use the resilient constrained learning technique [36] to

[2] https://github.com/shervinkhal/Constrained_Diffusion_Dual_Training

Table 1: Parameters of U-net Model used as noise predictor

# Res-Net layers per U-Net block	2
# Res-Net down/upsampling blocks	6
# Output channels for U-Net blocks	(128, 128, 256, 256, 512, 512)

adjust the thresholds $\{b_i\}_{i=1}^m$ during training. In essence, the resilient constrained learning adds a constraint relaxation cost to the loss and relaxes the thresholds by updating them through gradient descent each time we update the dual variable. It can further be shown theoretically that an equivalent formulation of resilience is achieved by adding a quadratic regularizer of the dual variable into the loss objective and setting the constraint thresholds to be zero, i.e., $\widetilde{b}_i = 0$. This is the approach we used in our experiments since it has fewer hyperparameters. We note that the only difference between Algorithm 2 and Algorithm 3 is the additional term in the dual variable update step (line 9 of Algorithm 3).

Algorithm 3 Resilient Constrained Diffusion Models via Dual Training

1: **Input**: total diffusion steps T, diffusion parameter α_t, total dual iterations H, number of primal descent steps per dual update N, dual step size η_d, primal step size η_p, constraint step size η_c, initial model parameters $\theta(0)$, constraint relaxation cost γ.
2: **Initialize**: $\lambda(1) = 0$.
3: **for** $h = 1, \cdots, H$ **do**
4: **for** $n = 1, \cdots, N$ **do**
5: $\theta_1 = \theta(h-1)$.
6: $$\theta_{n+1} = \theta_n - \eta_p \nabla_\theta \left(\widehat{\mathbb{E}}_{x_0 \sim q,\, t,\, x_t} \left[\|\widehat{\epsilon}_{\theta_n}(x_t, t) - \epsilon_0\|^2 \right] + \sum_{i=1}^m \lambda_i \widehat{\mathbb{E}}_{x_0 \sim q^i,\, t,\, x_t} \left[\|\widehat{\epsilon}_{\theta_n}(x_t, t) - \epsilon_0\|^2 \right] \right).$$
7: $\theta(h) = \theta_{N+1}$.
8: **end for**
9: Update the dual variable
$$\lambda_i(h+1) = \left[\lambda_i(h) + \eta_d \left(\widehat{\mathbb{E}}_{x_0 \sim q^i,\, t,\, x_t} \left[\left\|\widehat{\epsilon}_{\theta(h)}(x_t, t) - \epsilon_0\right\|^2 \right] - \widetilde{b}^i(h) - 2\gamma\lambda_i(h) \right) \right]_+ \text{ for all } i.$$
10: **end for**

Model architecture. We use a time-conditioned U-net model as is common in image diffusion tasks for all three datasets. The time conditioning is done by adding a positional embedding of the time to the input image. The parameters of the model are summarized in Table 1. The fifth downsampling block and the corresponding upsampling block are Res-Net blocks with spatial self-attention.

Hyperparameters. We summarize the important hyperparameters in our experiments in Table 2. In the unconstrained models that we train for comparison, we use the same hyperparameters as the constrained version, disregarding the parameters related to the dual and relaxation updates. For models trained on Image-Net, when training the constrained models, we initialized to the parameters of the unconstrained model to make training times shorter.

Hyperparameter sensitivity. We remark the sensitivity of the dual training algorithm to the number of dual iterations, primal/dual batch sizes, and primal/dual learning rates.

- **Number of dual iterations:** In our implementation this shows up as the number of primal GD steps per dual update, N. Experimentally, we have observed that as long as N is greater than 1, the results are not sensitive to this value. Additionally, the dual updates add a negligible computational overhead. Hence, updating the dual nearly as many times as we update model parameters doesn't reduce training efficiency.
- **Primal/dual batch sizes:** We have included results of training a constrained model on an unbalanced subset of MNIST, using different primal/dual batch sizes (See Table 3.) The results suggest that for the minority class experiments, when the ratio between Primal and Dual batch sizes is larger, the model performs better (lower FID and more evenly distributed

https://doi.org/10.52202/079017-0836

Table 2: Hyperparameter values used in the main experiments. MC denotes Minority Class experiments and FT denotes Fine-Tuning experiments. - denotes that resilience was not used for the experiment.

	MNIST MC	Celeb-A MC	Image-Net MC	MNIST FT
#training epochs ($N \times H$)	250	1000	2000	500
# primal steps per dual step (N)	2	2	2	2
Primal batch size	128	256	128	256
Dual batch size	64	128	64	256
Primal learning rate	0.0001	0.0001	5e-5	5e-6
Dual learning rate	0.1	0.1	0.05	1e5
Resilience Relaxation cost	0.09	0.005	0.025	-
main dataset size	31000	12500	2000	200
constraint dataset(s) size	5000	500	64	-

samples). This is in line with the heuristic we used in the included experiments in the paper where we chose the batch sizes such that the ratio of primal to dual batch size is close to size ratio of entire dataset to constraint datasets (which are much smaller). Howver for the fine-tuning task, the batch sizes did not seem to affect the final result as much.

- **Primal/dual learning rate:** For the primal learning rate, we followed the best practice used to train standard diffusion models. For the dual learning rate η , we refer to Theorem 8 in the paper, showing a smaller error bound for smaller η while slowing convergence. In practice, as long as $\eta \leq 1$, we observed that the model converges to similar results reliably.

Efficiency of constrained diffusion: We note that the complexity of sampling from our constrained diffusion model does not increase with the number of constraints, as our trained diffusion model functions like a standard diffusion model to generate samples. Importantly, we remark that training our constrained diffusion model has comparable efficiency to training standard diffusion models detailed next.

The additional computational cost of our dual-based training (Algorithm 1) arises from: (i) updating the dual variables; (ii) updating the diffusion model in the primal update.

- **Cost of updating the dual variables:** We note that our dual-based training has the same number of dual variables as the number of constraints. Thus, the cost for the dual update is linear in the number of constraints. To update each dual variable, we can directly use the ELBO loss over the batches sampled from each constrained dataset (already computed for the Lagrangian). Therefore, the cost of updating dual variables is negligible.
- **Cost of updating the diffusion model in the primal update:** We note that the primal update trains a standard diffusion model based on the Lagrangian with updated dual variables. In our experiments, this primal training often requires as few as 2-3 updates per dual update. Thus, when training our constrained model, we can train for the same number of epochs as an unconstrained model but update the dual variables after every few primal steps. As a result, training our constrained diffusion model is almost as efficient as training standard unconstrained models.

The only concern we encountered regarding efficiency is that batches need to be sampled from every constrained dataset at each step to estimate the Lagrangian. This introduces a small GPU memory overhead that increases with additional constraints. However, this is somewhat mitigated by the fact that constrained datasets are often much smaller than the original dataset, allowing us to choose a smaller batch size for the constrained datasets without degrading performance (see discussion on batch sizes in hyperparameter sensitivity section).

Table 3: Constrained model trained on MNIST with different Primal/Dual Batch sizes

(Primal batch size, Dual batch size)	(64, 16)	(64, 64)	(128, 16)	(128, 64)
FID score	**16.6**	20.6	16.7	20.24

FID scores. As a quantitative means of evaluating our constrained diffusion models, we use the FID (Frechet Inception Distance) as a metric to gauge the quality of the samples generated by diffusion models. The FID score was first introduced in [33] in form of

$$d^2_{\text{FID}}((m, C), (m_w, C_w)) \;=\; \|m - m_w\|_2^2 \;+\; \text{Tr}(C + C_w - 2(CC_w)^{1/2}) \tag{29}$$

where m and C represent mean and variance, respectively, of the distribution of the features of the data samples which have been extracted by an inception model [69]. Similarly, m_w and C_w represent the mean and variance of some reference distribution that we are computing the distance to. In our experiment, we compute the FID scores by generating 15000 samples from the diffusion model we are evaluating and comparing them to a balanced version of the original dataset. In the experiment with MNIST, this is the actual dataset. In the experiments with Celeb-A and Image-Net, since there is an imbalance in the original dataset, we consider a balanced subset of each with an equal number of samples from each class as reference. We use the clean-FID library [58] for standard computation of the FID scores.

We note that our FID scores are somewhat larger compared to typical baselines in the literature. This is expected as a consequence of our experimental setup. We train both the unconstrained and constrained models, on a biased subset of the dataset wherein some of the classes have significantly fewer samples than the rest. We then compute the FID scores for these models compared to the actual dataset itself which is unbiased (i.e., every class has the same number of samples). These FID scores approximate how close the learned distribution of the model trained on biased data, is to the underlying unbiased distribution.

This setup contrasts with existing results in the literature, where the FID is computed with respect to unbiased data, and the models are also trained on unbiased data. Therefore, it is expected that such models will achieve better FID scores than constrained or unconstrained models trained with biased data. Our purpose in reporting the FIDs was not to compare them to existing results (as such a comparison would be uninformative) but to demonstrate that, when trained on biased data, the constrained model achieves better FID scores than the unconstrained model.

Compute resources. We run all experiments on two NVIDIA RTX 3090-Ti GPUs in parallel. The amount of GPU memory used was 16 Gigabytes per GPU. For experiments with MNIST and Celeb-A datasets, training each model took between 2-3 hours. This increased to 7-8 hours for latent diffusion models trained for the Image-Net experiments.

Assets and libraries. We use the PyTorch [59] and Diffusers [73] Python libraries for training our constrained diffusion models, and Adam with decoupled weight decay [52] as an optimizer. The accelerate library [30] is used for the parallelization of the training processes across multiple GPUs. For classifiers used in evaluating the generated samples of the models, we use the following pretrained models accessible on the Huggingface model database: A Vision transformer-based classifier for MNIST digits with $\%99.5$ validation accuracy (https://huggingface.co/farleyknight-org-username/vit-base-mnist). A classifier for images of male/female faces with $\%98.6$ validation accuracy (https://huggingface.co/cledoux42/GenderNew_v002). For classifying the image-net data, a zero-shot classifier based on a CLIP model (https://huggingface.co/openai/clip-vit-base-patch32) was used. The Autoencoder for the latent diffusion model was the stable diffusion VAE with KL regularization found on (https://huggingface.co/stabilityai/sd-vae-ft-mse).

https://doi.org/10.52202/079017-0836

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We summarize our constrained diffusion models in Section 3, optimality guarantees of unparametrized constrained diffusion models in Section 3.1, optimality guarantees of parametrized constrained diffusion models and training algorithms in Section 4, and experimental results in Section 5.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We mention two potential limitations of this work as several future directions in Section 6.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We provide all assumptions, lemmas, and theorems in Sections 3 and 4, and provide proof details in Appendix B.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We summarize our experimental results in Section 5, and provide implementation details of experiments in Appendix C, together with additional experimental results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: A link to the code for replicating our main experiments has been provided in Appendix C. The datasets are open source machine learning datasets that are accessible online.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide experimental details in Appendix C including the hyperparameters used for each experiment. Our training details can be found in the code in supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

Justification: While the histogram plots that showcase our main results do not include error bars, we do provide Frechet Distance metrics (that are a measure of statistical distance between sample distributions) which emphasize our results.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We include compute details in Appendix C.

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and fully comply with it during the preparation of this paper.

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The potential impacts of the work are discussed in Section 1.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We release no models and the supplemental code for training image generation model, trains the model on MNIST and Celeb-A datasets neither of which have potential for misuse.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The libraries and assets used have been noted and credited in Appendix C.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: [NA]

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: [NA]

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

https://doi.org/10.52202/079017-0836

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: [NA]