2026 American Control Conference (ACC)
May 27-29, 2026. New Orleans, Louisiana, USA

Fine-tuning diffusion models via stochastic control: entropy regularization and beyond

Wenpin Tang and Fuzhong Zhou

Abstract— **This paper aims to develop and provide a rigorous treatment to entropy regularized fine-tuning in the context of continuous-time diffusion models, which was proposed by Uehara et al. (arXiv:2402.15194). The idea is to use stochastic control for sample generation, where the entropy regularizer is introduced to mitigate reward collapse. We also show how the analysis can be extended to fine-tuning with a general f-divergence regularizer. Numerical experiments on large-scale text-to-image models – Stable Diffusion v1.5 are conducted to validate our approach.**

I. Introduction

Diffusion models [13], [25], [26] are a promising generative approach to produce high-quality samples, which have been observed to outperform adversarial generative nets in image and audio synthesis [7], [16], and underpin recent success in text-to-image creators such as DALL·E2 [22] and Stable Diffusion [23], and the text-to-video generator Sora [19]. Though diffusion models can capture intricate and high-dimensional data distributions, they may suffer from sources of bias or fairness concerns [18], and the training process (especially for the large models) requires considerable time and effort.

There is growing interest in improving diffusion models in terms of generated sample quality, as well as controllability. One straightforward approach is to fine-tune the sampler customized for a specific task using the pretrained (diffusion) model as a base model. For instance, in image/video generation, we aim to fine-tune diffusion models to enhance the aesthetic quality and prevent distorted contents. With the emergence of human-interactive platforms such as ChatGPT, there is substantial demand to align generative models with user/human preference or feedback. Recent works [3], [9], [10], [39] proposed to fine-tune diffusion models by reinforcement learning (RL), and [36] by direct preference optimization. In these works, the reward functions are learned statistical models, e.g., an aesthetic reward in image generation is a ranking model fit to the true aesthetic preferences of human raters. Refer to [4], [33], [38] for reviews on preference tuning of diffusion models.

The above methods fine-tune the diffusion model to generate samples with high nominal rewards. However, they may lead to *reward collapse* or *hacking* [1], [27], a phenomenon referring to overfitting the reward (trained by e.g., a limited number of human ratings.) In other words, the diffusion model is fine-tuned with respect to some human-rated score which may fail to generalize. Moreover, exploiting the reward exclusively also harms *diversity*, a criterion that is at the heart of generative modeling.

In order to mitigate reward collapse/hacking and enhance diversity, [34] proposed to add an entropy regularizer with respect to the pretrained model in the loss objective. This yields the *entropy-regularized fine-tuning*, which is an exponential tilting of that generated by the pretrained model and can be viewed as a "soft" *diffusion guidance* [7], [14], [29]. A stochastic control approach was developed to emulate this novel distribution. In comparison with the standard control framework where the initial distribution is fixed, both the control and the initial distribution are decision variables. The resulting problem is to "decouple" these two variables by first solving a standard control problem, followed by finding the optimal initial distribution. To the best of our knowledge, this is the first time that stochastic control is used for sample generation. In the concurrent work [8], stochastic control is also used for fine-tuning diffusion models, but without learning the initial distribution (instead, it requires cautious model selection.) The idea of adding an entropy regularizer also appeared in [39] for fine-tuning the diffusion model via continuous RL, and in [15] for constraint optimization via the diffusion

Department of Industrial Engineering and Operations Research, Columbia University. wt2319@columbia.edu, fz2329@columbia.edu

979-8-3315-9381-0//$31.00 ©2026 AACC

model. In a different context, (entropy regularized) stochastic control was proposed to solve non-convex problems [11], [30]. as an alternative to the simulated annealing algorithm.

The purpose of this paper is two-fold. First, we provide a rigorous treatment to the theory of entropy-regularized fine-tuning proposed in [34]. The paper is mostly self-contained. Second, we extend to the problem of fine-tuning regularized by general f-divergence, and show how the analysis carries over. The remainder of the paper is organized as follows. In Section II, we recall background of diffusion models. The entropy-regularized fine-tuning is studied in Section III, and the extension to fine-tuning regularized by f-divergence is given in Section IV. We conduct numerical experiments in Section V. Concluding remarks are summarized in Section VI. Due to space constraints, all the proofs are reported in `https://arxiv.org/abs/2403.06279`.

Notations: The notation $X \sim p(\cdot)$ means that the random variable is distributed according to $p(\cdot)$. We write $\mathbb{E}_p(X)$ for the expectation of $X \sim p(\cdot)$. For $p(\cdot)$ and $q(\cdot)$ two probability distributions, $D_{TV}(p(\cdot), q(\cdot)) := \sup_A |p(A) - q(A)|$ is the total variation distance between $p(\cdot)$ and $q(\cdot)$, and $D_{KL}(p(\cdot), q(\cdot)) := \int \log \frac{dp}{dq} dp$ is the KullbackLeilber (KL) divergence between $p(\cdot)$ and $q(\cdot)$.

II. Diffusion models

In this section, we review diffusion models that will be used as the pretrained models in fine-tuning. We follow the presentation in [32].

The goal of diffusion models is to generate new samples that resemble the target data, while maintain a certain level of diversity. Diffusion modeling relies on a forward-backward procedure: the forward process is a stochastic differential equation (SDE):

$$dX_t = b(t, X_t)dt + \sigma(t)dW_t, \quad X_0 \sim p_{\text{data}}(\cdot), \quad (1)$$

where $\{W_t\}$ is d-dimensional Brownian motion, and $b : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma : \mathbb{R}_+ \to \mathbb{R}_+$. Some conditions on $b(\cdot, \cdot)$, $\sigma(\cdot)$ are required so that (1) is well-defined (see [28]). We assume for simplicity that the distribution of X_t has a suitably smooth density, and denote by $p(t, x) := \mathbb{P}(X_t \in dx)/dx$.

The key to the success of continuous-time diffusion models is that their time reversal has a rather tractable form, making it more convenient for methodological development. It is known [2], [12] that the distribution of the time reversal process $\overline{X}_t = X_{T-t}$ is governed by the SDE: $d\overline{X}_t = (-b(T-t, \overline{X}_t) + \sigma^2(T-t)\nabla \log p(T-t, \overline{X}_t))dt + \sigma(T-t)dB_t$, $\overline{X}_0 \sim p(T, \cdot)$, where $(B_t,\, t \geq 0)$ is a copy of d-dimensional Brownian motion.

There are, however, two twists in diffusion modeling. First, diffusion models aim to generate the target distribution from *noise*, and the noise should not depend on the target distribution. So instead of setting $\overline{X}_0 \sim p(T, \cdot)$, the backward process is initiated with some noise $p_{\text{noise}}(\cdot)$ as a proxy of $p(T, \cdot)$. The diffusion model is specified by the pair $(b(\cdot, \cdot), \sigma(\cdot))$, and existing examples include Ornstein-Ulenback processes [6], variance exploding (VE), variance preserving (VP) SDEs [26], and contractive diffusion models [31]. The choice of $p_{\text{noise}}(\cdot)$ depends on each specific model, and varies case by case.

Second, all but the term $\nabla \log p(T - t, \overline{X}_t)$ for the time-reversal process $(\overline{X}_t, 0 \leq t \leq T)$ are available. So we need to compute $\nabla \log p(t, x)$, known as *score function*. *Score matching* techniques allow to estimate the score function via function approximations. More precisely, we use a family of functions $\{s_\theta(t, x)\}_\theta$ (e.g., neural nets) to approximate the score $\nabla \log p(t, x)$. The most widely used approach is *denoising score matching* [35]: $\min_\theta \mathbb{E}_{t\sim \text{Unif}\,[0,T],\, X_0 \sim p_{data}} \Big[\mathbb{E}_{p(t,\cdot|X_0)} \Big| s_\theta(t, X_t) - \nabla \log p(t, X_t|X_0) \Big|^2 \Big]$.

Now with the (true) score function $\nabla \log p(t, x)$ being replaced with the score matching function $s_\theta(t, x)$, the backward process is set to:

$$dY_t = \underbrace{\big(-b(T-t, Y_t) + \sigma^2(T-t)s_\theta(T-t, Y_t)\big)}_{\overline{b}(t, Y_t)} dt + \sigma(T-t)dB_t, \quad Y_0 \sim p_{\text{noise}}(\cdot). \quad (2)$$

Here we use the symbol Y to represent the process using the score matching $s_\theta(\cdot, \cdot)$, which is distinguished from the process $\overline{X}$ (with the true score $\nabla \log p(t, x)$). The SDE (2) provides the generic form of the diffusion sampling, which serves as a pretrained model for further goal-directed fine-tuning. Denote by $Q_{[0,T]}(\cdot)$ the probability distribution (on the path space) of the process $(Y_t,\, 0 \leq t \leq T)$, and $Q_t(\cdot)$ its marginal distribution at time t. Write $Y_T \sim Q_T(\cdot) =: p_{\text{pre}}(\cdot)$.

III. Entropy-regularized fine-tuning and stochastic control

In this section, we consider the problem of entropy-regularized fine-tuning that aims to mitigate the reward collapse in the context of diffusion models. We provide a rigorous treatment for the closely related stochastic control problem, following the idea of [34]. As we will see, the arguments used in this section can be generalized to study the problem of fine-tuning with an f-divergence regularizer.

A. Entropy-regularized fine-tuning

Let's explain entropy-regularized fine-tuning using a pretrained diffusion model (2). The idea of fine-tuning is to calibrate a (possibly powerful) pretained model to some downstream task with a specific goal, which is captured by a reward function $r : \mathbb{R}^d \to \mathbb{R}_+$. An obvious way is to maximize $\mathbb{E}_{\widetilde{Q}_{[0,T]}}[r(Y_T)]$ over a class of fine-tuned diffusion models $\widetilde{Q}_{[0,T]}(\cdot)$. As mentioned in the introduction, this approach may generate samples that overfit the reward. In order to avoid the reward collapse, we consider:

$$p_{\text{ftune}}(\cdot) := \arg\max_{p} \mathbb{E}_p[r(Y)] - \alpha D_{KL}(p(\cdot), p_{\text{pre}}(\cdot)), \tag{3}$$

where the maximization is over all probability distribution on $\mathbb{R}^d$, and the hyperparameter $\alpha > 0$ controls the level of exploration relative to the pretrained model. The problem (3) is referred to as entropy-regularized fine-tuning of the pretrained model (2).

The problem (3) has a closed-form solution:

Lemma 3.1: Assume that $\int \exp\left(\frac{r(y)}{\alpha}\right) p_{\text{pre}}(y)dy < \infty$. We have:

$$p_{\text{ftune}}(y) = \frac{1}{C} \exp\left(\frac{r(y)}{\alpha}\right) p_{\text{pre}}(y), \tag{4}$$

where C is the normalizing constant.

At a high level, the target data with distribution $p_{\text{data}}(\cdot)$ is used to build the pretrained model that generates $p_{\text{pre}}(\cdot) \approx p_{\text{data}}(\cdot)$, and the pretrained model is fine-tuned via some reward function to produce $p_{\text{ftune}}(\cdot)$. Note that this differs from the setting of rejection sampling in such a way that we can only sample $p_{\text{pre}}(\cdot)$ and evaluate (usually in blackbox) via $r(\cdot)$ without knowing their exact forms.

As we will see in the next subsection, the main idea is to apply a (stochastic) control to the diffusion sampler (2) in order to generate $p_{\text{ftune}}(\cdot)$ in just one pass. The next result shows how the fine-tuned distribution (4) differs from $p_{\text{data}}(\cdot)$.

Proposition 3.2: We have:

$$D_{TV}(p_{\text{ftune}}(\cdot), p_{\text{data}}(\cdot)) \le D_{TV}(p_{\text{pre}}(\cdot), p_{\text{data}}(\cdot)) + \frac{1}{2}\sqrt{\mathbb{E}_{p_{\text{pre}}}\left[e^{\frac{2r(Y)}{\alpha}}\right]}. \tag{5}$$

Let's make some remarks on the bound (5): the first term in (5) quantifies the performance of the pretrained model, while the second term specifies the difference between the distribution of samples generated from the pretrained model and the fine-tuned distribution. Assuming that $\sigma(\cdot)$ is bounded away from zero and that $\mathbb{E}_{p(t,\cdot)}|\nabla \log p(t, X) - s_\theta(t, X)|^2 \le \varepsilon^2$ for all t, [5], [32] shows that $D_{TV}(p_{\text{pre}}(\cdot), p_{\text{data}}(\cdot)) \le D_{TV}(p_{\text{noise}}(\cdot), p(T, \cdot)) + \frac{\varepsilon}{2}\sqrt{T}$. If the reward $r(x)$ is bounded, say by $\overline{r} > 0$, then the second term is bounded from above by $\frac{1}{2}\exp\left(\frac{\overline{r}}{\alpha}\right)$. Thus, the bound becomes $D_{TV}(p_{\text{noise}}(\cdot), p(T, \cdot)) + \frac{\varepsilon}{2}\sqrt{T} + \frac{1}{2}e^{\frac{\overline{r}}{\alpha}}$. When $\alpha \to 0$, the bound increases to $+\infty$, partly explaining reward collapse.

B. Stochastic control

Now the goal is to emulate the fine-tuned distribution $p_{\text{ftune}}(\cdot)$ specified by (4). We follow the idea of [34] to lift the problem to the process level via stochastic control:

$$dY_t = (\overline{b}(t, Y_t) + u_t)dt + \overline{\sigma}(t)dB_t, \quad Y_0 \sim \nu(\cdot), \tag{6}$$

where $\overline{b}(\cdot, \cdot)$ is defined in (2), $\overline{\sigma}(t) := \sigma(T - t)$, and $(u_t,\, t \ge 0)$ is adapted. Here both u and $\nu(\cdot)$ are decision variables, and $\nu(\cdot)$ is not (necessarily) the noise distribution $p_{\text{noise}}(\cdot)$ as in (2).

Denote by $Q^{u,\nu}_{[0,T]}(\cdot)$ the probability distribution of the process $(Y_t,\, 0 \le t \le T)$ defined by (6), and $Q^{u,\nu}_t(\cdot)$ its marginal distribution at time t. Recall that $Q_{[0,T]}(\cdot)$ is the distribution of the pretrained model (2), so $Q_{[0,T]}(\cdot) = Q^{0,p_{\text{noise}}}_{[0,T]}(\cdot)$.

It is natural to consider the following objective for the stochastic control problem:

$$(u^*, \nu^*) := \arg\max_{(u,\nu)\in\mathcal{A}} \mathbb{E}_{Q^{u,\nu}_{[0,T]}}[r(Y_T)] - \alpha D_{KL}(Q^{u,\nu}_{[0,T]}(\cdot), Q_{[0,T]}(\cdot)), \tag{7}$$

where $\mathcal{A}$ is an admissible set for the decision variables that we will specify. As we will see, $Y_T \sim p_{\text{ftune}}(\cdot)$ under $Q^{u^*,\nu^*}_{[0,T]}(\cdot)$.

The result below elucidates the form of the objective (7), which is a consequence of Girsanov's theorem.

Proposition 3.3: Let

$$\mathcal{A} = \Bigg\{(u,\nu) : \nu \text{ has a smooth density,} \exp\left(-\int_0^t \frac{u_s}{\overline{\sigma}(s)} dB_s - \frac{1}{2}\int_0^t \left|\frac{u_s}{\overline{\sigma}(s)}\right|^2 ds\right) \text{ and } \int_0^t \frac{u_s}{\overline{\sigma}(s)} dB_s \text{ are } Q^{u,\nu}_{[0,T]}\text{-martingales}\Bigg\}.$$

Then for each $(u,\nu) \in \mathcal{A}$,

$$\begin{aligned}&\mathbb{E}_{Q^{u,\nu}_{[0,T]}}[r(Y_T)] - \alpha D_{KL}(Q^{u,\nu}_{[0,T]}(\cdot), Q_{[0,T]}(\cdot)) \\ &= \mathbb{E}_{Q^{u,\nu}_{[0,T]}}\left[r(Y_T) - \frac{\alpha}{2}\int_0^T \left|\frac{u_t}{\overline{\sigma}(t)}\right|^2 dt - \alpha \log\left(\frac{\nu(Y_0)}{p_{\text{noise}}(Y_0)}\right)\right].\end{aligned} \tag{8}$$

The $Q^{u,\nu}_{[0,T]}$-martingales condition allows to simplify the expression of $D_{KL}(Q^{u,\nu}_{[0,T]}(\cdot), Q_{[0,T]}(\cdot))$, bypassing the local martingale trap. Sufficient conditions are known for these local martingales to be (true) martingales. For instance, Novikov's condition $\mathbb{E}_{Q^{u,\nu}_{[0,T]}}\left[\exp\left(\frac{1}{2}\int_0^T \left|\frac{u_t}{\overline{\sigma}(t)}\right|^2 dt\right)\right] < \infty$ ensures the use of Girsanov's theorem, and also the $Q^{u,\nu}_{[0,T]}$-martingale property of $\int_0^t \frac{u_s}{\overline{\sigma}(s)} dB_s$.

By Proposition 3.3, the stochastic control problem (7) is to maximaize over $(u,\nu) \in \mathcal{A}$:

$$\int \mathbb{E}_{Q^{u,y}_{[0,T]}}\left[r(Y_T) - \frac{\alpha}{2}\int_0^T \left|\frac{u_t}{\overline{\sigma}(t)}\right|^2 dt\right]\nu(y)dy - \alpha D_{KL}(\nu(\cdot), p_{\text{noise}}(\cdot)), \tag{9}$$

where $Q^{u,y}_{[0,T]}(\cdot)$ denotes the probability distribution of the process $(Y_t,\, 0 \le t \le T)$ in (6) conditioned on $Y_0 = y$. It is easy to see that the two decision variables (u,ν) are separable in the problem (9). It boils down to the following two steps:

(1) Solve the stochastic control problem:

$$v^*(0,y) := \max_{u\in\mathcal{A}} \mathbb{E}_{Q^{u,y}_{[0,T]}}\left[r(Y_T) - \frac{\alpha}{2}\int_0^T \left|\frac{u_t}{\overline{\sigma}(t)}\right|^2 dt\right], \tag{10}$$

to get the optimal control $u^*(\cdot,\cdot)$.

(2) Given $v^*(0,y)$, solve for the optimal initial distribution $\nu^*(\cdot)$:

$$\max_{\nu} \mathbb{E}_{\nu}[v^*(0,Y)] - \alpha D_{KL}(\nu(\cdot), p_{\text{noise}}(\cdot)). \tag{11}$$

C. Solve the stochastic control problem

Here we study the stochastic control problem (7) (or equivalently (9)), following (10)–(11). We start with the problem (11) to find the optimal initial distribution $\nu^*(\cdot)$, the easier one.

Proposition 3.4: Let $\int \exp\left(\frac{v^*(0,y)}{\alpha}\right) p_{\text{noise}}(y)dy < \infty$. We have:

$$\nu^*(x) = \frac{1}{C'}\exp\left(\frac{v^*(0,y)}{\alpha}\right) p_{\text{noise}}(y), \tag{12}$$

where C' is the normalizing constant.

The assumption $\int \exp\left(\frac{v^*(0,y)}{\alpha}\right) p_{\text{noise}}(y)dy < \infty$ is equivalent to $\int \exp\left(\frac{r(y)}{\alpha}\right) p_{\text{pre}}(y)dy < \infty$ in Lemma 3.1.

Now we proceed to solving the problem (10). To this end, define the value-to-go:

$$v^*(t,y) := \max_{u\in\mathcal{A}_t} \mathbb{E}_{Q^{u,y}_{[t,T]}}\left[r(Y_T) - \frac{\alpha}{2}\int_t^T \left|\frac{u_s}{\overline{\sigma}(s)}\right|^2 ds\right], \tag{13}$$

where $Q^{u,y}_{[t,T]}(\cdot)$ denotes the probability distribution of $(Y_s,\, t \le s \le T)$ in (6) conditioned on $Y_t = y$, and $\mathcal{A}_t$ is the admissible set for u on $[t,T]$. The following result provides a verification theorem for the problem (13).

Proposition 3.5: Assume that there exists $V(t,y) \in C^{1,2}([0,T]\times\mathbb{R}^d)$ of at most polynomial growth in y, which solves the Hamilton-Jacobi equation:

$$\begin{aligned}&\frac{\partial v}{\partial t} + \frac{\overline{\sigma}^2(t)}{2}\Delta v + \overline{b}(t,y)\cdot\nabla v + \frac{\overline{\sigma}^2(t)}{2\alpha}|\nabla v|^2 = 0, \\ &v(T,y) = r(y),\end{aligned} \tag{14}$$

and that $\frac{\overline{\sigma}^2(t)}{\alpha}\nabla V(\cdot,\cdot) \in \mathcal{A}$. Then $v^*(t,y) = V(t,y)$, and the feedback control $u^*(t,y) = \frac{\overline{\sigma}^2(t)}{\alpha}\nabla V(t,y)$.

It is preferable to characterize the value function $v^*(t,y)$ as the viscosity solution to the Hamilton-Jacobi equation (14). Under suitable conditions on $r(\cdot)$, $\sigma(\cdot)$ and $\overline{b}(\cdot,\cdot)$, it can be shown that the equation (14) has a unique viscosity solution which is smooth. So Proposition 3.5 can be applied. We do not pursue this direction, and refer to [17] and [30, Section 5] for related discussions.

The purpose of the stochastic control (7) (or (9)) is to fine-tune the pretrained model for data generation. We consider the distribution of $(Y_t, 0 \le t \le T)$ under the optimal control $Q^{u^*,\nu^*}_{[0,T]}(\cdot)$.

Proposition 3.6: Let (u^*, ν^*) be the optimal solution to the problem (7) (or (9)), and let the assumptions of Proposition 3.4 and 3.5 hold. We have:

$$Q^{u^*,\nu^*}_{[0,T]}(dY_\bullet) = \frac{1}{C}\exp\left(\frac{r(Y_T)}{\alpha}\right) Q_{[0,T]}(dY_\bullet), \tag{15}$$

where $Y_\bullet = (Y_t,\, 0 \le t \le T)$. Consequently, the marginal distribution is

$$Q^{u^*,\nu^*}_t(y) = \frac{1}{C}\mathbb{E}_{Q^y_{[t,T]}}\left[\exp\left(\frac{r(Y_T)}{\alpha}\right)\right] Q_t(y), \tag{16}$$

where $Q^y_{[t,T]}(\cdot)$ denotes the probability distribution of $(Y_s,\, t \le s \le T)$ under $Q_{[t,T]}(\cdot)$ conditioned on $Y_t = y$.

By specializing (16) at $t = T$, we get $Q^{u^*,\nu^*}_T(\cdot) = p_{\text{ftune}}(\cdot)$. Indeed, the stochastic control problem (7) yields the fine-tuned distribution (4). The optimal feedback control is given by

$$u^*(t,y) = \overline{\sigma}^2(t)\nabla\left(\ln \mathbb{E}_{Q^y_{[t,T]}}\left[\exp\left(\frac{r(Y_T)}{\alpha}\right)\right]\right). \tag{17}$$

To sample the optimal initial distribution $\nu^*(\cdot)$, we consider another stochastic control problem:

$$dY_t = q_t dt + \sigma'(t) dB'_t, \quad Y_0 \sim p_{\text{fix}}(\cdot), \tag{18}$$

where q_t is the control variable, $\sigma' : \mathbb{R}_+ \to \mathbb{R}_+$, and the distribution $p_{\text{fix}}(\cdot)$ is set so that the process $dY'_t = \sigma'(t)dB'_t$, $Y'_0 \sim p_{\text{fix}}(\cdot)$ will generate $Y'_T \sim p_{\text{noise}}(\cdot)$ at time T. Denote by $P^q_{[0,T]}(\cdot)$ be the probability distribution of $(Y_t,\, 0 \le t \le T)$, defined by (18), and P^q_t its marginal distribution at time t. Solve the optimization problem for q^*:

$$\max_q \mathbb{E}_{P^q_{[0,T]}}\left[v^*(0,Y_0) - \frac{\alpha}{2}\int_0^T \left|\frac{q_t}{\sigma'(t)}\right|^2 dt\right], \tag{19}$$

and it is easy to see that $P^{q^*}_T(\cdot) = \nu^*(\cdot)$.

The following result quantifies the performance of generating $p_{\text{ftune}}(\cdot)$ using any approximation $(\widetilde{u}^*, \widetilde{\nu}^*)$.

Proposition 3.7: Define

$$\eta^2 = \max_{0\le t\le T} \mathbb{E}_{Q^{u^*,\nu^*}_t(\cdot)}|\widetilde{u}^*(t,Y) - u^*(t,Y)|^2.$$

We have:

$$D_{TV}(Q^{\widetilde{u}^*,\nu^*}_T(\cdot), p_{\text{ftune}}(\cdot)) \le \frac{\eta}{2}\sqrt{\int_0^T \frac{1}{\overline{\sigma}^2(t)}dt} + \sqrt{\frac{1}{2}D_{KL}(\widetilde{\nu}^*(\cdot), \nu^*(\cdot))}. \tag{20}$$

IV. Extension to regularization by f-divergence

In this section, we go beyond entropy regularization, and consider the problem of fine-tuning regularized by f-divergence. Let $f : \mathbb{R}_+ \to (-\infty, +\infty]$ be a convex function such that $f(x) < \infty$ for $x > 0$, $f(1) = 0$ and $f(0) = \lim_{t\to 0^+} f(t)$. Given two probability distribution $p(\cdot)$, $q(\cdot)$ on suitably nice space, the f-divergence between $p(\cdot)$ and $q(\cdot)$ is defined by $D_f(P,Q) = \int f\left(\frac{dp}{dq}\right)dq = \mathbb{E}_q\left[f\left(\frac{dp}{dq}\right)\right]$. The KL divergence corresponds to the choice $f(t) = t\ln t$. Other popular examples include forward KL $f(t) = -\ln t$, and γ-divergence $f(t) = \frac{t^{1-\gamma} - (1-\gamma)t - \gamma}{\gamma(\gamma-1)}$. See [20, Chapter 7] for details on the f-divergence.

Inspired by entropy-regularized fine-tuning (3), an obvious candidate is:

$$p^f_\star(\cdot) := \arg\max_p \mathbb{E}_p[r(Y)] - \alpha D_f(p(\cdot), p_{\text{pre}}(\cdot)), \tag{21}$$

where the maximization is over all probability distribution on $\mathbb{R}^d$. Similar to entropy-regularized fine-tuning (3), the problem (21) is solvable under some conditions on f. The following result was proved by [37] in a different but closely related context of direct preference optimization [21].

Proposition 4.1: Assume that $p_{\text{pre}}(\cdot)$ has full support on $\mathbb{R}^d$, and f' is strictly increasing, and 0 is not in the domain of f'. Also assume that there exists λ such that $\frac{r(\cdot)-\lambda}{\alpha}$ is in the range of f', and $\int (f')^{-1}\left(\frac{r(y)-\lambda}{\alpha}\right) p_{\text{pre}}(y)dy = 1$. We have:

$$p^f_\star(y) = (f')^{-1}\left(\frac{r(y)-\lambda}{\alpha}\right) p_{\text{pre}}(y). \tag{22}$$

It was claimed in [37, p.5] that the solution (22) to the (constrained) problem (21) can be written in the form $\frac{1}{C_f}(f')^{-1}\left(\frac{r(\cdot)}{\alpha}\right) p_{\text{pre}}(\cdot)$. This seems to be wrong because $\frac{r(\cdot)}{\alpha}$ may even not lie in the range of f'. In fact, λ in (22) plays the role of the normalizing constant, and can be factorized out only for entropy regularization where $(f')^{-1}(t) = e^{t-1}$. Sampling the distribution (22) is generally difficult due to the complex dependence in the normalizing constant λ.

As an alternative, we can first solve the problem (21) with no constraint, and then normalize it to a probability distribution (which was claimed in [37] to be the "solution".) This leads to the following fine-tuning proposal. Recall that $r : \mathbb{R}^d \to \mathbb{R}_+$ is assumed to be nonnegative.

Definition 4.2: Assume that $(f')^{-1}(\cdot)$ is well-defined on the range of $\frac{r(\cdot)}{\alpha}$, or $(f')^{-1}(\cdot)$ is not but $(-f')^{-1}(\cdot)$ is well-defined on the range of $\frac{r(\cdot)}{\alpha}$, which we denote by $(f'_{\pm})^{-1}(\cdot)$. Also assume that $\int (f'_{\pm})^{-1}\left(\frac{r(y)}{\alpha}\right) p_{\text{pre}}(y)dy < \infty$. Define

$$p^f_{\text{ftune}}(y) := \frac{1}{C_f}(f'_{\pm})^{-1}\left(\frac{r(y)}{\alpha}\right) p_{\text{pre}}(y), \qquad (23)$$

where C_f is the normalizing constant. We abuse by calling (23) (instead of (21)) fine-tuning regularized by f-divergence.

Let's illustrate Definition 4.2 with examples.

• KL: $f'(t) = \ln t + 1$, so $p^f_{\text{ftune}}(y) \propto \exp(\frac{r(y)}{\alpha}) p_{\text{pre}}(y)$. The definitions (21) and (23) agree.

• Forward KL: $f'(t) = -\frac{1}{t}$ so $(f'_{\pm})^{-1}(t) = (-f')^{-1}(t) = \frac{1}{t}$. We have $p^f_{\text{ftune}}(y) \propto \alpha p_{\text{pre}}(y)/r(y)$.

• γ-divergence: $f'(t) = (1 - t^{-\gamma})/\gamma$. So if $r(\cdot)$ is bounded from below by α/γ, we get $(f'_{\pm})^{-1}(t) = (-f')^{-1}(t) = (\gamma t - 1)^{-\frac{1}{\gamma}}$ and $p^f_{\text{ftune}}(y) \propto \left(\frac{\gamma\, r(y)}{\alpha} - 1\right)^{-\frac{1}{\gamma}} p_{\text{pre}}(y)$; if $r(\cdot)$ is bounded from above by α/γ, we get $(f'_{\pm})^{-1}(t) = (f')^{-1}(t) = (1 - \gamma t)^{-\frac{1}{\gamma}}$ and $p^f_{\text{ftune}}(y) \propto \left(1 - \frac{\gamma\, r(y)}{\alpha}\right)^{-\frac{1}{\gamma}} p_{\text{pre}}(y)$.

The advantage of the definition (23) is that the normalizing constant is factored out in $(f'_{\pm})^{-1}$, so the machinery in Section III can be applied to generate $p^f_{\text{ftune}}(\cdot)$ by only modifying the reward function. Let

$$r_f(y) := \ln\left((f'_{\pm})^{-1}\left(\frac{r(y)}{\alpha}\right)\right), \qquad (24)$$

so $p^f_{\text{ftune}}(y) = \frac{1}{C_f}\exp(r_f(y)) p_{\text{pre}}(y)$. Following the stochastic control setting in Section III, we consider the problem of maximizing over $(u, \nu) \in \mathcal{A}$:

$$\int \mathbb{E}_{Q^{u,y}_{[0,T]}}\left[r_f(Y_T) - \frac{1}{2}\int_0^T \left|\frac{u_t}{\overline{\sigma}(t)}\right|^2 dt\right]\nu(y)dy - D_{KL}(\nu(\cdot), p_{\text{noise}}(\cdot)), \qquad (25)$$

where the parameter α is included in $r_f(\cdot)$, so it does not appear in the regularizer $D_{KL}(\nu(\cdot), p_{\text{noise}}(\cdot))$.

Here we use the entropy regularizer instead of f-divergence, i.e., $D_f(Q^{u,\nu}_{[0,T]}(\cdot), Q_{[0,T]}(\cdot))$ in the stochastic control problem. This is because it is difficult to simplify $D_f(Q^{u,\nu}_{[0,T]}(\cdot), Q_{[0,T]}(\cdot))$ due to lack of the chain rule (for general f-divergence), which will incur nonlocal terms. So in the problem (25), the "regularization by f-divergence" is reflected in the reward function $r_f(\cdot)$. As a result, Propositions 3.4–3.6 are easily adapted to solve the control problem (25). We summarize the results as follows.

Corollary 4.3: Define the value-to-go $v^*_f(t,y) := \max_{u\in\mathcal{A}_t} \mathbb{E}_{Q^{u,y}_{[t,T]}}\left[r_f(Y_T) - \frac{1}{2}\int_t^T \left|\frac{u_s}{\overline{\sigma}(s)}\right|^2 ds\right]$.

(1) Assume that there exists $V(t,y) \in C^{1,2}([0,T]\times\mathbb{R}^d)$ of at most polynomial growth in y, which solves the Hamilton-Jacobi equation:

$$\frac{\partial v}{\partial t} + \frac{\overline{\sigma}^2(t)}{2}\Delta v + \overline{b}(t,y)\cdot\nabla v + \overline{\sigma}^2(t)|\nabla v|^2 = 0,$$
$$v(T,y) = r_f(y),$$

and that $\overline{\sigma}^2(t)\nabla V(\cdot,\cdot) \in \mathcal{A}$. Then $v^*_f(t,y) = V(t,y)$, and the feedback control $u^*_f(t,y) = \overline{\sigma}^2(t)\nabla V(t,y)$.

(2) The optimal initial distribution is given by

$$\nu^*_f(x) = \frac{1}{C_f}\exp\left(v^*_f(0,y)\right) p_{\text{noise}}(y). \qquad (26)$$

(3) Let (u^*_f, ν^*_f) be the optimal solution to the problem (25), specified by (1) and (2). The distribution of the process under the optimal control is:

$$Q^{u^*_f,\nu^*_f}_{[0,T]}(dY_\bullet) = \frac{1}{C_f}\exp\left(r_f(Y_T)\right) Q_{[0,T]}(dY_\bullet), \qquad (27)$$

and in particular, $Q^{u^*_f,\nu^*_f}_T(\cdot) = p^f_{\text{ftune}}(\cdot)$.

Similar to the discussions in Section III, the optimal control problem is given by

$$u^*_f(t,y) = \overline{\sigma}^2(t)\nabla\left(\ln \mathbb{E}_{Q^y_{[t,T]}}\left[\exp\left(r_f(Y_T)\right)\right]\right), \qquad (28)$$

The optimal initial distribution $\nu^*_f(\cdot)$ is sampled by solving the control problem for q^*_f:

$$\max_q \mathbb{E}_{P^q_{[0,T]}}\left[v^*_f(0,Y_0) - \frac{1}{2}\int_0^T \left|\frac{q_t}{\sigma'(t)}\right|^2 dt\right], \qquad (29)$$

with $P^{q^*_f}_T(\cdot) = \nu^*_f(\cdot)$.

V. Numerical experiments

In this section, we apply our stochastic control approach for fine-tuning large-scale text-to-image models – Stable Diffusion v1.5 (SD v1.5). Specifically, we consider three different types of regularization as in Section IV: KL divergence, Forward KL divergence, and γ-divergence. The exploration parameter α is evaluated at the levels $\{0.1, 1, 10\}$, while the parameter γ is tested at the values $\{0.25, 0.5, 1\}$.

We follow the same experimental setup as described in [34]. The reward function for fine-tuning is the LAION Aesthetics Predictor V2 model [24], which is a linear MLP applied to CLIP embeddings, designed to approximate human aesthetic preferences. All methods are evaluated with respect to three key metrics: (1) *Reward*: the average aesthetic score assigned to the generated images. (2) *Divergence*: the entropy between the sample distribution of the fine-tuned model and that of the pretrained model, which quantifies how much the fine-tuned outputs deviate from those from the baseline model. (3) *Background ornamentation (B.O.)*: the presence of excessive decorative or non-essential visual structures in the background. This phenomenon is a form of reward collapse, where the model exploits imperfections in the reward function to artificially increase scores without improving image quality.

Table I reports the performance of different regularized fine-tuning methods under the three metrics, and Fig. 1 illustrates the outputs of the fine-tuned models with $\alpha = 10$. All experiments are conducted on four NVIDIA L40S GPUs.

Method	α	γ	Reward	Div.	B.O.
KL	10		8.36 ± 0.24	0.10	Heavy
Forward	10		8.95 ± 0.17	0.10	
Gamma	10	0.25	8.57 ± 0.27	0.10	
Gamma	10	0.5	8.73 ± 0.28	0.10	
Gamma	10	1	8.55 ± 0.39	0.25	Moderate
KL	1		8.30 ± 0.22	0.31	
Forward	1		8.50 ± 0.23	0.09	Heavy
Gamma	1	1	8.49 ± 0.20	0.11	Moderate
KL	0.1		8.13 ± 0.31	0.57	Moderate
Forward	0.1		8.45 ± 0.23	0.10	
Gamma	0.1	0.25	8.30 ± 0.28	0.27	Heavy
Gamma	0.1	0.5	8.73 ± 0.23	0.17	Moderate
Gamma	0.1	1	8.65 ± 0.24	0.10	Moderate

TABLE I: Evaluation of fine-tuning SD v1.5.

Our main observations are summarized as follows:

• *Performance*: Forward KL and γ-divergence consistently achieve higher mean reward values than KL divergence, with Forward KL at $\alpha = 10$ yielding the best overall performance.

• *Divergence*: Forward KL and γ-divergence maintain lower divergence from the pretrained output distribution across all settings, whereas KL divergence exhibits large drift under small α.

• *Image quality*: Forward KL and γ-divergence substantially reduce background ornamentation compared to KL divergence.

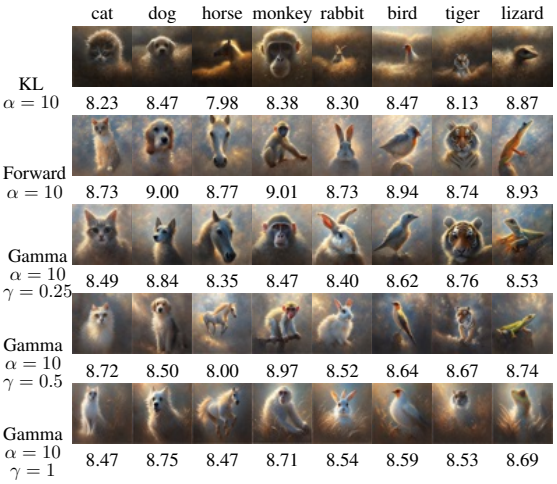

Fig. 1: Generated images of the fine-tuned models with $\alpha = 10$.

• *Robustness*: The performance of γ-divergence is relatively insensitive to changes in α, while larger α values improve performance for both Forward KL and γ-divergence. Moreover, larger α is effective in suppressing background ornamentation.

VI. Conclusion

In this paper, we give a rigorous treatment to the theory of fine-tuning by entropy-regularized stochastic control, which was recently proposed by [34] in the context of continuous-time diffusion models. We also generalize to the setting where an f-divergence regularizer is introduced. Numerical results are conducted on large-scale text-to-image models, which suggest using Forward KL or γ-divergence for regularization, with a large exploration parameter α.

There are several directions to extend this work. We provide an entropy-regularized approach to emulate the fine-tuned distribution (23) regularized by f-divergence. It is interesting to know whether it can be sampled by directly regularizing the controlled process by f-divergence. Also it remains unknown whether the fine-tuned distribution (21) can be generated by solving some stochastic control problem.

ACKNOWLEDGMENT

We thank Haoxian Chen, Minshuo Chen and Hanyang Zhao for various pointers to the literature. We thank David Yao and Xun Yu Zhou for stimulating discussions. W. Tang gratefully acknowledges financial support through NSF grant DMS-

2206038, the Columbia Innovation Hub grant, and the Tang Family Assistant Professorship. F. Zhou acknowledges financial support through NSF grant DMS-2206038, and the support from a Columbia-CityU/HK collaborative project led by InnoHK Initiative, The Government of the HKSAR and the AIFT Lab.

REFERENCES

[1] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané. Concrete problems in AI safety. 2016. arXiv:1606.06565.

[2] B. D. O. Anderson. Reverse-time diffusion equation models. *Stochastic Process. Appl.*, 12(3):313–326, 1982.

[3] K. Black, M. Janner, Y. Du, I. Kostrikov, and S. Levine. Training diffusion models with reinforcement learning. In *ICLR*, 2024.

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

[5] 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 *ICLR*, 2023.

[6] V. De Bortoli, J. Thornton, J. Heng, and A. Doucet. Diffusion Schrödinger bridge with applications to score-based generative modeling. In *Neurips*, volume 34, pages 17695–17709, 2021.

[7] P. Dhariwal and A. Nichol. Diffusion models beat GANs on image synthesis. In *Neurips*, volume 34, pages 8780–8794, 2021.

[8] C. Domingo-Enrich, M. Drozdzal, B. Karrer, and R. T. Chen. Adjoint matching: Fine-tuning flow and diffusion generative models with memoryless stochastic optimal control. In *ICLR*, 2025.

[9] Y. Fan and K. Lee. Optimizing DDPM sampling with shortcut fine-tuning. 2023. arXiv:2301.13362.

[10] Y. Fan, O. Watkins, Y. Du, H. Liu, M. Ryu, C. Boutilier, P. Abbeel, M. Ghavamzadeh, K. Lee, and K. Lee. DPOK: Reinforcement learning for fine-tuning text-to-image diffusion models. In *Neurips*, 2023.

[11] X. Gao, Z. Q. Xu, and X. Y. Zhou. State-dependent temperature control for Langevin diffusions. *SIAM J. Control Optim.*, 60(3):1250–1268, 2022.

[12] U. G. Haussmann and E. Pardoux. Time reversal of diffusions. *Ann. Probab.*, 14(4):1188–1205, 1986.

[13] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Neurips*, volume 33, pages 6840–6851, 2020.

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

[15] L. Kong, Y. Du, W. Mu, K. Neklyudov, V. De Bortol, H. Wang, D. Wu, A. Ferber, Y.-A. Ma, and C. P. Gomes. Diffusion models as constrained samplers for optimization with unknown constraints. 2024. arXiv:2402.18012.

[16] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *ICLR*, 2021.

[17] N. V. Krylov. *Nonlinear elliptic and parabolic equations of the second order*, volume 7 of *Mathematics and its Applications (Soviet Series)*. D. Reidel Publishing Co., Dordrecht, 1987.

[18] A. S. Luccioni, C. Akiki, M. Mitchell, and Y. Jernite. Stable bias: Analyzing societal representations in diffusion models. In *Neurips*, volume 36, 2023.

[19] OpenAI. Sora: Creating video from text. 2024. Available at `https://openai.com/sora`.

[20] Y. Polyanskiy and Y. Wu. *Information Theory: From Coding to Learning*. Cambridge University Press, 2025.

[21] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. In *Neurips*, volume 36, 2023.

[22] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. 2022. arXiv:2204.06125.

[23] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.

[24] C. Schuhmann. LAION – Aesthetics. 2022. Available at `https://laion.ai/blog/laion-aesthetics/`.

[25] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, volume 32, pages 2256–2265, 2015.

[26] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021.

[27] Z. Song, T. Cai, J. D. Lee, and W. J. Su. Reward collapse in aligning large language models. 2023. arXiv:2305.17608.

[28] D. W. Stroock and S. R. S. Varadhan. *Multidimensional diffusion processes*, volume 233 of *Grundlehren der Mathematischen Wissenschaften*. Springer-Verlag, 1979.

[29] W. Tang and R. Xu. A stochastic analysis approach to conditional diffusion guidance. 2025. Working paper. Available at `https://www.columbia.edu/~wt2319/CDG.pdf`.

[30] W. Tang, Y. P. Zhang, and X. Y. Zhou. Exploratory HJB equations and their convergence. *SIAM J. Control Optim.*, 60(6):3191–3216, 2022.

[31] W. Tang and H. Zhao. Contractive diffusion probabilistic models. 2024. arXiv:2401.13115.

[32] W. Tang and H. Zhao. Score-based diffusion models via stochastic differential equations. *Stat. Surv.*, 19:28–64, 2025.

[33] M. Uehara, Y. Zhao, T. Biancalani, and S. Levine. Understanding reinforcement learning-based fine-tuning of diffusion models: A tutorial and review. 2024. arXiv:2407.13734.

[34] 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. 2024. arXiv:2402.15194.

[35] P. Vincent. A connection between score matching and denoising autoencoders. *Neural Comput.*, 23(7):1661–1674, 2011.

[36] B. Wallace, M. Dang, R. Rafailov, L. Zhou, A. Lou, S. Purushwalkam, S. Ermon, C. Xiong, S. Joty, and N. Naik. Diffusion model alignment using direct preference optimization. 2023. arXiv:2311.12908.

[37] C. Wang, Y. Jiang, C. Yang, H. Liu, and Y. Chen. Beyond reverse KL: Generalizing direct preference optimization with diverse divergence constraints. In *ICLR*, 2024.

[38] G. I. Winata, H. Zhao, A. Das, W. Tang, D. D. Yao, S.-X. Zhang, and S. Sahu. Preference tuning with human feedback on language, speech, and vision tasks: A survey. *J. Artif. Intell. Res.*, 82:2595–2661, 2025.

[39] H. Zhao, H. Chen, J. Zhang, D. D. Yao, and W. Tang. Score as action: Fine-tuning diffusion generative models by continuous-time reinforcement learning. 2025. arXiv:2502.01819. To appear in ICML.