FINE-TUNING OF DIFFUSION MODELS VIA STOCHASTIC CONTROL: ENTROPY REGULARIZATION AND BEYOND

WENPIN TANG

Abstract. This paper aims to develop and provide a rigorous treatment to the problem of entropy regularized fine-tuning in the context of continuous-time diffusion models, which was recently proposed by Uehara et al. (arXiv:2402.15194, 2024). 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 involving a general f-divergence regularizer.

Key words: Constrained optimization, diffusion models, entropy-regularization, f-divergence, fine-tuning, stochastic control, stochastic differential equations.
AMS 2020 Mathematics Subject Classification: 60J60, 65C30, 93E20.

1. Introduction

Diffusion models [17, 33, 35] have emerged as a promising generative approach to produce high-quality samples, which is observed to outperform generative adversarial nets on image and audio synthesis [9, 23], and underpins the recent success in text-to-image creators such as DALL·E2 [30] and Stable Diffusion [31], and the text-to-video generator Sora [26]. Though diffusion models can capture intricate and high-dimensional data distributions, they may suffer from sources of bias or fairness concerns [25], and the training process (especially for the aforementioned 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 work [2, 11, 12] proposed to fine-tune diffusion models by reinforcement learning (RL), and [44] by direct preference optimization. In these work, the reward functions are commonly learned statistical models, e.g. an aesthetic reward in image generation is a ranking model fit to the true aesthetic preferences of human raters.

The above methods allow to fine-tune the diffusion model to generate samples with high nominal rewards. However, they may lead to *catastrophic forgetting* or *reward collapse* [36], a phenomenon referring to overfitting the reward (that is 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 and enhance diversity, [41] 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* [9, 18]. A stochastic control approach was developed to emulate

Date: March 20, 2024.

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. The idea of adding an entropy regularizer also appears in [40, Section 7.3], [47] for fine-tuning the diffusion model via continuous RL, and in [22] for constraint optimization via the diffusion model. In a different context, (entropy regularized) stochastic control was proposed to solve non-convex problems [14, 38], as an alternative to the simulated annealing algorithm.

The purpose of this paper is to elaborate and give a rigorous treatment to the theory of entropy-regularized fine-tuning proposed in [41]. The paper is mostly self-contained. Quantitative results are provided. We also 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 2, we recall background of diffusion models. The entropy-regularized fine-tuning is studied in Section 3, and the extension to fine-tuning regularized by f-divergence is reported in Section 4. We conclude with Section 5.

Notations: Below we collect a few notations that will be used throughout.

- $\mathbb{R}_+$ denotes the set of nonnegative real numbers.
- For x, y vectors, denote by $x \cdot y$ the inner product between x and y, and $|x|$ is the Euclidean (L^2) norm of x.
- For $p(\cdot)$ a probability distribution on $\mathbb{R}^d$, we assume that it has a density $p(x)$, and write $p(dx) = p(x)dx$.
- The notation $X \sim p(\cdot)$ means that the random variable is distributed according to $p(\cdot)$. We write $\mathbb{E}_{p(\cdot)}(X)$ 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 $q(\cdot)$.

2. Diffusion models

In this section, we provide background on diffusion models that will be used as the pretrained models in fine-tuning. We follow closely the presentation in [40].

The goal of diffusion models is to generate new samples (e.g. images, video, text) that resemble the target data, while maintain a certain level of diversity. Diffusion modeling relies on a forward-backward procedure:

- *Forward deconstruction*: start from the target distribution $X_0 \sim p_{\text{data}}(\cdot)$, the model gradually adds noise to transform the signal into noise $X_0 \to X_1 \to \cdots \to X_n \sim p_{\text{noise}}(\cdot)$.
- *Backward reconstruction*: start with the noise $X_n \sim p_{\text{noise}}(\cdot)$, and reverse the forward process to recover the signal from noise $X_n \to X_{n-1} \to \cdots X_0 \sim p_{\text{data}}(\cdot)$.

Here we consider the diffusion model in continuous time, governed by the stochastic differential equation (SDE):

$$dX_t = b(t, X_t)dt + \sigma(t)dW_t, \quad X_0 \sim p_{\text{data}}(\cdot), \tag{2.1}$$

where $(W_t,\, t \ge 0)$ is d-dimensional Brownian motion, and $b : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma : \mathbb{R}_+ \to \mathbb{R}_+$ are model parameters. Some conditions on $b(\cdot, \cdot)$, $\sigma(\cdot)$ are required so that the SDE (2.1) is well-defined (see [20, Chapter 5], [37]). In the sequel, we assume for simplicity that the target data distribution $p_{\text{data}}(\cdot)$ has a suitably smooth density, and write $p_{\text{data}}(dx) = p_{\text{data}}(x)dx$.

Fix $T > 0$ as the time horizon. Denote by $p(t, \cdot)$ the probability density of X_t. It is known [1, 16] that the distribution of the time reversal process $\overline{X}_t = X_{T-t}$ is governed by the SDE:

$$d\overline{X}_t = \left(-b(T-t, \overline{X}_t) + \sigma^2(T-t)\nabla \log p(T-t, \overline{X}_t)\right) dt + \sigma(T-t) dB_t, \quad \overline{X}_0 \sim p(T, \cdot),$$

where $(B_t,\, t \geq 0)$ is (a copy of) d-dimensional Brownian motion. Thus, the process recovers $\overline{X}_T \sim p_{\text{data}}(\cdot)$ at time T.

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 $p(T, \cdot)$:

$$d\overline{X}_t = \left(-b(T-t, \overline{X}_t) + \sigma^2(T-t)\nabla \log p(T-t, \overline{X}_t)\right) dt + \sigma(T-t) dB_t, \quad \overline{X}_0 \sim p_{\text{noise}}(\cdot). \tag{2.2}$$

The diffusion model is specified by the pair $(b(\cdot, \cdot), \sigma(\cdot))$, and existing examples include Ornstein-Ulenback processes [8], variance exploding (VE) SDEs, variance preserving (VP) SDEs [35], and contractive diffusion models [39]. The choice of $p_{\text{noise}}(\cdot)$ depends on each specific model, and varies case by case.

Second, all but the term $\nabla \log p(T-t, \overline{X}_t)$ in (2.2) are available. So in order to implement the SDE (2.2), we need to compute $\nabla \log p(t, x)$, known as *Stein's score function.* Recently developed *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 $\nabla \log p(t, x)$. by solving the stochastic optimization problem:

$$\operatorname{argmin}_\theta \mathbb{E}_{p(t,\cdot)} |s_\theta(t, X) - \nabla \log p(t, X)|^2 = \operatorname{argmin}_\theta \mathbb{E}_{p(t,\cdot)}[|s_\theta(t, X)|^2 + 2\, \nabla \cdot s_\theta(t, X)]. \tag{2.3}$$

Note that the left side of (2.3) still involves the unknown score function, which can be recast into a admissible optimization problem on the right side. The equality (2.3) hinges on an integration by parts argument, which was discovered by [19] in the context of estimating statistical models with unknown normalizing constant. In practice, more scalable variants of (2.3) are used to estimate the score function (e.g. sliced score matching [34] and denoising score matching [42].) See [40, Section 4] for a review.

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:

$$dY_t = \underbrace{\left(-b(T-t, Y_t) + \sigma^2(T-t) s_\theta(T-t, Y_t)\right)}_{\overline{b}(t, Y_t)} dt + \sigma(T-t) dB_t, \quad Y_0 \sim p_{\text{noise}}(\cdot). \tag{2.4}$$

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 $\nabla \log p(t, x)$) defined by (2.2). The SDE (2.4) 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

$$Y_T \sim Q_T(\cdot) =: p_{\text{pre}}(\cdot). \tag{2.5}$$

3. 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 [41]. As we will see, the arguments used in this section can be generalized to study the problem of fine-tuning involving an f-divergence regularizer.

3.1. **Entropy-regularized fine-tuning.** Let's explain entropy-regularized fine-tuning using a pretrained diffusion model (2.4). The idea of fine-tuning is to calibrate a (possibly powerful) pretained model to some downstream task with a specific goal, e.g. the aesthetic quality of generated images. Such a goal is captured by a reward function $r : \mathbb{R}^d \to \mathbb{R}_+$, and 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, the phenomenon known as the reward collapse (or catastrophic forgetting). In order to avoid the reward collapse and utilize the pretrained model, it is natural to consider the optimization problem:

$$p_{\text{ftune}}(\cdot) := \text{argmax}_p\, \mathbb{E}_{p(\cdot)}[r(Y)] - \alpha D_{KL}(p(\cdot), p_{\text{pre}}(\cdot)), \tag{3.1}$$

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 (3.1) is referred to as entropy-regularized fine-tuning of the pretrained model (2.4).

It is easily seen that the problem (3.1) has a closed-form solution.

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

$$p_{ftune}(y) = \frac{1}{C} \exp\left(\frac{r(y)}{\alpha}\right) p_{pre}(y), \tag{3.2}$$

where $C := \int \exp\left(\frac{r(y)}{\alpha}\right) p_{pre}(y)dy$ *is the normalizing constant.*

Proof. This follows from the duality formula for the variational problem in the theory of large deviations [10]. □

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 $p_{\text{ftune}}(\cdot)$. If $r(\cdot)$ is bounded, then we can use von Neumann's rejection algorithm [43] to generate $p_{\text{ftune}}(\cdot)$ from the diffusion samplers $p_{\text{pre}}(\cdot)$. The average number of iterations to generate the distribution (3.2) is $\sup_y \frac{1}{C} \exp\left(\frac{r(y)}{\alpha}\right)$. If $r(\cdot)$ is unbounded, then von Neumann's algorithm cannot be applied. In this case, we need to use an analog of von Neumann's algorithm – the filling scheme [4, 32] to generate $p_{\text{pre}}(\cdot)$ (see also [27, p.78-81].) However, the expected number of iterations to generate the fine-tuned distribution is infinite.

As we will see in the next subsection, the main idea is to apply a (stochastic) control to the diffusion sampler (2.4) in order to generate $p_{\text{ftune}}(\cdot)$ in just one pass. The next result shows how the fine-tuned distribution (3.2) differs from $p_{\text{data}}(\cdot)$.

Proposition 3.2. *We have:*

$$D_{TV}(p_{ftune}(\cdot), p_{data}(\cdot)) \le D_{TV}(p_{pre}(\cdot), p_{data}(\cdot)) + \frac{1}{2}\sqrt{\mathbb{E}_{p_{pre}(\cdot)}\left[e^{\frac{2r(Y)}{\alpha}}\right]}. \tag{3.3}$$

Before proving the proposition, let's make some remarks on the two terms on the right side of (3.3):

- The term $D_{TV}(p_{\text{pre}}(\cdot), p_{\text{data}}(\cdot))$ quantifies the performance of the pretrained model, i.e. how well (or how bad) the diffusion model recovers the target data distribution $p_{\text{data}}(\cdot)$. Assuming that $\sigma(\cdot)$ is bounded away from zero, and under the blackbox assumption that $\mathbb{E}_{p(t,\cdot)}|\nabla \log p(t,X) - s_\theta(t,X)|^2 \le \varepsilon^2$ for all t, [7, 40] shows that

$$D_{TV}(p_{\text{pre}}(\cdot), p_{\text{data}}(\cdot)) \le D_{TV}(p_{\text{noise}}(\cdot), p(T,\cdot)) + \varepsilon\sqrt{T}/2. \tag{3.4}$$

  The term $D_{TV}(p_{\text{noise}}(\cdot), p(T,\cdot))$ can further be quantified under each specific model. As an example, for VP (with $b(t,x) = -\frac{1}{2}\beta(t)x$ and $\sigma(t) = \sqrt{\beta(t)}$), we have $D_{TV}(p_{\text{noise}}(\cdot), p(T,\cdot)) \le$

$e^{-\frac{1}{2}\int_0^T \beta(t)dt}\sqrt{\mathbb{E}_{p_{\mathrm{data}}(\cdot)}|X|^2/2}$. So the bound (3.3) specializes to:

$$D_{TV}(p_{\mathrm{ftune}}(\cdot), p_{\mathrm{data}}(\cdot)) \le e^{-\frac{1}{2}\int_0^T \beta(t)dt}\sqrt{\mathbb{E}_{p_{\mathrm{data}}(\cdot)}|X|^2/2} + \frac{\varepsilon}{2}\sqrt{T} + \frac{1}{2}\sqrt{\mathbb{E}_{p_{\mathrm{pre}}(\cdot)}\left[e^{\frac{2r(Y)}{\alpha}}\right]}. \tag{3.5}$$

- The second term on the right side of (3.3) quantifies the difference between the distribution $p_{\mathrm{pre}}(\cdot)$ of samples generated from the pretrained model and the fine-tuned distribution $p_{\mathrm{ftune}}(\cdot)$. If the reward $r(x)$ is bounded, say by $\overline{r} > 0$, then the term is bounded from above by $\frac{1}{2}\exp\left(\frac{\overline{r}}{\alpha}\right)$. In this case, the bound (3.5) for VP becomes:

$$D_{TV}(p_{\mathrm{ftune}}(\cdot), p_{\mathrm{data}}(\cdot)) \le e^{-\frac{1}{2}\int_0^T \beta(t)dt}\sqrt{\mathbb{E}_{p_{\mathrm{data}}(\cdot)}|X|^2/2} + \frac{\varepsilon}{2}\sqrt{T} + \frac{1}{2}e^{\frac{\overline{r}}{\alpha}}. \tag{3.6}$$

When $\alpha \to 0$, the bound increases to $+\infty$, which gives an explanation of the reward collapse.

Proof of Proposition 3.2. We have:

$$D_{KL}(p_{\mathrm{ftune}}(\cdot), p_{\mathrm{pre}}(\cdot)) = -\mathbb{E}_{p_{\mathrm{pre}}(\cdot)}\left[\frac{r(Y)}{\alpha}\right] + \log \mathbb{E}_{p_{\mathrm{pre}}(\cdot)}[e^{r(Y)/\alpha}]. \tag{3.7}$$

Note that $x \to \log x + \frac{x^2}{2}$ is convex on $[1, +\infty)$. By Jensen's inequality,

$$\log \mathbb{E}_{p_{\mathrm{pre}}(\cdot)}[e^{r(Y)/\alpha}] + \frac{1}{2}\left(\mathbb{E}_{p_{\mathrm{pre}}(\cdot)}[e^{r(Y)/\alpha}]\right)^2 \le \mathbb{E}_{p_{\mathrm{pre}}(\cdot)}\left(\frac{r(Y)}{\alpha} + \frac{1}{2}e^{2r(Y)/\alpha}\right). \tag{3.8}$$

Combining (3.7) and (3.8) yields

$$D_{KL}(p_{\mathrm{ftune}}(\cdot), p_{\mathrm{pre}}(\cdot)) \le \frac{1}{2}\mathrm{Var}_{p_{\mathrm{pre}}(\cdot)}(e^{r(Y)/\alpha}) \le \frac{1}{2}\mathbb{E}_{p_{\mathrm{pre}}}(e^{2r(Y)/\alpha}).$$

Further by Pinsker's inequality (see [28, Theorem 7.10]), we get $D_{TV}(p_{\mathrm{ftune}}(\cdot), p_{\mathrm{pre}}(\cdot)) \le \frac{1}{2}\sqrt{\mathbb{E}_{p_{\mathrm{pre}}(\cdot)}\left[e^{\frac{2r(Y)}{\alpha}}\right]}$. The bound (3.3) follows by the triangle inequality. □

3.2. **Stochastic control.** Now the goal is to emulate the fine-tuned distribution $p_{\mathrm{ftune}}(\cdot)$ specified by (3.2). We follow the idea of [41] to lift the problem to the process via stochastic control:

$$dY_t = (\overline{b}(t, Y_t) + u_t)dt + \sigma(T-t)dB_t, \quad Y_0 \sim \nu(\cdot), \tag{3.9}$$

where $\overline{b}(\cdot,\cdot)$ is defined in (2.4), 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 $p_{\mathrm{noise}}(\cdot)$ as in (2.4).

Denote by $Q^{u,\nu}_{[0,T]}(\cdot)$ the probability distribution of the process $(Y_t,\, 0 \le t \le T)$ defined by (3.9), 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 (2.4), so

$$Q_{[0,T]}(\cdot) = Q^{0,p_{\mathrm{noise}}}_{[0,T]}(\cdot).$$

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

$$(u^*, \nu^*) := \mathrm{argmax}_{(u,\nu)\in\mathcal{A}}\, \mathbb{E}_{Q^{u,\nu}_{[0,T]}(\cdot)}[r(Y_T)] - \alpha D_{KL}(Q^{u,\nu}_{[0,T]}(\cdot), Q_{[0,T]}(\cdot)), \tag{3.10}$$

where $\mathcal{A}$ is an admissible set for the decision variables that we will specify. As we will see, $Y_T \sim p_{\mathrm{ftune}}(\cdot)$ under $Q^{u^*,\nu^*}_{[0,T]}(\cdot)$ (which may not hold for the regularization by general f-divergence.)

The following result elucidates the form of the objective (3.10).

Proposition 3.3. *Let*

$$\mathcal{A} = \Bigg\{(u,\nu) : \nu \textit{ has a smooth density, } \exp\left(-\int_0^t \frac{u_s}{\sigma(s)}dB_s - \frac{1}{2}\int_0^t \left|\frac{u_s}{\sigma(s)}\right|^2 ds\right) \textit{ and } \int_0^t \frac{u_s}{\sigma(s)}dB_s \textit{ are } Q^{u,\nu}_{[0,T]}\textit{-martingales}\Bigg\}. \tag{3.11}$$

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

$$\begin{aligned}
&\mathbb{E}_{Q^{u,\nu}_{[0,T]}(\cdot)}[r(Y_T)] - \alpha D_{KL}(Q^{u,\nu}_{[0,T]}(\cdot), Q_{[0,T]}(\cdot)) \\
&\qquad = \mathbb{E}_{Q^{u,\nu}_{[0,T]}(\cdot)}\left[r(Y_T) - \frac{\alpha}{2}\int_0^T \left|\frac{u_t}{\sigma(t)}\right|^2 dt - \alpha \log\left(\frac{\nu(Y_0)}{p_{noise}(Y_0)}\right)\right].
\end{aligned} \tag{3.12}$$

Proof. For $(u,v) \in \mathcal{A}$, we have $\exp\left(-\int_0^t \frac{u_s}{\sigma(s)} dB_s - \frac{1}{2}\int_0^t \left|\frac{u_s}{\sigma(s)}\right|^2 ds\right)$ is a $Q^{u,\nu}_{[0,T]}$-martingale. So by Girsanov's theorem ([20, Chapter 3, Theorem 5.1]),

$$\frac{dQ^{u,\nu}_{[0,T]}}{dQ_{[0,T]}}(Y) = \frac{\nu(Y_0)}{p_{\mathrm{noise}}(Y_0)} \exp\left(\int_0^T \frac{u_t}{\sigma(t)} dB_t + \frac{1}{2}\int_0^T \left|\frac{u_t}{\sigma(t)}\right|^2 dt\right).$$

As a result,

$$\begin{aligned}
D_{KL}(Q^{u,\nu}_{[0,T]}(\cdot), Q_{[0,T]}(\cdot)) &= \mathbb{E}_{Q^{u,\nu}_{[0,T]}(\cdot)}\left[\log \frac{dQ^{u,\nu}_{[0,T]}}{dQ_{[0,T]}}\right] \\
&= \mathbb{E}_{Q^{u,\nu}_{[0,T]}(\cdot)}\left[\log\left(\frac{\nu(Y_0)}{p_{\mathrm{noise}}(Y_0)}\right) + \int_0^T \frac{u_t}{\sigma(t)} dB_t + \frac{1}{2}\int_0^T \left|\frac{u_t}{\sigma(t)}\right|^2 dt\right] \\
&= \mathbb{E}_{Q^{u,\nu}_{[0,T]}(\cdot)}\left[\log\left(\frac{\nu(Y_0)}{p_{\mathrm{noise}}(Y_0)}\right) + \frac{1}{2}\int_0^T \left|\frac{u_t}{\sigma(t)}\right|^2 dt\right],
\end{aligned}$$

where the last equation is due to the fact that $\int_0^t \frac{u_s}{\sigma(s)} dB_s$ is a $Q^{u,\nu}_{[0,T]}$-martingale. This yields the identity (3.12). □

The condition that $\exp\left(-\int_0^t \frac{u_s}{\sigma(s)} dB_s - \frac{1}{2}\int_0^t \left|\frac{u_s}{\sigma(s)}\right|^2 ds\right)$ and $\int_0^t \frac{u_s}{\sigma(s)} dB_s$ are $Q^{u,\nu}_{[0,T]}$-martingales 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 $\mathbb{E}_{Q^{u,\nu}_{[0,T]}(\cdot)}\left[\exp\left(\frac{1}{2}\int_0^T \left|\frac{u_t}{\sigma(t)}\right|^2 dt\right)\right] < \infty$ ensures the use of Girsanov's theorem, and also the $Q^{u,\nu}_{[0,T]}$-martingale property of $\int_0^t \frac{u_s}{\sigma(s)} dB_s$.

By Proposition 3.3, the stochastic control problem (3.10) is rewritten as:

$$\begin{aligned}
(u^*, \nu^*) :&= \operatorname{argmax}_{(u,\nu)\in\mathcal{A}} \mathbb{E}_{Q^{u,\nu}_{[0,T]}(\cdot)}\left[r(Y_T) - \frac{\alpha}{2}\int_0^T \left|\frac{u_t}{\sigma(t)}\right|^2 dt - \alpha\log\left(\frac{\nu(Y_0)}{p_{\mathrm{noise}}(Y_0)}\right)\right] \\
&= \operatorname{argmax}_{(u,\nu)\in\mathcal{A}} \int \mathbb{E}_{Q^{u,y}_{[0,T]}(\cdot)}\left[r(Y_T) - \frac{\alpha}{2}\int_0^T \left|\frac{u_t}{\sigma(t)}\right|^2 dt\right] \nu(y)dy - \alpha D_{KL}(\nu(\cdot), p_{\mathrm{noise}}(\cdot)),
\end{aligned} \tag{3.13}$$

where $Q^{u,y}_{[0,T]}(\cdot)$ denotes the probability distribution of the process $(Y_t,\, 0 \le t \le T)$ in (3.9) conditioned on $Y_0 = y$. It is easy to see that the two decision variables (u,ν) are separable in the problem (3.13). It boils down to the following two steps:

(1) Solve the (standard) stochastic control problem:

$$v^*(0,y) := \max_{u\in\mathcal{A}} \mathbb{E}_{Q^{u,y}_{[0,T]}(\cdot)}\left[r(Y_T) - \frac{\alpha}{2}\int_0^T \left|\frac{u_t}{\sigma(t)}\right|^2 dt\right], \tag{3.14}$$

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

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

$$\nu^*(\cdot) := \operatorname{argmax}_\nu \mathbb{E}_{\nu(\cdot)}[v^*(0,Y)] - \alpha D_{KL}(\nu(\cdot), p_{\text{noise}}(\cdot)). \tag{3.15}$$

3.3. **Solve the stochastic control problem.** Here we study the stochastic control problem (3.10) (or equivalently (3.13)), following (3.14)–(3.15). We start with the problem (3.15) to find the optimal initial distribution $\nu^*(\cdot)$, the easier one.

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

$$\nu^*(x) = \frac{1}{C'} \exp\left(\frac{v^*(0,y)}{\alpha}\right) p_{noise}(y), \tag{3.16}$$

where $C' = \int \exp\left(\frac{v^*(0,y)}{\alpha}\right) p_{noise}(y)dy$ *is the normalizing constant.*

Proof. The proof is the same as Lemma 3.1, by replacing $r(y)$ with $v^*(0,y)$ and $p_{\text{pre}}(y)$ with $p_{\text{noise}}(y)$. □

As we will see, 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 (3.14). To this end, define the value-to-go:

$$v^*(t,y) := \max_{u \in \mathcal{A}_t} \mathbb{E}_{Q^{u,y}_{[t,T]}(\cdot)} \left[r(Y_T) - \frac{\alpha}{2} \int_t^T \left| \frac{u_s}{\sigma(s)} \right|^2 ds \right], \tag{3.17}$$

where $Q^{u,y}_{[t,T]}(\cdot)$ denotes the probability distribution of $(Y_s,\, t \le s \le T)$ in (3.9) 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 (3.17).

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:*

$$\frac{\partial v}{\partial t} + + \frac{\sigma^2(t)}{2} \Delta v + \bar{b}(t,y) \cdot \nabla v + \frac{\sigma^2(t)}{2\alpha} |\nabla v|^2 = 0, \quad v(T,y) = r(y), \tag{3.18}$$

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

Proof. Let's first derive the Hamilton-Jacobi equation (3.18). Dynamic program shows that $v^*(t,y)$ solves (in some sense) that Hamilton-Jacobi-Bellman equation:

$$\max_u \left\{ \frac{\partial v}{\partial t} + \frac{\sigma^(t)}{2} \Delta v + (\bar{b}(t,y) + u) \cdot \nabla v - \frac{\alpha |u|^2}{2\sigma^2(t)} \right\} = 0. \tag{3.19}$$

The maximum is attained at $u^*(t,y) = \frac{\sigma^2(t)}{\alpha} \nabla v(t,y)$ (if admissible). Plugging it back into (3.19) yields the equation (3.18). By standard verification argument (see [46]), if a smooth function $V(t,y)$ solves the Hamilton-Jacobi equation (3.18) with $U(t,y) := \frac{\sigma^2(t)}{\alpha} \nabla V(t,y)$ admissible, then $v^* = V$ and $u^* = U$. □

It is preferable to characterize the value function $v^*(t,y)$ as the viscosity solution to the Hamilton-Jacobi equation (3.18). Under suitable conditions on $r(\cdot)$, $\sigma(\cdot)$ and $\bar{b}(\cdot,\cdot)$, it can be shown that the equation (3.18) has a unique viscosity solution which is smooth. So Proposition 3.5 can be applied. We do not pursue this direction, and refer to [24] and [38, Section 5] for related discussions.

The purpose of the stochastic control (3.10) (or (3.13)) is to fine-tune the pretrained model for data generation. To conclude this section, 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* (3.10) *(or* (3.13)*), and let the assumptions of Proposition 3.4 and 3.5 hold. We have:*

$$Q^{u^*,\nu^*}_{[0,T]}(dY_\bullet) = \frac{1}{C}\exp\left(\frac{r(Y_T)}{\alpha}\right) Q_{[0,T]}(dY_\bullet), \tag{3.20}$$

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

$$Q^{u^*,\nu^*}_t(y) = \frac{1}{C}\mathbb{E}_{Q^y_{[t,T]}(\cdot)}\left[\exp\left(\frac{r(Y_T)}{\alpha}\right)\right] Q_t(y), \quad 0 \le t \le T, \tag{3.21}$$

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 $Y_t = y$.

Proof. Recall that $Q^{u^*,y}_{[0,T]}(\cdot)$ denotes the probability distribution of (3.9) conditioned on $Y_0 = y$. We first show that

$$Q^{u^*,y}_{[0,T]}(y_\bullet) = \frac{\exp\left(\frac{r(Y_T)}{\alpha}\right) Q^y_{[0,T]}(Y_\bullet)}{\mathbb{E}_{Q^y_{[0,T]}(\cdot)}\left[\exp\left(\frac{r(Y_T)}{\alpha}\right)\right]}. \tag{3.22}$$

Note that

$$\begin{aligned}
&D_{KL}\left(Q^{u^*,y}_{[0,T]}(Y_\bullet), \frac{\exp\left(\frac{r(Y_T)}{\alpha}\right) Q^y_{[0,T]}(Y_\bullet)}{\mathbb{E}_{Q^y_{[0,T]}(\cdot)}\left[\exp\left(\frac{r(Y_T)}{\alpha}\right)\right]}\right) \\
&= D_{KL}(Q^{u^*,y}_{[0,T]}(Y_\bullet), Q^y_{[0,T]}(Y_\bullet)) - \mathbb{E}_{Q^y_{[0,T]}(\cdot)}\left\{\frac{r(Y_T)}{\alpha} - \log \mathbb{E}_{Q^y_{[0,T]}(\cdot)}\left[\exp\left(\frac{r(Y_T)}{\alpha}\right)\right]\right\} \\
&= \mathbb{E}_{Q^y_{[0,T]}(\cdot)}\left[\frac{1}{2}\int_0^T \left|\frac{u^*(t,Y_t)}{\sigma(t)}\right|^2 dt - \frac{r(Y_T)}{\alpha}\right] + \log \mathbb{E}_{Q^y_{[0,T]}(\cdot)}\left[\exp\left(\frac{r(Y_T)}{\alpha}\right)\right] \\
&= -v^*(0,y) + \log \mathbb{E}_{Q^y_{[0,T]}(\cdot)}\left[\exp\left(\frac{r(Y_T)}{\alpha}\right)\right].
\end{aligned} \tag{3.23}$$

where the first equation follows from the chain rule for KL divergence, and the second equation is from Girsanov's theorem. Recall that $v^*(t,y)$ solves the Hamilton-Jacobi equation (3.18). Let $\mathcal{V}(t,y) := \exp(v^*(t,y)/\alpha)$. By Itô's formula, $\mathcal{V}(t,y)$ solves the linear equation:

$$\frac{\partial \mathcal{V}}{\partial t} + \frac{\sigma^2(t)}{2}\Delta\mathcal{V} + \bar{b}(t,y)\cdot\mathcal{V} = 0, \quad \mathcal{V}(T,y) = \exp\left(\frac{r(y)}{\alpha}\right).$$

The Feynman-Kac formula ([20, Chapter 4, Theorem 4.2]) yields:

$$\exp\left(\frac{v^*(t,y)}{\alpha}\right) = \mathbb{E}_{Q^y_{[t,T]}(\cdot)}\left[\exp\left(\frac{r(Y_T)}{\alpha}\right)\right]. \tag{3.24}$$

By setting $t = 0$ in (3.24) and injecting it into (3.23), we get $D_{KL}\left(Q^{u^*,y}_{[0,T]}(y_\bullet), \frac{\exp\left(\frac{r(Y_T)}{\alpha}\right) Q^y_{[0,T]}(Y_\bullet)}{\mathbb{E}_{Q^y_{[0,T]}(\cdot)}\left[\exp\left(\frac{r(Y_T)}{\alpha}\right)\right]}\right) = 0$ so (3.22) is proved.

Next by Proposition 3.4, we have:

$$\nu^*(y) = \frac{1}{C'}\exp\left(\frac{v^*(0,y)}{\alpha}\right) p_{\text{noise}}(y) = \frac{1}{C'}\mathbb{E}_{Q^y_{[0,T]}(\cdot)}\left[\exp\left(\frac{r(Y_T)}{\alpha}\right)\right] p_{\text{noise}}(y). \tag{3.25}$$

where the second equation, again, follows (3.24). Furthermore,

$$C' := \int \exp\left(\frac{v^*(0,y)}{\alpha}\right) p_{\text{noise}}(y)dy = \mathbb{E}_{Q_{[t,T]}(\cdot)}\left[\exp\left(\frac{r(Y_T)}{\alpha}\right)\right] = \mathbb{E}_{p_{\text{pre}}(\cdot)}\left[\exp\left(\frac{r(Y)}{\alpha}\right)\right] =: C. \tag{3.26}$$

Combining (3.22), (3.25) and (3.26) yields (3.20). The identity (3.21) is obtained by marginalizing over t. □

By specializing (3.21) at $t = T$, we get $Q_T^{u^*,\nu^*}(\cdot) = p_{\text{ftune}}(\cdot)$. Indeed, the stochastic control problem (3.10) yields the fine-tuned distribution (3.2). The equality (3.26) shows that the assumptions in Lemma 3.1 and Proposition 3.4 are equivalent.

In [41], the optimal control problem (3.14) or (3.17) is solved by neural ODEs/SDEs [6, 21]. Here we offer some other thoughts. By (3.24) and Proposition 3.5, the optimal feedback control is given by

$$u^*(t,y) = \frac{\sigma^2(t)}{\alpha}\nabla v^*(t,y) = \sigma^2(t)\nabla\left(\ln \mathbb{E}_{Q^y_{[t,T]}(\cdot)}\left[\exp\left(\frac{r(Y_T)}{\alpha}\right)\right]\right). \tag{3.27}$$

By "simply" exchanging ln and $\mathbb{E}_{Q^y_{[t,T]}(\cdot)}$, we get

$$\widetilde{u}^*(t,y) := \frac{\sigma^2(t)}{\alpha}\nabla\,\mathbb{E}_{Q^y_{[t,T]}(\cdot)}[r(Y_T)] \overset{?}{\approx} u^*(t,y), \tag{3.28}$$

where the term $\nabla\,\mathbb{E}_{Q^y_{[t,T]}(\cdot)}[r(Y_T)]$ can further be computed by classifier guidance [9]. An alternative approach to compute $\nabla\,\mathbb{E}_{Q^y_{[t,T]}(\cdot)}[r(Y_T)]$ is by Malliavin calculus [13] (that is well-studied in the mathematical finance literature.) To be more precise,

$$\nabla\,\mathbb{E}_{Q^y_{[t,T]}(\cdot)}[r(Y_T)] = \mathbb{E}_{Q^y_{[t,T]}(\cdot)}[\nabla\, r(Y_T)\cdot Z_T], \tag{3.29}$$

where $(Z_s,\, t \le s \le T)$ (formally) solves:

$$dZ_s = \nabla\overline{b}(s,Y_s)Z_s ds, \quad Z_t = I. \tag{3.30}$$

See [5, 15] for a rigorous treatment. However, it is not clear how good/bad the approximation (3.27) is generally.

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

$$dY_t = q_t dt + \sigma'(t)dB'_t, \quad Y_0 \sim p_{\text{fix}}(\cdot), \tag{3.31}$$

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 (3.31)., and P^q_t its marginal distribution at time t. Solve the optimization problem:

$$q^* := \operatorname{argmax}_q \mathbb{E}_{P^q_{[0,T]}(\cdot)}\left[v^*(0,Y_0) - \frac{\alpha}{2}\int_0^T \left|\frac{q_t}{\sigma'(t)}\right|^2 dt\right], \tag{3.32}$$

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

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

Proposition 3.7. *Define* $\eta^2 := \max_{0\le t\le T}\mathbb{E}_{Q_t^{u^*,\nu^*}(\cdot)}|\widetilde{u}^*(t,Y) - u^*(t,Y)|^2$. *We have:*

$$D_{TV}(Q_T^{\widetilde{u}^*,\nu^*}(\cdot), p_{ftune}(\cdot)) \le \frac{\eta}{2}\sqrt{\int_0^T \frac{1}{\sigma^2(t)}dt} + \sqrt{\frac{1}{2}D_{KL}(\widetilde{\nu}^*(\cdot),\nu^*(\cdot))}. \tag{3.33}$$

Proof. By the data processing inequality, we get

$$\begin{aligned} D_{KL}(p_{\text{ftune}}(\cdot), Q_T^{\widetilde{u}^*,\widetilde{\nu}^*}(\cdot)) &= D_{KL}(Q_T^{u^*,\nu^*}(\cdot), Q_T^{\widetilde{u}^*,\widetilde{\nu}^*}(\cdot)) \\ &\le D_{KL}(Q_{[0,T]}^{u^*,\nu^*}(\cdot), Q_{[0,T]}^{\widetilde{u}^*,\widetilde{\nu}^*}(\cdot)) \\ &= \frac{1}{2}\mathbb{E}_{Q_{[0,T]}^{u^*,\nu^*}(\cdot)}\left(\int_0^T \left|\frac{u^*(t,Y_t) - \widetilde{u}^*(t,Y_t)}{\sigma(t)}\right|^2 dt\right) + D_{KL}(\widetilde{\nu}^*(\cdot), \nu^*(\cdot)) \\ &\le \frac{\eta^2}{2}\int_0^T \frac{1}{\sigma^2(t)} dt + D_{KL}(\widetilde{\nu}^*(\cdot), \nu^*(\cdot)). \end{aligned}$$

Applying Pinsker's inequality yields (3.33). □

4. 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:

$$D_f(P, Q) := \int f\left(\frac{dp}{dq}\right) dq = \mathbb{E}_{q(\cdot)}\left[f\left(\frac{dp}{dq}\right)\right]. \tag{4.1}$$

The KL divergence corresponds to the choice $f(t) = t \ln t$. Other popular examples include:

- Forward KL: $f(t) = -\ln t$.
- γ-divergence: $f(t) = \frac{t^{1-\gamma} - (1-\gamma)t - \gamma}{\gamma(\gamma-1)}$.
- Total variation distance: $f(t) = \frac{1}{2}|t - 1|$.

The γ-divergence becomes forward KL when $\gamma \to 1$, and KL when $\gamma \to 0$. So the KL divergence is also referred to as the reverse KL. See [28, Chapter 7] for an introduction to the f-divergence.

Inspired by entropy-regularized fine-tuning (3.1), an obvious candidate is:

$$p_\star^f(\cdot) := \text{argmax}_p\, \mathbb{E}_{p(\cdot)}[r(Y)] - \alpha D_f(p(\cdot), p_{\text{pre}}(\cdot)), \tag{4.2}$$

where the maximization is over all probability distribution on $\mathbb{R}^d$. Similar to entropy-regularized fine-tuning (3.1), the problem (4.2) is solvable under some conditions on f. The following result was proved by [45] in a different but closely related context of direct preference optimization [29]. For ease of reference, we record the proof.

Proposition 4.1. *Assume that $p_{pre}(\cdot)$ 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 $\int (f')^{-1}\left(\frac{r(y)-\lambda}{\alpha}\right) p_{pre}(y)dy = 1$. We have:*

$$p_\star^f(y) = (f')^{-1}\left(\frac{r(y) - \lambda}{\alpha}\right) p_{pre}(y). \tag{4.3}$$

Proof. Define the Lagrangian

$$\mathcal{L}(p, \lambda, \alpha) := \mathbb{E}_{p(\cdot)}[r(Y)] - \alpha\mathbb{E}_{p_{\text{pre}}(\cdot)}\left[f\left(\frac{p(Y)}{p_{\text{pre}}(Y)}\right)\right] - \lambda\left(\int \pi(y)dy - 1\right) + \mathbb{E}_{p(\cdot)}[\kappa(Y)],$$

where p is the primal variable, and λ, $\kappa(y)$ are dual variables. By the KKT condition (see [3, Section 5.5]), we get:

- First order condition: $\nabla_p \mathcal{L}(p, \lambda, \alpha) = 0$, which yields

$$r(y) - \alpha f'\left(\frac{p(y)}{p_{\text{pre}}(y)}\right) - \lambda + \kappa(y) = 0, \quad \text{for all } y. \tag{4.4}$$

- Primal feasibility: $\int p(y)dy = 1$ and $p(y) \geq 0$.
- Dual feasibility: $\kappa(y) \geq 0$ for all y.
- Complementary slackness: $\kappa(y)p(y) = 0$ for all y.

Since f' is strictly increasing (invertible), we obtain by (4.4):

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

Because $p_{\text{pre}}(y) > 0$ and 0 is not in the domain of f', we have $p(y) > 0$ for all y. Thus, the complementary slackness implies $\kappa(y) = 0$. The constant λ is determined by the primal condition. This yield the formula 4.3. □

It was claimed in [45, p.5] that the solution (4.3) to the (constrained) problem (4.2) 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 (4.3) 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 (4.3) is generally difficult due to the complex dependence in the normalizing constant λ.

As an alternative, we can first solve the problem (4.2) with no constraint, and then normalize it to a probability distribution (which was claimed in [45] 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 $\int (f'_\pm)^{-1}\left(\frac{r(y)}{\alpha}\right)p_{pre}(y)dy < \infty$. Define*

$$p^f_{ftune}(y) = \frac{1}{C_f}(f'_\pm)^{-1}\left(\frac{r(y)}{\alpha}\right)p_{pre}(y), \tag{4.5}$$

where $C_f := \int (f'_\pm)^{-1}\left(\frac{r(y)}{\alpha}\right)p_{pre}(y)dy$ is the normalizing constant. We abuse by calling (4.5) *(instead of* (4.2)*) fine-tuning regularized by f-divergence.*

Let's illustrate Definition 4.2 with examples.

- KL divergence: $f'(t) = \ln t + 1$. So $p^f_{\text{ftune}}(y) \propto \exp\left(\frac{r(y)}{\alpha}\right)p_{\text{pre}}(y)$, and the definitions (4.2) and (4.5) 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 $p^f_{\text{ftune}}(y) \propto \left(1 - \frac{\gamma\, r(y)}{\alpha}\right)^{-\frac{1}{\gamma}} p_{\text{pre}}(y)$
- Total variation distance: $(f'_\pm)^{-1}$ is not well-defined.

The advantage of the definition (4.5) is that the normalizing constant is factored out in $(f'_\pm)^{-1}$, so the machinery in Section 3 can be applied to generate $p^f_{\text{ftune}}(\cdot)$ by only modifying the reward function.

Let

$$r_f(y) := \ln\left((f'_\pm)^{-1}\left(\frac{r(y)}{\alpha}\right)\right), \tag{4.6}$$

so $p^f_{\text{ftune}}(y) = \frac{1}{C_f}\exp(r_f(y))p_{\text{pre}}(y)$. The following result is a corollary of Proposition 3.2 to quantify the difference between the fine-tuned distribution (4.5) and the original data distribution $p_{\text{data}}(\cdot)$.

Corollary 4.3. *Assume that* $\kappa := \inf_y r_f(y) > -\infty$. *We have:*

$$D_{TV}(p^f_{ftune}(\cdot), p_{data}(\cdot)) \le D_{TV}(p_{pre}(\cdot), p_{data}(\cdot)) + \frac{e^{-\kappa}}{2}\sqrt{\mathbb{E}_{p_{pre}(\cdot)}\left[\left((f'_\pm)^{-1}\left(\frac{r(Y)}{\alpha}\right)\right)^2\right]}. \tag{4.7}$$

Proof. It suffices to note that $x \to \log x + \frac{e^{-2\kappa}x^2}{2}$ is convex on $[e^\kappa, \infty)$. The same argument in Proposition 3.2 shows:

$$D_{KL}(p^f_{\text{ftune}}(\cdot), p_{\text{pre}}(\cdot)) \le \frac{e^{-2\kappa}}{2}\mathbb{E}_{p_{\text{pre}}(\cdot)}[e^{2r_f(Y)}] = \frac{e^{-2\kappa}}{2}\mathbb{E}_{p_{\text{pre}}(\cdot)}\left[\left((f'_\pm)^{-1}\left(\frac{r(Y)}{\alpha}\right)\right)^2\right].$$

This yields the bound (4.7). □

Following the stochastic control setting in Section 3, we consider the problem:

$$\begin{aligned}(u^*, \nu^*) : &= \text{argmax}_{u,\nu)\in\mathcal{A}}\, \mathbb{E}_{Q^{u,\nu}_{[0,T]}(\cdot)}[r_f(Y_t)] - D_{KL}(Q^{u,\nu}_{[0,T]}(\cdot), Q_{[0,T]}(\cdot)) \\ &= \text{argmax}_{(u,\nu)\in\mathcal{A}} \int \mathbb{E}_{Q^{u,y}_{[0,T]}(\cdot)}\left[r_f(Y_T) - \frac{1}{2}\int_0^T \left|\frac{u_t}{\sigma(t)}\right|^2 dt\right]\nu(y)dy - D_{KL}(\nu(\cdot), p_{\text{noise}}(\cdot)),\end{aligned} \tag{4.8}$$

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 (4.8), the "regularization by f-divergence" is reflected in the reward function $r_f(\cdot)$. As a consequence, Propositions 3.4–3.6 are easily adapted to solve the control problem (4.8). We summarize the results in the following corollary.

Corollary 4.4. *Define the value-to-go:*

$$v^*_f(t,y) := \max_{u\in\mathcal{A}_t} \mathbb{E}_{Q^{u,y}_{[t,T]}(\cdot)}\left[r_f(Y_T) - \frac{1}{2}\int_t^T \left|\frac{u_s}{\sigma(s)}\right|^2 ds\right]. \tag{4.9}$$

(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:*

$$\frac{\partial v}{\partial t} + +\frac{\sigma^2(t)}{2}\Delta v + \overline{b}(t,y)\cdot\nabla v + \sigma^2(t)|\nabla v|^2 = 0, \quad v(T,y) = r_f(y), \tag{4.10}$$

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

(2) The optimal initial distribution is given by

$$\nu^*_f(x) = \frac{1}{C_f}\exp\left(v^*_f(0,y)\right)p_{noise}(y). \tag{4.11}$$

(3) Let (u_f^*, ν_f^*) *be the optimal solution to the problem* (4.8)*, specified by (1) and (2). The distribution of the process under the optimal control is:*

$$Q_{[0,T]}^{u_f^*,\nu_f^*}(dY_\bullet) = \frac{1}{C_f} \exp\left(r_f(Y_T)\right) Q_{[0,T]}(dY_\bullet), \tag{4.12}$$

and in particular, $Q_T^{u_f^*,\nu_f^*}(\cdot) = p_{ftune}^f(\cdot)$.

Similar to the discussions in Section 3, the optimal control problem can be solved by neural ODEs/SDEs, or be approximated by

$$\begin{aligned} u_f^*(t,y) &= \sigma^2(t)\nabla\left(\ln \mathbb{E}_{Q^y_{[t,T]}(\cdot)}\left[\exp\left(r_f(Y_T)\right)\right]\right) \\ &\approx \sigma^2(t)\nabla\left(\mathbb{E}_{Q^y_{[t,T]}(\cdot)}\left[\ln\left((f'_\pm)^{-1}\left(\frac{r(Y)}{\alpha}\right)\right)\right]\right). \end{aligned} \tag{4.13}$$

The optimal initial distribution $\nu_f^*(\cdot)$ is sampled by solving the control problem:

$$q_f^* := \operatorname{argmax}_q \mathbb{E}_{P^q_{[0,T]}(\cdot)}\left[v_f^*(0,Y_0) - \frac{1}{2}\int_0^T \left|\frac{q_t}{\sigma'(t)}\right|^2 dt\right], \tag{4.14}$$

with $P_T^{q_f^*}(\cdot) = \nu_f^*(\cdot)$.

5. 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 [41] in the context of continuous-time diffusion models. We also generalize to the setting where an f-divergence regularizer is introduced.

There are several directions to extend this work. First, we proved a verification theorem for the value function (Proposition 3.5). It is possible to provide a full characterization using the theory of viscosity solution. Second, we provide an entropy-regularized approach to emulate the fine-tuned distribution (4.5) 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 (4.2) can be generated by solving some stochastic control problem. Finally, the fine-tuned distribution (4.5) and the stochastic control approach (4.8) can be applied to real data such as image synthesis and protein sequence generation.

Acknowledgement: We thank Haoxian Chen and Hanyang Zhao for various pointers to the literature. We thank Xun Yu Zhou for stimulating discussions. W. Tang gratefully acknowledges financial support through NSF grants DMS-2113779 and DMS-2206038, and through a start-up grant at Columbia University.

References

[1] B. D. O. Anderson. Reverse-time diffusion equation models. *Stochastic Process. Appl.*, 12(3):313–326, 1982.
[2] K. Black, M. Janner, Y. Du, I. Kostrikov, and S. Levine. Training diffusion models with reinforcement learning. In *ICLR*, 2024.
[3] S. Boyd and L. Vandenberghe. *Convex optimization.* Cambridge University Press, Cambridge, 2004.
[4] R. V. Chacon and D. S. Ornstein. A general ergodic theorem. *Illinois J. Math.*, 4:153–160, 1960.
[5] N. Chen and P. Glasserman. Malliavin Greeks without Malliavin calculus. *Stochastic Process. Appl.*, 117(11):1689–1723, 2007.
[6] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. Neural ordinary differential equations. In *Neurips*, volume 31, page 6572–6583, 2018.
[7] 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.

[8] 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.
[9] P. Dhariwal and A. Nichol. Diffusion models beat GANs on image synthesis. In *Neurips*, volume 34, pages 8780–8794, 2021.
[10] M. D. Donsker and S. R. S. Varadhan. Asymptotic evaluation of certain Markov process expectations for large time. IV. *Comm. Pure Appl. Math.*, 36(2):183–212, 1983.
[11] Y. Fan and K. Lee. Optimizing DDPM sampling with shortcut fine-tuning. 2023. arXiv:2301.13362.
[12] 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.
[13] E. Fournié, J.-M. Lasry, J. Lebuchoux, P.-L. Lions, and N. Touzi. Applications of Malliavin calculus to Monte Carlo methods in finance. *Finance Stoch.*, 3(4):391–412, 1999.
[14] 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.
[15] E. Gobet and R. Munos. Sensitivity analysis using Itô-Malliavin calculus and martingales, and application to stochastic optimal control. *SIAM J. Control Optim.*, 43(5):1676–1713, 2005.
[16] U. G. Haussmann and E. Pardoux. Time reversal of diffusions. *Ann. Probab.*, 14(4):1188–1205, 1986.
[17] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Neurips*, volume 33, pages 6840–6851, 2020.
[18] J. Ho and T. Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.
[19] A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*, 6:695–709, 2005.
[20] I. Karatzas and S. E. Shreve. *Brownian motion and stochastic calculus*, volume 113 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, second edition, 1991.
[21] P. Kidger, J. Foster, X. C. Li, and T. Lyons. Efficient and accurate gradients for neural SDEs. In *Neurips*, volume 34, pages 18747–18761, 2021.
[22] L. Kong, Y. Du, W. Mu, K. Neklyudov, V. De Bortol, H. Wang, D. Wu, A. Ferber, Y.-A. Ma, C. P. Gomes, et al. Diffusion models as constrained samplers for optimization with unknown constraints. 2024. arXiv:2402.18012.
[23] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *ICLR*, 2021.
[24] 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.
[25] A. S. Luccioni, C. Akiki, M. Mitchell, and Y. Jernite. Stable bias: Analyzing societal representations in diffusion models. In *Neurips*, volume 36, 2023.
[26] OpenAI. Sora: Creating video from text. 2024. Available at `https://openai.com/sora`.
[27] J. Pitman and W. Tang. Patterns in random walks and Brownian motion. In *In memoriam Marc Yor—Séminaire de Probabilités XLVII*, volume 2137 of *Lecture Notes in Math.*, pages 49–88. Springer, Cham, 2015.
[28] Y. Polyanskiy and Y. Wu. *Information Theory: From Coding to Learning*. 2023. Available at `http://www.stat.yale.edu/~yw562/teaching/itbook-export.pdf`.
[29] 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.
[30] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. 2022. arXiv:2204.06125.
[31] 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.
[32] H. Rost. Markoff-Ketten bei sich füllenden Löchern im Zustandsraum. *Ann. Inst. Fourier (Grenoble)*, 21(1):253–270, 1971.
[33] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, volume 32, pages 2256–2265, 2015.
[34] Y. Song, S. Garg, J. Shi, and S. Ermon. Sliced score matching: A scalable approach to density and score estimation. In *UAI*, volume 35, pages 574–584, 2020.
[35] 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.
[36] Z. Song, T. Cai, J. D. Lee, and W. J. Su. Reward collapse in aligning large language models. 2023. arXiv:2305.17608.
[37] D. W. Stroock and S. R. S. Varadhan. *Multidimensional diffusion processes*, volume 233 of *Grundlehren der Mathematischen Wissenschaften*. Springer-Verlag, 1979.
[38] W. Tang, Y. P. Zhang, and X. Y. Zhou. Exploratory HJB equations and their convergence. *SIAM J. Control Optim.*, 60(6):3191–3216, 2022.
[39] W. Tang and H. Zhao. Contractive diffusion probabilistic models. 2024. arXiv:2401.13115.

[40] W. Tang and H. Zhao. Score-based diffusion models via stochastic differential equations–a technical tutorial. 2024. arXiv:2402.07487.

[41] 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.

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

[43] J. von Neumann. Various techniques used in connection with random digits. *Applied Math Series*, 12(36-38):1, 1951.

[44] 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.

[45] 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.

[46] J. Yong and X. Y. Zhou. *Stochastic controls*, volume 43 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, 1999. Hamiltonian systems and HJB equations.

[47] H. Zhao, J. Zhang, X. D. Gu, D. D. Yao, and W. Tang. Score as action: tuning diffusion models by continuous reinforcement learning. 2024+. In preparation.

Department of Industrial Engineering and Operations Research, Columbia University.

Email address: `wt2319@columbia.edu`