

814 Contents

815	A Functionals and Derivation of Gradients of First-order Variations	22
816	A.1 Overview of utilities and divergences in Table 1	22
817	A.2 A brief tutorial on first variation derivation	22
818	A.3 Derivation of gradients of first-order variation for functionals in Table 1	23
819	B Proof for Theorem 5.1	26
820	C Proof for Theorem 5.2	27
821	D Detailed Example of Algorithm Implementation	29
822	D.1 Implementation of ENTROPYREGULARIZEDCONTROLSOLVER	29
823	D.2 Discussion: computational complexity and cost of FDC	29
824	E Experimental Details	31
825	E.1 Used computational resources	31
826	E.2 Experiments in Illustrative Settings	31
827	E.3 Real-World Experiments	31

A Functionals and Derivation of Gradients of First-order Variations

A.1 Overview of utilities and divergences in Table 1

In the following, we report the missing details for the functionals presented within Table 1, and discuss some possible applications.

Manifold Exploration and Generative Model De-biasing As mentioned within Sec. 3, maximization of the entropy functional has been recently introduced as a principled objective for manifold exploration [12]. Moreover, we wish to point out that it can be interpreted also from the viewpoint of de-biasing a prior generative model to re-distribute more uniformly its density while preserving a certain notion of support, e.g., via sufficient KL-divergence regularization.

Risk-averse and Novelty-seeking reward maximization A definition of q_β^r can be found below, explanations of these utilities can be found in Sec. 1, and experimental illustrative examples are provided in Sec. 6.

Optimal Experiment Design The task of Optimal Experimental Design (OED) [7] involves choosing a sequence of experiments so as to minimize some uncertainty metric for an unknown *quantity of interest* $f : \mathcal{X} \rightarrow \mathbb{R}$, where \mathcal{X} is the set of all possible experiments. From a probabilistic standpoint, an optimal design may be viewed as a probability distribution over \mathcal{X} , prescribing how frequently each experiment should be performed to achieve maximal reduction in uncertainty about f [46]. This problem has been recently studied in the case where f is an element of a reproducing kernel Hilbert space (RKHS), i.e., $f \in \mathcal{H}_k$, induced by a known kernel $k(x, x') = \Phi(x)^\top \Phi(x')$ where $x, x' \in \mathcal{X}$ [38]. Given this setting, one might aim to acquire information about f according to different *criteria* captured by the scalarization function $s(\cdot)$ [39]. In particular, in Table 1, we report three illustrative choices for s :

- D-design: $\log \det(\cdot)$ (Information)
- A-design: $-\text{Tr}(\cdot)$ (Parameter error)
- E-design: $\lambda_{\max}(\cdot)$ (Worst projection error)

as reported in previous work [Table 1 39].

Diverse Mode Discovery This objective corresponds to a re-interpretation of the Diverse Skill Discovery objective introduced in the context of Reinforcement Learning [59]. Consider the case where it is given a discrete and finite set \mathcal{S} of symbols interpretable as latent variables, which can be leveraged to (exactly or approximately) perform conditional generation. This objective captures the task of assuring maximal diversity, in terms of KL divergence between the different conditional components, represented as $p^{\pi, k}$ with $k \in \mathcal{S}$.

Log-barrier constrained generation This formulation can be found within the General Utilities RL literature [61]. In particular, here we show the case where constraints are enforced via a log-barrier function, namely $\log(\cdot)$. Nonetheless, the functional presented in Table 1 remains meaningful for general penalty functions.

Optimal transport distances OT distances within Table 1 and their relative notation are introduced below in the context of their first variation computation.

Maximum Mean Discrepancy Here k denotes a positive-definite kernel, which measures similarity between two points in sample space. Moreover, μ_p denotes a kernel mean embedding of distribution p [37]. In terms of applications, choosing a proper kernel k could render possible to preserve specific structure of the initial pre-trained model that would be otherwise lost via KL regularization.

A.2 A brief tutorial on first variation derivation

In this work, we focus on the functionals that are Fréchet differentiable: Let V be a normed spaces. Consider a functional $F : V \rightarrow \mathbb{R}$. There exists a linear operator $A : V \rightarrow \mathbb{R}$ such that the following

873 limit holds

$$\lim_{\|h\|_V \rightarrow 0} \frac{|F(f+h) - F(f) - A[h]|}{\|h\|_V} = 0. \quad (15)$$

874 We further assume that V admits certain structure such that every element in its dual space (the space
875 of bounded linear operator on V) admits some compact representation. For example, when V is the
876 set of compact-supported continuous bounded functions, there exists a unique positive Borel measure
877 μ with the same support, which can be identified as the linear functional. We denote this element as
878 $\delta F[f]$ such that $\langle \delta F[f], h \rangle = A[h]$. Sometimes we also denote it as $\frac{\delta F}{\delta f}$. We will refer to $\delta F[f]$ as
879 the first-order variation of F at f .

880 In this section, we briefly review strategies for deriving the first-order variation of two broad classes
881 of functionals: those defined in closed form with respect to the density (e.g., expectation and entropy)
882 and those defined via variational formulations (e.g., CVaR, Wasserstein distance, and MMD).

883 • **Category 1: Functional defined in a closed form w.r.t. the density.** For this class of functionals,
884 the first-order variations can typically be computed using its definition and chain rule.

885 With definition (15) in mind, we can try to calculate the first-order variation of the mean functional.
886 Consider a continuous and bounded function $r : \mathbb{R}^d \rightarrow \mathbb{R}$ and a probability measure μ on \mathbb{R}^d .
887 Consider the functional $F(\mu) = \int r(x)\mu(x)dx$. We have

$$|F(\mu + \delta\mu) - F(\mu) - \langle r, \delta\mu \rangle| = 0. \quad (16)$$

888 We therefore obtain $\delta F[\mu] = r$ for all μ . We will compute the first-order variations for other
889 functionals in the next subsection.

890 • **Category 2: Functionals defined through a variational formulation.** Another important subclass
891 of functionals considered in this paper is the ones defined via a variational problem

$$F[f] = \sup_{g \in \Omega} G[f, g], \quad (17)$$

892 where Ω is a set of functions or vectors independent of the choice of f , and g is optimized over the
893 set Ω . We will assume that the maximizer $g^*(f)$ that reaches the optimal value for $G[f, \cdot]$ is unique
894 (which is the case for the functionals considered in this project). It is known that one can use the
895 Danskin's theorem (also known as the envelope theorem) to compute

$$\frac{\delta F[f]}{\delta f} = \partial_f G[f, g^*(f)], \quad (18)$$

896 under the assumption that F is differentiable [36].

897 A.3 Derivation of gradients of first-order variation for functionals in Table 1

898 • **Risk-Averse Optimization (Category 2)** Recall that $q_\beta^r(p^\pi) = \sup\{v \in \mathbb{R} | F_Z(v) \leq \beta\}$, where
899 the random variable Z is defined as $Z = r(x)$ with $x \sim p^\pi(x)$. From [49], we have

$$\text{CVaR}_\beta^r(p^\pi) = \mathbb{E}[r(x) | r(x) \leq q_\beta^r(p^\pi)] = \beta \inf_{\zeta} \left\{ \zeta + \frac{1}{\beta} \mathbb{E}[\min\{r(x) - \zeta, 0\}] \right\}.$$

900 Moreover, we have ζ^* that solves the above optimization problem is exactly $\zeta^* = q_\beta^r(p^\pi)$. By
901 Danskin's theorem, one has (in a weak sense)

$$\frac{\delta \text{CVaR}_\beta^r(p^\pi)}{\delta p^\pi} = \beta \min\{r(x) - q_\beta^r(p^\pi), 0\}. \quad (19)$$

902 • **Risk-Seeking Optimization (Category 2)** Recall that $q_\beta^r(p^\pi) = \sup\{v \in \mathbb{R} | F_Z(v) \leq \beta\}$, where
903 the random variable Z is defined as $Z = r(x)$ with $x \sim p^\pi(x)$. From [49], we have

$$\text{SQ}_\beta^r(p^\pi) = \mathbb{E}[r(x) | r(x) \geq q_\beta^r(p^\pi)] = (1 - \beta) \inf_{\zeta} \left\{ \zeta + \frac{1}{1 - \beta} \mathbb{E}[\max\{r(x) - \zeta, 0\}] \right\}.$$

904 Moreover, we have ζ^* that solves the above optimization problem is exactly $\zeta^* = q_\beta^r(p^\pi)$. By
905 Danskin's theorem, one has (in a weak sense)

$$\frac{\delta \text{SQ}_\beta^r(p^\pi)}{\delta p^\pi} = (1 - \beta) \max\{r(x) - q_\beta^r(p^\pi), 0\}. \quad (20)$$

APPLICATION	FUNCTIONAL $\mathcal{F} / \mathcal{D}$	FIRST-ORDER VARIATION	DENSITY CONTROL	
			CONVEX	GENERAL
REWARD OPTIMIZATION [14, 56]	$\mathbb{E}_{x \sim p^\pi} [r(x)]$	r	✓	✓
MANIFOLD EXPLORATION GEN. MODEL DE-BIASING	$\mathcal{H}(p^\pi) := -\mathbb{E}_{x \sim p^\pi} [\log p^\pi(x)]$	$-1 - \log p^\pi$	✓	✓
RISK-AVERSE OPTIMIZATION	$\text{CVaR}_\beta^r(p^\pi) := \mathbb{E}_{x \sim p^\pi} [r(x) \mid r(x) \leq q_\beta^r(p^\pi)]$	$\beta \min\{r(x) - q_\beta^r(p^\pi), 0\}$	✓	✓
	$\mathbb{E}_{x \sim p^\pi} [r(x)] - \text{Var}(p^\pi)$	$r(x) - (r(x)^2 - 2\mathbb{E}_{x \sim p^\pi} [r(x)]r(x))$	✗	✓
RISK-SEEKING OPTIMIZATION	$\text{SQ}_\beta^r(p^\pi) := \mathbb{E}_{x \sim p^\pi} [r(x) \mid r(x) \geq q_\beta^r(p^\pi)]$	$(1 - \beta) \max\{r(x) - q_\beta^r(p^\pi), 0\}$	✗	✓
OPTIMAL EXPERIMENT DESIGN	$s(\mathbb{E}_{x \sim p^\pi} [\Phi(x)\Phi(x)^\top - \lambda \mathbb{I}])$	SEE EQUATION (30)	✓	✓
	$s(\cdot) \in \{\log \det(\cdot), -\text{Tr}(\cdot)^{-1}, -\lambda_{\max}(\cdot)\}$			
DIVERSE MODES DISCOVERY	$-\mathbb{E}_z [D_{KL}(p^{\pi,z} \parallel \mathbb{E}_k p^{\pi,k})]$	SEE EQUATION (32)	✗	✓
LOG-BARRIER CONSTRAINED GENERATION	$\mathbb{E}_{x \sim p^\pi} [r(x)] - \beta \log \langle p^\pi, c \rangle - C$	SEE EQUATION (31)	✓	✓
KULLBACK–LEIBLER DIVERGENCE	$D_{KL}(p^\pi \parallel p^{pre}) = \int p^\pi(x) \log \frac{p^\pi(x)}{p^{pre}(x)} dx$	$1 + \log p^\pi - \log p^{pre}$	✓	✓
RÉNYI DIVERGENCES	$D_\beta(p^\pi \parallel p^{pre}) := \frac{1}{\beta-1} \log \int (p^\pi(x))^\beta (p^{pre}(x))^{1-\beta} dx$	$\frac{\beta}{\beta-1} \left(\int \left(\frac{p}{q}\right)^\beta dq(x) \right)^{-1} \left(\frac{p}{q}\right)^{\beta-1}$	✓	✓
OPTIMAL TRANSPORT DISTANCES	$W_p(p^\pi \parallel p^{pre}) := \inf_{\gamma \in \Gamma(p^\pi, p^{pre})} \mathbb{E}_{(x,y) \sim \gamma} [d(x,y)^p]^{\frac{1}{p}}$	SEE EQUATION (29)	✓	✓
MAXIMUM MEAN DISCREPANCY	$\text{MMD}_k(p^\pi, p^{pre}) := \ \mu_{p^\pi} - \mu_{p^{pre}}\ , \mu_p := \mathbb{E}_{x \sim p} [k(x, \cdot)]$	$\arg \max_{\phi \in \mathcal{H}} \langle \phi, p^\pi - p^{pre} \rangle$	✓	✓

Table 2: Examples of practically relevant utilities \mathcal{F} (blue) and divergences \mathcal{D} (orange), and their first-order variations.

- 906 • **Rényi Divergence (Category 1)** Recall the definition of Rényi Divergence

$$D_\beta(p \parallel q) = \frac{1}{\beta-1} \log \int \left(\frac{p}{q}\right)^\beta dq(x). \quad (21)$$

- 907 We ignore higher-order terms like $O((\delta p)^2)$.

$$D_\beta(p + \delta p \parallel q) - D_\beta(p \parallel q) = \frac{1}{\beta-1} \log \frac{\int \left(\frac{p+\delta p}{q}\right)^\beta dq(x)}{\int \left(\frac{p}{q}\right)^\beta dq(x)} \quad (22)$$

$$= \frac{1}{\beta-1} \log \frac{\int \left(\frac{p}{q}\right)^\beta + \beta \left(\frac{p}{q}\right)^{\beta-1} \frac{\delta p}{q} dq(x)}{\int \left(\frac{p}{q}\right)^\beta dq(x)} \quad (23)$$

$$= \frac{1}{\beta-1} \log 1 + \frac{\int \beta \left(\frac{p}{q}\right)^{\beta-1} \frac{\delta p}{q} dq(x)}{\int \left(\frac{p}{q}\right)^\beta dq(x)} \quad (24)$$

$$= \frac{1}{\beta-1} \frac{\int \beta \left(\frac{p}{q}\right)^{\beta-1} \frac{\delta p}{q} dq(x)}{\int \left(\frac{p}{q}\right)^\beta dq(x)} \quad (25)$$

908

$$\frac{\delta}{\delta p} R_\beta(p, q) = \frac{\beta}{\beta-1} \left(\int \left(\frac{p}{q}\right)^\beta dq(x) \right)^{-1} \left(\frac{p}{q}\right)^{\beta-1} \quad (26)$$

- 909 • **Optimal transport and Wasserstein-p distance (Category 2)** Consider the optimal transport
910 problem

$$\text{OT}_c(u, v) = \inf_\gamma \left\{ \int \int c(x, y) d\gamma(x, y) : \int \gamma(x, y) dx = u(y), \int \gamma(x, y) dy = v(x) \right\} \quad (27)$$

911 where

$$\Gamma = \left\{ \gamma : \int \gamma(x, y) dx = u(y), \int \gamma(x, y) dy = v(x) \right\}$$

912 It admits the following equivalent dual formulation

$$\text{OT}_c(u, v) = \sup_{f, g} \left\{ \int f du + \int g dv : f(x) + g(y) \leq c(x, y) \right\} \quad (28)$$

913 By taking $c(x, y) = \|x - y\|^p$, we recover $\text{OT}_c(u, v) = W_p(u, v)^p$. Let f^* and g^* be the solution
914 to the above dual optimization problem. From the Danskin's theorem, we have

$$\frac{\delta}{\delta u} W_p(u, v)^p = f^*. \quad (29)$$

915 In the special case of $p = 1$, we know that $g^* = -f^*$ (note that the constraint can be equivalently
916 written as $\|\nabla f\| \leq 1$), in which case f^* is typically known as the critic in the WGAN framework.

917 • **Optimal Experiment Design. (Category 1)** We take $s(M) = \log \det(M)$ as example. By chain
918 rule, we have

$$\delta F[p^\pi] = \text{Tr} \left[\left(\mathbb{E}_{x \sim p^\pi} [\Phi(x) \Phi(x)^\top - \lambda \mathbb{I}] \right)^{-1} (\Phi(x) \Phi(x)^\top - \lambda \mathbb{I}) \right]. \quad (30)$$

919 • **Log-Barrier Constrained Generation. (Category 1)** By chain rule, we obtain

$$\delta F[p^\pi] = r - \frac{\beta c}{\langle p^\pi, c \rangle - C}. \quad (31)$$

920 • **Diverse modes discovery. (Category 1)** By chain rule, we obtain

$$\begin{aligned} \frac{\delta F}{\delta p^{\pi, z}} &= -\frac{\delta}{\delta p^{\pi, z}} \mathbb{E}_z \left[\int p^{\pi, z} \log p^{\pi, z} dx - \int p^{\pi, z} \log (\mathbb{E}_k[p^{\pi, k}]) dx \right] \\ &= -\mathbb{E}_z \left[\frac{\delta}{\delta p^{\pi, z}} \left(\int p^{\pi, z} \log p^{\pi, z} dx \right) - \frac{\delta}{\delta p^{\pi, z}} \left(\int p^{\pi, z} \log (\mathbb{E}_k[p^{\pi, k}]) dx \right) \right] \\ &= -\mathbb{E}_z \left[\log p^{\pi, z} + 1 - \log (\mathbb{E}_k[p^{\pi, k}]) - \frac{p^{\pi, z}}{\mathbb{E}_k[p^{\pi, k}]} \right] \end{aligned} \quad (32)$$

921 • **Entropy. (Category 1)** As a first example, consider the entropy functional $\mathcal{F}(p) = -\int p \log p, dx$.
922 By the definition of the first-order variation, we have $\frac{\delta \mathcal{F}}{\delta p}(p) = -1 - \log p$, and therefore $\nabla \frac{\delta \mathcal{F}}{\delta p}(p) =$
923 $-\nabla \log p$. This gradient term can be effectively estimated using standard score approximations;
924 see [12].

B Proof for Theorem 5.1

Theorem 5.1 (Convergence guarantee of Flow Density Control with concave functionals). *Given Assumptions 5.1, fine-tuning a pre-trained model π^{pre} via FDC (Algorithm 1) with $\eta_k = L \forall k \in [K]$, leads to a policy π inducing a marginal distribution p_1^π such that:*

$$\mathcal{G}(p_1^*) - \mathcal{G}(p_1^\pi) \leq \frac{L-l}{K} D_{KL}(p_1^* \| p_1^{pre}) \quad (12)$$

where $p_1^* := p_1^{\pi^*}$ is the marginal distribution induced by the optimal policy $\pi^* \in \arg \max_\pi \mathcal{G}(p_1^\pi) := \mathcal{F}(p_1^\pi) - \alpha \mathcal{D}(p_1^\pi \| p_1^{pre})$.

Proof. We prove this result using the framework of relative smoothness and relative strong convexity introduced in Section 5.

The analysis is based on the classical mirror descent framework under relative properties [33]. For notational simplicity, we let $\mu_k := p_T^{\pi_k}$, and fix an arbitrary reference density $\mu \in \mathbb{P}(\Omega_{pre})$. To better align the notation of our theory with existing literature, we will proceed with the *convex* functional $\tilde{\mathcal{G}} := -\mathcal{G}$ below.

We begin by showing the following inequality:

$$\tilde{\mathcal{G}}(\mu_k) \leq \tilde{\mathcal{G}}(\mu_{k-1}) + \langle \delta \tilde{\mathcal{G}}(\mu_{k-1}), \mu_k - \mu_{k-1} \rangle + LD_{\mathcal{Q}}(\mu_k, \mu_{k-1}) \quad (33)$$

$$\leq \tilde{\mathcal{G}}(\mu_{k-1}) + \langle \delta \tilde{\mathcal{G}}(\mu_{k-1}), \mu - \mu_{k-1} \rangle + LD_{\mathcal{Q}}(\mu, \mu_{k-1}) - LD_{\mathcal{Q}}(\mu, \mu_k). \quad (34)$$

The first inequality follows from the L -smoothness of \mathcal{G} relative to \mathcal{Q} as defined in Definition 1. The second inequality uses the three-point inequality of the Bregman divergence [33, Lemma 3.1] with $\phi(\mu) = \frac{1}{L} \langle \delta \mathcal{G}(\mu_{k-1}), \mu - \mu_{k-1} \rangle$, $z = \mu_{k-1}$, and $z^+ = \mu_k$.

Next, using the l -strong concavity of \mathcal{G} relative to \mathcal{Q} , again from Definition 1, we obtain:

$$\tilde{\mathcal{G}}(\mu_k) \leq \tilde{\mathcal{G}}(\mu) + (L-l)D_{\mathcal{Q}}(\mu, \mu_{k-1}) - LD_{\mathcal{Q}}(\mu, \mu_k). \quad (35)$$

By recursively applying the above inequality and using the monotonicity of $\mathcal{G}(\mu_k)$ along with the non-negativity of the Bregman divergence, we obtain [33]:

$$\sum_{k=1}^K \left(\frac{L}{L-l} \right)^k \left(\tilde{\mathcal{G}}(\mu_k) - \tilde{\mathcal{G}}(\mu) \right) \leq LD_{\mathcal{Q}}(\mu, \mu_0) - L \left(\frac{L}{L-l} \right)^K D_{\mathcal{Q}}(\mu, \mu_K) \leq LD_{\mathcal{Q}}(\mu, \mu_0). \quad (36)$$

Letting

$$\frac{1}{C_K} := \sum_{k=1}^K \left(\frac{L}{L-l} \right)^k, \quad (37)$$

and rearranging terms, we arrive at the convergence rate:

$$\tilde{\mathcal{G}}(\mu_K) - \tilde{\mathcal{G}}(\mu) \leq C_K LD_{\mathcal{Q}}(\mu, \mu_0) = \frac{lD_{\mathcal{Q}}(\mu, \mu_0)}{\left(1 + \frac{l}{L-l}\right)^K - 1}. \quad (38)$$

Finally, the convergence rate stated in the theorem follows by observing that $\left(1 + \frac{l}{L-l}\right)^K \geq 1 + \frac{Kl}{L-l}$.

□

C Proof for Theorem 5.2

To establish our main convergence result, we introduce two additional technical assumptions that are satisfied in virtually all practical settings:

Assumption C.1 (Support Compatibility). *We assume that the support of $p_T^{\pi_k}$ is contained in a fixed compact domain $\tilde{\Omega}$ for all k , and that for some j , we have $\text{supp}(p_j^{\pi_k}) = \tilde{\Omega}$.*

Assumption C.2 (Precompactness). *The sequence $\{\delta\mathcal{H}(p_T^{\pi_k})\}_k$ is precompact in the topology induced by the L_∞ norm.*

We are now ready to present the full proof. For the reader's convenience, we restate the theorem:

Theorem 5.2 (Convergence guarantee of Flow Density Control for general functionals). *Given the Robbins-Monro step-size rule: $\sum_k \gamma_k = \infty$, $\sum_k \gamma_k^2 < \infty$, under Assumption 5.2 and technical assumptions (see Appendix C), the sequence of marginal densities p_1^k induced by the iterates π_k of Algorithm 1 converges weakly to a stationary point \tilde{p}_1 of \mathcal{G} almost surely, formally: $p_1^k \rightharpoonup \tilde{p}_1$ a.s..*

Proof. To facilitate readability, we begin with an outline of the key steps.

Proof Outline The main idea is to relate the discrete iterates $\{p_T^k\}_{k \in \mathbb{N}}$ produced by Algorithm 1 to a continuous-time dynamical system. Let us define the initial dual variable as:

$$h_0 = \delta\mathcal{H}(p_{pre}) = -\log p_{pre},$$

and consider the following gradient flow:

$$\begin{cases} \dot{h}_t = \delta\mathcal{G}(p_t), \\ \dot{p}_t = \delta(-\mathcal{H})^*(h_t), \end{cases} \quad (\text{MF})$$

where $(-\mathcal{H})^*(h) = \log \int_{\Omega} e^h$ is the Fenchel dual of the negative entropy functional [25, 22].

To connect this with our algorithm, we construct a continuous-time interpolation of the dual iterates $h^k = \delta\mathcal{H}(p_T^{\pi_k})$. Define the effective time $\tau^k = \sum_{r=0}^k \alpha_r$, and let the interpolated process $h(t)$ be given by:

$$h(t) = h^k + \frac{t - \tau^k}{\tau + 1^k - \tau^k} (\tau + 1^k - h^k). \quad (\text{Int})$$

Intuitively, our convergence result follows if two conditions hold:

Informal Assumption 1 (Closeness to Continuous-Time Flow). *The interpolated process $h(t)$ asymptotically follows the dynamics of (MF) as $k \rightarrow \infty$.*

Informal Assumption 2 (Convergence of the Flow). *The trajectories of (MF) converge to a stationary point of \mathcal{G} .*

To formalize this, we invoke the stochastic approximation framework of [5]. Let \mathcal{Z} be the space of integrable functions on Ω , and let Θ denote the flow of (MF). We define:

Definition 2 (Asymptotic Pseudotrajectory (APT)). *We say $h(t)$ is an asymptotic pseudotrajectory (APT) of (MF) if for all $T > 0$,*

$$\lim_{t \rightarrow \infty} \sup_{0 \leq h \leq T} \|h(t+h) - \Theta_h(h(t))\|_\infty = 0.$$

If $h(t)$ is a precompact APT, then [5] show:

Theorem C.1 (APT Limit Set Theorem). *Let $h(t)$ be a precompact APT for the flow (MF). Then, almost surely, the limit set of $h(t)$ is contained in the set of internally chain-transitive (ICT) points of (MF).*

The proof of our result follows from two claims:

1. The iterates $\{h^k\}$ generate a precompact APT under Assumptions C.1 and 5.2.

979 2. The ICT set of (MF) consists only of stationary points of \mathcal{G} .

980 The second claim holds because (MF) is a gradient flow—specifically, the spherical
981 Hellinger–Kantorovich flow [35]. By Sard’s theorem and standard results in dynamical systems [5],
982 the ICT set must consist of stationary points.

983 For the first claim, Assumptions C.1 and C.2 ensure that the interpolated process is well-defined and
984 precompact, while Assumption 5.2 allows us to apply standard stochastic approximation arguments
985 [27]. We conclude the proof by applying Theorem C.1. \square

D Detailed Example of Algorithm Implementation

D.1 Implementation of ENTROPYREGULARIZEDCONTROLSOLVER

To ensure completeness, below we provide pseudocode for one concrete realization of a ENTROPYREGULARIZEDCONTROLSOLVER as in Eq. (8) using a first-order optimization routine. In particular, we describe exactly the version employed in Sec. 6, which builds on the Adjoint Matching framework [14], casting linear fine-tuning as a stochastic optimal control problem and tackling it via regression.

Let u^{pre} be the initial, pre-trained vector field, and $u^{finetuned}$ its fine-tuned counterpart. We also use $\bar{\alpha}$ to refer to the accumulated noise schedule from [23] effectively following the flow models notation introduced by Adjoint Mathing [14, Sec. 5.2]. The full procedure is in Algorithm 2.

Algorithm 2 ENTROPYREGULARIZEDCONTROLSOLVER (Adjoint Matching [14]) based implementation

- 1: **Input:** N : number of iterations, u^{pre} : pre-trained flow vector field, η regularization coefficient as in Eq. (8), h : step size, ∇f : reward function gradient, m batch size
- 2: **Init:** $u^{finetuned} := u^{pre}$ with parameter θ
- 3: **for** $n = 0, 1, 2, \dots, N - 1$ **do**
- 4: Sample m trajectories $\{X_t\}_{t=1}^T$ via memoryless noise schedule [14], e.g.,

sample $\epsilon_t \sim \mathcal{N}(0, I)$, $X_0 \sim \mathcal{N}(0, I)$, then:

$$X_{t+h} = X_t + h \left(2v_{\theta}^{finetuned}(X_t, t) - \frac{\bar{\alpha}_t}{\alpha_t} X_t \right) + \sqrt{h}\sigma(t)\epsilon_t$$

Use reward gradient:

$$\tilde{a}_T = -\frac{1}{\eta} \nabla f(X_1)$$

For each trajectory, solve the lean adjoint ODE, see [14, Eq. 38-39], from $t = 1$ to 0, e.g.,:

$$\tilde{a}_{t-h} = \tilde{a}_t + h \tilde{a}_t^\top \nabla_{X_t} \left(2u^{pre}(X_t, t) - \frac{\bar{\alpha}_t}{\alpha_t} X_t \right)$$

Where X_t and \tilde{a}_t are computed without gradients, i.e., $X_t = \text{stopgrad}(X_t)$, $\tilde{a}_t = \text{stopgrad}(\tilde{a}_t)$. For each trajectory compute the Adjoint Matching objective [14, Eq. 37]:

$$\mathcal{L}_{\theta} = \sum_{t=0}^{1-h} \left\| \frac{2}{\sigma(t)} \left(u_{\theta}^{finetuned}(X - t, t) - u^{pre}(X_t, t) \right) + \sigma(t) \tilde{a}_t \right\|$$

Compute the gradient $\nabla_{\theta} \mathcal{L}(\theta)$ and update θ .

5: **end for**

6: **output:** Fine-tuned noise predictor $u_{\theta}^{finetuned}$

D.2 Discussion: computational complexity and cost of FDC

Flow Density Control (see Algorithm 1) is a sequential fine-tuning scheme, which performs K iterations of a base fine-tuning oracle, as shown in Algorithm 1. Typically, as for the case of Adjoint Matching [14], which is contextualized in Algorithm 2, the inner oracle also performs N iterations to solve the classic fine-tuning problem. As a consequence, at first glance, this lead to FDC having a computational complexity scaling linearly in K the one of classic fine-tuning. Nonetheless, this does not seem to capture well the practical computational cost. In particular, we wish to point out the two following observations:

- As discussed for the molecular design experiment in Sec. 6 and further in Appendix E, the FDC scheme might work well even with a very approximate oracle to solve the entropy-regularized control problem at each iteration.
- For many real-world problems a very small number of iterations K might be sufficient to approximate the non-linear functional sufficiently well and hence obtain useful fine-tuned

1008 models. This is shown in text-to-image bridge design experiment in Sec. 6 and in Appendix
1009 E. In this case, merely $K = 2$ iterations of FDC lead to promising results.

1010 E Experimental Details

1011 E.1 Used computational resources

1012 We run all experiments on a single Nvidia H100 GPU.

1013 E.2 Experiments in Illustrative Settings

1014 **Shared experimental setup.** For all illustrative experiments we utilize Adjoint Matching (AM) [14]
1015 for the entropy-regularized fine-tuning solver in Algorithm 1. Moreover, the stochastic gradient steps
1016 within the AM scheme are performed via an Adam optimizer.

1017 **Risk-averse reward maximization for better worst-case validity or safety.** In this experiment,
1018 we execute FDC for $K = 2$ iterations with a total of 1000 gradient steps within each iteration, AM
1019 solver (within the FDC scheme) with learning rate of $2e^{-2}$, $\alpha = 10^9$, and $\eta = 10$. Meanwhile,
1020 the AM baseline, is run for 1000 gradient steps with $\alpha = 0.2857$, and learning rate of $1e^{-5}$. The
1021 resulting CVaR is computed via the standard torch quantile method. The values of β reported in the
1022 main paper effectively refers to the value of $1 - \beta$.

1023 **Novelty-seeking reward maximization for discovery.** We run FDC for $K = 2$ iterations with a
1024 total of 1000 gradient steps within each iteration, AM solver (within the FDC scheme) with learning
1025 rate of $3e^{-6}$, $\alpha = 10^5$, and $\eta = 0.625$, and 8000 samples are used to estimate the first variation
1026 gradient as explained in Appendix A. Meanwhile, the AM baseline, is run for 1000 gradient steps
1027 with $\alpha = 0.333$, and learning rate of $1e^{-5}$. The resulting SQ is computed via the standard torch
1028 quantile method.

1029 **Reward maximization regularized via optimal transport distance.** Within this experiment,
1030 we present two runs of FDC, namely FDC-A and FDC-B, compared against AM. Both FDC-A and
1031 FDC-B have been run for $K = 6$ iterations of FDC, with $\alpha = 0.1$, AM oracle learning rate of $1e^{-6}$,
1032 $\eta = 6.666$. Both their discriminators to solve the dual OT problem as presented in Appendix A and
1033 mentioned within Sec. 4, have been learned via a simple MLP architecture with 800 gradient steps,
1034 by enforcing the 1-Lip. condition via the standard gradient penalty technique with regularization
1035 strength of $\lambda_{GP} = 10.0$ and learning rate of $1e^{-4}$. In particular, FDC-A is based on the distance
1036 defined, for two 2-dimensional points $x = (x_1, x_2)$ and $y = (y_1, y_2)$ by:

$$d_A(x, y) = \sqrt{(x_1 - y_1)^2 + (K(x_2 - y_2))^2}$$

1037 Analogously, FDC-B leverages d_B defined as:

$$d_B(x, y) = \sqrt{(K(x_1 - y_1))^2 + (x_2 - y_2)^2}$$

1038 Where $K = 7$ in both cases. On the other hand, the AM baseline is run for 1000 gradient steps with
1039 learning rate of $1e^{-3}$ and $\alpha = 1.538$.

1040 **Conservative manifold exploration.** We ran FDC for $K = 50$ iterations and 2500 gradient steps
1041 in total with $\eta = 10$ and $\alpha = 0.0, 0.01, 0.1, 0.5, 1.0$. We set the AM learning rate to $2e^{-4}$ and sample
1042 trajectories of length 400 for computing the AM loss.

1043 E.3 Real-World Experiments

1044 **Molecular design for single-point energy minimization.** In this experiment FDC is run for
1045 $K = 10$ iterations, with merely 2 gradient steps at each iteration (i.e., the AM oracle is very
1046 approximate), AM learning rate of $1e^{-4}$, $\eta = 0.01$ and $\alpha = 0$. Meanwhile, the AM baseline is run
1047 for 240 gradient steps with $\alpha = 0.0045$.

1048 **Text-to-image bridge designs conservative exploration.** For this experiment we ran FDC on
1049 a single Nvidia H100 GPU, with $K = 2$, $\eta = 200$, $\alpha = 0.001$ and a 100 gradient steps in total.
1050 Similarly to previous work, we tuned the vector field resulting from applying classifier-free guidance
1051 with guidance scale $w = 8$ in SD-1.5.

References

- [1] Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022.
- [2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017.
- [3] Pierre-Cyril Aubin-Frankowski, Anna Korba, and Flavien Léger. Mirror descent with relative smoothness in measure spaces, with application to sinkhorn and em. *Advances in Neural Information Processing Systems*, 35:17263–17275, 2022.
- [4] Anas Barakat, Ilyas Fatkhullin, and Niao He. Reinforcement learning with general utilities: Simpler variance reduction and large state-action space. In *International Conference on Machine Learning*, pages 1753–1800. PMLR, 2023.
- [5] Michel Benaïm. Dynamics of stochastic approximation algorithms. In *Seminaire de probabilités XXXIII*, pages 1–68. Springer, 2006.
- [6] Camille Bilodeau, Wengong Jin, Tommi Jaakkola, Regina Barzilay, and Klavs F Jensen. Generative models for molecular discovery: Recent advances and challenges. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 12(5):e1608, 2022.
- [7] Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, pages 273–304, 1995.
- [8] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *arXiv preprint arXiv:2303.04137*, 2023.
- [9] Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking. *arXiv preprint arXiv:2210.01776*, 2022.
- [10] Riccardo De Santi, Federico Arangath Joseph, Noah Liniger, Mirco Mutti, and Andreas Krause. Geometric active exploration in markov decision processes: the benefit of abstraction. *arXiv preprint arXiv:2407.13364*, 2024.
- [11] Riccardo De Santi, Manish Prajapat, and Andreas Krause. Global reinforcement learning: Beyond linear and convex rewards via submodular semi-gradient methods. *arXiv preprint arXiv:2407.09905*, 2024.
- [12] Riccardo De Santi, Marin Vlastelica, Ya-Ping Hsieh, Zebang Shen, Niao He, and Andreas Krause. Provable maximum entropy manifold exploration via diffusion models. In *ICLR 2025 Workshop on Deep Generative Model in Machine Learning: Theory, Principle and Efficacy*.
- [13] Alexander Decruyenaere, Heidelinde Dehaene, Paloma Rabaey, Johan Decruyenaere, Christiaan Polet, Thomas Demeester, and Stijn Vansteelandt. Debiasing synthetic data generated by deep generative models. *Advances in Neural Information Processing Systems*, 37:41539–41576, 2024.
- [14] Carles Domingo-Enrich, Michal Drozdal, Brian Karrer, and Ricky TQ Chen. Adjoint matching: Fine-tuning flow and diffusion generative models with memoryless stochastic optimal control. *arXiv preprint arXiv:2409.08861*, 2024.
- [15] Ian Dunn and David Ryan Koes. Mixed continuous and categorical flow matching for 3d de novo molecule generation. *ArXiv*, pages arXiv–2404, 2024.
- [16] Pavel Dvurechensky and Jia-Jie Zhu. Analysis of kernel mirror prox for measure optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2350–2358. PMLR, 2024.
- [17] Jesse Farebrother, Matteo Pirota, Andrea Tirinzoni, Rémi Munos, Alessandro Lazaric, and Ahmed Touati. Temporal difference flows. *arXiv preprint arXiv:2503.09817*, 2025.

- [18] Marvin Friede, Christian Hölzer, Sebastian Ehlert, and Stefan Grimme. dxtb—an efficient and fully differentiable framework for extended tight-binding. *The Journal of Chemical Physics*, 161(6), 2024.
- [19] Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022.
- [20] Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. In *International Conference on Machine Learning*, 2019.
- [21] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, 2021.
- [22] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis*. Springer Science & Business Media, 2004.
- [23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [24] Emiel Hoogeboom, Victor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pages 8867–8887. PMLR, 2022.
- [25] Ya-Ping Hsieh, Chen Liu, and Volkan Cevher. Finding mixed nash equilibria of generative adversarial networks. In *International Conference on Machine Learning*, pages 2810–2819. PMLR, 2019.
- [26] Yanwei Jia and Xun Yu Zhou. Policy evaluation and temporal-difference learning in continuous time and space: A martingale approach. *Journal of Machine Learning Research*, 23(154):1–55, 2022.
- [27] Mohammad Reza Karimi, Ya-Ping Hsieh, and Andreas Krause. Sinkhorn flow as mirror flow: A continuous-time framework for generalizing the sinkhorn algorithm. In *International Conference on Artificial Intelligence and Statistics*, pages 4186–4194. PMLR, 2024.
- [28] Flavien Léger. A gradient descent perspective on sinkhorn. *Applied Mathematics & Optimization*, 84(2):1843–1855, 2021.
- [29] Yingzhen Li and Richard E Turner. Rényi divergence variational inference. *Advances in neural information processing systems*, 29, 2016.
- [30] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.
- [31] Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky TQ Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. Flow matching guide and code. *arXiv preprint arXiv:2412.06264*, 2024.
- [32] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.
- [33] Haihao Lu, Robert M Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.
- [34] Panayotis Mertikopoulos, Ya-Ping Hsieh, and Volkan Cevher. A unified stochastic approximation framework for learning in games. *Mathematical Programming*, 203(1):559–609, 2024.
- [35] Alexander Mielke and Jia-Jie Zhu. Hellinger-kantorovich gradient flows: Global exponential decay of entropy functionals. *arXiv preprint arXiv:2501.17049*, 2025.
- [36] Paul Milgrom and Ilya Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70(2):583–601, 2002.

- [37] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.
- [38] Mojmír Mutný. *Modern Adaptive Experiment Design: Machine Learning Perspective*. PhD thesis, ETH Zurich, 2024.
- [39] Mojmír Mutný, Tadeusz Janik, and Andreas Krause. Active exploration via experiment design in Markov chains. In *International Conference on Artificial Intelligence and Statistics*, 2023.
- [40] Mirco Mutti, Riccardo De Santi, Piersilvio De Bartolomeis, and Marcello Restelli. Challenging common assumptions in convex reinforcement learning. *Advances in Neural Information Processing Systems*, 35:4489–4502, 2022.
- [41] Mirco Mutti, Riccardo De Santi, Piersilvio De Bartolomeis, and Marcello Restelli. Convex reinforcement learning in finite trials. *Journal of Machine Learning Research*, 24(250):1–42, 2023.
- [42] Mirco Mutti, Riccardo De Santi, and Marcello Restelli. The importance of non-markovianity in maximum state entropy exploration. In *International Conference on Machine Learning*, pages 16223–16239. PMLR, 2022.
- [43] Arkadij Semenovič Nemirovskij and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.
- [44] Kushagra Pandey, Jaideep Pathak, Yilun Xu, Stephan Mandt, Michael Pritchard, Arash Vahdat, and Morteza Mardani. Heavy-tailed diffusion models. *arXiv preprint arXiv:2410.14171*, 2024.
- [45] Manish Prajapat, Mojmír Mutný, Melanie N Zeilinger, and Andreas Krause. Submodular reinforcement learning. *arXiv preprint arXiv:2307.13372*, 2023.
- [46] Friedrich Pukelsheim. *Optimal design of experiments*. SIAM, 2006.
- [47] Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data*, 1(1):1–7, 2014.
- [48] R Tyrrell Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal of banking & finance*, 26(7):1443–1471, 2002.
- [49] R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.
- [50] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.
- [51] Marta Skreta, Lazar Atanackovic, Avishek Joey Bose, Alexander Tong, and Kirill Neklyudov. The superposition of diffusion models using the $\hat{\pi}$ density estimator. *arXiv preprint arXiv:2412.17762*, 2024.
- [52] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.
- [53] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.
- [54] Wenpin Tang. Fine-tuning of diffusion models via stochastic control: entropy regularization and beyond. *arXiv preprint arXiv:2403.06279*, 2024.
- [55] Lenart Treven, Jonas Hübötter, Bhavya Sukhija, Florian Dorfler, and Andreas Krause. Efficient exploration in continuous-time model-based reinforcement learning. *Advances in Neural Information Processing Systems*, 36:42119–42147, 2023.

- 457 [56] Masatoshi Uehara, Yulai Zhao, Kevin Black, Ehsan Hajiramezanali, Gabriele Scalia,
458 Nathaniel Lee Diamant, Alex M Tseng, Tommaso Biancalani, and Sergey Levine. Fine-
459 tuning of continuous-time diffusion models as entropy-regularized control. *arXiv preprint*
460 *arXiv:2402.15194*, 2024.
- 461 [57] Masatoshi Uehara, Yulai Zhao, Kevin Black, Ehsan Hajiramezanali, Gabriele Scalia,
462 Nathaniel Lee Diamant, Alex M Tseng, Sergey Levine, and Tommaso Biancalani. Feedback
463 efficient online fine-tuning of diffusion models. *arXiv preprint arXiv:2402.16359*, 2024.
- 464 [58] Haoran Wang, Thaleia Zariphopoulou, and Xun Yu Zhou. Reinforcement learning in continuous
465 time and space: A stochastic control approach. *Journal of Machine Learning Research*,
466 21(198):1–34, 2020.
- 467 [59] Tom Zahavy, Brendan O’Donoghue, Guillaume Desjardins, and Satinder Singh. Reward is
468 enough for convex mdps. *Advances in Neural Information Processing Systems*, 34:25746–25759,
469 2021.
- 470 [60] Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Sasha
471 Shysheya, Jonathan Crabbé, Lixin Sun, Jake Smith, et al. Mattergen: a generative model for
472 inorganic materials design. *arXiv preprint arXiv:2312.03687*, 2023.
- 473 [61] Junyu Zhang, Alec Koppel, Amrit Singh Bedi, Csaba Szepesvari, and Mengdi Wang. Variational
474 policy gradient method for reinforcement learning with general utilities. In H. Larochelle,
475 M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information*
476 *Processing Systems*, volume 33, pages 4572–4583. Curran Associates, Inc., 2020.
- 477 [62] Hanyang Zhao, Haoxian Chen, Ji Zhang, David D Yao, and Wenpin Tang. Scores as actions: a
478 framework of fine-tuning diffusion models by continuous-time reinforcement learning. *arXiv*
479 *preprint arXiv:2409.08400*, 2024.