2021 60th IEEE Conference on Decision and Control (CDC)
December 13-15, 2021. Austin, Texas

2021 60th IEEE Conference on Decision and Control (CDC) | 978-1-6654-3659-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/CDC45484.2021.9683159

Sampling, variational Bayesian inference, and conditioned stochastic differential equations

Todd P. Coleman and Maxim Raginsky

***Abstract*— We consider the problem of sampling and statistical inference in probabilistic generative models, where the latent object is a finite-dimensional diffusion process. In general, it is difficult to obtain exact expressions for the log-likelihood, so one has to resort to so-called variational inference, where a change of measure is used to come up with a tractable upper bound. We first show, using W. Fleming's logarithmic transformation, that the problem of constructing a variational approximation to the log-likelihood can be interpreted as an optimal control problem, where the choice of a variational approximation amounts to adding a drift to the original diffusion. We then analyze this class of control problems using the formalism of conditioned stochastic differential equations due to F. Baudoin. We discuss the relation of this problem to entropic optimal transport and to the stochastic maximum principle.**

I. Introduction

The term 'probabilistic generative model' refers to any process by which a sample from a target probability measure μ on $\mathbb{R}^n$ is produced by applying a deterministic transformation G to a sample W from a fixed probability measure P on some latent space. Typically, P is relatively simple, such as the canonical Gaussian measure on $\mathbb{R}^d$, and the mapping G has some internal parameters that can be tuned to ensure that the pushforward measure G_*P is (approximately) equal to μ. Thus, if some class $\mathcal{G}$ of admissible transformations is given (e.g., those implementable by a feedforward neural net with some constraints on width, depth, or weights), then we seek $G \in \mathcal{G}$ with the best trade-off between fidelity (i.e., how close G_*P is to μ) and complexity (e.g., how far G is from the identity map if $d = n$).

In modern machine learning applications, where the target measure μ is typically supported on a space of high dimensionality, it is customary to use so-called *implicit* generative models, where closed-form expressions for μ or G_*P are unavailable, but one can readily perform some optimization procedure, such as gradient descent, over the internal parameters of G. For example, in deep latent Gaussian models [1], [2], X is generated recursively as

$$V_0 = W_0, \quad V_j = g_j(V_{j-1}, W_j, \theta_j),\ j = 1, \ldots, \ell, \quad X = V_\ell$$

where $W_0, \ldots, W_\ell$ are independent Gaussian random vectors and $g_j(\cdot, \cdot, \theta_j)$ is a given sequence of transformations with tunable internal parameters θ_j. The overall transformation can then be written as $X = G(W; \theta)$ for a suitable mapping $G(\cdot; \cdot)$, where $W = (W_0, \ldots, W_\ell)$ is the latent Gaussian random vector and $\theta = (\theta_1, \ldots, \theta_\ell)$ is the vector of parameters. Then, given a suitable description of the target distribution μ (e.g., via independent and identically distributed samples), one can attempt to approximate it by the pushforward measure $G(\cdot; \theta)_*P$, where P is the probability law of W, with an appropriate choice of θ.

In this paper, we will consider a generalization of models of this type, where the role of the latent object is played by the d-dimensional standard Brownian motion $W = (W_t)_{t \in [0,1]}$, and G is a well-behaved map from the space of continuous paths $C([0,1]; \mathbb{R}^d)$ into $\mathbb{R}^n$. These models, recently introduced under the name of *neural stochastic differential equations* [3], [4], are attractive due to their expressiveness (i.e., ability to generate samples from a broad class of target probability measures), and can be trained efficiently using gradient descent with backpropagation [5]. While both sampling and inference in such models can be viewed through the lens of optimal stochastic control of diffusion processes [3], our goal here is to explore these control-theoretic aspects further. In particular, we examine the structure of optimal controls via the complementary perspectives of dynamic programming [6] and the stochastic maximum principle [7], as well as outline an approach to the construction of suboptimal yet computationally tractable controls inspired by the work of Beneš [8] on finite-dimensional nonlinear filters.

II. A finite-dimensional analogue

Some of the underlying ideas can already be seen in the simpler finite-dimensional setting. Let P be the canonical Gaussian measure on $\mathbb{R}^d$, and let a smooth function $F : \mathbb{R}^d \to \mathbb{R}^n$ be given. Let μ_0 denote the pushforward measure F_*P, such that, for any bounded measurable function $h : \mathbb{R}^n \to \mathbb{R}$,

$$\mathbf{E}[h(X)] = \int_{\mathbb{R}^n} h(x)\mu_0(\mathrm{d}x) = \int_{\mathbb{R}^d} h \circ F(w)P(\mathrm{d}w).$$

Now let some target probability measure μ on $\mathbb{R}^n$ be given. Then we have the following (cf. also Proposition 3 in [9]):

Proposition 1. *There exists a unique probability measure P^μ on $\mathbb{R}^d$, such that:*

1) $\mathbf{E}^\mu[h(W)|X] = \mathbf{E}[h(W)|X]$ *for any bounded measurable* $h : \mathbb{R}^d \to \mathbb{R}$.
2) $F_*P^\mu = \mu$.

Explicitly, P^μ can be disintegrated as $P^\mu(A) = \int_{\mathbb{R}^n} P^x(A)\mu(\mathrm{d}x)$, where P^x denotes the (regular) conditional

This research was supported in part by the NSF CAREER award CCF-1254041, by the Illinois Institute for Data Science and Dynamical Systems (iDS^2), an NSF HDR TRIPODS institute, under award CCF-1934986, by NSF grant NSF BCS-1932619, by NIH grants NIH 1R21NR018558 and NIH 1R03MH120406, and by the ARO MURI ARO-W911NF-15-1-0479.

T.P. Coleman is with Stanford University, email: toddcol@stanford.edu. M. Raginsky is with the University of Illinois, Urbana-Champaign, email: maxim@illinois.edu

probability distribution of W given $X = x$. Moreover, if $\mu \ll \mu_0$, then $\frac{\mathrm{d}P^\mu}{\mathrm{d}P} = \frac{\mathrm{d}\mu}{\mathrm{d}\mu_0} \circ F$.

The key property here is that both P^μ and P have the same conditional distribution given $X = x$.

In addition, P^μ is *minimal* in the following sense (cf. also Proposition 6 in [9]): For a convex function $\varphi : \mathbb{R}_+ \to \mathbb{R}$, let Γ_φ^μ denote the set of all Borel probability measures Q on $\mathbb{R}^d$, such that $Q \ll P$, $\varphi(\frac{\mathrm{d}Q}{\mathrm{d}P}) \in L^1(P)$, and $F_*Q = \mu$.

Proposition 2. *Assume that $\mu \ll \mu_0$ and $\varphi(\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0}) \in L^1(\mu_0)$. Then $P^\mu \in \Gamma_\varphi^\mu$, and*

$$\inf_{Q \in \Gamma_\varphi^\mu} D_\varphi(Q\|P) = D_\varphi(P^\mu\|P) = D_\varphi(\mu\|\mu_0),$$

*where $D_\varphi(Q\|P) = \int_{\mathbb{R}^d} \varphi(\frac{\mathrm{d}Q}{\mathrm{d}P})\,\mathrm{d}P$ is the φ-*divergence *between P and Q [10].*

In other words, P^μ is a minimal 'modification' of P, under which the nominal pushforward measure $\mu_0 = F_*P$ is 'transported' to the given target $\mu = F_*P^\mu$. Different choices of φ are possible — for example, if we take $\varphi(u) = u^2 - 1$, then $D_\varphi(Q\|P) = \int_{\mathbb{R}^d}[(\frac{\mathrm{d}Q}{\mathrm{d}P})^2 - 1]\,\mathrm{d}P$ is the variance of $\frac{\mathrm{d}Q}{\mathrm{d}P}$ under P; if $\varphi(u) = -\log u$, then $D_\varphi(Q\|P) = D(P\|Q)$, the usual relative entropy (Kullback–Leibler divergence) between P and Q. Now, if we take $\varphi(u) = u \log u$, then $D_\varphi(Q\|P) = D(Q\|P)$, which gives another optimality criterion for P^μ:

Proposition 3. *Suppose $\mu \ll \mu_0$. For any Borel probability measure Q on $\mathbb{R}^d$, such that $Q \ll P$ and $\log(\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0} \circ F) \in L^1(Q)$, define the* free energy

$$\mathsf{F}(Q) := D(Q\|P) - \int_{\mathbb{R}^d} \mathrm{d}Q \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0} \circ F\right).$$

Then $\mathsf{F}(Q) \geq 0$, with equality if and only if $Q = P^\mu$.

Consequently, if $Q \ll P$ and $F_*Q = \mu$, then $D(Q\|P) \geq D(\mu\|\mu_0)$, with equality if and only if $Q = P^\mu$.

The free-energy formulation of Prop. 3 allows us to consider computationally tractable relaxations to the problem of finding P^μ. For instance, we may restrict Q to pushforward measures of the form g_*P where the mappings $g : \mathbb{R}^d \to \mathbb{R}^d$ come from a given class $\mathcal{G}$, and obtain $\hat{P}^\mu = \hat{g}^\mu_* P$, where $\hat{g}^\mu$ minimizes the free energy $\mathsf{F}(g_*P)$ over $\mathcal{G}$. For instance, $\mathcal{G}$ could consist of all affine maps of the form $g(w) = Aw + b$ with nonsingular $A \in \mathbb{R}^{d\times d}$ and $b \in \mathbb{R}^d$, in which case g_*P would be the nondegenerate Gaussian measure with mean b and covariance matrix AA^T, and the relative entropy $D(g_*P\|P)$ is available in closed form. Other possibilities would be to let $\mathcal{G}$ consist of multilayer feedforward neural nets parameterized by the weights of their neurons [2] or of all functions of the form $g(u) = \nabla\psi(u)$ for convex differentiable $\psi : \mathbb{R}^d \to \mathbb{R}$ [11]. We now show that this formulation naturally covers both sampling and inference.

A. Sampling

Here, the objective is to generate a sample X from a given target distribution μ on $\mathbb{R}^n$ by applying a suitable deterministic transformation to a random element W of the 'latent space' $\mathbb{R}^d$. In the context of the above formalism, the procedure of sampling from μ consists of the minimum-entropy modification $P \mapsto P^\mu$, followed by the pushforward $P^\mu \mapsto F_*P^\mu = \mu$. Here, $F : \mathbb{R}^d \to \mathbb{R}^n$ is a given 'nominal' map, for which the conditional measures P^x are readily available. By Prop. 1, the conditional measures $P^{\mu,x}$ coincide with P^x. Variational relaxations will then result in *approximate* sampling, i.e., $\hat{P}^\mu$ may differ from P^μ.

B. Bayesian inference

In Bayesian inference [12], we have a random couple (X, Y) taking values in a product space $\mathbb{R}^n \times \mathsf{Y}$ according to a given joint distribution Π. We let π and $\tilde{\pi}$ denote the marginal distributions of X and Y, respectively; to keep things simple, we assume that $\Pi \ll \pi \otimes \tilde{\pi}$, and that the associated Radon–Nikodym derivative $q := \frac{\mathrm{d}\Pi}{\mathrm{d}(\pi\otimes\tilde{\pi})}$ is such that $\int_{\mathbb{R}^n} q(x, y)\pi(\mathrm{d}x) \in (0, \infty)$ everywhere. Then the process of Bayesian inference entails mapping the 'prior' distribution π to the 'posterior' distribution $\pi(\cdot|y)$ given the 'evidence' $Y = y$ via the abstract Bayes' formula

$$\pi(A|y) = \frac{\int_A \exp(-H(x, y))\pi(\mathrm{d}x)}{\int_{\mathbb{R}^n} \exp(-H(x, y))\pi(\mathrm{d}x)}, \tag{1}$$

where A ranges over all Borel sets in $\mathbb{R}^n$, and $H := -\log q$. The mapping $\pi \mapsto \pi(\cdot|y)$ in (1) is nonlinear and often computationally expensive. However, using the above formalism, we can envision the following alternative procedure. Suppose that the prior π can be expressed as a pushforward F_*P of the canonical Gaussian measure P on $\mathbb{R}^d$ by some mapping $F : \mathbb{R}^d \to \mathbb{R}^n$. Then, for every $y \in \mathsf{Y}$, we can write $\pi(\cdot|y) = F_*P^{\pi(\cdot|y)}$, where $\mathrm{d}P^{\pi(\cdot|y)} = \left(\frac{\mathrm{d}\pi(\cdot|y)}{\mathrm{d}\pi} \circ F\right)\mathrm{d}P$ is the minimum-entropy modification of P among all Q, such that $F_*Q = \pi(\cdot|y)$. With regards to the free energy objective, we note that, since

$$\log \frac{\mathrm{d}\pi(\cdot|y)}{\mathrm{d}\pi} = -H(\cdot, y) + K(y) \tag{2}$$

for some function $K(\cdot)$ of y only, the computation of P^μ involves only the negative log-likelihood H:

$$P^\mu = \arg\min_{Q \ll P} \mathsf{G}(Q; y), \tag{3}$$

where

$$\mathsf{G}(Q; y) := D(Q\|P) + \int_{\mathbb{R}^d} H(F(w), y)Q(\mathrm{d}w)$$

is the *variational free energy* functional that also depends on the evidence y. This is, essentially, the variational formulation of Bayesian inference due to Mitter and Newton [12]; note that the advantage of minimizing $\mathsf{G}(\cdot; y)$ is that we do not need to know the normalizing constant $e^{K(y)}$ in (2), which is often difficult to compute. By the same token, we can consider computationally tractable relaxations of (3) — e.g., we can restrict the minimization over Q to Gaussian measures whose mean vectors $m(y)$ and covariance matrices $M(y)M(y)^T$ are parametrized by some sufficiently rich class of functions $m : \mathsf{Y} \to \mathbb{R}^d$ and $M : \mathsf{Y} \to \mathbb{R}^{d\times r}$ of the evidence y [2].

III. Generative models and variational approximation in Wiener space

We now move on to the main subject of this paper, namely the case when the latent object W is not a d-dimensional Gaussian random vector, but rather the standard d-dimensional Brownian motion $W = (W_t)_{t\in[0,1]}$. More precisely, we work with the Wiener space $(\Omega, (\mathcal{F}_t)_{t\in[0,1]}, (W_t)_{t\in[0,1]}, \mathbf{P})$, where:

- $\Omega = C([0,1];\mathbb{R}^d)$ is the space of continuous functions $\omega : [0,1] \to \mathbb{R}^d$;
- $(W_t)_{t\in[0,1]}$ is the coordinate process, $W_t(\omega) := \omega(t)$;
- $(\mathcal{F}_t)_{t\in[0,1]}$ is the natural filtration of $W_\cdot$;
- $\mathbf{P}$ is the Wiener measure, under which $W_\cdot$ is the standard d-dimensional Brownian motion on $[0,1]$.

Let $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ be a Polish space equipped with its Borel σ-algebra, and let a measurable map $F : \Omega \to \mathsf{X}$ be given. As before, we will denote by μ_0 the pushforward measure $F_*\mathbf{P}$, under which, for any bounded measurable $h : \mathsf{X} \to \mathbb{R}$,

$$\int_{\mathsf{X}} h(x)\mu_0(\mathrm{d}x) = \int_\Omega h \circ F(\omega)\mathbf{P}(\mathrm{d}\omega).$$

For the most part, we will focus on the special case when $\mathsf{X} = \mathbb{R}^n$ and $F(\omega) = g \circ Z_1(\omega)$, where $g : \mathbb{R}^d \to \mathbb{R}^n$ is a sufficiently well-behaved map and $(Z_t)_{t\in[0,1]}$ is a Markov diffusion process driven by W starting at $Z_0 = 0$, i.e.,

$$Z_t = \int_0^t b(Z_s, s)\,\mathrm{d}s + W_t, \quad t \in [0,1] \tag{4}$$

for some time-varying drift $b : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$; e.g., if $n = d$, g is the identity map on $\mathbb{R}^d$, and $b \equiv 0$, then $F(\omega) = \omega(1)$ and $\mu_0 = P$, the canonical Gaussian measure on $\mathbb{R}^d$.

As before, let a target probability measure μ on X be given, and assume for simplicity that $\mu \ll \mu_0$. In full analogy to Prop. 1, the probability measure $\mathbf{P}^\mu$ specified by

$$\mathrm{d}\mathbf{P}^\mu = \left(\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0} \circ F\right)\mathrm{d}\mathbf{P} \tag{5}$$

is the unique probability measure on $(\Omega, (\mathcal{F}_t)_{t\in[0,1]})$, such that the regular conditional distributions $\mathbf{P}(\cdot|X = x)$ and $\mathbf{P}^\mu(\cdot|X = x)$ coincide and $\mu = F_*\mathbf{P}^\mu$ [9, Prop. 3]. The measure $\mathbf{P}^\mu$ is also minimal in the same sense as in Prop. 2: for any convex $\varphi : \mathbb{R}_+ \to \mathbb{R}$, $\mathbf{P}^\mu = \arg\min_{\mathbf{Q}\in\Gamma^\mu_\varphi} D_\varphi(\mathbf{Q}\|\mathbf{P})$, where Γ^μ_φ is the set of all probability measures on $(\Omega, (\mathcal{F}_t)_{t\in[0,1]})$ that are absolutely continuous w.r.t. $\mathbf{P}$, $\varphi(\frac{\mathrm{d}\mathbf{P}}{\mathrm{d}\mathbf{Q}}) \in L^1(\mathbf{P})$, and $F_*\mathbf{Q} = \mu$ [9, Prop. 6]. In the terminology of Baudoin [9], the pair (F, μ) is a *conditioning*, and the goal is to construct an Itô process representation of the path-space measure $\mathbf{P}^\mu$. This representation is referred to in [9] as the *conditioned SDE*. The key feature of the conditioned SDE representation is that, instead of reweighting the Wiener measure $\mathbf{P}(\mathrm{d}w)$ by $\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0}(F(\omega))$ as in (5), we can generate samples from μ by applying to $\omega \sim \mathbf{P}$ the composition $G = F \circ M$, where $M : \Omega \to \Omega$ is the *Itô map* of the conditioned SDE [13, Sec. 5.2], i.e., a progressively measurable map such that $M_*\mathbf{P} = \mathbf{P}^\mu$. Thus, for any bounded measurable $h : \mathsf{X} \to \mathbb{R}$, we have[1]

$$\int_{\mathsf{X}} h(x)\mu(\mathrm{d}x) = \int_\Omega h \circ F(\omega)\mathbf{P}^\mu(\mathrm{d}\omega) = \int_\Omega h \circ G(\omega)\mathbf{P}(\mathrm{d}\omega).$$

In this context, the main result of interest is the Wiener space analogue of Prop. 3 (cf. [15] for a more general result):

Proposition 4. *For any probability measure* $\mathbf{Q}$ *on* $(\Omega, (\mathcal{F}_t)_{t\in[0,1]})$ *which is absolutely continuous w.r.t.* $\mathbf{P}$ *and* $\log(\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0} \circ F) \in L^1(\mathbf{Q})$*, define the free energy*

$$\mathsf{F}(\mathbf{Q}) := D(\mathbf{Q}\|\mathbf{P}) - \int_\Omega \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0} \circ F\right)\mathrm{d}\mathbf{Q}. \tag{6}$$

Then $\mathsf{F}(\mathbf{Q}) \geq 0$*, with equality if and only if* $\mathbf{Q} = \mathbf{P}^\mu$.

The significance of this result stems from the fact that, by the Cameron–Martin–Girsanov theory [13, Prop. 3.9.13], all such $\mathbf{Q}$ can be obtained by adding an adapted drift process U to $\mathbf{P}$ [16], [17]. Using this representation, we can (abusing the notation slightly) lift the free energy functional (6) to the space of all such processes U by letting

$$\mathsf{F}(U) := \mathbf{E}\left[\frac{1}{2}\int_0^1 \|U_t\|^2\,\mathrm{d}t - \log\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0}\left(F\left(W_\cdot + \int_0^\cdot U_t\,\mathrm{d}t\right)\right)\right], \tag{7}$$

where the expectation is w.r.t. the Wiener measure $\mathbf{P}$; then $\mathsf{F}(U) \geq 0$, with equality achieved uniquely by the drift process U^μ that generates $\mathbf{P}^\mu$. A similar free energy functional $\mathsf{G}(U; y)$ can be constructed for Bayesian inference, where the prior π can be expressed as the pushforward $F_*\mathbf{P}$. In both cases, variational approximation amounts to restricting attention to a suitable structured collection of candidate drift processes that may not contain the optimal drift U^μ.

IV. Optimal stochastic control formulation

We now focus on the special case of the above problem that can be addressed using the theory of controlled diffusions. Specifically, we let $F(\omega) = g \circ W_1(\omega)$, where $g : \mathbb{R}^d \to \mathbb{R}^n$ is a given map — here, we take $b \equiv 0$ in (4). Thus, $\mu_0 = g_*P$, where P is the canonical Gaussian measure on $\mathbb{R}^d$. The problem of minimizing the free energy (7) can now be stated as follows: Consider the set of all feedback controls, i.e., measurable functions $u : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$. To each such control u, we associate a diffusion process $(Z^u_t)_{t\in[0,1]}$ governed by the Itô SDE

$$\mathrm{d}Z^u_t = u(Z^u_t, t)\,\mathrm{d}t + \mathrm{d}W_t, \qquad Z^u_0 = 0,\ 0 \leq t \leq 1. \tag{8}$$

We then seek a control u^* that would minimize the total cost

$$J(u) := \mathbf{E}^u\left[\frac{1}{2}\int_0^1 \|u(Z^u_t, t)\|^2\,\mathrm{d}t - \log\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0}(g(Z^u_1))\right] \tag{9}$$

[1] A similar implementation of pathwise reweighting by an Itô map of a diffusion process was proposed by Ezawa, Klauder, and Shepp [14].

where $\mathbf{E}^u[\cdot]$ denotes expectation w.r.t. the process law of $(Z_t^u)_{t\in[0,1]}$. This is a finite-horizon optimal stochastic control problem with running cost $c(z,u) = \frac{1}{2}\|u\|^2$ and terminal cost $\psi(z) = -\log \frac{\mathrm{d}\mu}{\mathrm{d}\mu_0}(g(z))$.

A. A dynamic programming solution

A standard dynamic programming argument [6] shows that, if such a control u^* exists, the process $U_t^* = u^*(Z_t^{u^*}, t)$ minimizes $\mathsf{F}(U)$ over all admissible drift processes U. Thus, $J(u^*) = \mathsf{F}(U^*) = 0$, so that the optimal controlled process $(Z_t^{u^*})_{t\in[0,1]}$ will have the law $\mathbf{P}^\mu$, and $g(Z_1^{u^*})$ will have the law μ. Indeed, we can characterize the optimal control u^* explicitly. The following result from our earlier work [3] is a synthesis of results of Pavon [18] and Dai Pra [19]:

Theorem 5. *Let Q_t, $t \geq 0$, denote the Euclidean heat semigroup, i.e., for any bounded measurable $h : \mathbb{R}^d \to \mathbb{R}$ and any $z \in \mathbb{R}^d$,*

$$\mathsf{Q}_t h(z) := \mathbf{E}[h(z + \sqrt{t}W)], \qquad W \sim P.$$

Then the optimal control u^* has the form

$$u^*(z,t) = \nabla_z \log \left[\mathsf{Q}_{1-t} \left(\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0} \circ g \right)(z) \right] \tag{10}$$

Remark 6. Following Lehec [17], we will refer to the optimal control (10) as the *Föllmer drift*, cf. also [16], [18]–[20].

An analogous argument works for the case when $F(\omega) = g \circ Z_1(\omega)$, where $(Z_t)_{t\in[0,1]}$ is the diffusion process (4) starting at $Z_0 = 0$. In this instance, we consider controlled diffusions

$$\mathrm{d}Z_t^u = \big(b(t, Z_t^u) + u(Z_t^u, t)\big)\,\mathrm{d}t + \mathrm{d}W_t$$

on $[0,1]$ with $Z_0^u = 0$, and seek a control u to minimize

$$J(u) = \mathbf{E}^u \left[\frac{1}{2} \int_0^1 \|u(Z_t^u, t)\|^2\,\mathrm{d}t - \log \frac{\mathrm{d}\mu}{\mathrm{d}\mu_0}(g(Z_1^u)) \right]. \tag{11}$$

Theorem 7. *The control that minimizes (11) is given by*

$$u^*(z,t) = \nabla_z \log \mathbf{E}\left[\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0}(g(Z_1)) \middle| Z_t = z \right], \tag{12}$$

where the (conditional) expectations are w.r.t. the uncontrolled diffusion process $(Z_t)_{t\in]0,1}$.

Note that the computation of the optimal control in (12) requires knowledge of the transition densities of the uncontrolled process $Z_t = Z_t^0$. When we consider Bayesian inference instead of sampling, Theorem 5 still applies, but now the optimal control depends on the evidence y:

$$u^*(z,t;y) = \nabla_z \log \mathsf{Q}_{1-t} e^{-H(g(z),y)}. \tag{13}$$

Due to the presence of the gradient w.r.t. z in (10) and (13), it suffices to know $\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0}$ or $e^{-H(\cdot,y)}$ up to a multiplicative constant (which in the latter case may depend on y).

B. Relation to the stochastic maximum principle

The problem of minimizing (9) subject to (8) can be phrased in terms of the stochastic version of Pontryagin's maximum principle [7]. We have the running cost $c(z,u) = \frac{1}{2}\|u\|^2$, the terminal cost $-\log f(z)$ where $f = \frac{\mathrm{d}\mu}{\mathrm{d}\mu_0} \circ g$ (for sampling) or $f = e^{-H(g(\cdot),y)}$ (for Bayesian inference), and the *Hamiltonian* $\mathcal{H} : \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^{d\times d} \to \mathbb{R}$, given by

$$\mathcal{H}(z,u,p,M) = p^T u + \frac{1}{2}\mathrm{tr}M - \frac{1}{2}\|u\|^2.$$

Observe that

$$\max_{u\in\mathbb{R}^d} \mathcal{H}(z,u,p,M) = \frac{1}{2}\|p\|^2 + \frac{1}{2}\mathrm{tr}M$$

is achieved uniquely by $u^*(z,p,M) = p$. Here, $p \in \mathbb{R}^d$ and $M \in \mathbb{R}^{d\times d}$ are the adjoint variables. We now consider the following forward-backward SDE:

$$\mathrm{d}Z_t = u^*(Z_t, P_t, M_t)\,\mathrm{d}t + \mathrm{d}W_t \tag{14a}$$
$$\mathrm{d}P_t = M_t\,\mathrm{d}W_t \tag{14b}$$

for $t \in [0,1]$, where the processes (Z_t), (P_t), and (M_t) are all adapted to $(\mathcal{F}_t)_{t\in[0,1]}$, with (Z_t) subject to initial condition $Z_0 = 0$ and (P_t) subject to terminal condition $P_1 = \nabla_z \log f(Z_1)$. The stochastic maximum principle (cf. Thm. 3.2 in [7]) then states that the existence of these processes is a necessary condition for the optimality of (U_t^*) with $U_t^* = u^*(Z_t, P_t, M_t)$ for our stochastic control problem.

We now show that the Föllmer drift of Theorem 5 can be characterized in this way. First, since $u^*(z,p,M) = p$, we can rewrite the forward SDE (14a) as $\mathrm{d}Z_t = P_t\,\mathrm{d}t + \mathrm{d}W_t$, where, using (14b), P_t can be expressed as

$$P_t = P_0 + \int_0^t M_s\,\mathrm{d}W_s.$$

Consider now the SDE

$$\mathrm{d}Z_t = \nabla_z \log \mathsf{Q}_{1-t} f(Z_t)\,\mathrm{d}t + \mathrm{d}W_t, \quad t \in [0,1],\ Z_0 = 0$$

and let $P_t := \nabla_z \log \mathsf{Q}_{1-t} f(Z_t)$ and $M_t := \nabla_z^2 \log \mathsf{Q}_{1-t} f(Z_t)$. A simple application of Itô's lemma to (the components of) the Föllmer drift $(z,t) \mapsto \nabla_z \log \mathsf{Q}_{1-t} f(z)$ yields the Itô representation

$$\begin{aligned} P_t &= P_0 + \int_0^t \nabla_z^2 \log \mathsf{Q}_{1-s} f(Z_s)\,\mathrm{d}W_s \\ &= P_0 + \int_0^t M_s\,\mathrm{d}W_s, \end{aligned}$$

where $P_0 = \frac{\int_{\mathbb{R}^d} w f(w) e^{-\|w\|^2/2}\,\mathrm{d}w}{\int_{\mathbb{R}^d} f(w) e^{-\|w\|^2/2}\,\mathrm{d}w}$ and $P_1 = \nabla_z \log f(Z_1)$. Moreover, since the Hamiltonian $\mathcal{H}$ is concave in z and u, the stochastic maximum principle is also a sufficient condition for optimality (see, e.g., Thm 5.2 in [7]) when the terminal cost $-\log f(z)$ is convex, i.e., when the function $\frac{\mathrm{d}\mu}{\mathrm{d}\mu_0} \circ g$ or $e^{-H(g(\cdot),y)}$ is log-concave. (Of course, the optimality of the controls given in Theorems 5 and 7 already follows from a dynamic programming argument.)

V. Tractable variational approximations

While Sec. IV presents an exact solution to the problem of constructing minimum-entropy modifications of $\mathbf{P}$ for sampling or for Bayesian inference, it is often desirable to work with computationally tractable controls that do not minimize (9) or (11). In this section, we outline one approach to choosing such variational approximations by adopting an ingenious idea of Beneš [8]. In a nutshell, the point is to choose a parametric family of drifts u in such a way that the computation of the expected cost $J(u)$ can be reduced to tractable integration w.r.t. the Wiener measure.

Let $(Z_t)_{t\in[0,1]}$ be a d-dimensional diffusion process governed by the Itô SDE

$$\mathrm{d}Z_t = f(Z_t)\,\mathrm{d}t + \mathrm{d}W_t, \qquad t \in [0,1]\; Z_0 = 0.$$

We wish to compute the expected value $\mathbf{E}[\varphi(Z_1)]$ for some test function $\varphi : \mathbb{R}^d \to \mathbb{R}$, e.g., either $-\log \frac{\mathrm{d}\mu}{\mathrm{d}\mu_0} \circ g$ (for sampling) or $H(g(\cdot), y)$ (for Bayesian inference). For simplicity, we will illustrate the key ideas first in one dimension (i.e., $d = 1$) and then discuss the general case.

Theorem 8. *Suppose that the drift $f : \mathbb{R} \to \mathbb{R}$ is differentiable and satisfies the nonlinear differential equation*

$$f'(z) + f^2(z) = az^2 + bz + c \tag{15}$$

for some $a \geq 0$ and $b, c \in \mathbb{R}$. Then

$$\begin{aligned}&\mathbf{E}[\varphi(Z_1)]\\ &= \frac{1}{\sqrt{2\pi K_{22}}} \int_{\mathbb{R}} \varphi(z) \exp\left\{ \int_0^z \left[f(v) + \sqrt{a}v\right] \mathrm{d}v \right.\\ &\quad \left. + \frac{1}{2}K_{22} - \frac{1}{2K_{22}}(z - K_{22})^2 - \frac{1}{2}(c + \sqrt{a}) \right\} \mathrm{d}z,\end{aligned}$$

where $K \in \mathbb{R}^{2\times 2}$ is the covariance matrix, at time $t = 1$, of the two-dimensional Gaussian process

$$\mathrm{d}Y_t = -\sqrt{a}Y_t\,\mathrm{d}t + \mathrm{d}W_t, \quad \mathrm{d}V_t = -\frac{b}{2}Y_t\,\mathrm{d}t$$

with $(Y_0, V_0) = (0, 0)$.

As an example of f satisfying the condition (15), we have the affine functions $f(z) = az + b$ and the tanh nonlinearity $f(z) = a\tanh(az + b)$.

The extension to multiple dimensions ($d > 1$) proceeds along the same lines as in [8]. Namely, consider the d-dimensional diffusion process

$$\mathrm{d}Z_t = \nabla g(Z_t)\,\mathrm{d}t + \mathrm{d}W_t, \qquad t \in [0,1],\; Z_0 = 0 \tag{16}$$

where g is a twice differentiable solution of the equation

$$\nabla^2 g(z) + \|\nabla g(z)\|^2 = z^T M z + m^T z + c \tag{17}$$

for some symmetric, positive semidefinite $M \in \mathbb{R}^{d\times d}$, $m \in \mathbb{R}^d$, and $c \in \mathbb{R}$. We can then compute the density of Z_1 explicitly. To that end, diagonalize M as $TMT^T = \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$, where $TT^T = T^T T = I_d$, and define d independent two-dimensional Gaussian processes $(Y_t^{(i)}, V_t^{(i)})$

$$\mathrm{d}Y_t^{(i)} = -\sqrt{\lambda_i}Y_t^{(i)}\,\mathrm{d}t + \mathrm{d}W_t^{(i)}, \quad \mathrm{d}V_t^{(i)} = -\frac{b_i}{2}\,\mathrm{d}Y_t^{(i)}\,\mathrm{d}t$$

with $(Y_0^{(i)}, V_0^{(i)}) = (0, 0)$, where $W^{(i)}$ are independent standard one-dimensional Brownian motions and $b = Tm$. Let $K^{(i)} \in \mathbb{R}^{2\times 2}$ denote the covariance matrix of $(Y_1^{(i)}, V_1^{(i)})$.

Theorem 9. *Suppose that the drift vector field ∇g in (16) satisfies (17). Then*

$$\begin{aligned}\mathbf{E}[\varphi(Z_1)] = \int_{\mathbb{R}^d} \varphi(z) \exp\Big\{ & g(z) + \frac{1}{2}z^T M z\\ & - \frac{1}{2}\sum_{i=1}^d \frac{((Tz)_i - K_{22}^{(i)})^2}{K_{22}^{(i)}} - \frac{1}{2}v^T K v\\ & - \frac{d}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^d K_{22}^{(i)} + \sum_{=1}^d \sqrt{\lambda_i} - \frac{c}{2}\Big\}\,\mathrm{d}z,\end{aligned}$$

where $K := \mathrm{diag}(K^{(1)}, \ldots, K^{(d)})$ and $v := (0, -1, 0, -1, \ldots, 0, -1)^T \in \mathbb{R}^{2d}$.

Beneš [8] gives a nice example of a class of functions satisfying (17): Let $h : \mathbb{R} \to \mathbb{R}$ be a bounded differentiable function, such that $h' + h^2 = r$, a constant. Then, for any traceless matrix $S \in \mathbb{R}^{d\times d}$ of full rank, let $f = \nabla g$ with $g(z) = \int_0^{\frac{1}{2}z^T S z} h(u)\,\mathrm{d}u$. Then the conditions of Theorem 9 will be met if $I_d + rS^T S$ is positive definite.

In both cases, we note that the same argument based on the Girsanov-type change of measure shows that the relative entropy $D(\mathbf{Q}\|\mathbf{P})$ can be computed (or approximated) using simple stochastic integration w.r.t. suitable Ornstein–Uhlenbeck processes. Of course, charaterizing the expressive capabilities of such variational approximations is an important question for further research.

VI. Relation to entropic optimal transport

Our construction of $\mathbf{P}^\mu$ is related to, but different from, the following problem [21]: Given μ_0 and μ, let $\Delta^{\mu_0,\mu}$ denote the set of all probability measures $\mathbf{Q} \ll \mathbf{P}$, such that $(W_0)_*\mathbf{Q} = \mu_0$ and $(W_1)_*\mathbf{Q} = \mu$, where $W_t : \Omega \to \mathbb{R}^d$ is the coordinate map $W_t(\omega) = \omega(t)$. We then wish to solve

$$\min_{\mathbf{Q}} D(\mathbf{Q}\|\mathbf{P}) \text{ subject to } \mathbf{Q} \in \Delta^{\mu_0,\mu}.$$

For simplicity, let us assume that both μ_0 and μ have densities m_0 and m w.r.t. the Lebesgue measure on $\mathbb{R}^d$. It is not hard to show, using the chain rule for the relative entropy, that the minimizer should be of the form

$$\mathbf{Q}(\cdot) = \int_{\mathbb{R}^d\times\mathbb{R}^d} q(w_0, w_1)\mathbf{P}_{w_0}^{w_1}(\cdot)\,\mathrm{d}w_0\,\mathrm{d}w_1, \tag{18}$$

where $\mathbf{P}_{w_0}^{w_1}(\cdot)$ is the Brownian bridge measure, i.e., the conditional probability law of the d-dimensional Brownian motion starting at w_0 at time $t = 0$ and conditioned to be at w_1 at $t = 1$, and where $q(\mathrm{d}w_0, \mathrm{d}w_1)\,\mathrm{d}w_0\,\mathrm{d}w_1$ is the coupling of μ_0 and μ that minimizes the entropic cost

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} q(w_0, w_1) \log \frac{q(w_0, w_1)}{p(w_0, w_1)}\,\mathrm{d}w_0\,\mathrm{d}w_1$$

where $p(w_0, w_1) = \frac{1}{(2\pi)^{d/2}} \exp(-\frac{1}{2}\|w_1 - w_0\|^2)$. We thus end up with the following minimization problem in the space of densities q on $\mathbb{R}^d \times \mathbb{R}^d$:

$$\min \Big\{ \frac{1}{2} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|w_0 - w_1\|^2 q(w_0, w_1)\, \mathrm{d}w_0\, \mathrm{d}w_1 + \int_{\mathbb{R}^d \times \mathbb{R}^d} q(w_0, w_1) \log q(w_0, w_1)\, \mathrm{d}w_0\, \mathrm{d}w_1 \Big\} \qquad (19)$$

subject to the constraint that q has the marginal densities m_0 and m. This is known as the *Schrödinger bridge problem* or the *entropic optimal transport problem* [19]–[23]. The latter name refers to the fact that, without the second (entropic) term in (19), we would end up with the classical Monge–Kantorovich optimal transportation problem with quadratic cost [11], and, indeed, if we were to replace the Wiener measure $\mathbf{P}$ with the probability law $\mathbf{P}_\varepsilon$ of the rescaled Brownian motion $(\sqrt{\varepsilon} W_t)_{t\in[0,1]}$ for some $\varepsilon > 0$, we would recover the optimal transport problem in the limit as $\varepsilon \to 0$ [24]. It can also be shown that the optimal probability measure $\mathbf{Q}$ in (18) is the law of a diffusion process started at $Z_0 \sim \mu_0$, whose drift can be characterized explicitly [19], [21]. A recent survey by Reich [23] describes an approach to Bayesian inference based on entropic optimal transport. By contrast, we treat both the nominal measure μ_0 and the target measure μ as images, under the same map $F : \Omega \to \mathbb{R}^d$, of the Wiener measure $\mathbf{P}$ and its minimum-entropy modification $\mathbf{P}^\mu$, and, moreover, $\mathbf{P}^\mu$ is the pushforward of $\mathbf{P}$ under the Itô map of a suitable conditioned SDE in the sense of Baudoin [9].

VII. Conclusion and Future Directions

We have outlined a unified approach to sampling and inference in probabilistic generative models, where the latent object is a finite-dimensional diffusion process. Using the theory of controlled diffusions, we were able to characterize optimal solutions and tractable variational relaxations. We close with a partial list of potential future research directions:

- Consider the case of multiple time instants, i.e., when $F(\omega) = g(\omega(t_1), \ldots, \omega_(t_N))$ for some $0 \le t_1 < \ldots < t_N \le 1$ and $g : (\mathbb{R}^d)^N \to \mathbb{R}^n$. This could arise, for instance, in the context of generative models for irregularly sampled time series.
- Consider dynamic causal inference [25], [26], which generalizes Bayesian inference to more than two processes. Similarly, posterior matching, which is dual to Bayesian inference in that it seeks to transform sample from the posterior to one from the prior, can be explored in this continuous-time setting [27].
- Consider the setting where the latent object is a process with jumps, such as a standard Poisson process, and path-space transformations amount to time rescaling.

References

[1] Diederik P. Kingma and Max Welling, "Auto-encoding Variational Bayes," in *International Conference on Learning Representations*, 2014.

[2] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *Proceedings of the 2014 International Conference on Machine Learning*, 2014, pp. 1278–1286.

[3] Belinda Tzen and Maxim Raginsky, "Theoretical guarantees for sampling and inference in generative models with latent diffusions," in *Proceedings of the 32nd Conference on Learning Theory*, 2019, vol. 99 of *PMLR*, pp. 3084–3114.

[4] Belinda Tzen and Maxim Raginsky, "Neural stochastic differential equations: deep latent Gaussian models in the diffusiono limit," arXiv:1905.09883, 2019.

[5] Xuechen Li, Ting-Kam Leonard Wong, Ricky T. Q. Chen, and David Duvenaud, "Scalable gradients for stochastic differential equations," in *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, 2020, vol. 208 of *PMLR*, pp. 3870–3882.

[6] Wendell H. Fleming and Raymond W. Rishel, *Deterministic and Stochastic Optimal Control*, Springer, 1975.

[7] Jiongmin Yong and Xun Yu Zhou, *Stochastic Controls: Hamiltonian Systems and HJB Equations*, Springer, 1999.

[8] Václav E. Beneš, "Exact finite-dimensional filters for certain diffusions with nonlinear drift," *Stochastics*, vol. 5, pp. 65–92, 1981.

[9] Fabrice Baudoin, "Conditioned stochastic differential equations: theory, examples and application to finance," *Stochastic Processes and Their Applications*, vol. 100, pp. 109–145, 2002.

[10] Friedrich Liese and Igor Vajda, "On divergences and informations in statistics and information theory," *IEEE Transactions on Information Theory*, vol. 52, no. 10, pp. 4394–4412, 2006.

[11] C. Villani, *Topics in Optimal Transportation*, vol. 58 of *Graduate Studies in Mathematics*, Amer. Math. Soc., Providence, RI, 2003.

[12] Sanjoy K. Mitter and Nigel J. Newton, "A variational approach to nonlinear estimation," *SIAM Journal on Control and Optimization*, vol. 42, no. 5, pp. 1813–1833, 2003.

[13] Klaus Bichteler, *Stochastic Integration with Jumps*, Cambridge University Press, 2002.

[14] Hiroshi Ezawa, John R. Klauder, and Lawrence A. Shepp, "A path space picture for Feynman–Kac averages," *Annals of Physics*, vol. 88, pp. 588–620, 1974.

[15] Joris Bierkens and Hilbert J. Kappen, "Explicit solution of relative entropy weighted control," *Systems & Control Letters*, vol. 72, pp. 36–43, 2014.

[16] Hans Föllmer, "An entropy approach to time reversal of diffusion processes," in *Stochastic Differential Systems (Marseille-Luminy, 1984)*, vol. 69 of *Lecture Notes in Control and Information Sciences*. Springer, 1985.

[17] Joseph Lehec, "Representation formula for the entropy and functional inequalities," *Annales de l'Institut Henri Poincaré - Proobabilités et Statistiques*, vol. 49, no. 3, pp. 885–899, 2013.

[18] Michele Pavon, "Stochastic control and nonequilibrium thermodynamical systems," *Applied Mathematics and Optimization*, vol. 19, pp. 187–202, 1989.

[19] Paolo Dai Pra, "A stochastic control approach to reciprocal diffusion processes," *Applied Mathematics and Optimization*, vol. 23, pp. 313–329, 1991.

[20] Benton Jamison, "The Markov processes of Schrödinger," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 32, no. 4, pp. 323–331, 1975.

[21] Yongxin Chen, Tryphon T. Georgiou, and Michele Pavon, "On the relation between optimal transport and Schrödinger bridges: a stochastic control viewpoint," *Journal of Optimization Theory and Applications*, vol. 169, pp. 671–691, 2016.

[22] Erwin Schrödinger, "Über die Umkehrung der Naturgesetze," *Sitzung der Preuss. Akad. Wissen., Berlin Phys. Math.*, vol. 144, 1931.

[23] Sebastian Reich, "Data assimilation: The Schrödinger perspective," *Acta Numerica*, vol. 28, no. 635-711, 2019.

[24] Toshio Mikami, "Monge's problem with a quadratic cost by the zero-noise limit of h-path processes," *Probability Theory and Related Fields*, vol. 129, pp. 245–260, 2004.

[25] Sanggyun Kim, Christopher J Quinn, Negar Kiyavash, and Todd P Coleman, "Dynamic and succinct statistical analysis of neuroscience data," *Proceedings of the IEEE*, vol. 102, no. 5, pp. 683–698, 2014.

[26] Gabriel Schamberg and Todd P Coleman, "Measuring sample path causal influences with relative entropy," *IEEE Transactions on Information Theory*, vol. 66, no. 5, pp. 2777–2798, 2020.

[27] Rui Ma and Todd P Coleman, "Generalizing the posterior matching scheme to higher dimensions via optimal transportation," in *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2011, pp. 96–102.