Diffusion Models and SDEs

Lecture 3:

Girsanov Theorem, KL Divergence, Half Bridges, FK- Formula

Reminder

Last Lecture

The optimal predictor of X as a function of Y (Hilbert projection)

$$\underset{f-\text{is measurable}}{\arg\min} \quad \mathbb{E}\left(X - f(Y)\right)^2$$

Is given by the conditional expectation:

$$f^*(Y) = \mathbb{E}[X|Y]$$

Yonatan’s Question - Tractable Score matching loss

Last Lecture – Song Score Matching Objective

$$s^* = \underset{s-\text{is measurable}}{\arg\min} \; \mathbb{E}\left[\int_0^T \left|\left|\nabla \ln p_{t|0}(X_t|X_0) - s(t, X_t)\right|\right|^2 \mathrm{d}t\right]$$

Yonatan’s Question - Tractable Score matching loss

Last Lecture – Song Score Matching Objective

$$s^* = \underset{s-\text{is measurable}}{\arg\min} \; \mathbb{E}\left[\int_0^T \left|\left|\nabla \ln p_{t|0}(X_t|X_0) - s(t, X_t)\right|\right|^2 \mathrm{d}t\right]$$

$$s^*(t,x) = \mathbb{E}_{X_0|X_t}[\nabla \ln p_{t|0}(X_t|X_0)|X_t = x]$$

Yonatan's Question - Tractable Score matching loss

Last Lecture – Song Score Matching Objective

$$s^* = \underset{s-\text{is measurable}}{\arg\min} \ \mathbb{E}\left[\int_0^T \left|\left|\nabla \ln p_{t|0}(X_t|X_0) - s(t, X_t)\right|\right|^2 \mathrm{d}t\right]$$

$$s^*(t,x) = \mathbb{E}_{X_0|X_t}[\nabla \ln p_{t|0}(X_t|X_0)|X_t = x]$$

$$s^*(t,x) = \int p_{0|t}(x_0|x)\nabla \ln p_{t|0}(x|x_0)\mathrm{d}x_0$$

Yonatan's Question - Tractable Score matching loss

Last Lecture – Song Score Matching Objective

$$s^* = \underset{s-\text{is measurable}}{\arg\min}\ \mathbb{E}\left[\int_0^T \left|\left|\nabla \ln p_{t|0}(X_t|X_0) - s(t, X_t)\right|\right|^2 \mathrm{d}t\right]$$

$$s^*(t,x) = \mathbb{E}_{X_0|X_t}[\nabla \ln p_{t|0}(X_t|X_0)|X_t = x]$$

$$s^*(t,x) = \int p_{0|t}(x_0|x)\nabla \ln p_{t|0}(x|x_0)\mathrm{d}x_0$$

$$s^*(t,x) = \int \frac{p_{t|0}(x|x_0)p_0(x_0)}{p_t(x)}\nabla \ln p_{t|0}(x|x_0)\mathrm{d}x_0$$

Yonatan’s Question - Tractable Score matching loss

Last Lecture – Song Score Matching Objective

$$s^*(t,x) = \int \frac{p_{t|0}(x|x_0)p_0(x_0)}{p_t(x)} \nabla \ln p_{t|0}(x|x_0) \mathrm{d}x_0$$

Yonatan's Question - Tractable Score matching loss

Last Lecture – Song Score Matching Objective

$$s^*(t,x) = \int \frac{p_{t|0}(x|x_0)p_0(x_0)}{p_t(x)} \nabla \ln p_{t|0}(x|x_0) \mathrm{d}x_0$$

$$s^*(t,x) = \frac{1}{p_t(x)} \int p_0(x_0) \nabla p_{t|0}(x|x_0) \mathrm{d}x_0$$

Yonatan’s Question - Tractable Score matching loss

Last Lecture – Song Score Matching Objective

$$s^*(t,x) = \int \frac{p_{t|0}(x|x_0)p_0(x_0)}{p_t(x)} \nabla \ln p_{t|0}(x|x_0)\mathrm{d}x_0$$

$$s^*(t,x) = \frac{1}{p_t(x)} \int p_0(x_0) \nabla p_{t|0}(x|x_0)\mathrm{d}x_0$$

$$s^*(t,x) = \frac{1}{p_t(x)} \nabla \int p_0(x_0) p_{t|0}(x|x_0)\mathrm{d}x_0$$

Yonatan’s Question - Tractable Score matching loss

Last Lecture – Song Score Matching Objective

$$s^*(t,x) = \int \frac{p_{t|0}(x|x_0)p_0(x_0)}{p_t(x)} \nabla \ln p_{t|0}(x|x_0)\mathrm{d}x_0$$

$$s^*(t,x) = \frac{1}{p_t(x)} \int p_0(x_0)\nabla p_{t|0}(x|x_0)\mathrm{d}x_0$$

$$s^*(t,x) = \frac{1}{p_t(x)} \nabla \int p_0(x_0)p_{t|0}(x|x_0)\mathrm{d}x_0$$

$$s^*(t,x) = \frac{1}{p_t(x)} \nabla p_t(x) = \nabla_x \ln p_t(x)$$

Novikovs Condition

Regularity assumption for IS with diffusions

We say a stochastic process Theta(t) satisfies Novikovs condition if:

$$\mathbb{E}\left[\exp\left(\frac{1}{2}\int_0^T ||\Theta(t)||^2 \mathrm{d}t\right)\right] < \infty$$

We will require similar conditions relating to the coeficients of diffusions in order to be able to do importance sampling. However very recent results in diffusion gen modelling have managed to extend the theory without assuming this condition.

Girsanov Theorem I

General Statement

Given Novikovs condition and a Brownian motion in the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ follows that:

$$B_t = W_t + \int_0^t \Theta(t) \mathrm{d}s$$

Is a Brownian motion in the probability space $(\Omega, \mathcal{F}, \mathbb{Q})$. Where

$$\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} = \exp\left(- \int_0^T \Theta(t)^\top \mathrm{d}W_t - \frac{1}{2} \int_0^T ||\Theta(t)||^2 \mathrm{d}t \right)$$

Girsanovs Theorem - Corollary

General Statement

Given the SDE

$$\mathrm{d}W_t^\sigma = \sigma(W_t^\sigma, t)\mathrm{d}W_t$$

With probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then it follows that:

$$B_t = W_t - \int_0^t \mu(W_s, s)\sigma^{-1}(W_s, s)\mathrm{d}s$$

Is a Brownian motion in the probability space $(\Omega, \mathcal{F}, \mathbb{Q})$. Where

$$\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} = \exp\left(\int_0^T \sigma^{-1}(W_t^\sigma, t)\mu(W_t^\sigma, t)^\top \mathrm{d}W_t - \frac{1}{2}\int_0^T \sigma^{-2}(W_t^\sigma, t)||\mu(W_t^\sigma, t)||^2 \mathrm{d}t\right)$$

Girsanovs Theorem - Corollary

General Statement

Furthermore, we have that

$$\begin{aligned} \mathrm{d}W_t^\sigma &= \sigma(W_t^\sigma, t)(\mathrm{d}B_t + \sigma^{-1}\mu(W_t^\sigma, t)\mathrm{d}t) \\ &= \mu(W_t^\sigma, t)\mathrm{d}t + \sigma(W_t^\sigma, t)\mathrm{d}B_t \end{aligned}$$

Thus, in the space $(\Omega, \mathcal{F}, \mathbb{Q})$ the process W_t^σ weakly solves the SDE

$$\mathrm{d}X_t = \mu(X_t, t)\mathrm{d}t + \sigma(X_t, t)\mathrm{d}B_t$$

With:

$$\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} = \exp\left(\int_0^T \sigma^{-1}(W_t^\sigma, t)\mu(W_t^\sigma, t)^\top \mathrm{d}W_t - \frac{1}{2}\int_0^T \sigma^{-2}(W_t^\sigma, t)||\mu(W_t^\sigma, t)||^2 \mathrm{d}t\right)$$

Girsanovs Theorem – RND Corollary

Importance Sampling Again

Then we have that:

$$\mathbb{E}_{\mathbb{Q}}[f(X)] = \mathbb{E}_{\mathbb{P}}\left[\exp\left(\int_0^T \sigma_t^{-1}\mu_t^\top \mathrm{d}W_t - \frac{1}{2}\int_0^T \sigma_t^{-2}||\mu_t||^2 \mathrm{d}t\right) f(W^\sigma)\right]$$

Which is effectively the statement of the RN theorem, so it follows that

$$\frac{\mathrm{d}\mathbb{P}_X}{\mathrm{d}\mathbb{P}_{W^\sigma}}(W^\sigma) = \exp\left(\int_0^T \sigma_t^{-1}\mu_t^\top \mathrm{d}W_t - \frac{1}{2}\int_0^T \sigma_t^{-2}||\mu_t||^2 \mathrm{d}t\right)$$

Girsanovs Theorem – RND Corollary

Caveat !!

This result gives us the RND when evaluated on a sample from W^σ if instead we wanted to evaluate the RND on a sample from X we would have to apply Girsanovs theorem with a sign flip starting from the SDE solving X and transforming it to the law of W^σresulting in:

$$\frac{\mathrm{d}\mathbb{P}_X}{\mathrm{d}\mathbb{P}_{W^\sigma}}(X) = \exp\left(\int_0^T \sigma_t^{-1}\mu_t^\top \mathrm{d}W_t + \frac{1}{2}\int_0^T \sigma_t^{-2}||\mu_t||^2 \mathrm{d}t\right)$$

So, remember depending on what we take expectations with respect to the signs in the RND will change.

Optional bonus exercise with 1d Gaussians to be added to homework.

RNDs – General Result

Likelihood Ratio Between Diffusions

Given 2 SDEs (with the same initial condition X_0=Y_0=x):

$$\mathrm{d}X_t = \mu(X_t, t)\mathrm{d}t + \sigma(X_t, t)\mathrm{d}B_t \,, \quad \mathrm{d}Y_t = \rho(Y_t, t)\mathrm{d}t + \sigma(Y_t, t)\mathrm{d}B_t$$

satisfying all the conditions we have discussed. It follows that:

$$\frac{\mathrm{d}\mathbb{P}_X}{\mathrm{d}\mathbb{P}_Y}(X) = \exp\left(\int_0^T \sigma_t^{-1}(\mu_t - \rho_t)^\top \mathrm{d}W_t + \frac{1}{2}\int_0^T \sigma_t^{-2}||\mu_t - \rho_t||^2 \mathrm{d}t\right)$$

$$\frac{\mathrm{d}\mathbb{P}_X}{\mathrm{d}\mathbb{P}_Y}(Y) = \exp\left(\int_0^T \sigma_t^{-1}(\mu_t - \rho_t)^\top \mathrm{d}W_t - \frac{1}{2}\int_0^T \sigma_t^{-2}||\mu_t - \rho_t||^2 \mathrm{d}t\right)$$

KL- Divergence

Likelihood Ratio Between Diffusions

Remember (changing notation a bit P^f refers to the SDE with drift f)

$$D_{KL}(\mathbb{P}^{\mu}||\mathbb{P}^{\rho}) = \mathbb{E}_{X\sim\mathbb{P}^{\mu}}\left[\ln \frac{\mathrm{d}\mathbb{P}^{\mu}}{\mathrm{d}\mathbb{P}^{\rho}}(X)\right]$$

Now applying Girsanov's theorem (e.g. the corollaries we derived):

$$D_{KL}(\mathbb{P}^{\mu}||\mathbb{P}^{\rho}) = \mathbb{E}_{X\sim\mathbb{P}^{\mu}}\left[\int_0^T \sigma_t^{-1}(\mu_t - \rho_t)^\top \mathrm{d}W_t + \frac{1}{2}\int_0^T \sigma_t^{-2}||\mu_t - \rho_t||^2 \mathrm{d}t\right]$$

KL- Divergence

Likelihood Ratio Between Diffusions - Martingale

Remember the Ito integral is a Martingale (1st Lecture) and thus has 0 expectation resulting in:

$$D_{KL}(\mathbb{P}^{\mu}||\mathbb{P}^{\rho}) = \mathbb{E}_{X\sim\mathbb{P}^{\mu}}\left[\frac{1}{2}\int_0^T \sigma_t^{-2}||\mu_t - \rho_t||^2 \mathrm{d}t\right]$$

KL- Divergence – Score Matching

Likelihood Ratio Between Diffusions – OU time reversal

Remember the Ito integral is a Martingale (1st Lecture) and thus has 0 expectation resulting in:

$$D_{KL}(\mathbb{P}^{\mu}||\mathbb{P}^{\rho}) = \mathbb{E}_{X\sim\mathbb{P}^{\mu}}\left[\frac{1}{2}\int_0^T \sigma_t^{-2}||\mu_t - \rho_t||^2 \mathrm{d}t\right]$$

Now consider the case where X is the time reversal of an OU process and we can parametrize P^\rho as a score network SDE, which results in:

$$D_{KL}(\mathbb{P}^{\mu}||\mathbb{P}^{\rho}) = \mathbb{E}_{X\sim\mathbb{P}^{\mu}}\left[\frac{1}{2}\int_0^T \sigma_{T-t}^2||\nabla \ln p_{T-t} - s_{T-t}^{\rho}||^2 \mathrm{d}t\right]$$

KL- Divergence – Score Matching

Likelihood Ratio Between Diffusions – OU time reversal

$$D_{KL}(\mathbb{P}^{\mu}||\mathbb{P}^{\rho}) = \mathbb{E}_{X\sim\mathbb{P}^{\mu}}\left[\frac{1}{2}\int_0^T \sigma_{T-t}^2||\nabla \ln p_{T-t} - s_{T-t}^{\rho}||^2 \mathrm{d}t\right]$$

Now remember we can sample X_t via sampling Z_{T-t} where Z_t is the original (non reversed) noising OU process thus we have:

$$D_{KL}(\mathbb{P}^{\mu}||\mathbb{P}^{\rho}) = \mathbb{E}_{Z\sim\mathbb{P}^{\mu}}\left[\frac{1}{2}\int_0^T \sigma_t^2||\nabla \ln p_t - s_t^{\rho}||^2 \mathrm{d}t\right]$$

Same mean squared error objective as in Song et al. 2021 !

Chain Rule – Disintegration Theorem

The chain rule is a little bit more complicated for path measures

$$\mathbb{P}(A_0 \times A_{(0,T]}) = \int_{A_0} \mathbb{P}_{\cdot|0}(A_{(0,T]}|x)\mathrm{d}\mathbb{P}_0(x)$$

Which under certain regularity assumptions (which SDEs satisfies) implies

$$\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(\cdot) = \frac{\mathrm{d}\mathbb{P}_{\cdot|0}(|x)}{\mathrm{d}\mathbb{Q}_{\cdot|0}(|x)}\frac{\mathrm{d}\mathbb{P}_0}{\mathrm{d}\mathbb{Q}_0}(x)$$

Sometimes written as

$$\mathrm{d}\mathbb{P} = \mathrm{d}\mathbb{P}_{\cdot|0}(|x)\mathrm{d}\mathbb{P}_0(x)$$

Chain Rule – Disintegration Theorem

The chain rule is a little bit more complicated for path measures

$$\mathbb{P}(A_0 \times A_{(0,T]}) = \int_{A_0} \mathbb{P}_{\cdot|0}(A_{(0,T]}|x)\mathrm{d}\mathbb{P}_0(x)$$

Which under certain regularity assumptions (which SDEs satisfies) implies

$$\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(\cdot) = \frac{\mathrm{d}\mathbb{P}_{\cdot|0}(|x)}{\mathrm{d}\mathbb{Q}_{\cdot|0}(|x)}\frac{\mathrm{d}\mathbb{P}_0}{\mathrm{d}\mathbb{Q}_0}(x)$$

Sometimes written as

$$\mathrm{d}\mathbb{P} = \mathrm{d}\mathbb{P}_{\cdot|0}(|x)\mathrm{d}\mathbb{P}_0(x)$$

Half Bridges – Constrained KL minimisation

Constrained Optimisation

$$\mathbb{P}^* = \underset{\mathbb{P}\,:\,\text{s.t. }\mathbb{P}_T = \pi}{\arg\min} \; D_{KL}(\mathbb{P}||\mathbb{P}^\rho)$$

Then

$$\mathrm{d}\mathbb{P}^* = \mathrm{d}\mathbb{P}^\rho \frac{\mathrm{d}\pi}{\mathrm{d}\mathbb{P}^\rho}$$

Half Bridges – Constrained KL minimisation

Unconstrained Formulation – Stochastic Control

$$\mathbb{P}^* = \arg\min_{\mathbb{P}} D_{KL}(\mathbb{P}^{\mu}||\mathbb{P}^*)$$
$$= \arg\min_{\mathbb{P}} D_{KL}(\mathbb{P}^{\mu}||\mathbb{P}^{\rho}) - \mathbb{E}\left[\ln \frac{\mathrm{d}\pi}{\mathrm{d}\mathbb{P}^{\rho}_T}\right]$$

$$\mathbb{P}^{\mu}_0 = \mathbb{P}^*_0$$

Now applying Girsanovs Theorem (Stochastic Control Objective)

$$\arg\min_{\mu} \mathbb{E}_{X\sim\mathbb{P}^{\mu}}\left[\frac{1}{2}\int_0^T \sigma_t^{-2}||\mu_t - \rho_t||^2 \mathrm{d}t\right] - \mathbb{E}\left[\ln \frac{\mathrm{d}\pi}{\mathrm{d}\mathbb{P}^{\rho}_T}\right]$$

Feynman - Kac Formula

PDE Solving via MC – Path Integral

Consider the linear Parabolic PDE

$$v_0(x) = \phi(x)$$

$$\partial_t v_t(x) = -\sum_{i=1}^{d} \mu_i(t, x_i)\partial_{x_i} v_t(x) - \sum_{i,j=1}^{d} [\sigma\sigma^\top]_{ij}(t, x)\partial_{x_i,x_j} v_t(x)] + v_t(x)V(x, t) - f(x, t)$$

Then subject to Lip conditions it follows that

$$v_t(x) = \mathbb{E}_{X\sim Q}\left[\int_t^T e^{-\int_t^s V(X_s,s)\mathrm{d}r} f(X_s, s)\mathrm{d}s + e^{-\int_t^T V(X_r,r)\mathrm{d}r}\phi(X_T)\right]$$

with

$$\mathrm{d}X_t = \mu(X_t, t)\mathrm{d}t + \sigma(X_t, t)\mathrm{d}W_t$$