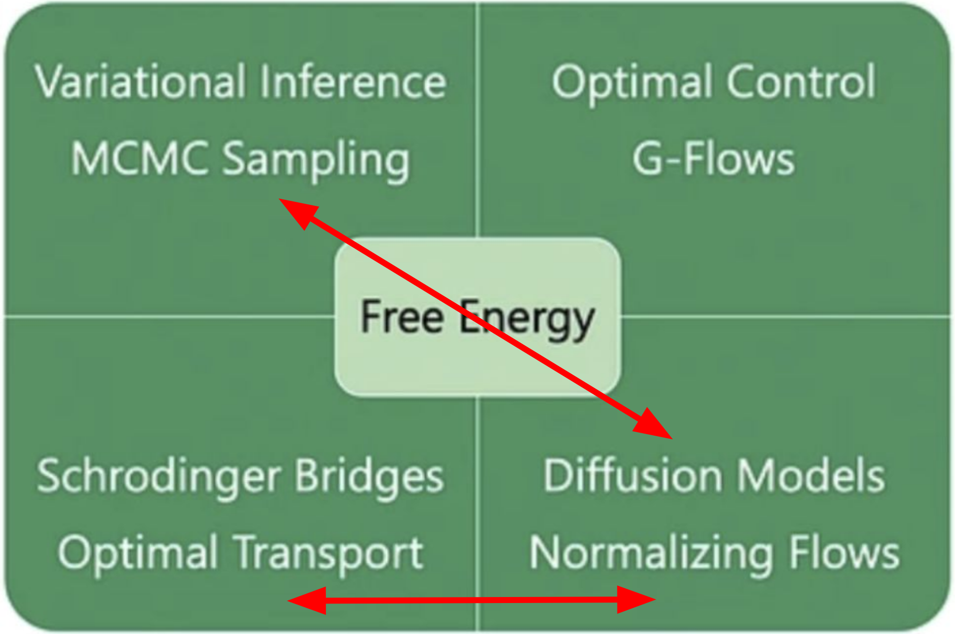
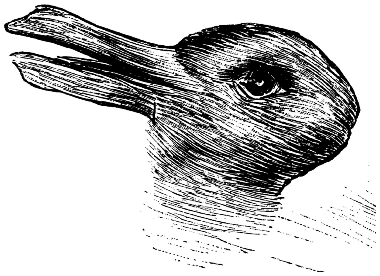
Two Perspectives on Stochastic RWs

Diffusion models v.s. Score function

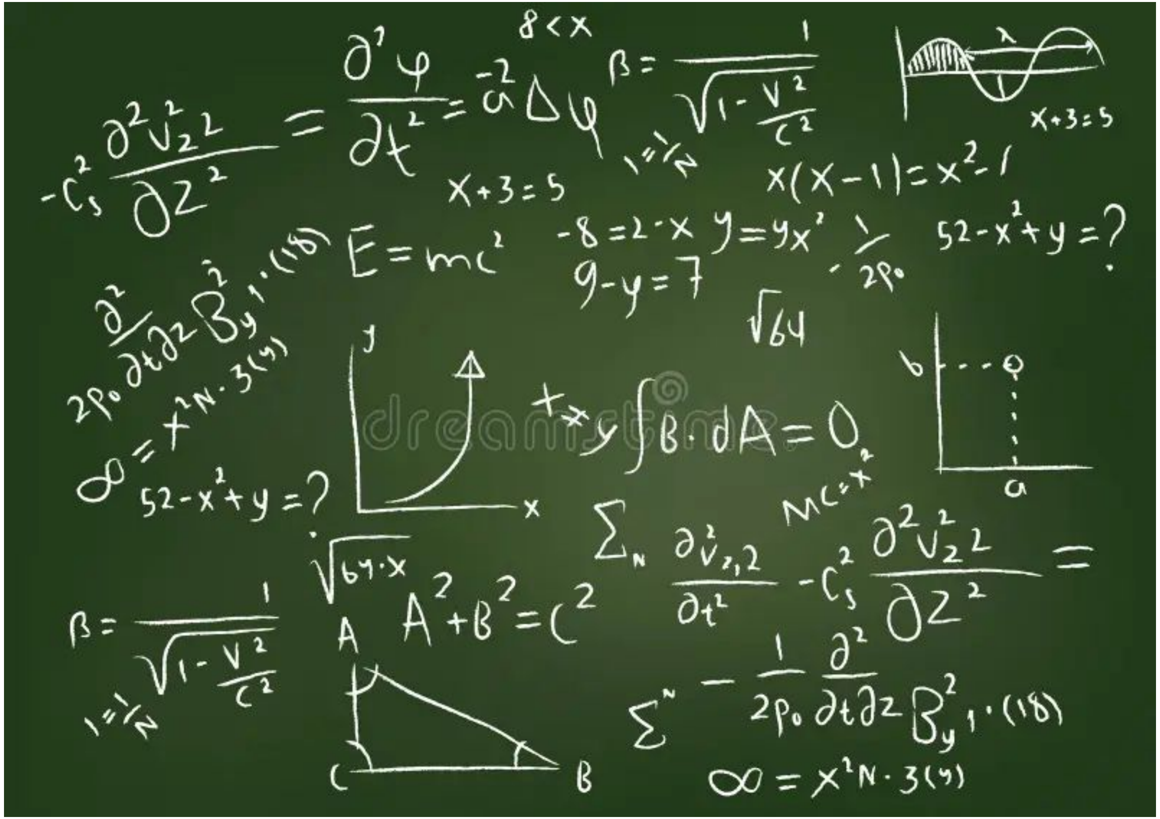

Variational Inference
MCMC Sampling
Optimal Control
G-Flows
Free Energy
Schrodinger Bridges
Optimal Transport
Diffusion Models
Normalizing Flows

Annoying Things about Probability Distributions



- Scoring
- Sampling

Too Many Equations! Which ones matter?



This one.

$$X_{t+1} = X_t + \epsilon f(X_t) + \sqrt{\epsilon} Z$$

$$Z, X_0 \sim \mathcal{N}(0, I)$$

This is simple, flexible recipe for generating a probability distribution.

Take out the noise term - becomes a (discretized) ODE/Normalizing Flow
Take out the drift term - (discrete) brownian motion!
Take out the X_t term - it's basically a noisy RNN

This one.

$$X_{t+1} = X_t + \epsilon f(X_t) + \sqrt{\epsilon} Z$$

$$Z, X_0 \sim \mathcal{N}(0, I)$$

This is simple, flexible recipe for generating a probability distribution.

X_t	- Adds Stability c.f. RNNs v.s. ResNets
$\epsilon f(X_t)$	- Trainability (actually shapes the resulting distribution)
$\sqrt{\epsilon} Z$	- Stochasticity

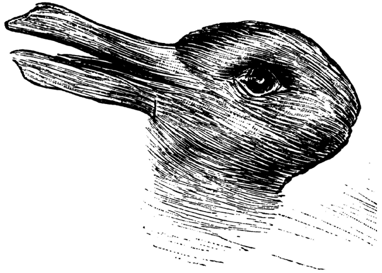
Can Generate Good(-ish) Distributions

$$X_{t+1} = X_t + \epsilon f(X_t) + \sqrt{\epsilon} Z$$

Unlike all those ingredients separately, it can actually model complex datasets like image data quite well.

…if we pick train the model to pick f correctly,

Any ideas on how to train it?



$$X_{t+1} = X_t + \epsilon f(X_t) + \sqrt{\epsilon} Z$$

Two things to notice about it.

1. It's a discretization of an SDE
   (and we can analyze the corresponding Fokker-Plank)
   Leads to Score function learning.

2. It's a Markov Chain
   (and can be the inverse of another, simpler markov chain)
   Leads to Diffusion models.

Variational Inference Recap

$$\ln p(x) + \ln p(z|x) = \ln p(x|z) + \ln p(z)$$

$$-\ln p(x) + -\ln \frac{p(z|x)}{q_\phi(z;x)} = -\ln \frac{p(x|z)}{q_\phi(z;x)} + -\ln \frac{p(z)}{q_\phi(z;x)} + -\ln q_\phi(z;x)$$

Variational Inference Recap

$$\mathbb{E}_{q_\phi(z;x)}\left[-\ln p(x) + -\ln \frac{p(z|x)}{q_\phi(z;x)}\right]$$

=

$$\mathbb{E}_{q_\phi(z;x)}\left[-\ln \frac{p(x|z)}{q_\phi(z;x)} + -\ln \frac{p(z)}{q_\phi(z;x)} + -\ln q_\phi(z;x)\right]$$

The Core of Variational Inference

$$-\ln p_\theta(x) + KL(q_\phi(z;x)||p_\theta(z|x))$$

=

$$KL(q_\phi(z;x)||p_\theta(z)) + \mathbb{E}\left[-\ln \frac{p(x|z)}{q_\phi(z;x)}\right] + H(q_\phi(z;x))$$

In the Case of a Diffusion Model

$$-\ln p_\theta(x_0)+KL(q_\phi(x_1,x_2,...,x_T;x_0)||p_\theta(x_1,x_2,...,x_T|x_0))$$

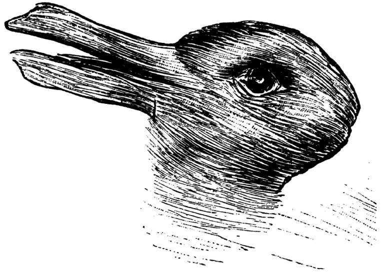
=

$$KL(q_\phi(x_1,x_2,...,x_T;x_0)||p_\theta(x_1,x_2,...,x_T))$$

$$+\mathbb{E}_{q_\phi(x_1,x_2,...,x_T;x_0)}\left[-\ln\frac{p_\theta(x_0|x_1,x_2,...,x_T)}{q_\phi(x_1,x_2,...,x_T;x_0)}\right]$$

$$+H(q_\phi(x_1,x_2,...,x_T;x_0))$$

The Actual Insight of the Paper



$$\mathbf{x}_t = \sqrt{1-\beta_t}\mathbf{x}_{t-1} + \sqrt{\beta_t}\boldsymbol{z}_t$$

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t$$

$$\alpha_t = 1 - \beta_t \qquad \bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$$

The Actual Insight of the Paper

$$D_{\mathrm{KL}}(p\,||\,q) = \frac{1}{2}\left(\log\frac{|\Sigma_q|}{|\Sigma_p|} - d + \mathrm{tr}\left(\Sigma_q^{-1}\Sigma_p\right) + (\mu_q - \mu_p)^T\Sigma_q^{-1}(\mu_q - \mu_p)\right)$$

$$||\mu_p - \mu_q||^2 + C$$

$$\mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\frac{\beta_t^2}{2\sigma_t^2\alpha_t(1-\bar{\alpha}_t)}\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)\right\|^2\right]$$

Sampling With an SDE

Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \mathrm{Uniform}(\{1, \ldots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$$
6: **until** converged

Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \ldots, 1$ **do**
3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

Fokker Plank to Guide Choice of SDEs

$$dX_t = a(X_t, t)\, dt + b(X_t, t)\, dW_t$$

$$\frac{\partial p(x,t)}{\partial t} = -\frac{\partial}{\partial x}\big(a(x,t)p(x,t)\big) + \frac{1}{2}\frac{\partial^2}{\partial x^2}\big(b(x,t)^2 p(x,t)\big)$$

Fokker Plank to Guide Choice of SDEs

$$\frac{\partial p(x,t)}{\partial t} = -\frac{\partial}{\partial x}\Big(\frac{1}{p_\infty(x)}\frac{\partial p_\infty(x)}{\partial x}p(x,t)\Big)+\frac{\partial^2}{\partial x^2}\big(p(x,t)\big)$$

$$\frac{\partial p(x,t)}{\partial t} = -\frac{\partial}{\partial x}\Big(\frac{p(x,t)}{p_\infty(x)}\Big)\frac{\partial p_\infty}{\partial x}-\frac{p(x,t)}{p_\infty(x)}\frac{\partial^2}{\partial x^2}\big(p_\infty(x)\big)+\frac{\partial^2}{\partial x^2}\big(p(x,t)\big)$$

$$\frac{\partial p(x,t)}{\partial t} = 0 \qquad \text{if} \qquad p(x,t) = p_\infty(x)$$