確率数理要論(12)

機械学習応用
- 勾配ランジュバン動力学
- 拡散モデル

GLD

- 勾配ランジュバン動力学

  Stochastic Gradient Langevin Dynamics (SGLD)

  非凸最適化：

$$\min_{x\in\mathbb{R}^d} L(x) = \min_{x\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \ell_i(x)$$

(非凸でも良い)

  サンプリング：

$$X \sim \exp(-\beta L(X))/Z_{\beta,L}$$

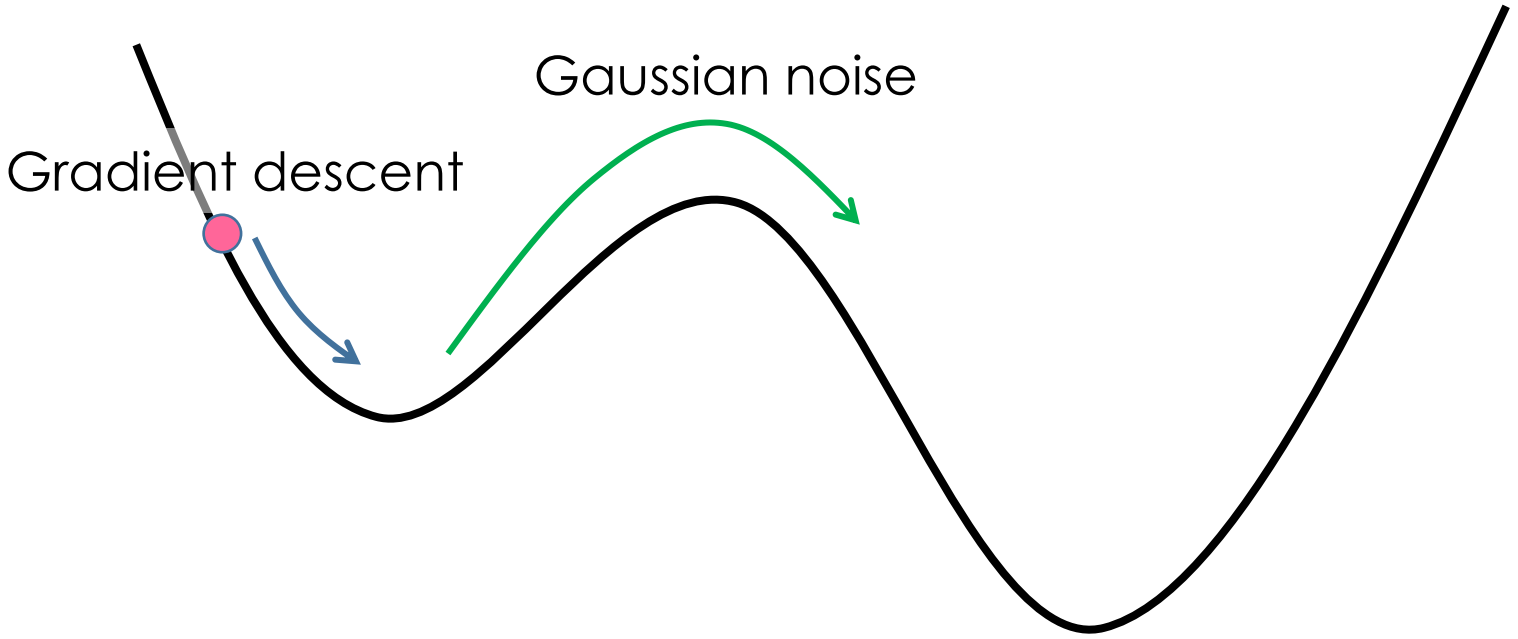
β: 逆温度パラメータ

$$\mathrm{d}X_t = -\nabla L(X_t)\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}B_t$$

(勾配ランジュバン動力学)

定常分布： $\pi \propto \exp(-\beta L(X))$

[Gelfand and Mitter (1991); Borkar and Mitter (1999); Welling and Teh (2011)]


Gaussian noise
Gradient descent

GLDのFokker-Planck方程式

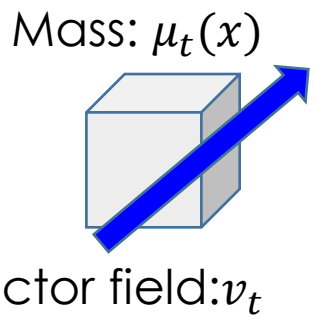
$$\mathrm{d}X_t = -\nabla L(X_t)\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}B_t$$

μ_t ：X_tの分布 (密度関数であると考えてもらって問題ない)

Fokker-Planck方程式

$$\nabla \cdot [\mu_t \nabla L] = \sum_{j=1}^{d} \partial_i [\mu_t \partial_i L]$$

$$\partial_t \mu_t = \frac{1}{\beta}\Delta_x \mu_t + \nabla \cdot [\mu_t \nabla L]$$



次のように解釈できる:

$$\partial_t \mu_t = \nabla \cdot \left[\underbrace{\left(\frac{1}{\beta}\nabla \log(\mu_t) + \nabla L \right)}_{-v_t \text{とおく}} \mu_t \right] = -\nabla \cdot [v_t \mu_t]$$

[連続の方程式]

連続の方程式

「連続の方程式」 $\dfrac{\partial \rho_t}{\partial t} = -\nabla \cdot (v_t \rho_t)$

この方程式の意味

$$\frac{\mathrm{d}}{\mathrm{d}t}\int f(w)\mathrm{d}\rho_t(w) = \int (\nabla f(w))^\top v_t(w)\mathrm{d}\rho_t(w)$$

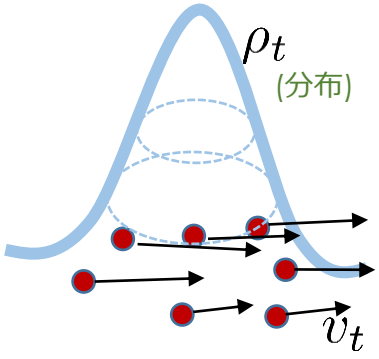
$(= -\int f(w)\mathrm{d}[\nabla \cdot (v_t\rho_t)])$ ($\forall f$: コンパクトサポート，C^∞-級)

- ベクトル場v_tで生成される写像をT_tとする: $\dfrac{\mathrm{d}T_t}{\mathrm{d}t}(w) = v_t\big(T_t(w)\big)$.
- ρ_tは写像$T_t: R^d \to R^d$によるρ_0の押し出し：$\rho_t = T_{t\#}\rho_0$.
  つまり，$w \sim \rho_0$に対する$T_t(w)$の分布がρ_t.

($t = 0$で導出: $T_0 = I$ (恒等写像))

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\int f(w)\mathrm{d}\rho_t(w) &= \frac{\mathrm{d}}{\mathrm{d}t}\int f(T_t(w))\mathrm{d}\rho_0(w)\\ &= \int \nabla f(T_0(w))^\top \frac{\mathrm{d}T_t(w)}{\mathrm{d}t}\mathrm{d}\rho_0(w)\\ &= \int \nabla f(w)^\top v_0(w)\mathrm{d}\rho_0(w)\end{aligned}$$

[連続の方程式]



連続の方程式

「連続の方程式」 $\dfrac{\partial \rho_t}{\partial t} = -\nabla \cdot (v_t \rho_t)$

この方程式の意味

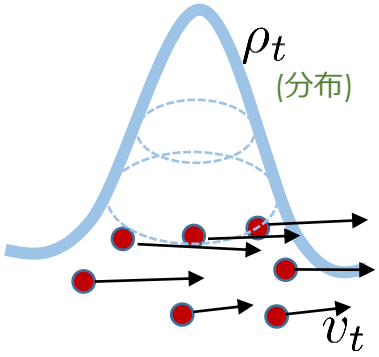
$$\frac{\mathrm{d}}{\mathrm{d}t}\int f(w)\mathrm{d}\rho_t(w) = \int (\nabla f(w))^\top v_t(w)\mathrm{d}\rho_t(w)$$

$(= -\int f(w)\mathrm{d}[\nabla \cdot (v_t\rho_t)])$ ($\forall f$: コンパクトサポート，C^∞-級)

- ベクトル場v_tで生成される写像をT_tとする: $\dfrac{\mathrm{d}T_t}{\mathrm{d}t}(w) = v_t\big(T_t(w)\big)$.
- ρ_tは写像$T_t: R^d \to R^d$によるρ_0の押し出し：$\rho_t = T_{t\#}\rho_0$.
  つまり，$w \sim \rho_0$に対する$T_t(w)$の分布がρ_t.

(一般のt)

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\int f(w)\mathrm{d}\rho_t(w) &= \frac{\mathrm{d}}{\mathrm{d}t}\int f(T_t(w))\mathrm{d}\rho_0(w)\\ &= \int \nabla f(T_t(w))^\top \frac{\mathrm{d}T_t(w)}{\mathrm{d}t}\mathrm{d}\rho_0(w)\\ &= \int \nabla f(T_t(w))^\top v_t(T_t(w))\mathrm{d}\rho_0(w)\\ &= \int \nabla f(w)^\top v_t(w)\mathrm{d}\rho_t(w). \quad \text{(連続の方程式)}\end{aligned}$$



定常分布

$$\partial_t \mu_t = \nabla \cdot \left[\left(\frac{1}{\beta} \nabla \log(\mu_t) + \nabla L \right) \mu_t \right] = -\nabla \cdot [v_t \mu_t]$$

定常分布: $\partial_t \mu_t = 0 \Rightarrow v_t = 0$ (分布がこれ以上動かない)

$$\frac{1}{\beta} \nabla \log(\mu^*) + \nabla L = 0 \quad \Rightarrow \quad \mu^*(x) \propto \exp(-L(x))$$

実は，これは以下の目的関数を最小化するWasserstein勾配流である:

$$\mu^* = \arg\min_{\mu \in \mathcal{P}} \int L(x) \mathrm{d}\mu(x) + \frac{1}{\beta} \mathrm{Ent}(\mu) =: \mathcal{L}(\mu)$$

$$(\mathrm{Ent}(\mu) = \int \log(\mu) \mathrm{d}\mu)$$

➡ 確かにこの最適解は定常分布と等しい:

$$\mu^*(x) \propto \exp(-\beta L(x))$$

Wasserstein勾配流

$$\beta\mathcal{L}(\mu) = \int \beta L(x)\mathrm{d}\mu(x) + \mathrm{Ent}(\mu)$$

$$\mu^*(x) \propto \exp(-\beta L(x))$$

$$= \int -\log(\mu^*)\mathrm{d}\mu + \int \log(\mu)\mathrm{d}\mu + (\mathrm{const.})$$

以下，無視

$$= \int \log\left(\frac{\mu}{\mu^*}\right)\mathrm{d}\mu = \mathrm{KL}(\mu||\mu^*)$$

連続の方程式 $\mu_t = -\nabla\cdot[v_t\mu_t]$ に従っているなら

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(\mu_t||\mu^*) = \frac{\mathrm{d}}{\mathrm{d}t}\int \log\left(\frac{\mu_t(x)}{\mu^*(x)}\right)\mu_t(x)\mathrm{d}x$$

$$= \int \log\left(\frac{\mu_t(x)}{\mu^*(x)}\right)\partial_t\mu_t(x)\mathrm{d}x + \int \frac{\partial_t\mu_t(x)}{\cancel{\mu_t(x)}}\cancel{\mu_t(x)}\mathrm{d}x$$

$$= \int \log\left(\frac{\mu_t(x)}{\mu^*(x)}\right)\nabla\cdot(-v_t\mu_t(x))\mathrm{d}x$$

$$= -\int \langle v_t, \nabla\log(\mu^*) - \nabla\log(\mu_t)\rangle\mathrm{d}\mu_t$$

Wasserstein勾配流

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(\mu_t||\mu^*) = -\int \langle v_t, \nabla \log(\mu^*) - \nabla \log(\mu_t)\rangle \mathrm{d}\mu_t$$

特に

$$v_t = -\left(\frac{1}{\beta}\nabla \log(\mu_t) + \nabla L\right) = \frac{1}{\beta}\left(\nabla \log(\mu^*) - \nabla \log(\mu_t)\right) \quad \text{(GLD)}$$

は最急降下方向で，以下が成り立つ．

$$\partial_t \mu_t = \nabla \cdot \left[\underbrace{\left(\frac{1}{\beta}\nabla \log(\mu_t) + \nabla L\right)}_{=: -v_t} \mu_t\right]$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{KL}(\mu_t||\mu^*) = -\frac{1}{\beta}\int \|\nabla \log(\mu^*) - \nabla \log(\mu_t)\|^2 \mathrm{d}\mu_t$$
$$= -\frac{1}{\beta} I(\mu_t||\mu^*)$$

定常分布μ^*からのKL-divを最小化するWasserstein勾配流

Fisher divergence:

$$I(\mu||\nu) = \int \|\nabla \log(\nu) - \nabla \log(\mu)\|^2 \mathrm{d}\mu$$

Wasserstein距離について

参考

μ, ν:距離空間$(\mathcal{X}, c)$上の確率測度(通常$\mathcal{X}$はPoland空間)

$$W_p(\mu,\nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} c(x,y)^p \mathrm{d}\pi(x,y) \right)^{1/p}$$

$\Pi(\mu,\nu)$: 周辺分布がμ,νである$\mathcal{X}\times\mathcal{X}$上の同時分布の集合

周辺分布を固定した同時分布の中で最小化

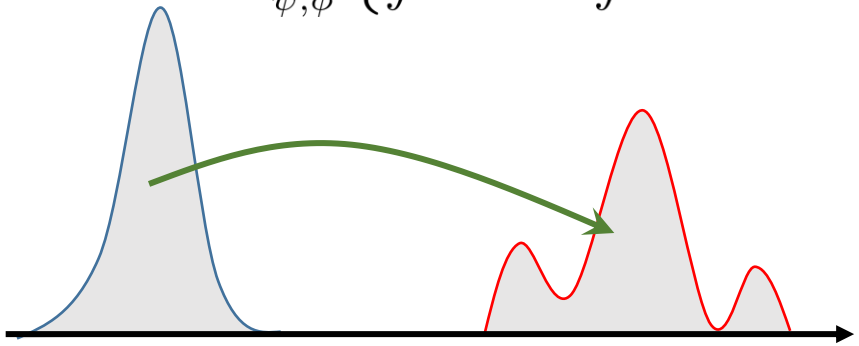
$$(\mathcal{X} = \mathbb{R}^d:\ c(x,y) = \|x-y\|)$$

- 分布のサポートがずれていてもwell-defined
- 底空間の距離が反映されている

  ※KL-divergenceは距離が反映されない.

(双対表現: Kantorovich双対)

$$\inf_{\pi\in\Pi(\mu,\nu)} \int c(x,y)^p \mathrm{d}\pi(x,y) = \sup_{\psi,\phi} \left\{ \int \psi \mathrm{d}\mu + \int \phi \mathrm{d}\nu \mid \psi(x)+\phi(y) \le c(x,y)^p \right\}$$



「輸送距離」とも言われる

接ベクトル

参考

定理

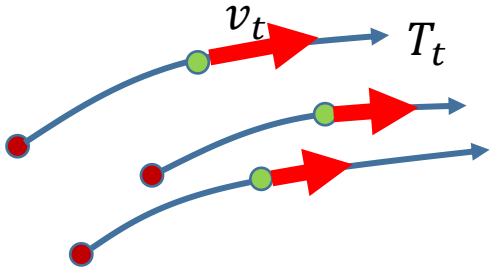
- $\rho_t = T_{t\#}\rho_0$
- $\frac{\mathrm{d}T_t}{\mathrm{d}t}(w) = v_t\big(T_t(w)\big)$
- あるφ$_t$を用いて$v_t = \nabla\phi_t$と書けるとする.

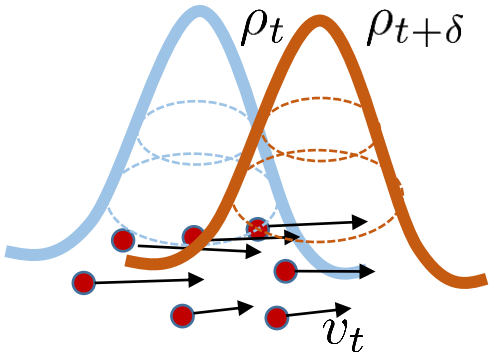
この時，以下が成り立つ:

$$\lim_{\delta\to 0}\frac{W_2(\rho_{t+\delta}, (\mathrm{id}+\delta v_t)_{\#}\rho_t)}{\delta} = 0$$

詳細は以下を参照:

Ambrosio, Gigli, and Savaré. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics. ETH Zürich. Birkhäuser Basel, 2008.





輸送写像

参考

Brenierの定理

ρ_0, ρ_1が確率密度関数を持つ時，以下が成り立つ:

$$W_2^2(\rho_0, \rho_1) = \inf_{T:T_{\#}\rho_0=\rho_1} \mathbb{E}_{X\sim\rho_0}[\|X - T(X)\|^2]$$

- Infを達成する写像T^*が存在する.
- しかも，ある凸関数ψが存在して$T^*(x) \in \partial\psi(x)$と書ける.
- このT^*を最適輸送写像という.

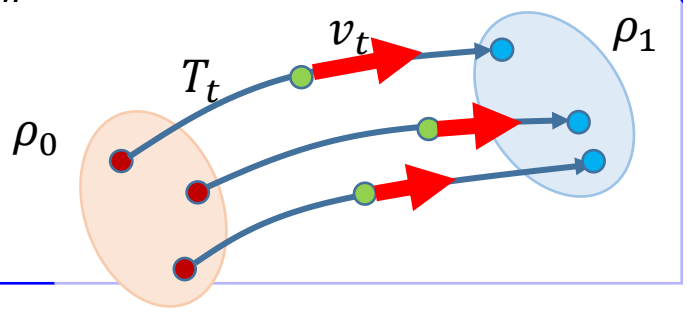
Benamou-Brenier formula (連続の方程式とW_2距離の関係):

同条件のもと

$$W_2^2(\rho_0, \rho_1) = \inf_{\{v_t\}_t} \int_0^1 \|v_t\|^2_{L_2(\rho_t)} \mathrm{d}t$$

ただし，infはρ_0からρ_1へ連続の方程式で“繋ぐ”全ての速度ベクトル場v_tに関して取る.

- $\rho_t = T_{t\#}\rho_0$
- $\frac{\mathrm{d}T_t}{\mathrm{d}t}(w) = v_t\big(T_t(w)\big)$

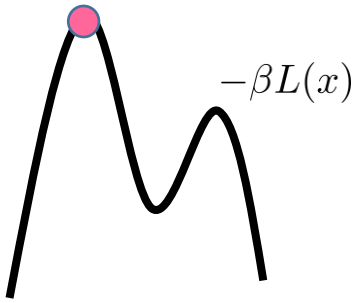


定常分布

$$\mathrm{d}X_t = -\nabla L(X_t)\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}B_t$$

適当な条件のもとGLDの定常分布は以下で与えられる:

$$\pi_\infty(\mathrm{d}x) \propto \exp(-\beta L(x))\mathrm{d}x$$

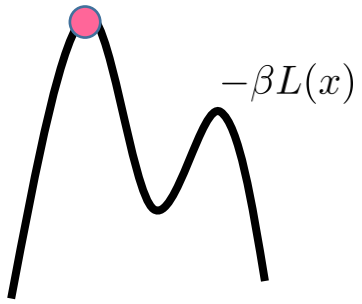


- βが大きければ，定常分布は最適解周りに集中する.
- 定常分布からのサンプリングにも使える (というかこっちが本来の目的).

対数ソボレフ不等式と幾何的エルゴード性

定常分布

$$\pi_\infty(\mathrm{d}x) \propto \exp(-\beta L(x))\mathrm{d}x$$



例：
- 二次関数+有界関数
- Weak Morse型関数

対数ソボレフ不等式 (π_∞の性質):

任意の(π_∞に対して絶対連続な)確率分布 $\mathrm{d}\nu = f\ \mathrm{d}\pi_\infty$ に対し,

$$\int f\log(f)\mathrm{d}\pi_\infty \leq \frac{c_{\mathrm{LS}}}{2}\int \frac{\|\nabla f\|^2}{f}\mathrm{d}\pi_\infty$$

$$(D(\nu||\pi_\infty) \leq \frac{c_{\mathrm{LS}}}{2} I(\nu||\pi_\infty))$$

KL-div

$$D(\mu||\nu) = \int \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\mathrm{d}\mu$$

Fisher-div

$$I(\mu||\nu) = \int \left\|\nabla \log \frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right\|^2 \mathrm{d}\mu$$

➡ **幾何的エルゴード性**

ρ_t: X_tの周辺分布

$$D(\rho_t||\pi_\infty) \leq \exp(-2t/c_{\mathrm{LS}})D(\rho_0||\pi_\infty)$$

定常分布へKL-divergenceの意味で指数オーダーの収束

[Bakry, Gentil, and Ledoux: *Analysis and Geometry of Markov Diffusion Operators*. Springer, 2014. Th. 5.2.1]

対数ソボレフ不等式の十分条件

強凸な場合: $L(x)$ がμ-強凸 ⇒ $c_{\mathrm{LS}} \leq 1/(\mu\beta)$

[Bakry and Émery, 1985]

$$L(x) = h(x) + \lambda_1 \|x\|^2$$

Bounded perturbation lemma (Hollley-Stroock):

$$|h(x)| \leq B \ (\forall x) \quad \Rightarrow \quad c_{\mathrm{LS}} \leq \frac{1}{2\lambda_1 \beta} \exp(4\beta B)$$

[R. Holley and D. Stroock. Logarithmic sobolev inequalities and stochastic Ising models. Journal of statistical physics, 46(5-6):1159–1194, 1987.]

- **散逸的 (dissipative):**

$$\exists m > 0,\ b \geq 0,\ \langle x, \nabla L(x) \rangle \geq m\|x\|^2 - b$$

- 平滑性:

$$\exists M,\ \|\nabla L(x) - \nabla L(y)\| \leq M\|x - y\|$$

$$\Rightarrow \quad c_{\mathrm{LS}} \leq \frac{2m^2 + 8M^2}{m^2 M\beta} + \left(\frac{6M(d+\beta)+2}{m}\right) e^{O(\beta+d)}$$

[Raginsky, Rakhlin and Telgarsky, 2017]

離散時間ダイナミクスの収束レート

過程: L はM-平滑: $\exists M,\ \|\nabla L(x) - \nabla L(y)\| \leq M\|x - y\|$

定理 [Vempala and Wibisono, 2019]

ν_k: Marginal distribution of X_k (discrete time dynamics)

$$D(\nu_k||\pi_\infty) \lesssim \exp(-k\eta/c_{\mathrm{LS}})D(\nu_0||\pi_\infty) + 8c_{\mathrm{LS}}dM^2\eta$$

定理 (informal)

散逸性と平滑性の条件のもと (and other technical condition),

$$\mathrm{E}[L(X_k)] - L(X^*) \lesssim \exp\left(-ck\eta/c_{\mathrm{LS}}\right) + c_{c_{\mathrm{LS}},\beta,d}\eta + \frac{d\log(\beta+1)}{\beta}$$

- $\exp\left(-ck\eta/c_{\mathrm{LS}}\right)$: 幾何的エルゴード性
- $c_{c_{\mathrm{LS}},\beta,d}\eta$: 時間離散化の誤差
- $\frac{d\log(\beta+1)}{\beta}$: $E_{\pi_\infty}[L(X)] - L(X^*)$ 定常分布が最適解まわりにどれだけ集中しているか

where $c, c_{C_{LS},\beta,d} > 0$ are constants.

[Raginsky, Rakhlin and Telgarsky, 2017; Xu, Chen, Zou, and Gu, 2018; Erdogdu, Mackey and Shamir, 2018]

- 逆温度パラメータβが十分大きければ，目的関数が非凸でも最適解の近くに到達できる.
- ただし，一般には対数ソボレフ不等式はβに指数的に依存することに注意.
  （そうでない場合もある：強凸目的関数，Weak Morse関数）

拡散モデル

DALL·E/DALL·E 2

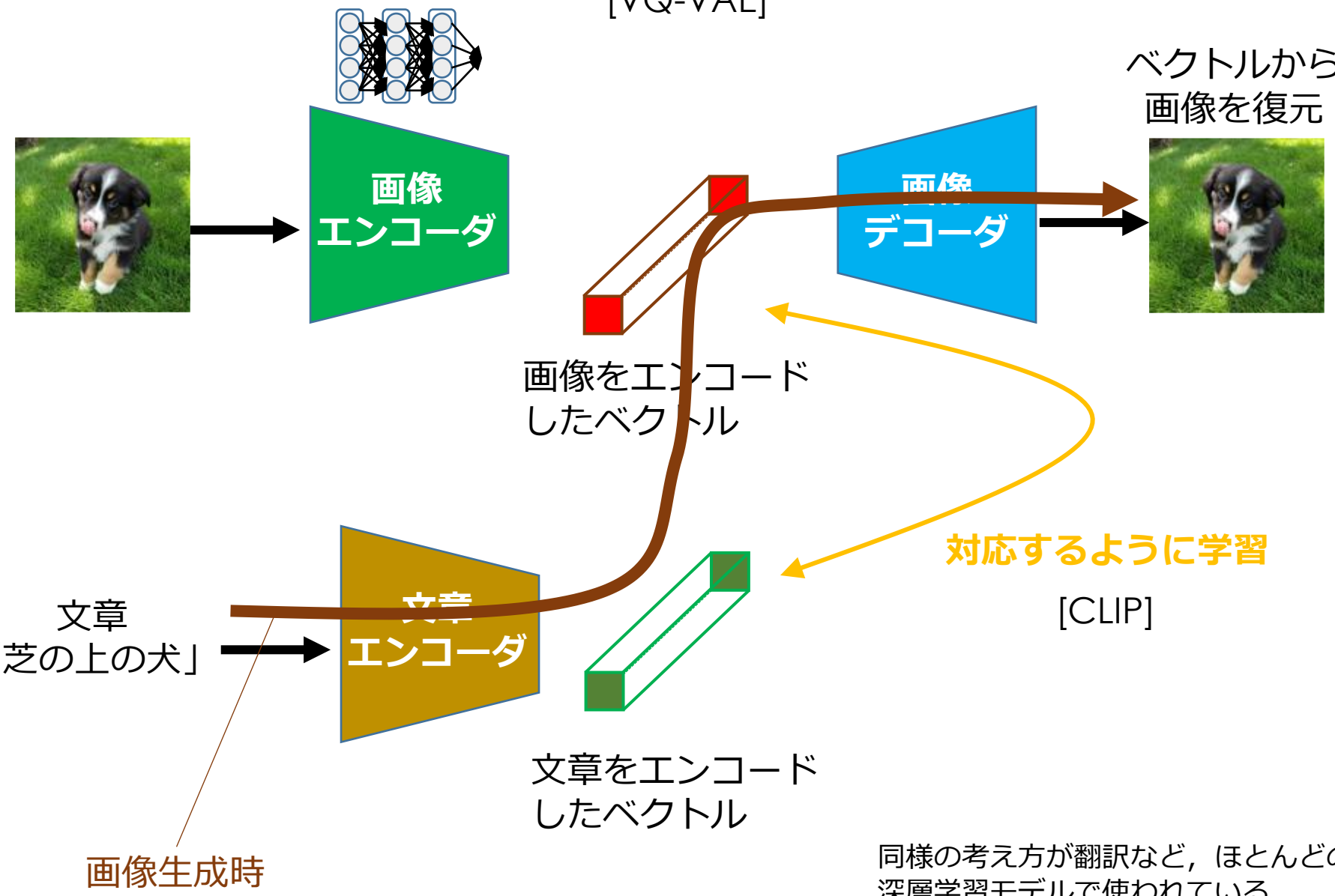
文章による説明から画像を生成するモデル

「An astronaut riding a horse in a photorealistic style」

DALL·E: [Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, Ilya Sutskever: Zero-Shot Text-to-Image Generation. ICML2021.]
DALL·E2:[Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, Mark Chen: Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv:2204.06125]

「Teddy bears shopping for groceries in the style of ukiyo-e」

特徴ベクトルの役割：DALL·Eで説明



同様の考え方が翻訳など，ほとんどの深層学習モデルで使われている．

他の作例

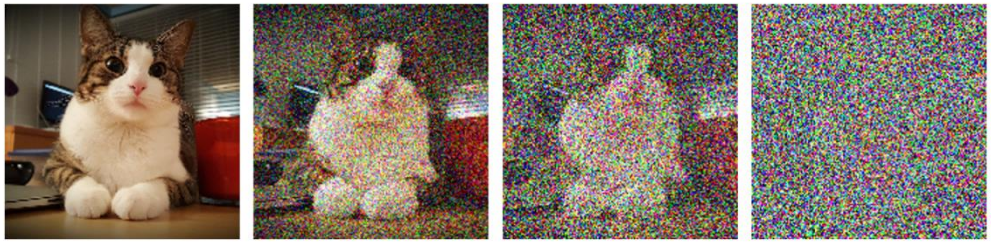
Jason Allen "Théâtre D'opéra Spatial“ generated by **Midjourney**. Colorado State Fair’s fine art competition, 1st prize in digital art category

Generated by NovelAI

デコーダー：拡散モデル (GLIDE)

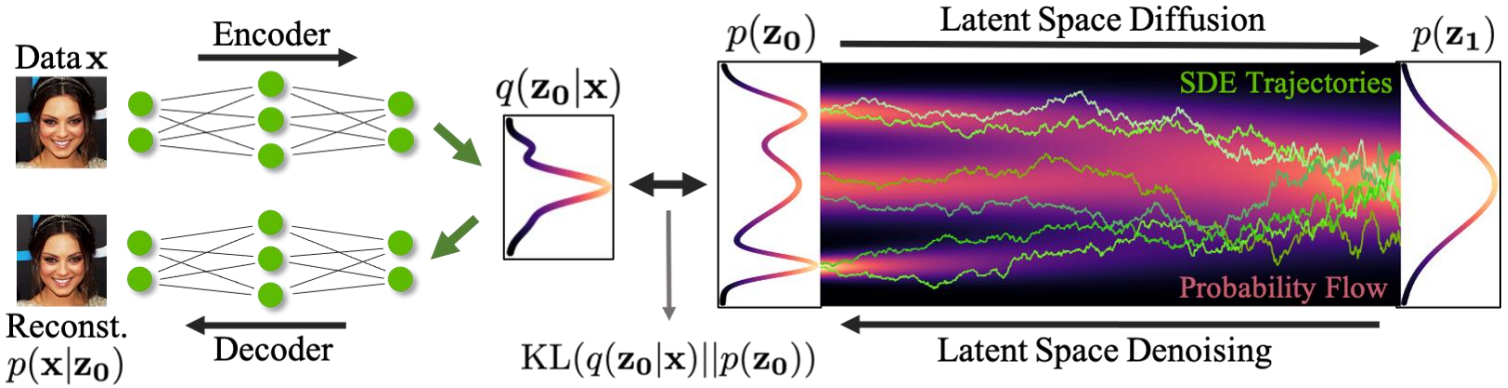
順モデル：所望の分布からノイズに変換

$$\mathrm{d}X_t = f(x,t)\mathrm{d}t + \mathrm{d}B_t$$



$$\mathrm{d}X_t = [f(x,t) - \nabla \log(p_t(X_t))]\mathrm{d}t + \mathrm{d}B_t$$

逆モデル：ノイズから所望の分布へ変換



- 順過程：ノイズの分布が標準正規分布の場合

標準正規分布へ向かう勾配ランジュバン動力学

$$\mu_\infty \propto \exp\left(-\frac{\|x\|^2}{2}\right) = \exp(-L(x))$$

$$\Rightarrow L(x) = \frac{\|x\|^2}{2}$$

$$\Rightarrow \mathrm{d}X_t = -X_t \mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$

（OU-過程）

Fokker-Planck方程式：$\partial_t \mu_t = \Delta \mu_t + \nabla \cdot (x \mu_t)$

$$\partial_t \mu_t = \frac{1}{\beta}\Delta_x \mu_t + \nabla \cdot [\mu_t \nabla L]$$

- 逆過程 (reverse SDE)：

$\tilde{t}$を逆向き時間として ($\infty \to 0$へ向かう)

$$\mathrm{d}X_{\tilde{t}} = (X_{\tilde{t}} + 2\nabla \log(\mu_{\tilde{t}}(x)))\mathrm{d}\tilde{t} + \sqrt{2}\mathrm{d}B_{\tilde{t}}$$

$b(x) = -x = -\nabla L(x)$ として次のページで導出.

逆向きSDEのFK-方程式

$$\partial_s p(s,x|t,y) = \partial_s \left[p(t,y|s,x)\frac{p(s,x)}{p(t,y)} \right]$$

($s < t$を想定)

$$= \partial_s p(t,y|s,x)\frac{p(s,x)}{p(t,y)} + p(t,y|s,x)\frac{\partial_s p(s,x)}{p(t,y)}$$

$$= [-b^\top \nabla_x p(t,y|s,x) - \Delta_x p(t,y|s,x)]\frac{p(s,x)}{p(t,y)}$$

$$+ p(t,y|s,x)\frac{-\nabla_x \cdot (bp(s,x)) + \Delta_x p(s,x)}{p(t,y)}$$

$$b^\top \nabla_x p(t,y|s,x)\frac{p(s,x)}{p(t,y)} = b^\top \nabla_x p(s,x|t,y) - p(t,y|s,x)\frac{b^\top \nabla_x p(s,x)}{p(t,y)}$$

$$\Delta_x p(t,y|s,x)\frac{p(s,x)}{p(t,y)} = \Delta_x p(s,x|t,y) - 2\nabla_x^\top p(t,y|s,x)\frac{\nabla_x p(s,x)}{p(t,y)} - p(t,y|s,x)\frac{\Delta_x p(s,x)}{p(t,y)}$$

$$= \Delta_x p(s,x|t,y) - 2\nabla_x^\top p(s,x|t,y)\nabla_x \log(p(s,x))$$

$$+ 2p(s,x|t,y)\|\nabla_x p(s,x)\|^2 - p(t,y|s,x)\frac{\Delta_x p(s,x)}{p(t,y)}$$

$$= \Delta_x p(s,x|t,y) - 2\nabla_x^\top p(s,x|t,y)\nabla_x \log(p(s,x))$$

$$- 2p(s,x|t,y)\Delta_x \log(p(s,x)) + p(t,y|s,x)\frac{\Delta_x p(s,x)}{p(t,y)}$$

$$p(t,y|s,x)\frac{-\nabla_x \cdot (bp(s,x))}{p(t,y)} = -(\nabla_x \cdot b + b^\top \nabla_x \log(p(s,x)))p(s,x|t,y)$$

まとめると，

$$\partial_s p(s,x|t,y) = -\nabla_x \cdot [(b - 2\nabla_x \log(p(s,x)))p(s,x|t,y)] - \Delta_x p(s,x|t,y)$$

時間を反転させて，$\mathrm{d}\tilde{s} \leftarrow -\mathrm{d}s$とすると，

$$\partial_{\tilde{s}} p(\tilde{s},x|t,y) = \nabla_x \cdot [(b - 2\nabla_x \log(p(\tilde{s},x)))p(\tilde{s},x|t,y)] + \Delta_x p(\tilde{s},x|t,y)$$

これはドリフト項が

$$-\Big(b - 2\nabla_x \log\big(p(s,x)\big)\Big) = x + 2\nabla_x \log(\mu_s(x))$$

の拡散過程の前向き方程式に他ならない．

$\tilde{s} \to 0$とすることで，時刻0における分布を得ることができる．
つまり，ドリフト項をデータから推定し，逆過程を走らせることでデータの分布からのサンプリングができるようになる．