確率過程の基礎と拡散モデルの数理

安田 勇輝

2025 年 10 月 9 日
Ver 1.0.1

目次

1 はじめに

デノイジング拡散確率モデル (以下，拡散モデル [1, 2]) はその推論精度の高さから現在盛んに応用されている．拡散モデルは，ノイズに多数回の変換を施すことで多様なデータを生成する．この変換は拡散過程と呼ばれ，確率微分方程式 (Stochastic Differential Equation; SDE) で数学的に記述される [3]．この SDE はニューラルネットワークで近似され，SDE の多数回の時間ステップの積分が，ニューラルネットワークによる微小変換の積み重ねとなる．この積み重ねにより，非常に深いニューラルネットワークが実効的に構築され，高精度な推論が実現される [4]．しかし，拡散モデルの表式は複雑そうに見え，実装も簡単には見えない．

一般に，ノイズに駆動されるランダム系の時間発展を扱う分野を「確率過程」や「確率解析」と呼ぶ [5, 6]．SDE はこのランダムな時間発展の表現の一形式であり，確率解析の中心を担う．確率解析の一般的な枠組みから眺めることで，拡散モデルの数理が浮き彫りとなり，その理解は大きく進む [7, 8]．しかし，確率過程は伝統的に数理ファイナンスでよく用いられてきたため，その説明は数学的であり，機械学習の研究者にとって必ずしも分かり易くない．そこで，このノートでは数学的な厳密さではなく，直感的な意味に重きをおき，確率過程の基礎を説明する．そして，確率解析の知識に基づき，拡散モデルの原理を説明する．

まず，なぜ微分方程式をデータ生成に利用するのかを議論し，今後の流れを説明する．拡散モデルの登場以前から，微分方程式を用いたデータ生成は研究されてきた．例えば，正規化流 (normalizing flow[9]) や Neural ODE[10] などの枠組みが挙げられる．これらの枠組みでは，ガウス分布など既知の量からサンプリングしたベクトル $\boldsymbol{x}$ に変換を施すことで，望みのデータを得る．データが複雑になるほど，この変換は多様で学習が困難になると想像できる．そこで変換を分割し，(比較的単純な) 微小変換の積み重ねで記述することを考える．この積み重ねは，微分方程式の数値積分に類似している．例えば，$d\boldsymbol{x}/dt = \boldsymbol{f}(\boldsymbol{x}, t)$ を数値的に解く際，$\boldsymbol{x}_{t+\Delta t} = \boldsymbol{x}_t + \boldsymbol{f}(\boldsymbol{x}_t, t)\Delta t$ を反復計算する．この関数 $\boldsymbol{f}$ をニューラルネットワークで表現すれば，ニューラルネットワークによる多数回の変換が微分方程式の数値積分に対応する．ここで Δt は変換の大きさを制御する．また，$\boldsymbol{f}$ を t へ陽に依存させることで，回数を重ねると変換自体も変わる非定常性を考慮できる．拡散モデルはこの変換を SDE により記述する枠組みである．

では，SDE で変換を記述するにはどうすればよいか？明らかに，SDE の時間変化項 (上の $\boldsymbol{f}$ に相当) の推定が重要となる．また，既知の量を望みのデータに変換するため，その「既知の量」の設定も重要となる．拡散モデルは，この既知の量を正規ノイズに設定する．変換を構成するために，まず拡散モデルは任意のデータを正規ノイズへ時間発展させる変換を考える．この変換は SDE で記述される．推論時は，この SDE を逆回しすることで，ノイズから望みのデータを生成する．

ここまでの説明でいくつかの疑問が出てくるかもしれない．

疑問 (i)　任意のデータを正規ノイズへ時間発展させるには SDE をどう設定するか？
疑問 (ii)　この SDE を時間について逆回しする方法はなにか？
疑問 (iii)　そもそも SDE をどうやってデータから決定 (または学習) するか？

このノートの目的はこれら三つの疑問に答えることにある．疑問 (i) と (iii) は同じこと，つまり SDE の設定方法について述べており，何も違いがないように見える．しかし，実は任意のデータを正規ノイズへ収束させる SDE は，データに依存せず事前に設定できる [1, 2]．この事前に決めた SDE を逆回しするときに，スコア関数と呼ばれる量が必要となり [3, 11, 12]，このスコア関数をデータから推定する．まず第 2 章で，確率解析の基礎を説明して疑問 (i) に答える．次の第 3 章は，導いた SDE を時間について逆回しする方法を説明し，

スコア関数を導入することで疑問 (ii) に答える．そして，第 4 章で，スコア関数を推定するための学習方法について説明し [2, 13]，疑問 (iii) に答える．第 5 章では，拡散モデルの応用として条件付き生成問題を説明する [3, 14].

2　確率過程の基礎

この章では，任意のデータを正規ノイズへ収束させる以下の SDE (1) を導入し，この SDE の解の集まり (アンサンブル) の時間発展を考える．アンサンブルは確率分布で表現され，確率分布の時間発展は Fokker-Planck 方程式 (Fokker-Planck Equation; FPE) により記述される [5, 6]．この FPE を利用して，次の第 3 章で SDE を時間について逆回しする．

この章の主役は以下の SDE である [5, 6]．まずこれを考える動機を説明する．

$$\mathrm{d}x = -ax\ \mathrm{d}t + b\ \mathrm{d}W \tag{1}$$

ここで x は実変数，a と b は正定数，そして $\mathrm{d}W$ は Wiener 過程の微分である．この $\mathrm{d}W$ が確率的ノイズの役割を果たす．もしこのノイズがない場合，式 (1) は $\mathrm{d}x = -ax\ \mathrm{d}t$ となり，適当な初期条件を使って $x_t = x_0 e^{-at}$ とかける．下付き添字は時刻を表す意味で用いる．この解は十分時間が経ったあと，初期条件に依存せずに $x_t \to 0$ に収束する．ノイズ $b\,\mathrm{d}W$ を加えることで，0 ではなく標準正規ノイズ ϵ へ収束させる．「標準」というのは平均 0 かつ分散 1 の正規分布またはそれに従う変数を指す．疑問 (i) の答えは「強い摩擦力 $-ax$ を掛けることで変数 x を指数的に減衰させ，初期条件 x_0 の項を速やか 0 にし，ノイズの効果を支配的にする」となる．これを理解することが，この章の目的の一つである．

式 (1) は，実数 x に関するものであり，画像でいうと 1 ピクセルの画素値に対応する．こう説明すると，画像は画素値の集まり「ベクトル $\boldsymbol{x}$」だからこの議論は理想化されている，と感じるかもしれない．しかし，拡散モデル [1, 2] は，各ピクセルに式 (1) を独立に適用し，画素値を独立に減衰させる．そのため，以下の議論は画像などのベクトル変数にも適用できる．画素間の相関構造 (つまり画像データの構造) は，スコア関数を通して考慮される．拡散モデルは，画像の構造を無視して正規ノイズへ緩和させるという，一見すると乱暴な操作を行う．

2.1　確率微分方程式 (SDE)

式 (1) の理解には，ノイズ $\mathrm{d}W$ の性質を定義する必要がある．そこで，この式を時間に関して離散化し，数値的に解くことを考える．この考察を通して，ノイズ $\mathrm{d}W$ の統計性を明らかにする．

式 (1) において $b = 1$ として，Euler 法で離散化すれば，以下の漸化式を得る．

$$x_{k+1} = x_k + (-ax_k)\Delta t + \xi_k \tag{2}$$

添字 k は時間ステップを表し，ξ_k は時刻 k のノイズを表す．このノイズ ξ_k が $\mathrm{d}W$ の離散化版である．各時刻のノイズ ξ_k $(k = 1, \ldots, N)$ を疑似乱数を使って事前に与えれば，上の漸化式を用いて，x_0 から $x_1, x_2, \ldots,$ と後の時刻の x_k が求められる．

では，ξ としてどのようなノイズが妥当だろうか？このようなノイズは「我々の観測できない細かい運動 (微小擾乱) の効果が合算され，目に見える変化量として現れている」と考えられる．すると，微小擾乱の「和」を考えることで中心極限定理が働き，ノイズの確率分布は正規分布に近づくと期待される．ここでは，ノイズ

にバイアスが無いとして，正規分布の平均値 $\mathbb{E}[\xi_k]$ は 0 と仮定する．ここで問題になるのは分散 $\mathbb{E}[\xi_k^2]$ の大きさである．

ノイズの分散 $\mathbb{E}[\xi_k^2]$ として，まず思いつくのは $(\Delta t)^2$ である．すると，標準偏差は Δt となり，微分方程式の離散化として良さそうに見える $[\xi_k = O(\Delta t)]$．この示唆は，時間ステップを細かくするにつれ，ノイズの振幅が小さくなることを意味する．この仮定は妥当に見え，「時間ステップが短くなると合算される微小擾乱の数が減り，その効果の和であるノイズが減少する」という直感に合う．

しかし，ノイズの分散 $\nu^2 = \mathbb{E}[\xi_k^2]$ を $(\Delta t)^2$ に設定するのは数学的に正しくない．この事実を確認するために，極限 $\Delta t \to 0$ の時，ある時間区間におけるノイズの分散を考える．今，各時間ステップのノイズは独立かつ同分布に従うと仮定する．この分布として，平均 0 かつ分散 ν^2 の正規分布 $\mathcal{N}(0, \nu^2)$ を選ぶ．各正規分布が独立の時，その分散は単純な足し算となる．今，$0 \le t \le 1$ の範囲を N 分割すれば $\Delta t = 1/N$ である．この時，ノイズ ξ_k $(k = 1, \ldots, N)$ は互いに独立だから，分散の和は以下となる．

$$\mathbb{E}\left[\sum_k \xi_k^2\right] = \sum_k \mathbb{E}[\xi_k^2] = \sum_k \nu^2 = N\nu^2 \tag{3}$$

この分散の合計が $N \to \infty$ (つまり $\Delta t \to 0$) で有限であるためには，$\nu^2 = 1/N$ とすれば良い．つまり，$\nu^2 = \Delta t$ となる．これ以外のスケール，例えば $\nu^2 = (\Delta t)^2$ を選ぶと，分散が発散する．分散が発散する時，数値解も発散するため，$\nu^2 = (\Delta t)^2$ の設定は不適である．

発展的な話題になるが (この段落は読み飛ばしてもよい)，この議論が可能なのはノイズの分散，すなわち二次モーメントに限られる．まず，奇数次のモーメントは正規分布の対称性より 0 になる．さらに，正規分布の場合，任意の偶数次のモーメントは分散で記述される．今，$\nu^2 = \Delta t$ のため，四次以上のモーメントは分散よりも小さなオーダー $o(\Delta t)$ となる．結果として，分散を $\nu^2 = \Delta t$ とすれば，高次のモーメントは 0 に収束する．そのため，正規分布を考える場合，スケールに関する議論対象として分散が妥当と分かる．

以上をまとめると，ノイズ ξ_k は分散 Δt の正規分布 $\mathcal{N}(0, \Delta t)$ に従うと仮定すればよい．この時，ノイズ自体のオーダーは $\xi_k = O(\sqrt{\Delta t})$ である．時間ステップを細かくする際，ノイズの振幅も $\sqrt{\Delta t}$ に応じて小さくする．もし $\xi_k = O(\sqrt{\Delta t})$ でないと分散が収束せず，$\Delta t \to 0$ の極限で数値解が発散してしまう．

標準正規分布に従う確率変数 $\epsilon \sim \mathcal{N}(0, 1)$ を用いると，離散化した微分方程式は以下となる．

$$x_{k+1} = x_k + (-a x_k)\Delta t + b\epsilon_k \sqrt{\Delta t} \tag{4}$$

ここで b を式 (1) に合わせて明記した．この漸化式を使えば，ランダムなノイズに駆動される x_t の数値解が得られる．まず，前もって標準正規分布から N 個の独立なサンプル ϵ_k $(k = 1, \ldots, N)$ を得ておく．このノイズと初期値 x_0 があれば，上の漸化式に従い順番に $x_1, x_2, \ldots$ と数値解が求まる．

確率微分方程式 (Stochastic Differential Equation; SDE) は，上の離散式 (4) を $\Delta t \to 0$ とすることで得られる．

一変数の線形確率微分方程式 (線形 SDE)

$$\mathrm{d}x = -ax\,\mathrm{d}t + b\,\mathrm{d}W \tag{1}$$

ここで，$\mathrm{d}x$ は $x_{i+1} - x_i$ に，$\mathrm{d}W$ は $\epsilon_i\sqrt{\Delta t}$ に対応する．パラメータ a と b は正である．

この $\mathrm{d}W$ 自体を積分して得られる時系列を W_t と表す．別の言い方をすると，$a = 0$ かつ $b = 1$ の時の x_t が W_t に相当する．この W_t は Wiener 過程と呼ばれる [5].

SDE 全体を $\mathrm{d}t$ で割って表記することもある．

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -ax + b\,\eta \tag{5}$$

ここで，$\eta = \mathrm{d}W/\mathrm{d}t$ とした．この η は白色正規ノイズ (あるいはホワイト Gauss ノイズ) と呼ばれる．

白色正規ノイズ $\mathrm{d}W/\mathrm{d}t$ は，あくまで形式的なものである．というのも，W_t は 100% の確率で滑らかでないためである [5]．つまり，あらゆる時刻で無限に小さくギザギザしており，滑らかになる (微分可能になる) ことはない．100% と断っているのは，ランダムなノイズの積分値を考えているが「万に一の確率でも W_t が滑らかにならない」ことを意味する．ちなみに，W_t は 100% の確率で連続である [5]．つまり，不連続なジャンプは存在しない．

2.2 Wiener 過程の微分 $\mathrm{d}W$ の公式

確率過程において最も大事な定理が $\mathrm{d}W$ の積の公式である [5, 6]．実際の応用では，この定理を計算ルールとして利用する．

Wiener 過程の微分に関する公式

$$(\mathrm{d}W_t)^n = \begin{cases} \mathrm{d}t & (n=2) \\ 0 & (n \geq 3) \end{cases} \tag{6}$$

$$\mathrm{d}W_{t_1}\mathrm{d}W_{t_2} = \begin{cases} \mathrm{d}t & (t_1 = t_2) \\ 0 & (\text{otherwise}) \end{cases} \tag{7}$$

ここで微分 $\mathrm{d}W_t$ がどの時刻のものかを明確にするため，$\mathrm{d}W_t := W_{t+\mathrm{d}t} - W_t$ という記号を用いた．以下では，時刻を明記する必要がない場合，$\mathrm{d}W$ のように t を省略することもある

上の定理で大事な点は，期待値を取っていないことである [5, 6]．期待値を取れば，上の式は容易に理解できる．まず，$\Delta W = \epsilon\sqrt{\Delta t}$ だった (ϵ は標準正規分布 $\mathcal{N}(0,1)$ に従う)．ΔW の分散は Δt に等しく，三次以上のモーメントは Δt よりも小さなオーダーとなる．また，時間ステップが異なればノイズは独立であり，結果として，異時刻間の共分散はゼロになる．この時，$\Delta t \to 0$ とし，$\mathrm{d}t$ の一次まで残せば，上の公式が得られる．ここでの大事なポイントは，これらの式が期待値を取らずとも成立する点にある [5, 6].

ここでは簡単にその事実を確認する (この段落を読み飛ばしても後の議論には影響しない)．期待値を取らなくても成立するということは，$(\Delta W)^2$ のランダムな実現値と定数である期待値のズレが無視できることを意味する．このズレを分散で評価する．$(\Delta W)^2$ の期待値は $\mathbb{E}[(\Delta W)^2] = \Delta t$ であるため，以下の式を評価する．

$$\begin{aligned}
\mathbb{E}\left[\left[(\Delta W)^2 - \Delta t\right]^2\right] &= \mathbb{E}\left[(\Delta W)^4\right] - 2\Delta t\,\mathbb{E}\left[(\Delta W)^2\right] + (\Delta t)^2 && (8)\\
&= \mathbb{E}\left[(\Delta W)^4\right] - (\Delta t)^2 && (9)\\
&= (\Delta t)^2\mathbb{E}\left[\epsilon^4\right] - (\Delta t)^2 && (10)\\
&= 3(\Delta t)^2 - (\Delta t)^2 && (11)\\
&= 2(\Delta t)^2 && (12)\\
&= o(\Delta t) && (13)
\end{aligned}$$

ここで，標準正規分布の四次モーメントの表式 $\mathbb{E}\left[\epsilon^4\right] = 3$ を利用した．$(\Delta W)^2$ の分散は $(\Delta t)^2$ のオーダーであり，$\Delta t \to 0$ の極限で，速やかに 0 に収束する [つまり $o(\Delta t)$]．この結果は，$(\Delta W)^2$ の確率分布が細まり，$(\Delta W)^2$ が平均値 Δt を取る定数のように振舞うことを示唆する．

ここで注意を述べておくと，一次モーメント (つまり $\mathrm{d}W$) は $\Delta t \to 0$ の極限で 0 にならない．そのオーダーは $\sqrt{\mathrm{d}t}$ であり，$\mathrm{d}t$ より大きなオーダーである．ただし，期待値を取れば $\langle dW \rangle = 0$ となる．ここで，統計平均 (つまり $\mathbb{E}[\,]$) を $\langle\,\rangle$ で表した．この結果は，離散式版 $\langle \Delta W \rangle = \langle \epsilon \rangle \sqrt{\Delta t} = 0$ から理解できる．

式 (6) に従うと x_t に依存する関数 $f(x_t)$ の微分が計算できる．これは，伊藤の公式と呼ばれ，確率過程の中で最も有名な式の一つである [5]．$\mathrm{d}x = -ax\,\mathrm{d}t + b\,\mathrm{d}W$ を使うと以下が導かれる．

$$\begin{aligned}
\mathrm{d}f &= \frac{\mathrm{d}f}{\mathrm{d}x}\mathrm{d}x + \frac{1}{2}\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}(\mathrm{d}x)^2 + O((\mathrm{d}x)^3) && (14)\\
&= \frac{\mathrm{d}f}{\mathrm{d}x}\left[-ax\,\mathrm{d}t + b\,\mathrm{d}W\right] + \frac{1}{2}\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}\left[-ax\,\mathrm{d}t + b\,\mathrm{d}W\right]^2 + O((\mathrm{d}x)^3) && (15)\\
&= \frac{\mathrm{d}f}{\mathrm{d}x}\left[-ax\,\mathrm{d}t + b\,\mathrm{d}W\right] + \frac{1}{2}\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}\left[(ax)^2(\mathrm{d}t)^2 - 2abx(\mathrm{d}t\,\mathrm{d}W) + b^2\mathrm{d}t\right] + O((\mathrm{d}x)^3) && (16)\\
&= \left[\frac{\mathrm{d}f}{\mathrm{d}x}(-ax) + \frac{1}{2}\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}b^2\right]\mathrm{d}t + \frac{\mathrm{d}f}{\mathrm{d}x}(b\,\mathrm{d}W) + o(\mathrm{d}t) && (17)
\end{aligned}$$

$o(\mathrm{d}t)$ は $\mathrm{d}t$ より小さなオーダーを表し，$\mathrm{d}t\,\mathrm{d}W$ に比例する項などを含む．この小さなオーダーの項は $\mathrm{d}t \to 0$ で 0 となり無視される．$(\mathrm{d}W)^2 = \mathrm{d}t$ のせいで，$f(x)$ の変分を考える際，その二階微分まで考える必要がある．通常の微分則では，$\mathrm{d}f = f'\mathrm{d}x$ のように一階微分まで考えれば良い．このように $\mathrm{d}W$ の特異性のために通常の微分則は成立しない．この点が確率過程と微分積分学の計算の最も大きな違いである．

2.3 線形 SDE の解析解

この節では $\mathrm{d}W$ の公式 (6) と (7) を用いて，一変数の線形 SDE の解析解を導く [5]．以下に SDE を再掲載する．この解析解およびその統計性は，拡散モデルの実装に直接利用されるため [2, 3]，この節は丁寧に式展開をする．

$$\mathrm{d}x = -ax\,\mathrm{d}t + b\,\mathrm{d}W \tag{1}$$

上の SDE の解は以下のように求められる．順番に式変形を行い，解を得る．

$$\begin{aligned}
\mathrm{d}x + ax\,\mathrm{d}t &= b\,\mathrm{d}W && (18)\\
e^{at}\mathrm{d}x + axe^{at}\mathrm{d}t &= be^{at}\mathrm{d}W && (19)\\
\mathrm{d}\left(xe^{at}\right) &= be^{at}\mathrm{d}W && (20)\\
x_t e^{at} - x_0 &= \int_0^t be^{at'}\mathrm{d}W_{t'} && (21)\\
x_t &= e^{-at}x_0 + \int_0^t be^{-a\left(t-t'\right)}\mathrm{d}W_{t'} && (22)
\end{aligned}$$

ここで下付き添字で時刻を陽に表わした．各項の意味を解釈する．まず第一項 $e^{-at}x_0$ は初期条件 x_0 が摩擦 $-ax$ により t と共に指数的に減衰する効果を表す．これにより，任意のデータ (任意の x_0) を 0 に緩和させられる．第二項は正規ノイズ $\mathrm{d}W$ の積算効果を表す．重み $be^{-a\left(t-t'\right)}$ は，過去ほど小さくなり，現在時刻 t に近づくに連れ，重みは 1 に近づく．正規ノイズ $\mathrm{d}W$ は正規分布 $\mathcal{N}(0, \mathrm{d}t)$ に従うため，正規分布の再生性により，正規ノイズの和も正規分布に従う．つまり，式 (19) は，時間とともに減衰する初期条件と，正規ノイズの和となる．この表式 (22) が疑問 (i) の答えであり，任意のデータを正規ノイズへと収束させられる．

式 (22) の統計性を調べる．まず平均は $\mathbb{E}[\mathrm{d}W] = 0$ より

$$\mathbb{E}\left[x_t\right] = e^{-at}\mathbb{E}\left[x_0\right] \tag{23}$$

となる．平均値も解析解 (22) と同様に，初期条件の効果 $\mathbb{E}\left[x_0\right]$ が t と共に指数的に減衰する．

次に二次モーメントは

$$\mathbb{E}\left[x_t^2\right] = e^{-2at}\mathbb{E}\left[x_0^2\right] + \mathbb{E}\left[\int_0^t b^2 e^{-a(t-t_1)}e^{-a(t-t_2)}dW_1 dW_{t_2}\right] \tag{24}$$

となる．初期条件 x_0 はノイズ $\mathrm{d}W$ と独立であるため，期待値の計算は x_0 と $\mathrm{d}W$ の項を分けて行える．具体的に，計算途中で交差項 $\mathbb{E}[x_0\,\mathrm{d}W]$ が出てくるが，$\mathbb{E}[x_0]\,\mathbb{E}[\mathrm{d}W]$ となり $\mathbb{E}[\mathrm{d}W] = 0$ より交差項は 0 となる．式 (24) の第二項はさらに以下のように変形される．

$$\mathbb{E}\left[\int_0^t b^2 e^{-a(t-t_1)}e^{-a(t-t_2)}\mathrm{d}W_{t_1}\mathrm{d}W_{t_2}\right] = \int_0^t b^2 e^{-a(t-t_1)}e^{-a(t-t_2)}\delta\left(t_1 - t_2\right)\mathrm{d}t_1\mathrm{d}t_2 \tag{25}$$

$$= \int_0^t b^2 e^{-a(2t-2t_1)}\mathrm{d}t_1 \tag{26}$$

$$= \frac{b^2}{2a}\left[e^{-a(2t-2t_1)}\right]_0^t \tag{27}$$

$$= \frac{b^2}{2a}\left[1 - e^{-2at}\right] \tag{28}$$

ここで式 (7) を用いて $\mathbb{E}[\mathrm{d}W_{t_1}\mathrm{d}W_{t_2}] = \delta(t_1 - t_2)\mathrm{d}t_1\mathrm{d}t_2$ とした ($\delta(x)$ は Dirac のデルタ関数)．これにより t_2 に関する積分が実行できる．結果として，二次モーメントは以下となる．

$$\mathbb{E}\left[x_t^2\right] = \frac{b^2}{2a} + e^{-2at}\left(\mathbb{E}\left[x_0^2\right] - \frac{b^2}{2a}\right) \tag{29}$$

二次のモーメントについても，解析解や平均値と同様に，初期条件の効果が指数的に減衰する．しかし，二乗されているため，その減衰スピードは e^{-2at} と早い．また，平均値と異なり，0 ではなく有限の値 $b^2/(2a)$ に収束する．この平衡状態は，SDE ($\mathrm{d}x = -ax\ \mathrm{d}t + b\ \mathrm{d}W$) において，摩擦 $-ax\ \mathrm{d}t$ とノイズ $b\ \mathrm{d}W$ が釣り合った状態に対応し，平均値は 0 かつ分散は $b^2/(2a)$ となる．

以上をまとめると以下となる．

一変数の線形 SDE の解

一変数の線形 SDE ($\mathrm{d}x = -ax\ \mathrm{d}t + b\ \mathrm{d}W$) の解は以下となる

$$x_t = e^{-at}x_0 + \int_0^t be^{-a\left(t-t'\right)}\mathrm{d}W_{t'} \tag{22}$$

$$= e^{-at}x_0 + \sigma_t\epsilon \tag{30}$$

$$\sigma_t^2 := \frac{b^2}{2a}\left[1 - e^{-2at}\right] \tag{31}$$

ただし ϵ は標準正規ノイズであり，正規分布 $\mathcal{N}(0,1)$ に従う．

上で説明したように，正規ノイズの積分 (式 (22) の第二項) は再生性により正規ノイズになるため，式 (22) から (30) への書き換えが可能となる．そして，この積分されたノイズの平均は 0，分散は式 (28) から与えられ，その結果から σ_t を式 (31) と定義した．以上から，線形 SDE を解くことで，任意の初期条件 x_0 を正規ノイズ $\sigma\epsilon$ へと緩和させる解を得た．これが疑問 (i) への答えとなる．

2.4 Fokker-Planck 方程式 (FPE)

確率過程の記述は大きく二つに分かれる [5, 6]．一つが時系列 (パス) を記述する方法で SDE を使う．SDE を解くことで，あるノイズが与えられた時，対応する系の解 (パス) が一本求まる．上の例でいうと，x_t のある数値解をパスと呼ぶ．このパスは，ノイズに依存するためランダムであり，試行毎に異なる．もう一つの見方が，パスの集合 (アンサンブル) を記述する方法で Fokker-Planck 方程式 (FPE) を使う．この FPE は SDE の逆回しする際に必要となるため [8, 12]，この節で導入する．

FPE は確率分布 $p(x,t)$ の時間発展を与える．$p(x,t)$ は，時刻 t における系の状態 x の確率 (正確には確率密度) を表す．SDE と異なり，FPE の予測にはランダム性がない．FPE は SDE で記述されるランダムなパス全ての集合を確率分布で記述する．全ての実現可能性が網羅されているため，その確率分布の時間発展は決定論的である．別の言い方をすれば，FPE は系のランダムさ自体 (あるいはランダム性の法則) を記述し，そのランダムさ自体は決定論的に振る舞う．サイコロの例で言えば，出る目が SDE の解 (パス) に対応し，出る目の確率分布が FPE の解に対応する．次に出る目を 100% の精度で予測することは出来ないが，その目の分布は 1/6 の一様分布であり確定している．確率分布を時間発展するようにした時，その発展が FPE により記述される．

FPE を上の一変数の線形 SDE に対して導出する [6]．これからの式展開のため，上で導出した式を再掲載する．

$$\mathrm{d}x = (-ax)\,\mathrm{d}t + b\,\mathrm{d}W \tag{1}$$

$$\mathrm{d}f = \left[\frac{\mathrm{d}f}{\mathrm{d}x}(-ax) + \frac{1}{2}\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}b^2\right]\mathrm{d}t + \left(\frac{\mathrm{d}f}{\mathrm{d}x}b\right)\mathrm{d}W \tag{17}$$

式 (1) は線形 SDE そのものであり，式 (17) は第 2.2 節で導いた伊藤の公式である．ここで $f(x)$ は任意の関数を表す．

FPE の導出方針は，$f(x)$ のアンサンブル平均の時間変化を考え，その変化を $p(x,t)$ の変化と，パスの変化の両方で記述することにある [6]．まず，アンサンブル平均を定義し，その時間変化を考える．

$$\langle f(x)\rangle = \int f(x)p(x,t)\,\mathrm{d}x \tag{32}$$

$$\langle f(x)\rangle\,|_{t+\Delta t} - \langle f(x)\rangle\,|_t = \int f(x)\frac{\partial p(x,t)}{\partial t}\Delta t\,\mathrm{d}x \tag{33}$$

ポイントは, この変化をパスレベルで見ることにある．ちょうど流体力学でいう Euler 的な見方と Lagrange 的な見方に対応する．上の $\langle f(x)\rangle$ の時間変化の記述は Euler 的なものである．x を場の変数と考えて，その場の変化を記述する $p(x,t)$ を時間微分している．もう一つの Lagrange 的な見方では $\langle f(x)\rangle = \langle f(x_t)\rangle$ と各パス x_t を粒子の軌跡とみなし，その軌跡に沿って $f(x_t)$ の変化を考える．この軌跡は SDE で記述される．

この全微分 $\mathrm{d}f/\mathrm{d}t$ は伊藤の公式 (17) で書ける．

$$\langle f(x)\rangle|_{t+\Delta t} - \langle f(x)\rangle|_t = \Delta t \left\langle \frac{\mathrm{d}}{\mathrm{d}t} f(x(t)) \right\rangle \tag{34}$$

$$= \Delta t \left\langle \frac{\mathrm{d}f}{\mathrm{d}x}\frac{\mathrm{d}x}{\mathrm{d}t} + \frac{1}{2}\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}\frac{(\mathrm{d}x)^2}{\mathrm{d}t} \right\rangle \tag{35}$$

$$= \Delta t \left\langle \left[\frac{\mathrm{d}f}{\mathrm{d}x}(-ax) + \frac{1}{2}\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}b^2\right]\frac{\mathrm{d}t}{\mathrm{d}t} \right\rangle + \Delta t \left\langle \frac{\mathrm{d}f}{\mathrm{d}x} b \frac{\mathrm{d}W}{\mathrm{d}t} \right\rangle \tag{36}$$

$$= \Delta t \left\langle \frac{\mathrm{d}f}{\mathrm{d}x}(-ax) + \frac{1}{2}\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}b^2 \right\rangle \tag{37}$$

$$= \Delta t \int \left[\frac{\mathrm{d}f}{\mathrm{d}x}(-ax) + \frac{1}{2}\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}b^2\right] p(x,t)\,\mathrm{d}x \tag{38}$$

$$= \Delta t \int \left[\frac{\partial}{\partial x}[axp(x,t)] + \frac{b^2}{2}\frac{\partial^2}{\partial x^2}p(x,t)\right] f(x)\,\mathrm{d}x \tag{39}$$

第二行目で見慣れない記号 $(\mathrm{d}x)^2/\mathrm{d}t$ が出てきている．$\mathrm{d}W$ の $O(\sqrt{\mathrm{d}t})$ の特異性により，$\mathrm{d}x$ の二乗のオーダーまで考える必要がある．第三行目で伊藤の公式 (17) を代入し，期待値を取っている．第四行目では，$\mathrm{d}W$ の期待値 $\langle \mathrm{d}W/\mathrm{d}t\rangle$ が 0 になることを利用した．そして，第五行目で $p(x,t)$ を利用した積分で式を書き換え，部分積分により，$\mathrm{d}/\mathrm{d}x$ を p の側に移した (適当な境界条件を仮定)．$f(x)$ は任意関数だから，これを Dirac のデルタ関数に取れば，$f(x)$ の係数が式 (33) と (39) で等しくなり，以下の Fokker-Planck 方程式 (FPE) が得られる．FPE の解釈のために，SDE も再掲載する．

一変数の線形 SDE と対応する FPE

$$\mathrm{d}x = -ax\,\mathrm{d}t + b\,\mathrm{d}W \tag{1}$$

$$\frac{\partial p(x,t)}{\partial t} = \frac{\partial}{\partial x}[axp(x,t)] + \frac{b^2}{2}\frac{\partial^2}{\partial x^2}p(x,t) \tag{40}$$

FPE の解釈を行う．式 (40) の第一項は，決定論的な発展 $\mathrm{d}x = -ax\,\mathrm{d}t$ に伴う $p(x,t)$ の移流と理解できる．第二項はノイズによる効果を表し，$p(x,t)$ が拡散する様子を表す．この拡散により $p(x,t)$ の幅が広がり，不確実性が増大する．この FPE の解は解析的に求められる．初期分布は一般に非正規分布であるが，時間が十分経つと正規分布に緩和し，その平均と分散は式 (23) と (29) を通して与えられる．各パス x_t ごとに見ると正規ノイズへと緩和するが，それを $p(x,t)$ でみると正規分布への緩和になる．この正規分布への緩和は「摩擦 $-ax$ による不確実性の減衰と，ノイズ $b\,\mathrm{d}W$ による不確実性の増大が釣り合うため」と解釈できる．

2.5 多変数への拡張

拡散モデルの理解には必ずしも必要ないが，様々な文献で目にする機会も多いと思うので，多変数 (つまりベクトル $\boldsymbol{x}$) へ SDE と FPE を拡張する．この節の内容は読み飛ばしても，後の理解に影響しないようにノートは構成してある．興味が出たらこの節を読めば良い．なおこの節でのみ，下付き添字でベクトルの成分を表す (この節を除き，下付き添字は時刻を表す)．

まず，SDE は N 元連立方程式となる．

N 変数の線形 SDE

$$\mathrm{d}x_i = -\sum_j A_{ij}x_j\ \mathrm{d}t + \sum_j B_{ij}\ \mathrm{d}W_j, \quad (i = 1, \ldots, N) \tag{41}$$

この時の $\mathrm{d}\boldsymbol{W}$ に関する積の公式は以下となる．各 i の $\mathrm{d}W_i$ に対し，一変数の時と同じルールが成立する [式 (6) と (7)]．異なる i と j に対しては，$\mathrm{d}W_i$ と $\mathrm{d}W_j$ は独立となり，期待値を取らずとも任意の次数の積が 0 になる．

N 変数に対する FPE は，一変数の時と同様に導出できる．詳細な式展開のみを載せる．まず，任意関数 $f(\boldsymbol{x})$ に対するアンサンブル平均と，アンサンブル平均の時間変化を考える．

$$\langle f(\boldsymbol{x})\rangle\ |_t = \int \mathrm{d}\boldsymbol{x}\ p(\boldsymbol{x}, t) f(\boldsymbol{x}) \tag{42}$$

$$\langle f(\boldsymbol{x})\rangle\ |_{t+\Delta t} - \langle f(\boldsymbol{x})\rangle\ |_t = \int \mathrm{d}\boldsymbol{x}\ \Delta t \frac{\partial p(\boldsymbol{x}, t)}{\partial t} f(\boldsymbol{x}) \tag{43}$$

ここで，$x_i\ (i = 1, \ldots, N)$ をまとめて，ベクトル $\boldsymbol{x}$ で表した．次にパスレベルで時間変化を記述する．この時 SDE と伊藤の公式を用いる．

$$\langle f(\boldsymbol{x})\rangle\ |_{t+\Delta t} - \langle f(\boldsymbol{x})\rangle\ |_t = \Delta t \left\langle \frac{\mathrm{d}}{\mathrm{d}t} f(\boldsymbol{x}_t) \right\rangle \tag{44}$$

$$= \Delta t \left\langle \sum_i \frac{\partial f}{\partial x_i}\frac{\mathrm{d}x_i}{\mathrm{d}t} + \frac{1}{2}\sum_{i,\ j} \frac{\partial^2 f}{\partial x_i \partial x_j}\frac{\mathrm{d}x_i \mathrm{d}x_j}{\mathrm{d}t} \right\rangle \tag{45}$$

第二項の変形のために $\mathrm{d}\boldsymbol{W}$ の公式を用いる．

$$\mathrm{d}x_i \mathrm{d}x_j = \sum_{k,l} B_{ik}\mathrm{d}W_k\ B_{jl}\mathrm{d}W_l + o(\mathrm{d}t) \tag{46}$$

$$= \sum_{k,l} B_{ik}B_{jl}\delta_{kl}\ \mathrm{d}t + o(\mathrm{d}t) \tag{47}$$

$$= \sum_k B_{ik}B_{jk}\ \mathrm{d}t + o(\mathrm{d}t) \tag{48}$$

$$= \left(BB^{\mathrm{T}}\right)_{ij}\ \mathrm{d}t + o(\mathrm{d}t) \tag{49}$$

上付き添字 T は行列の転置を表し，δ_{ij} は Kronecker のデルタを表す．得られた式を (45) に代入すれば，以

下のように変形できる．

$$\langle f(x)\rangle\,|_{t+\Delta t} - \langle f(x)\rangle\,|_t = \Delta t \left\langle \sum_i \frac{\partial f}{\partial x_i} \left\{ -\sum_j A_{ij} x_j + \sum_j B_{ij} \frac{\mathrm{d}W_j}{\mathrm{d}t} \right\} \right\rangle \tag{50}$$

$$+ \Delta t \left\langle \sum_{ij} \frac{1}{2} \frac{\partial^2 f}{\partial x_i \partial x_j} \left(BB^{\mathrm{T}}\right)_{ij} \right\rangle \tag{51}$$

$$= \Delta t \left\langle \sum_{ij} \left\{ -\left(A_{ij} x_j\right) \frac{\partial f}{\partial x_i} + \frac{1}{2} \left(BB^{\mathrm{T}}\right)_{ij} \frac{\partial^2}{\partial x_i \partial x_j} f \right\} \right\rangle \tag{52}$$

$$= \Delta t \int \mathrm{d}\boldsymbol{x}\; p\left(\boldsymbol{x}, t\right) \left\{ \sum_{ij} \left(-A_{ij} x_j \frac{\partial f}{\partial x_i} + \frac{1}{2} \left(BB^{\mathrm{T}}\right)_{ij} \frac{\partial^2}{\partial x_i \partial x_j} f \right) \right\} \tag{53}$$

$$= \Delta t \int \mathrm{d}\boldsymbol{x}\; f(\boldsymbol{x}) \left\{ \sum_{ij} \frac{\partial}{\partial x_i} \left(A_{ij} x_j p(\boldsymbol{x}, t)\right) + \sum_{ij} \left(\frac{1}{2} \left(BB^{\mathrm{T}}\right)_{ij} \frac{\partial^2}{\partial x_i \partial x_j} p(\boldsymbol{x}, t) \right) \right\} \tag{54}$$

最後に，$f(\boldsymbol{x})$ の任意性を利用して， $f(\boldsymbol{x})$ を Dirac のデルタ関数に取れば，式 (43) と (54) の $f(\boldsymbol{x})$ の係数が等しくなり，以下が得られる．これが N 変数版の FPE である．

N 変数の線形 SDE に対する FPE

$$\frac{\partial p(\boldsymbol{x}, t)}{\partial t} = \sum_{ij} \frac{\partial}{\partial x_i} \left(A_{ij} x_j p(\boldsymbol{x}, t)\right) + \sum_{ij} \left[\frac{1}{2} \left(BB^{\mathrm{T}}\right)_{ij} \frac{\partial^2}{\partial x_i \partial x_j} p(\boldsymbol{x}, t) \right] \tag{55}$$

3　逆 SDE の導出とスコア関数の導入

この章は疑問 (ii)「SDE を時間について逆回しする方法はなにか？」について答える．時間の向きに関して順方向に積分する場合，順 SDE (Forward SDE) や順過程 (Forward Process) と呼ぶ．そして，逆方向に積分する場合，逆 SDE (Reverse SDE or Backward SDE) や逆過程 (Reverse Process or Backward Process) と呼ぶ．順 SDE から逆 SDE を導出する過程で，スコア関数が導入される [8, 12]．

3.1　順過程から逆過程への変換

まずは SDE と FPE の対応関係を再掲載する．

$$\mathrm{d}x = -ax\;\mathrm{d}t + b\;\mathrm{d}W \tag{1}$$

$$\frac{\partial p(x,t)}{\partial t} = \frac{\partial}{\partial x}[axp(x,t)] + \frac{b^2}{2}\frac{\partial^2}{\partial x^2}p(x,t) \tag{40}$$

FPE は，別名「移流拡散方程式」と呼ばれ，特に右辺第一項がない場合，拡散方程式と呼ばれる．拡散方程式は，熱が冷めていく様子や，物質が拡散していく様子を記述する式で，時間に関して不可逆である．実際，冷めてしまった飲み物が，何もせず放っておいて，逆に温まる様子を目にしたことはないはずである．つまり，FPE (40) をそのまま用いたとしても，決して逆過程を記述することはできない．方程式の形を変更する必要がある．ちなみに可逆な系も存在する．例えば，放物運動を記述する運動方程式は，初期条件を上手く変更すれば，方程式の形を変更することなく逆過程を記述できる．

FPE を変形する際に鍵になるのは，SDE との対応関係である．ノイズ項 $b\,\mathrm{d}W$ と拡散項 $b^2/2\,\partial_x^2 p$ が対応している．この対応関係を損なわないように，時間反転を行う．時間反転の操作は，時間軸の逆転で記述される．

$$t^{\mathrm{R}} := (T - t) \tag{56}$$

ここで T は正定数で，$t = 0$ から $t = T$ へと変化するとした．この時，t^{R} を用いれば，時間反転が記述できる (R は Reverse の意味)．実際，$t^{\mathrm{R}} = 0$ から $t^{\mathrm{R}} = T$ に変化する時，元の t は逆に $t = T$ から $t = 0$ へ減少している．この t^{R} で FPE を書き換える．まずは左辺の時間微分項を変形すると

$$\frac{\partial}{\partial t^{\mathrm{R}}} p = \frac{\partial t}{\partial t^{\mathrm{R}}} \frac{\partial}{\partial t} p = -\frac{\partial}{\partial t} p \tag{57}$$

となる．

拡散モデルの目的は，時間について逆回しした際に，正規分布をデータ分布へ変換することであった．これを各サンプルで見れば，正規ノイズがデータサンプルへと変換される．前者は FPE，後者は SDE の視点である．第 2.4 節で議論したように，順過程を記述する FPE (40) は，データ分布を正規分布へと変換する．つまり，この順過程を解とするように逆過程を構成できれば，望みの変換が得られる．言い換えると，「順過程の逆回し」が厳密解であることを保ったまま FPE を変形し，この解を記述する「逆過程」の方程式を導出する．

式 (57) の p が順 FPE (40) の解であるとすれば，以下のようになる．

$$\begin{aligned}
\frac{\partial}{\partial t^{\mathrm{R}}} p^{\mathrm{R}}(x, t^{\mathrm{R}}) &= -\frac{\partial}{\partial t} p(x, t) && (58)\\
&= -\frac{\partial}{\partial x}[axp(x, t)] - \frac{b^2}{2}\frac{\partial^2}{\partial x^2} p(x, t) && (59)\\
&= -\frac{\partial}{\partial x}[axp^{\mathrm{R}}(x, t^{\mathrm{R}})] - \frac{b^2}{2}\frac{\partial^2}{\partial x^2} p^{\mathrm{R}}(x, t^{\mathrm{R}}) && (60)
\end{aligned}$$

ここで

$$p^{\mathrm{R}}(x, t^{\mathrm{R}}) := p(x, t)|_{t=T-t^{\mathrm{R}}} \tag{61}$$

と定義した．順方向の FPE の解 $p(x, t)$ に $t = T - t^{\mathrm{R}}$ を代入し，得られる式を x と t^{R} の関数とみなしたものが $p^{\mathrm{R}}(x, t^{\mathrm{R}})$ となる．時間軸の取り方の違いのため両関数形は異なるが，その値は同一時刻では同じとなる．つまり，$p^{\mathrm{R}}(x, t^{\mathrm{R}} = s) = p(x, t = T - s)$ が 任意の s で成立する．

このままでは，式 (60) の拡散項がマイナスとなっており，SDE と FPE の対応関係が成り立たない (対応関係とは前ページの式 (1) と (40) のこと)．そこで，FPE と SDE の対応関係が成立するように，0 を足し算して，式を変形する [8]．

$$\begin{aligned}
\frac{\partial}{\partial t^{\mathrm{R}}} p^{\mathrm{R}} &= -\frac{\partial}{\partial x}[axp^{\mathrm{R}}] - \frac{b^2}{2}\frac{\partial^2}{\partial x^2} p^{\mathrm{R}} + \underbrace{\left[-\frac{c^2}{2}\frac{\partial^2}{\partial x^2} p^{\mathrm{R}} + \frac{c^2}{2}\frac{\partial^2}{\partial x^2} p^{\mathrm{R}}\right]}_{=0} && (62)\\
&= \frac{\partial}{\partial x}\left[-axp^{\mathrm{R}} - \left(\frac{b^2}{2} + \frac{c^2}{2}\right)\frac{\partial}{\partial x} p^{\mathrm{R}}\right] + \frac{c^2}{2}\frac{\partial^2}{\partial x^2} p^{\mathrm{R}} && (63)\\
&= \frac{\partial}{\partial x}\left[p^{\mathrm{R}}\left\{-ax - \frac{b^2 + c^2}{2p^{\mathrm{R}}}\frac{\partial}{\partial x} p^{\mathrm{R}}\right\}\right] + \frac{c^2}{2}\frac{\partial^2}{\partial x^2} p^{\mathrm{R}} && (64)\\
&= -\frac{\partial}{\partial x}\left[p^{\mathrm{R}}\left\{ax + \frac{b^2 + c^2}{2}\frac{\partial}{\partial x}\ln p^{\mathrm{R}}\right\}\right] + \frac{c^2}{2}\frac{\partial^2}{\partial x^2} p^{\mathrm{R}} && (65)
\end{aligned}$$

第一行で 0 を足し算することで，擬似的に正の符号の拡散項 $c^2/2\,\partial_x^2 p^{\mathrm{R}}$ が現れるようにした．これにより，FPE と SDE の対応関係が利用できる．式 (1) と (40) を見比べて，同じように対応させれば，逆 FPE と逆 SDE が得られる．

逆 SDE と対応する逆 FPE

$$\mathrm{d}x = \left[ax + \left(\frac{b^2+c^2}{2}\right)\frac{\partial}{\partial x}\ln p^{\mathrm{R}}\right]\,\mathrm{d}t^{\mathrm{R}} + c\,\mathrm{d}W^{\mathrm{R}} \tag{66}$$

$$\frac{\partial p^{\mathrm{R}}}{\partial t^{\mathrm{R}}} = -\frac{\partial}{\partial x}\left[p^{\mathrm{R}}\left\{ax + \left(\frac{b^2+c^2}{2}\right)\frac{\partial}{\partial x}\ln p^{\mathrm{R}}\right\}\right] + \frac{c^2}{2}\frac{\partial^2}{\partial x^2}p^{\mathrm{R}} \tag{67}$$

ここで $t^{\mathrm{R}} = T - t$ および $p^{\mathrm{R}}(x, t^{\mathrm{R}}) = p(x,t)|_{t=T-t^{\mathrm{R}}}$ と定義した (上では p^{R} の引数を省略)．パラメータ a, b は正であり，c は 0 以上の実定数である．他の文献 [3, 12] では逆 SDE (66) を t を用いて記述することもあるので，そちらのバージョンも載せておく．

$$\mathrm{d}x = \left[-ax - \left(\frac{b^2+c^2}{2}\right)\frac{\partial}{\partial x}\ln p\right]\,\mathrm{d}t + c\,\mathrm{d}W \tag{68}$$

Wiener 過程の微分 $\mathrm{d}W$ と $\mathrm{d}W^{\mathrm{R}}$ は同じ統計性を持ち，微小時間区間 $|\Delta t|$ に対して，平均 0 かつ分散 $|\Delta t|$ の正規分布に従う．

二つ注意を述べる．第一に，逆 SDE が式 (66) と (68) の形式で表現されている点であるが，これらは等価である．式 (66) は $t^{\mathrm{R}} = 0$ から $t^{\mathrm{R}} = T$ まで積分するので，離散形式で書けば常に $\Delta t^{\mathrm{R}} > 0$ である．一方，式 (68) は，時間を逆回しするため，$t = T$ から $t = 0$ まで積分する．そのため，常に $\Delta t < 0$ である．同様のことは，SDE 中の $\ln p^{\mathrm{R}}$ と $\ln p$ についても言える．例えば，$t^{\mathrm{R}} = 0.2$ の時点では，$\ln p^{\mathrm{R}}$ には $t^{\mathrm{R}} = 0.2$ を代入するが，同じ時点では $t = T - 0.2$ を $\ln p$ に代入する．代入後の値は全く同じであるが $\ln p^{\mathrm{R}}(x, t^{\mathrm{R}} = 0.2) = \ln p(x, t = T - 0.2)$，関数形と時間座標の値が異なるため，明確に区別している．

第二に，c の値の決定方法である．まず導出から明らかなように，この節での式変形は順 FPE の解 $p(x,t)$ を一貫して利用してきた．p^{R} は式 (61) で定義されたように，この解を時間反転させたものに過ぎない．式変形は恒等変形だから，任意の正定数 c に対して，式 (66) と (67) は成立する [8]．つまり，逆過程においてノイズの振幅 c は調整パラメータとなり，それは順過程の振幅 b とは別に取れる [3, 8, 12, 15]．更に興味深いことに，$c = 0$ としても上の式は成立する．つまり，確率的ノイズをなくし，逆過程を決定論的とすることも可能である [3, 8, 12, 15]．いずれの場合でも，p^{R} (と p) は順 FPE の解であり，対応する逆過程は正規分布をデータ分布へと変換する．SDE (66) を用いれば，ある正規ノイズをデータに変換でき，これによってデータ生成が実現される．

3.2 スコア関数を通したデータ生成の仕組み

導出した SDE と FPE を調べることで，なぜデータ生成が可能かを考察する．ここでは $a = b = c = 1$ とし，表式を簡単にする．SDE と FPE を再掲載すると

$$\mathrm{d}x = -x\,\mathrm{d}t + \mathrm{d}W \tag{69}$$

$$\frac{\partial p(x,t)}{\partial t} = \frac{\partial}{\partial x}[xp(x,t)] + \frac{1}{2}\frac{\partial^2 p(x,t)}{\partial x^2} \tag{70}$$

$$\mathrm{d}x = \left[x + \frac{\partial}{\partial x}\ln p^{\mathrm{R}}(x,t^{\mathrm{R}})\right]\,\mathrm{d}t^{\mathrm{R}} + \mathrm{d}W^{\mathrm{R}} \tag{71}$$

$$\frac{\partial p^{\mathrm{R}}(x,t^{\mathrm{R}})}{\partial t^{\mathrm{R}}} = -\frac{\partial}{\partial x}\left[p^{\mathrm{R}}(x,t^{\mathrm{R}})\left\{x + \frac{\partial}{\partial x}\ln p^{\mathrm{R}}(x,t^{\mathrm{R}})\right\}\right] + \frac{1}{2}\frac{\partial^2 p^{\mathrm{R}}(x,t^{\mathrm{R}})}{\partial x^2} \tag{72}$$

となる．前の二式が順過程を表わし，SDE (69) はデータを正規ノイズへと変換し，対応する FPE (70) は (非正規の) データ分布を正規分布へと変換する．後の二式が逆過程を表わし，SDE (71) は正規ノイズをデータへと変換し，対応する FPE (72) は正規分布をデータ分布へと変換する．逆過程は順過程を時間反転させただけであり，各時点の確率分布は順過程と逆過程で等しい $[p(x,t) = p^{\mathrm{R}}(x, T - t^{\mathrm{R}})]$．拡散モデルは，順 SDE (69) を通して，$\partial_x \ln p$ を学習し，逆 SDE (71) を積分することでノイズからデータを生成する．この $\partial_x \ln p$ はスコア関数と呼ばれる．

スコア関数を通してデータ生成が実現される仕組みを考察する．そのために，以下のポテンシャル $U(x,t)$ を導入する．

$$p(x,t) = \frac{1}{Z(t)}\exp\left[-U(x,t)\right] \tag{73}$$

$$Z(t) = \int \mathrm{d}x\ \exp\left[-U(x,t)\right] \tag{74}$$

このような変換は，実用上の多くの確率分布で可能である．実際 $p(x,t) = 0$ の領域を除けば，確率分布は常に正であるため，ポテンシャル $U(x,t)$ を定義できる．規格化のための $Z(t)$ は x に依存しないため，以下の議論では無視できる．式 (73) を見ると，p が高い地点 x において，U が低いことが分かる．つまり，ポテンシャルの低い地点と，確率の高い地点が対応する．後のために，ポテンシャル $U(x,t)$ を直接定義する形を載せる．また，p^{R} と同様に U^{R} を定義する．

$$U(x,t) = -\ln p(x,t) - \ln Z(t) \tag{75}$$

$$U^{\mathrm{R}}(x,t^{\mathrm{R}}) = U(x,t)|_{t=T-t^{\mathrm{R}}} \tag{76}$$

時間反転をさせた U^{R} に対しても同じ関係が成立し，p^{R} が高い地点 x において，U^{R} が低い．

この U^{R} を用いて，SDE (71) を書けば，

$$\mathrm{d}x = \left[x - \frac{\partial}{\partial x}U^{\mathrm{R}}(x,t^{\mathrm{R}})\right]\,\mathrm{d}t^{\mathrm{R}} + \mathrm{d}W^{\mathrm{R}} \tag{77}$$

となる．右辺の $x\,\mathrm{d}t^{\mathrm{R}}$ は順 SDE (69) の摩擦 $-x\,\mathrm{d}t$ を時間反転させただけであるから，以下の議論の本質ではない．本質的なのは，右辺第二項の $-\partial_x U^{\mathrm{R}}$ である．この系は典型的な勾配系 (Gradient System) であり [16]，ポテンシャル U^{R} の低い地点へと x を緩和させる効果を持つ．言い換えると，拡散モデルによる生成とは，各時刻において確率の高い方向への緩和を施し，正規ノイズからのデータ変換を実現している．ポテン

シャルの低い方向は，確率の高い方向に対応し，この方向に引っ張ることで実際に有り得そうなデータを生成する．

ここで二つの注意を述べる．まず，ポテンシャル $U^{\mathrm{R}}(x, t^{\mathrm{R}})$ が時間変化する点が重要である．初期時刻 $t^{\mathrm{R}} = 0$ では，確率分布は正規分布であるため，ポテンシャルは $U^{\mathrm{R}} = 0.5x^2$ と二次関数になる (つまり原点 $x = 0$ が最小値)．そして，時刻が進むにつれポテンシャルは変化し，最終時刻 $t^{\mathrm{R}} = T$ では，非正規のデータ分布に対応するポテンシャルとなる．このように，いきなり複雑なポテンシャルを用いるのではなく，二次関数から初めて段々と複雑なポテンシャルに変化させ，ポテンシャルの極小地点への緩和を逐次施すことで，最終的に複雑な構造を持つデータを生成する．

この「複雑な構造を持つ」データという語も注意が必要である．ここまでは，一変数系を考えてきたので，複雑なデータとは非正規分布を持つデータを指す．実際に逆 SDE はこの非正規分布を再現する．実応用では，この複雑性は画像などの空間構造を持ったデータを指すこともある．この複雑性は，データの集まりである確率分布を必ずしも指していない．つまり，ベクトル $\boldsymbol{x}$ に対して多次元データ生成がどのような機構で行われるのか，つまり成分間に相関のある $\boldsymbol{x}$ が (成分間が無相関な) 多次元正規ノイズから生成される過程を論じる必要がある．実は，これまでの議論を多次元にするのは容易であり，摩擦 $-ax$ やノイズ $b\,\mathrm{d}W$ はベクトルの各成分に独立に働くとすればよい．また a や b の値も全ての成分に対して同一値とすればよい．結果として，データの空間構造はスコア関数を通してのみ考慮される．例えば，順 SDE と逆 SDE の多次元版は

$$\mathrm{d}\boldsymbol{x} = -\boldsymbol{x}\,\mathrm{d}t + \mathrm{d}\boldsymbol{W} \tag{78}$$

$$\mathrm{d}\boldsymbol{x} = \left[\boldsymbol{x} - \boldsymbol{\nabla}_{\boldsymbol{x}} U^{\mathrm{R}}(\boldsymbol{x}, t^{\mathrm{R}})\right]\,\mathrm{d}t^{\mathrm{R}} + \mathrm{d}\boldsymbol{W}^{\mathrm{R}} \tag{79}$$

となる．順過程では，各成分に対し摩擦とノイズを独立に加えて，空間構造を破壊し，(各成分で独立な) 多次元正規ノイズへと変換する．逆過程は，この多次元正規ノイズからデータへの変換を行うが，その際，スコア関数を通して空間構造が復元される．

まとめると，スコア関数 $\partial_x \ln p$ を導入することで，FPE と SDE の形を変え，時間に関する逆回しを実現した．これは，拡散過程が非可逆で，同じ式の形では逆過程を記述できないことに起因する．そして，この逆 SDE によるデータ生成が，勾配系の時間積分に対応し，ポテンシャルの低い地点 (または確率の高い地点) への逐次緩和を表現する．次の第 4 章は，スコア関数を学習する方法を議論する．第 4 章で初めて，「データからの学習」という機械学習らしい話題へ突入する．

3.3 定常分布への緩和

ここでは，勾配系 (77) をさらに調べる．この節の内容は，後の議論に用いないため飛ばしてもよいが，ここで説明する技術は予測子–修正子 (Predictor–Corrector) などで用いられるため [3, 17]，実用上の基礎となる．

まず，逆 SDE と逆 FPE を再掲載する．議論を簡単にするために，$a = 0$ かつ $b = c = 1$ と設定した．また，この節では時間の方向を変えないため，添字 R を省いた．

$$\mathrm{d}x = \left[\frac{\partial}{\partial x} \ln p\right]\,\mathrm{d}t + \mathrm{d}W \tag{80}$$

$$\frac{\partial p}{\partial t} = -\frac{\partial}{\partial x}\left[p \frac{\partial}{\partial x} \ln p\right] + \frac{1}{2}\frac{\partial^2}{\partial x^2} p \tag{81}$$

ここでポテンシャル $-0.5\,V(x)$ で $\ln p$ を置き換えると

$$\mathrm{d}x = -\frac{1}{2}\frac{\mathrm{d}V(x)}{\mathrm{d}x}\,\mathrm{d}t + \mathrm{d}W \tag{82}$$

$$\frac{\partial p}{\partial t} = \frac{\partial}{\partial x}\left[\frac{p}{2}\frac{\mathrm{d}V(x)}{\mathrm{d}x}\right] + \frac{1}{2}\frac{\partial^2}{\partial x^2}p \tag{83}$$

となる．確率分布 $p(x,t)$ は時刻 t に依存しているが，置き換えたポテンシャル $V(x)$ は t に依存しない．この系を調べることで，ポテンシャル V が時間変化しないときの系の振る舞いを議論できる．

まず SDE (82) だが，これはノイズ $\mathrm{d}W$ がない場合を考えると容易に理解できる．第一項の $-0.5\,\mathrm{d}V/\mathrm{d}x$ によってポテンシャルの低い地点へと緩和され，極小地点に淘汰すると $\mathrm{d}V/\mathrm{d}x = 0$ となり，時間変化が止まる．ノイズがある場合も同様に変化し，最終的に極小地点周りでノイズにより揺らぐ状態となる．

次に FPE (83) であるが，これは以下のように書き換えられる．

$$\frac{\partial p}{\partial t} = \frac{1}{2}\frac{\partial}{\partial x}\left\{p\frac{\mathrm{d}V(x)}{\mathrm{d}x} + \frac{\partial}{\partial x}p\right\} \tag{84}$$

この系は，定常分布 $p^{\mathrm{steady}}(x)$ への緩和を記述する．

$$p^{\mathrm{steady}}(x) = \frac{\exp\left[-V(x)\right]}{Z} \tag{85}$$

実際，この定常分布を式 (84) の右辺に代入すれば 0 となり，時間変化しないことが確認できる．この定常分布は，ポテンシャルの低い点で確率の高い状態を表わし，ポテンシャルの極小点周りに x が集まり揺らいでいる状態に対応する．

拡散モデルは，ポテンシャル $U(x,t)$ が時間変化するため，上記のような定常分布へは到達しない．しかし，系のダイナミクスは同様であり，ポテンシャルの低い地点への逐次緩和を記述している．実装上は，離散化に伴う誤差などにより望みのデータ分布への緩和が精度よく実現できないかもしれない．そこで，時刻 t を一時停止させ，停止させた状態で擬似的に SDE (80) を解くことで，固定した時刻のポテンシャル $U(x,t)$ を用いて，固定した時刻の確率分布 $p(x,t)$ へ緩和させることを考える [3, 17]．これにより，$p(x,t)$ の時間変化を精度よく記述できるようになる．この技術は，予測子–修正子 (Predictor–Corrector) と呼ばれ [3, 17]，予測子が逆 SDE を解くことに相当し，そして修正子が時間を一時停止してポテンシャルへの緩和を記述することに相当する．

4　スコア関数の推定：デノイジングスコアマッチング (DSM)

この章から深層学習に入り，拡散モデルの学習方法を説明する．目標はスコア関数 $\partial_x \ln p(x,t)$ を推定可能な損失関数を導くことである [2, 13]．この損失関数を用いて，ニューラルネットワークを最適化すれば，スコア関数が推定できる．推定したスコア関数を代入することで逆 SDE が定義され，これを積分することでノイズからデータを生成できる [3]．この一連の手続きが，疑問 (iii)「そもそも SDE をどうやってデータから決定 (または学習) するか？」の答えとなる．

4.1 議論の準備

その準備のため，順 SDE の解と対応する確率分布を導入する．まず第 2.3 節の結果を再掲載する．第一式が順 SDE であり，第二式以降がその解を表す．

$$\mathrm{d}x_t = -ax_t\,\mathrm{d}t + b\,\mathrm{d}W_t \tag{1}$$

$$x_t = e^{-at}x_0 + \sigma_t\epsilon \tag{30}$$

$$\sigma_t^2 := \frac{b^2}{2a}\left[1 - e^{-2at}\right] \tag{31}$$

$$\epsilon \sim \mathcal{N}(0,1) \tag{86}$$

時刻 t に対する依存性を下付き添字で陽に書いた．x_0 が $t=0$ の初期条件を表わし，実装上は訓練データとなる．ノイズ ϵ は平均 0，分散 1 の正規分布に従う変数である (式 (86) がこの仮定を表す)．拡散モデルの学習はこのノイズ ϵ の推定問題となる [2, 3, 13]．この点をこの章では明らかにしていく．

議論の準備のために，条件付き確率 $p(x,t|x_0)$ を導入する．文献によっては $p_t(x_t|x_0)$ とも表現される．これは初期分布を $\delta(x-x_0)$ という点分布に制限したときの確率分布であり，順 FPE の解となる．この解の確率分布は容易に求められる．SDE を調べて，対応する確率分布を明らかにすればよい．初期条件を x_0 に固定した際，SDE の解は式 (30) で与えられる正規変数であるため，その確率分布は以下の正規分布となる．

$$p(x,t|x_0) = \frac{1}{\sqrt{2\pi\sigma_t^2}}\exp\left[-\frac{1}{2}\frac{\left(x-e^{-at}x_0\right)^2}{\sigma_t^2}\right] \tag{87}$$

(無条件の) 確率分布 $p(x,t)$ は，条件付き分布の積分で与えられる．

$$p(x,t) = \int \mathrm{d}x_0\; p(x,t|x_0)p(x_0) \tag{88}$$

ここで初期分布 $p(x,0)$ を $p(x_0)$ と表わした．今まで議論してきたことであるが，$p(x,t)$ は FPE の解である．これを異なる視点から見ることができる．FPE は線形方程式だから重ね合わせの原理が成り立つ．式 (88) は $p(x,t)$ が FPE の解である $p(x,t|x_0)$ の重ね合わせであることを表し，結果として $p(x,t)$ 自体も解となる．

データ分布 $p(x_0)$ が未知であるため，簡単にスコア関数の推定はできない．この点を説明する．まず，スコア関数 $\partial_x \ln p(x,t)$ をパラメータ $\boldsymbol{\theta}$ を持つニューラルネットワークで近似する．

$$\hat{s}_{\boldsymbol{\theta}}(x,t) \approx \frac{\partial}{\partial x}\ln p(x,t) \tag{89}$$

我々の目標は，この $\hat{s}_{\boldsymbol{\theta}}(x,t)$ を精度良く推定することである．そのためには，ニューラルネットワークの学習を行うための損失関数が必要となる．

素朴な考えとして，損失関数を $\hat{s}_{\boldsymbol{\theta}}(x,t)$ とスコア関数の差とすればよさそうに見える．

$$\mathcal{L}_{\boldsymbol{\theta}} = \frac{1}{T}\mathbb{E}\left[\left\|\hat{s}_{\boldsymbol{\theta}}(x,t) - \frac{\partial}{\partial x}\ln p(x,t)\right\|^2\right] \tag{90}$$

$$= \frac{1}{T}\int \mathrm{d}x\mathrm{d}t\; p(x,t)\left\|\hat{s}_{\boldsymbol{\theta}}(x,t) - \frac{\partial}{\partial x}\ln p(x,t)\right\|^2 \tag{91}$$

$$= \frac{1}{T}\int \mathrm{d}x_0\mathrm{d}x\mathrm{d}t\; p(x,t|x_0)p(x_0)\left\|\hat{s}_{\boldsymbol{\theta}}(x,t) - \frac{\partial}{\partial x}\ln p(x,t)\right\|^2 \tag{92}$$

実装上は，期待値の計算をモンテカルロサンプリングに変えれば良い．すなわち，i 番目の訓練データ $x_0^{(i)}$ に対して，時刻 $t^{(i)}$ と正規乱数 $\epsilon^{(i)}$ をサンプリングし (時刻は 0 から T の範囲の一様分布などからサンプリング)，$x_t = e^{-at}x_0 + \sigma_t\epsilon$ [式 (30)] に従い $x^{(i)}$ を生成する．そして，式 (92) の絶対値の中身を評価し，サンプル平均を取れば良い．

$$\mathcal{L}_{\boldsymbol{\theta}} \approx \frac{1}{N}\sum_i \left[\left\| \hat{s}_{\boldsymbol{\theta}}(x^{(i)}, t^{(i)}) - \frac{\partial}{\partial x}\ln p(x^{(i)}, t^{(i)}) \right\|^2\right] \tag{93}$$

ここで N をサンプルサイズとした．この式を評価するためにはスコア関数の具体形が必要になる．そこでスコア関数を式 (87) と (88) に従い求めてみる．

$$\begin{aligned}
\frac{\partial}{\partial x}\ln p\left(x,t\right) &= \frac{\partial_x p\left(x,t\right)}{p\left(x,t\right)} && (94)\\
&= \frac{\int \left[\partial_x p\left(x,t \mid x_0\right)\right] p\left(x_0\right) \mathrm{d}x_0}{p\left(x,t\right)} && (95)\\
&= \int \frac{\partial_x p\left(x,t \mid x_0\right)}{p\left(x,t \mid x_0\right)} \underbrace{\frac{p\left(x,t \mid x_0\right) p\left(x_0\right)}{p\left(x,t\right)}}_{=p(x_0|x_t)} \mathrm{d}x_0 && (96)\\
&= \int \left[\partial_{x_t}\ln p\left(x_t \mid x_0\right)\right] p\left(x_0|x_t\right) \mathrm{d}x_0 && (97)\\
&= \int \mathrm{d}x_0\, p\left(x_0|x_t\right) \left[\partial_{x_t}\ln p\left(x_t|x_0\right)\right] && (98)\\
&= \int \mathrm{d}x_0\, p\left(x_0|x_t\right) \left[-\frac{x_t - e^{-at}x_0}{\sigma_t^2}\right] && (99)
\end{aligned}$$

関数 $p(x,t|x_0)$ は，時刻 t の x に対する確率分布であるため，必要に応じて $p(x_t|x_0)$ と書いている．同様に式 (88) の $p(x,t)$ も $p(x_t)$ と書いている．第三行では Bayes の公式を利用し

$$p\left(x_0|x_t\right) := \frac{p\left(x_t \mid x_0\right) p\left(x_0\right)}{p(x_t)} = \frac{p\left(x,t \mid x_0\right) p\left(x_0\right)}{p\left(x,t\right)} \tag{100}$$

とした．最終行では，式 (87) を利用して，x_t に関する微分を実行している．結果として，条件付き期待値を利用すれば，スコア関数は以下となる．

$$\frac{\partial}{\partial x}\ln p\left(x,t\right) = -\frac{x_t - e^{-at}\mathbb{E}[x_0|x_t]}{\sigma_t^2} \tag{101}$$

時刻 t の状態 x_t は，データ x_0 をノイズにより乱すことで簡単に計算できる．しかし，ここで必要な量は，この x_t を固定したときの x_0 の平均値である．ノイズにより乱れた画像に対しクリーンな画像が複数ありうるように，x_t を固定したときには複数の x_0 があり得る．そして，候補となる x_0 を大量の訓練データから抽出し，平均を取ることで $\mathbb{E}[x_0|x_t]$ を求めるのは困難である．モンテカルロサンプリングの時のように，固定するのは x_t ではなく，x_0 であって欲しい．これを実現し，モンテカルロ計算により評価可能な形に持っていくのが，デノイジングスコアマッチング (DSM) である [2, 13].

4.2 損失関数の変形

この節および次の節にかけて DSM で用いる損失関数を導出する [2, 13]．その手がかりを得るために，式 (92) の損失関数 $\mathcal{L}_{\boldsymbol{\theta}}$ を展開してみる．

$$\mathcal{L}_{\boldsymbol{\theta}} = \int \left[\hat{s}_{\boldsymbol{\theta}}^2 - 2\hat{s}_{\boldsymbol{\theta}} \frac{\partial}{\partial x} \ln p(x,t) + \left(\frac{\partial}{\partial x} \ln p(x,t) \right)^2 \right] p(x,t) dx \tag{102}$$

式を簡単にするため，ニューラルネットワークによる近似スコア $\hat{s}_{\boldsymbol{\theta}}$ の引数は無視した．また，一時的に t の積分を省略している．第一項は，式 (93) の議論のように，モンテカルロサンプリングで評価できる．第二項は，上で見たように $\mathbb{E}[x_0|x_t]$ の計算が現れ，容易に評価できない．第三項は，ニューラルネットワークのパラメータ θ を含まないため，拡散モデルの学習という観点において無視できる．そこで難点となっている第二項に着目し，式変形を進める．

$$\begin{aligned}
\int \mathrm{d}x\, p(x,t) \left[\hat{s}_{\boldsymbol{\theta}} \frac{\partial}{\partial x} \ln p(x,t) \right] &= \int \mathrm{d}x\, p(x,t) \left[\hat{s}_{\boldsymbol{\theta}} \frac{1}{p(x,t)} \frac{\partial}{\partial x} p(x,t) \right] && (103)\\
&= \int \mathrm{d}x \left[\hat{s}_{\boldsymbol{\theta}} \frac{\partial}{\partial x} \int \mathrm{d}x_0\, p(x,t|x_0) p(x_0) \right] && (104)\\
&= \int \mathrm{d}x \mathrm{d}x_0\, p(x_0) \left[\hat{s}_{\boldsymbol{\theta}} \frac{\partial}{\partial x} p(x,t|x_0) \right] && (105)\\
&= \int \mathrm{d}x \mathrm{d}x_0\, p(x_0) \left[\hat{s}_{\boldsymbol{\theta}} \frac{p(x,t|x_0)}{p(x,t|x_0)} \frac{\partial}{\partial x} p(x,t|x_0) \right] && (106)\\
&= \int \mathrm{d}x \mathrm{d}x_0\, p(x,t|x_0) p(x_0) \left[\hat{s}_{\boldsymbol{\theta}} \frac{1}{p(x,t|x_0)} \frac{\partial}{\partial x} p(x,t|x_0) \right] && (107)\\
&= \int \mathrm{d}x \mathrm{d}x_0\, p(x,t|x_0) p(x_0) \left[\hat{s}_{\boldsymbol{\theta}} \frac{\partial}{\partial x} \ln p(x,t|x_0) \right] && (108)
\end{aligned}$$

この変形では，式 (87) と (88) を利用した．この変形により，困難であった $\partial_x \ln p(x,t)$ を評価する必要がなくなった．評価する必要があるのは，条件付き分布に対する表式 $\partial_x \ln p(x,t|x_0)$ であり，この項は式 (87) より陽に計算できる．

まとめると，以下の損失関数を用いれば良い．

$$\begin{aligned}
\mathcal{L}_{\boldsymbol{\theta}}^{\text{conditonal}} &= \frac{1}{T} \int \left[\hat{s}_{\boldsymbol{\theta}}^2 - 2\hat{s}_{\boldsymbol{\theta}} \frac{\partial}{\partial x} \ln p(x,t|x_0) + \left(\frac{\partial}{\partial x} \ln p(x,t|x_0) \right)^2 \right] p(x,t|x_0) p(x_0)\, \mathrm{d}x \mathrm{d}x_0 \mathrm{d}t && (109)\\
&= \frac{1}{T} \int \left\| \hat{s}_{\boldsymbol{\theta}} - \frac{\partial}{\partial x} \ln p(x,t|x_0) \right\|^2 p(x,t|x_0) p(x_0)\, \mathrm{d}x \mathrm{d}x_0 \mathrm{d}t && (110)
\end{aligned}$$

元の損失関数 $\mathcal{L}$ [式 (92) または (102)] と比較すると，ln の中身が条件付き分布 $p(x,t|x_0)$ に変わっており，容易に評価できるようになっている．両損失関数 $\mathcal{L}$ と $\mathcal{L}_{\boldsymbol{\theta}}^{\text{conditonal}}$ は，スコア関数 [$\partial_x \ln p(x,t)$ か $\partial_x \ln p(x,t|x_0)$] の二乗量が異なるため，等しくはない．しかし，パラメータ $\boldsymbol{\theta}$ に依存する項は全て等しくなり，結果として $\boldsymbol{\nabla}_{\boldsymbol{\theta}} \mathcal{L}$ と $\boldsymbol{\nabla}_{\boldsymbol{\theta}} \mathcal{L}_{\boldsymbol{\theta}}^{\text{conditonal}}$ は等しくなる．拡散モデルの学習に必要なのは，パラメータ $\boldsymbol{\theta}$ に関する微分量であるから，$\mathcal{L}_{\boldsymbol{\theta}}^{\text{conditonal}}$ を用いた最適化は，元の $\mathcal{L}$ による最適化と等価になり，結果として $\mathcal{L}_{\boldsymbol{\theta}}^{\text{conditonal}}$ を用いた学習を通し，スコア関数 $\partial_x \ln p(x,t)$ がニューラルネットワーク $\hat{s}_{\boldsymbol{\theta}}(x,t)$ により近似される [2, 3, 13].

この導出の肝となったのは，損失関数 $\mathcal{L}$ を二乗量で構成したことにある．これにより，交差項が 1 つ現れ，この交差項を式 (103) から (108) のように変形した．これにより，元の損失関数と等価な $\mathcal{L}_{\boldsymbol{\theta}}^{\text{conditonal}}$ が導出できた．損失関数 $\mathcal{L}$ を四乗量などで構成すれば，変形困難な交差項が現れ，同様の導出は難しくなってしまう．

4.3 ノイズの推定問題

損失関数 $\mathcal{L}_{\boldsymbol{\theta}}^{\text{conditonal}}$ は，条件付き確率の表式 (87) を用いることで，更に簡単になる．具体的に

$$\mathcal{L}_{\boldsymbol{\theta}}^{\text{conditonal}} = \frac{1}{T}\int \left\| \hat{s}_{\boldsymbol{\theta}}(x,t) + \frac{x - e^{-at}x_0}{\sigma_t^2} \right\|^2 p(x,t|x_0)p(x_0)\,\mathrm{d}x\mathrm{d}x_0\mathrm{d}t \tag{111}$$

となる．

実装を見据えて，この式をモンテカルロサンプリングにより表現する．ここで重要なポイントは，SDE の解から

$$x - e^{-at}x_0 = \sigma_t \epsilon \tag{112}$$

が成立する点にある．これにより，ノイズ ϵ が損失関数の中に陽に入ってくる．以上をまとめると，以下の損失関数が得られる [2, 13].

デノイジングスコアマッチング (DSM) の損失関数

$$\mathcal{L}_{\boldsymbol{\theta}}^{\text{DSM}} = \frac{1}{N}\sum_i \left\| \hat{s}_{\boldsymbol{\theta}}(x^{(i)}, t^{(i)}) + \frac{\epsilon^{(i)}}{\sigma_{t^{(i)}}} \right\|^2 \tag{113}$$

ここで，i 番目の訓練データ $x_0^{(i)}$ に対して，時刻 $t^{(i)}$ と標準正規乱数 $\epsilon^{(i)}$ を発生させ (時刻は 0 から T の範囲の一様分布などからサンプリング)，以下の式 (30) に従い $x^{(i)}$ を生成する．

$$x^{(i)} = e^{-at^{(i)}}x_0^{(i)} + \sigma_t^{(i)}\epsilon^{(i)} \tag{30}$$

そして，最後に N 個のサンプルに対して平均を取る．

このフレームワークは，デノイジングスコアマッチング (Denoising Score Matching; DSM) と呼ばれ [2, 3, 13]，ノイズ ϵ の推定問題を通して，スコア関数 $\hat{s}_{\boldsymbol{\theta}}$ の学習を行う．文献によっては，ノイズ ϵ の推定問題である点を強調した表式が使われている．そこでその表式も導く．まず，ニューラルネットワークによるノイズの推定値を $\hat{\epsilon}(x,t)$ とすれば

$$\hat{s}_{\boldsymbol{\theta}}(x,t) = -\frac{\hat{\epsilon}(x,t)}{\sigma_t} \tag{114}$$

となり，損失関数は

$$\mathcal{L}_{\boldsymbol{\theta}}^{\text{DSM}} = \frac{1}{N}\sum_i \frac{\left\|\hat{\epsilon}(x^{(i)}, t^{(i)}) - \epsilon^{(i)}\right\|^2}{\sigma_{t^{(i)}}^2} \tag{115}$$

となる．実用上は，分母の $\sigma_{t^{(i)}}^2$ を省いたり (または別の値 (重みと呼ばれる) に置き換えたり)，L2 損失ではなく L1 損失を用いて $\hat{\epsilon}(x,t)$ の学習を行うこともある．

なぜノイズの推定でスコア関数が得られるのかを考察する．そもそもスコア関数は厳密に式 (101) で与えられる．条件付き分布が正規であるため，指数の肩が微分され，得られた一次関数の期待値が取られている．

これは，もとをたどれば $p(x,t)$ が $p(x,t|x_0)$ の重み付き平均であることに起因する [式 (88)]．式 (101) の条件付き期待値 $E[x_0|x_t]$ は，スコア関数を評価している地点 x_t を固定した時，あり得る x_0 を抽出し，そこから計算した x_0 の平均値に等しい．あり得る x_0 が唯一の場合，分子が $x_t - e^{-at}x_0$ となり，これはノイズ $\sigma_t\epsilon$ に等しい．つまり，式 (101) の時点で，差分 $x_t - e^{-at}x_0$ がスコア関数を構成する様子が見えている．微分量とは一般に摂動に対する変化量であるが，条件付き分布が正規であるため，対数を取り微分した量が摂動そのものになっている．この摂動の条件付き平均がスコア関数であるため，その摂動 (つまりノイズ) を推定することが学習となる．

4.4 逆 SDE によるデータの生成

推定した $\hat{\epsilon}(x,t)$ (または $\hat{s}_{\boldsymbol{\theta}}$) によりデータ生成を行う際は，逆 SDE を数値積分すれば良い．式 (114) および $\hat{s}_{\boldsymbol{\theta}}$ がスコア関数 $\partial_x \ln p(x,t)$ の近似であることを思い出すと，逆 SDE (66) は以下のように近似できる．

$$\mathrm{d}x = \left[ax - \left(\frac{b^2+c^2}{2}\right)\frac{\hat{\epsilon}(x,t)}{\sigma_t}\right]\mathrm{d}t^{\mathrm{R}} + c\,\mathrm{d}W^{\mathrm{R}} \tag{116}$$

係数を無視すれば，各時刻で $-\hat{\epsilon}(x,t)\,\mathrm{d}t$ が作用しており，ノイズ $\hat{\epsilon}$ を差し引いている．この点を指して，拡散モデルによるデータ生成を「ノイズ除去」と表現することがあるが，これは必ずしも正確ではない．なぜなら，SDE (116) は新たなノイズ $c\,\mathrm{d}W^{\mathrm{R}}$ を加える操作を含むためである．数学的に，スコア関数を用いた逆 SDE の導出がまず存在し，そのスコア関数を DSM で推定することを通して，拡散モデルによるデータ生成は実現される．本質的なのは，スコア関数の存在であり，ノイズ除去ではない．

4.5 実装上の注意

このノートの主目的ではないが，拡散モデルの原理から理解できる実装上の注意に言及する．まず，SDE の解は $x = e^{-at}x_0 + \sigma_t\epsilon$ のように指数的な減衰を含む．そのため，時間刻み幅を十分細かく取る必要がある．等速直線運動のように変化率が変わらないならば，時間刻み幅はいくらでも大きくできる．しかし，指数減衰する x_t のように速度も急に変わる場合，その速度変化を表せるような刻み幅が要求される．刻み幅が小さくなり，結果として時間ステップ数が増大し，拡散モデルの推論には計算時間がかかる [18].

実装上，パラメータ a, b, c は事前に与える．逆 SDE のノイズ振幅 c は，順過程の振幅 b と同じにしばしば設定される [2]．残りの a と b であるが，例えば，データの分散が t について一定とすれば，どちらかは自動的に決まる [2]．分散の形を再掲載すると

$$\mathbb{E}\left[x_t^2\right] = \frac{b^2}{2a} + e^{-2at}\left(\mathbb{E}\left[x_0^2\right] - \frac{b^2}{2a}\right) \tag{29}$$

となる．ここで平均値 $\mathbb{E}[x_0]$ は 0 とし，二次モーメントを分散とみている．訓練データを標準化すれば $\mathbb{E}\left[x_0^2\right] = 1$ であるから (つまり標準化も大事)，$a = \beta/2$ かつ $b = \sqrt{\beta}$ と設定すればよい ($b^2 = 2a$)．ここで a と b をパラメータ β であらためて表現した．得られる SDE は

$$\mathrm{d}x = -\frac{\beta}{2}x\,\mathrm{d}t + \sqrt{\beta}\,\mathrm{d}W \tag{117}$$

となる．この β により指数的な緩和スピードが定まる．緩和時間スケールは $1/a = 2/\beta$ であるから，$t \in [0,1]$ とすれば， 例えば β は 10 程度に設定すればよい．ただし，この設定は β が定数の場合であり，実装時はこれを t の関数に取る [2]．この関数形はノイズスケジュールと呼ばれ，いくつかの形が提案されている [19].

5　条件付き生成問題

この章は，拡散モデルの応用として，条件付き生成問題を説明する．前章から引き続き，一変数系を考えるが，条件付き生成は空間構造を持つデータに対して適用できる [3, 14]．問題の本質を掴むために，あえて一変数系を取り上げる．なお，以下で説明する理論は条件付き生成問題を拡散モデルで解くための一手法であり，他にも様々な理論や手法が提案されている．

超解像やイメージ修復やデータ同化などの逆問題では，条付き分布を求めることが課題となる [3]．この条件 y は，低解像度画像や観測値やモデルパラメータなど何か固定したい量であり，その量を固定した際の変数 x の不確実性を条件付き確率分布 $p(x|y)$ で評価する．この $p(x|y)$ は事後分布と呼ばれ，Bayes の定理を使えば以下のように分解できる．

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} \tag{118}$$

$$p(y) := \int \mathrm{d}x\, p(y|x)p(x) \tag{119}$$

変数 x を推定する際，分母の $p(y)$ は x に依存しないため，多くの状況で無視できる．例えば，勾配法では確率分布そのものではなく，対数を取った量 $\ln p(x|y)$ を評価する．これは対数を取ることで分母 $p(y)$ を無視できる上，(通常は非常に小さくなる) 確率の計算でアンダーフローを起こさないためである．この時，勾配はスコア関数で書ける．

$$\frac{\partial}{\partial x} \ln p(x|y) = \frac{\partial}{\partial x} \ln p(y|x) + \frac{\partial}{\partial x} \ln p(x) \tag{120}$$

無条件スコア関数 $\partial_x \ln p(x)$ は拡散モデルの学習を通して推定できる．他方の項に出てくる $\ln p(y|x)$ は x が与えられた時の y の条件付き分布を表わし，尤度と呼ばれる．例えば，予測値 x を固定した際の観測値 y の不確実性を表す．尤度 $\ln p(y|x)$ は ($\ln p(x)$ と比べると) 正規分布など簡単な関数で与えられることがある．そのため，$\partial_x \ln p(x)$ を知っていれば，勾配全体 (120) の評価ができると期待される．特に実際の応用では，低解像度画像のサイズや観測点の位置などの y の仕様が変化し，結果として $\ln p(y|x)$ が変わることがある．このような仕様の変化に耐えるフレームワークを式 (120) の分解と拡散モデルにより構築していく [3, 14].

もし，無条件スコア関数 $\partial_x \ln p(x)$ を拡散モデルで学習しておけば，生成時に $\partial_x \ln p(y|x)$ を $\partial_x \ln p(x)$ に足せば，事後分布 $\ln p(x|y)$ に従うようなサンプルを生成できるかもしれない．これは雑な想像であるが，まとめると以下のようになる．学習時は今まで解説してきた通りに拡散モデルを学習させ，無条件スコア関数 $\partial_x \ln p(x)$ を推定する．そして，推論時は，逆 SDE に事後分布 $\ln p(x|y)$ を代入し，式 (120) の分解を使って，以下の SDE を解く ($a = b = c = 1$ とした).

$$\mathrm{d}x = \left[x + \frac{\partial}{\partial x} \left\{ \ln p^{\mathrm{R}}(x,t) + \ln p(y|x) \right\} \right] \mathrm{d}t^{\mathrm{R}} + \mathrm{d}W^{\mathrm{R}} \tag{121}$$

この時，新しく付け加えた項は x を y へ緩和させる効果を持つ．例えば，正規分布を考えれば

$$p(y|x) = \frac{1}{\sqrt{2\pi\lambda^2}} \exp\left[-\frac{1}{2} \frac{(y-x)^2}{\lambda^2} \right] \tag{122}$$

$$\frac{\partial}{\partial x} \ln p(y|x) = -\frac{x-y}{\lambda^2} \tag{123}$$

となり，微分 $\partial_x \ln p(y|x)$ から y への線形緩和項が出てくる．結果として，x は y に近づき，逆 SDE を積分した結果得られるサンプルは事後分布 $p(x|y)$ に従いそうに見える．しかし，これは数学的に正しくない．条件付き生成問題の考え方としては上記の説明でよいのだが，このままだと得られるサンプルは $p(x|y)$ に従わない．以下では，数学的に正しい理論へと修正を行う [3, 14].

5.1 尤度の構成方法

上記の議論が正しくない原因は，生成されたデータ x_0 に対して尤度を構成しているのにも関わらず，それを生成途中の x_t に対して適用している点にある [3]．生成途中では，x_t はノイズにより乱された状態となるが，その乱れ度合いは t により異なる．$t = T$ (つまり逆過程の始点 $t^{\mathrm{R}} = 0$) では完全な正規ノイズであるが，$t = 0$ (等価だが $t^{\mathrm{R}} = T$) ではデータ分布に従うサンプル x_0 ($x_0 \sim p(x_0)$) となる．つまり，ノイズによる乱れ具合に応じて，尤度 $p(y|x)$ を調整する必要がある．

別の論点として，事後分布 $p(x_t|y)$ が FPE に従うかを確認する必要がある．拡散モデルでは，データ分布 $p(x_0)$ を正規分布へと緩和させる順過程を FPE で記述し，それを時間に関して逆回しした解を逆 FPE で記述した．そして，対応する逆 SDE を考えることで，正規ノイズから各サンプルの生成を実現した．同じことがデータ分布 $p(x_0)$ を事後分布 $p(x_0|y)$ に置き換えて成立するかは自明ではない．

以上の点を議論するために，確率過程で慣習的に用いられる記法を利用する．今までは $p(x,t)$ のように時刻 t に関する確率分布を記述していたが，これを $p(x_t)$ と記述する．これにより，どの時刻の x を記述しているのかが明確になる．例えば，条件付き分布 $p(x,t|x_0)$ は以下となる．

$$p(x_t|x_0) = \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left[-\frac{1}{2}\frac{\left(x_t - e^{-at}x_0\right)^2}{\sigma_t^2}\right] \tag{124}$$

この式は式 (87) と等価である．この記法の難点として，$p(x_t|x_0)$ のように書いた時，関数 p 自体も t に陽に依存している点を明記できない．実際，式 (124) を見ると分かるように，e^{-at} や σ_t など，関数 p は t に陽に依存している．ただし記法として便利なため，確率過程の議論ではしばしば用いられる．特に，複数の時刻における結合分布を表現する際に便利となる．例えば，$p(x_{t_2}, x_{t_1}, x_{t_0})$ などである．この場合，関数 p は時刻 t_2, t_1, t_0 に陽に依存しうる．

この記法を用いて理論を構成し直す [3, 14]．まず問題設定を整理する．周辺分布 $p(x_t)$ は条件付き分布 $p(x_t|x_0)$ により与えられ，Bayes の定理を通して $p(x_0|x_t)$ を与える．

$$p(x_t) = \int \mathrm{d}x_0\, p(x_t|x_0)p(x_0) = \int \mathrm{d}x_0\, p(x_t, x_0) \tag{125}$$

$$p(x_t, x_0) = p(x_t|x_0)p(x_0) \tag{126}$$

$$p(x_0|x_t) = \frac{p(x_t, x_0)}{p(x_t)} = \frac{p(x_t|x_0)p(x_0)}{p(x_t)} \tag{127}$$

ここで条件付き分布 (124) は，初期分布をデルタ関数 $\delta(x - x_0)$ に取った際の順 FPE の解である (第 4.1 節)．FPE は線形方程式であるから重ね合わせの原理により $p(x_t)$ も解となる (第 4.1 節)．この確率分布 $p(x_t)$ は前章までは $p(x,t)$ と表されていた．この時，尤度 $p(y|x_0)$ を考える．この尤度は任意の確率分布である (つまり一般に非正規分布)．この時，我々の目的は事後分布 $p(x_0|y)$ の推定または事後分布に従うサンプルの生

成となる．

$$p(x_0|y) = \frac{p(y|x_0)p(x_0)}{p(y)} \tag{128}$$

$$p(y) := \int p(y|x_0)p(x_0)\mathrm{d}x_0 \tag{129}$$

さらに具体的に言えば，この事後分布 $p(x_0|y)$ に従うサンプル x_0 を無条件データ分布 $p(x_0)$ に対する拡散モデルから「追加の訓練を行うことなく」推定したい [3, 14]．

拡散モデルは x_t の発展を考えるため，$p(x_t|y)$ を評価してみる．今まで定義した量で，この任意の t に対する事後分布を表すと

$$p\left(x_t \mid y\right) = \int p\left(x_t \mid x_0\right) p\left(x_0 \mid y\right) \mathrm{d}x_0 \tag{130}$$

$$= \int p\left(x_t \mid x_0\right) \frac{p(y|x_0)p(x_0)}{p(y)} dx_0 \tag{131}$$

$$= \frac{1}{p(y)} \int p\left(y \mid x_0\right) p(x_t|x_0) p\left(x_0\right) \frac{p\left(x_t\right)}{p\left(x_t\right)} \mathrm{d}x_0 \tag{132}$$

$$= \frac{p\left(x_t\right)}{p(y)} \int p\left(y \mid x_0\right) \frac{p\left(x_t \mid x_0\right) p\left(x_0\right)}{p\left(x_t\right)} \mathrm{d}x_0 \tag{133}$$

$$= \frac{p\left(x_t\right)}{p(y)} \int p\left(y \mid x_0\right) p\left(x_0 \mid x_t\right) \mathrm{d}x_0 \tag{134}$$

$$= \frac{1}{p(y)}\, l\left(y \mid x_t\right) p\left(x_t\right) \tag{135}$$

ここで平均尤度 $l\left(y \mid x_t\right)$ を以下のように定義した．

$$l\left(y \mid x_t\right) = \int p\left(y \mid x_0\right) p\left(x_0 \mid x_t\right) \mathrm{d}x_0 \tag{136}$$

この章の導入の議論と異なり，平均尤度 $l\left(y \mid x_t\right)$ を用いて事後分布 $p(x_t|y)$ を評価している点が重要である．この関数 $l(y|x_t)$ は確率分布の定義を満たす．つまり，0 以上の量かつ y に関して積分すると 1 になる．この平均尤度 $l(y|x_t)$ により理論が正しく修正される．

特に $p\left(x_t \mid y\right)$ は FPE の解である．この事実は $p(x_t|x_0)$ が FPE を満たす事実から重ね合わせの原理により導かれる．式 (130) を見ると，$p\left(x_t \mid y\right)$ は $p(x_t|x_0)$ を重み $p(x_0|y)$ で重ね合わせたものである．さらにこの事実から $p(x_t|y)$ は正規分布に漸近することも分かる．十分に時間が経つと $p(x_t|x_0)$ が標準正規分布に漸近し，$p(x_t|x_0)$ から x_0 依存性が消える．結果，式 (130) の積分が実行できて，x_0 依存性が式全体から消えてしまう ($\int \mathrm{d}x_0\, p(x_0|y) = 1$)．以上から，$p\left(x_t \mid y\right)$ は FPE の解であり，その関数形は $l\left(y \mid x_t\right) p\left(x_t\right)$ と無条件分布 $p(x_t)$ との積で与えられる [式 (135)]．

対応する逆 SDE は，$p\left(x_t \mid y\right)$ を代入すると，以下で与えられる．

$$\mathrm{d}x_t = \left[ax_t + \left(\frac{b^2+c^2}{2}\right) \frac{\partial}{\partial x_t} \left\{\ln l(y|x_t) + \ln p^{\mathrm{R}}(x_t)\right\}\right] \mathrm{d}t^{\mathrm{R}} + c\, \mathrm{d}W_t^{\mathrm{R}} \tag{137}$$

無条件スコアを用いた逆 SDE (66) と比較すると，$\partial_{x_t} \ln l(y|x_t)$ が新しく追加されている．無条件スコア関数 $\partial_{x_t} \ln p^{\mathrm{R}}(x_t)$ を拡散モデルの学習により推定すれば，逆 SDE (137) に基づき，任意の $l(y|x_t)$ を用いたデータ生成が実現される．結果として得られるサンプル x_0 は事後分布 $p(x_0|y)$ に従う．

残りの問題は $l\left(y \mid x_t\right)$ の評価となる．この量は式 (136) で定義されるが，$p(x_0|x_t)$ の積分計算を含むため，簡単に得られない．この根本原因は $p(x_0|x_t)$ [式 (127)] の中に未知の量 $p(x_0)$ が含まれるためである．拡散

モデルはスコア関数 $\partial_x \ln p(x)$ を推定するが，確率分布 $p(x)$ の値は未知のままである．何らかの近似を用いて $l\left(y \mid x_t\right)$ を推定する必要がある．

5.2 尤度の近似方法

条件付き生成問題の原論文 [14] に従い，$l\left(y \mid x_t\right)$ を正規分布で近似する．前節 (第 5.1 節) の議論は，正規性を何も仮定しておらず，一切の近似を用いていないことに注意する．この節から近似を導入する．まず，式 (136) の $p\left(y \mid x_0\right)$ と $p\left(x_0 \mid x_t\right)$ を以下で与える．

$$p\left(y \mid x_0\right)=\frac{1}{\sqrt{2 \pi \lambda^2}} \exp \left[-\frac{1}{2} \frac{\left(y-x_0\right)^2}{\lambda^2}\right]=\mathcal{N}\left(y ; x_0, \lambda^2\right) \tag{138}$$

$$p\left(x_0 \mid x_t\right)=\mathcal{N}\left(x_0 ; \mathbb{E}\left[x_0 \mid x_t\right], \operatorname{Var}\left[x_0 \mid x_t\right]\right) \tag{139}$$

分布 $p\left(x_0 \mid x_t\right)$ を条件付き平均 $\mathbb{E}\left[x_0 \mid x_t\right]$ と条件付き分散 $\operatorname{Var}\left[x_0 \mid x_t\right]$ を持つ正規分布で近似している．この時，式 (136) に従い，$l\left(y \mid x_t\right)$ を計算できる．この計算は正規分布同士の畳み込み演算であるため，正規分布の再生性から簡単に実行される．

$$\begin{aligned}
l\left(y \mid x_t\right) &= \int p\left(y \mid x_0\right) p\left(x_0 \mid x_t\right) \mathrm{d} x_0 && (140)\\
&= \int \mathcal{N}\left(y ; x_0, \lambda^2\right) \mathcal{N}\left(x_0 ; \mathbb{E}\left[x_0 \mid x_t\right], \operatorname{Var}\left[x_0 \mid x_t\right]\right) \mathrm{d} x_0 && (141)\\
&= \frac{1}{2 \pi \sqrt{\lambda^2 \operatorname{Var}\left[x_0 \mid x_t\right]}} \int \mathrm{d} x_0 \exp \left[-\frac{1}{2} \frac{\left(x_0-y\right)^2}{\lambda^2}\right] \exp \left[-\frac{1}{2} \frac{\left(x_0-\mathbb{E}\left[x_0 \mid x_t\right]\right)^2}{\operatorname{Var}\left[x_0 \mid x_t\right]}\right] && (142)\\
&= \frac{1}{\sqrt{2 \pi\left(\lambda^2+\operatorname{Var}\left[x_0 \mid x_t\right]\right)}} \exp \left[-\frac{1}{2} \frac{\left(y-\mathbb{E}\left[x_0 \mid x_t\right]\right)^2}{\lambda^2+\operatorname{Var}\left[x_0 \mid x_t\right]}\right] && (143)\\
&= N\left(y ; \mathbb{E}\left[x_0 \mid x_t\right], \lambda^2+\operatorname{Var}\left[x_0 \mid x_t\right]\right) && (144)
\end{aligned}$$

ここで $\mathbb{E}\left[x_0 \mid x_t\right]$ と $\operatorname{Var}\left[x_0 \mid x_t\right]$ は x_0 に依存しない (x_0 から見れば定数として振る舞う) ことを利用した．以上より，y は平均 $\mathbb{E}\left[x_0 \mid x_t\right]$ かつ分散 $\lambda^2+\operatorname{Var}\left[x_0 \mid x_t\right]$ の正規分布に従う．

残りの問題は，条件付き平均 $\mathbb{E}\left[x_0 \mid x_t\right]$ と条件付き分散 $\operatorname{Var}\left[x_0 \mid x_t\right]$ の計算となる．条件付き平均は，以前導出した式 (101) から計算される．

$$\mathbb{E}\left[x_0 \mid x_t\right]=e^{a t}\left[x_t+\sigma_t^2 \frac{\partial}{\partial x_t} \ln p\left(x_t\right)\right] \tag{145}$$

無条件スコア関数を含むが，この量は DSM を通して拡散モデルにより推定されている．そのため，右辺の全ての量を求められる．

条件付き分散を得るためには，スコア関数の微分を評価すればよい．式展開は少し長くなるが，計算自体は

単純で，式 (124) と (127) を用いて条件付き分布 $p(x_t|x_0)$ で書き換えていく．

$$\frac{\partial^2}{\partial x_t^2}\ln p(x_t) = \frac{\partial}{\partial x_t}\left(\frac{\partial_{x_t} p(x_t)}{p(x_t)}\right) \tag{146}$$

$$= \frac{\partial_{x_t}^2 p(x_t)}{p(x_t)} - \frac{1}{p(x_t)^2}\left(\partial_{x_t} p(x_t)\right)^2 \tag{147}$$

$$= \frac{1}{p(x_t)}\partial_{x_t}^2 \int p\left(x_t \mid x_0\right) p\left(x_0\right) \mathrm{d}x_0 - \left[\frac{1}{p(x_t)}\partial_{x_t}\int p(x_t|x_0) p\left(x_0\right) \mathrm{d}x_0\right]^2 \tag{148}$$

$$= \frac{-1}{p(x_t)}\partial_{x_t}\int \left(\frac{x_t - e^{-at}x_0}{\sigma_t^2}\right) p\left(x_t \mid x_0\right) p\left(x_0\right) \mathrm{d}x_0 - \left[\int \left(\frac{x_t - e^{-at}x_0}{\sigma_t^2}\right) \frac{p\left(x_t \mid x_0\right) p(x_0)}{p(x_t)}\right]^2 \tag{149}$$

$$= \frac{-1}{\sigma_t^2}\int \frac{p\left(x_t \mid x_0\right) p(x_0)}{p(x_t)} \mathrm{d}x_0 + \int \left(\frac{x_t - e^{-at}x_0}{\sigma_t^2}\right)^2 \frac{p\left(x_t \mid x_0\right) p(x_0)}{p(x_t)} dx_0 - \left[\int \left(\frac{x_t - e^{-at}x_0}{\sigma_t^2}\right) p(x_0|x_t)\right]^2 \tag{150}$$

$$= \frac{-1}{\sigma_t^2}\int p(x_0|x_t) \mathrm{d}x_0 + \int \left(\frac{x_t - e^{-at}x_0}{\sigma_t^2}\right)^2 p(x_0|x_t) dx_0 - \left[\int \left(\frac{x_t - e^{-at}x_0}{\sigma_t^2}\right) p(x_0|x_t)\right]^2 \tag{151}$$

$$= \frac{-1}{\sigma_t^2} + \frac{x_t^2 - 2x_t e^{-at}\mathbb{E}[x_0|x_t] + e^{-2at}\left(\mathbb{E}[x_0|x_t]\right)^2}{\sigma_t^4} - \frac{x_t^2 - 2x_t e^{-at}\mathbb{E}[x_0|x_t] + e^{-2at}\mathbb{E}[(x_0)^2|x_t]}{\sigma_t^4} \tag{152}$$

$$= -\frac{1}{\sigma_t^2} + \frac{e^{-2at}}{\sigma_t^4}\mathrm{Var}\left[x_0 \mid x_t\right] \tag{153}$$

以上より，条件付き分散 $\mathrm{Var}\left[x_0 \mid x_t\right]$ をスコア関数の微分で表せた．

$$\mathrm{Var}\left[x_0 \mid x_t\right] = e^{2at}\left[\sigma_t^2 + \sigma_t^4 \frac{\partial^2}{\partial x_t^2}\ln p(x_t)\right] \tag{154}$$

スコア関数の微分は未知量であるが，例えば，自動微分を用いることでスコア関数から推定できる．あるいは，未知であるスコア関数の微分をハイパーパラメータとみなす近似も可能である．

今までの結果をまとめると，対数平均尤度は以下で与えられる

$$\ln l(y|x_t) = -\frac{1}{2}\frac{\left(y - \mathbb{E}[x_0|x_t]\right)^2}{\lambda^2 + \mathrm{Var}[x_0|x_t]} + \mathrm{const} \tag{155}$$

$$= -\frac{1}{2}\frac{\left[y - e^{at}\left\{x_t + \sigma_t^2 \partial_{x_t}\ln p(x_t)\right\}\right]^2}{\lambda^2 + \mathrm{Var}[x_0|x_t]} + \mathrm{const} \tag{156}$$

これを逆 SDE (137) に代入すれば，無条件スコア関数の推定結果から追加の学習を行うことなく，条件付き分布 $p(x_0|y)$ に従うサンプルを生成できる．その際，無条件スコア関数の高階微分を求める必要があるが，例えば，それらはハイパーパラメータとして与えるか，自動微分を使って計算できる．

6 おわりに

このノートは確率過程の基礎を説明し，その知識に基づいて拡散モデルの数理を解説した．第 1 章の疑問とそれらへの回答を再掲載すると

疑問 (i)　任意のデータを正規ノイズへ時間発展させるには SDE をどう設定するか？

回答 (i)　線形摩擦を含む SDE を設定し，任意のデータを指数関数的に減衰させる

疑問 (ii)　この SDE を時間について逆回しする方法はなにか？
回答 (ii)　スコア関数を導入することで，順過程と同じ確率分布を解とする逆過程を構成する

疑問 (iii)　そもそも SDE をどうやってデータから決定 (または学習) するか？
回答 (iii)　デノイジングスコアマッチングによりノイズの推定問題を通してスコア関数の近似を得る

このノートで説明していない点について疑問を持つかもしれない．それらの解決のために参考文献をあげておく．指数関数的な減衰をコントロールするパラメータ a と b の設定方法は色々と提案されている [19]．そもそも $a = 0$ として，摩擦項を含まない SDE も利用可能である [3]．逆 SDE のノイズ振幅 c は自由に設定可能だが [12, 15]，それを設定する指針も提案されている [20]．そもそも SDE の変換を等速直線運動のように単純にする方法は盛んに研究されている [21]．データ生成を正規分布ではなく任意の分布から始める手法も盛んに研究されている [22]．画像の空間構造を加味してノイズへと緩和させる方法も提案されている [23]．このノートでは解説できなかったが，ノイズは正則化の効果を持ち，学習時に有効となりうる [24]．この正則化が，ノイズなしの常微分方程式ではなく，ノイズありの確率微分方程式を学習に利用する一つの理由である．

参考文献

[1] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, Vol. 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, Lille, France, 07–09 Jul 2015. PMLR.

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, Vol. 33, pp. 6840–6851. Curran Associates, Inc., 2020.

[3] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[4] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 45, No. 4, pp. 4713–4726, 2023.

[5] Crispin Gardiner. *Stochastic Methods*. Springer Series in Synergetics. Springer Berlin, Heidelberg, 2009.

[6] Naoto Shiraishi. *An Introduction to Stochastic Thermodynamics: From Basic to Advanced*, Vol. 212. Springer Nature, 2023.

[7] 岡野原大輔. 拡散モデル. 岩波書店, 2023.

[8] 橋本幸士（編）. 学習物理学入門. 朝倉書店, 2024.

[9] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*,

Vol. 37 of *Proceedings of Machine Learning Research*, pp. 1530–1538, Lille, France, 07–09 Jul 2015. PMLR.

[10] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, p. 6572–6583, Red Hook, NY, USA, 2018. Curran Associates Inc.

[11] Brian D.O. Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, Vol. 12, No. 3, pp. 313–326, 1982.

[12] Yuji Hirono, Akinori Tanaka, and Kenji Fukushima. Understanding diffusion models by feynman's path integral. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[13] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, Vol. 23, No. 7, pp. 1661–1674, 2011.

[14] François Rozet and Gilles Louppe. Score-based data assimilation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[15] Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

[16] Steven H Strogatz. *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. CRC press, second edition edition, 2018.

[17] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, Vol. 35, pp. 26565–26577. Curran Associates, Inc., 2022.

[18] Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, Vol. 34, pp. 17695–17709. Curran Associates, Inc., 2021.

[19] Zhehao Guo, Jiedong Lang, Shuyu Huang, Yunfei Gao, and Xintong Ding. A comprehensive review on noise control of diffusion model, 2025.

[20] Yifan Chen, Mark Goldstein, Mengjian Hua, Michael Samuel Albergo, Nicholas Matthew Boffi, and Eric Vanden-Eijnden. Probabilistic forecasting with stochastic interpolants and föllmer processes. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, Vol. 235 of *Proceedings of Machine Learning Research*, pp. 6728–6756. PMLR, 21–27 Jul 2024.

[21] Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport, 2024.

[22] Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos Theodorou, Weili Nie, and Anima Anandkumar. I^2SB: Image-to-image schrödinger bridge. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th In-*

ternational Conference on Machine Learning, Vol. 202 of *Proceedings of Machine Learning Research*, pp. 22042–22062. PMLR, 23–29 Jul 2023.

[23] Artan Sheshmani, Yi-Zhuang You, Baturalp Buyukates, Amir Ziashahabi, and Salman Avestimehr. Renormalization group flow, optimal transport and diffusion-based generative model, 2024.

[24] Tianrong Chen, Guan-Horng Liu, and Evangelos Theodorou. Likelihood training of schrödinger bridge using forward-backward SDEs theory. In *International Conference on Learning Representations*, 2022.