拡散モデル（Diffusion Model）を理解するために

20B01392 松本侑真

2023 年 9 月 20 日

概要

近年、機械学習を応用した生成系 AI がブームである。生成系 AI を使うことで、プロンプトに入力したテキストデータから画像データなどを出力することができる。拡散モデルを用いた有名な生成系 AI の Stable Diffusion では、無料でプロンプトに入力した文字（例：かわいい犬）から、かわいい犬の画像を出力することができる。このような生成系 AI で用いられる拡散モデルについて理解するための基礎を説明する。

目次

1 生成モデルでは何を行っているのか 2

2 ボルツマンマシン 2
- 2.1 Kullback-Leibler 情報量 . . . 3
- 2.2 例：ガウス分布の KL 情報量と学習 . . . 4
- 2.3 最尤法（=KL 最小化） . . . 5
- 2.4 高次元化 . . . 5
- 2.5 よくあるボルツマンマシン . . . 6
- 2.6 勾配法と代理関数 . . . 7

3 隠れ変数を導入してリッチなモデルへ 7

4 制限ボルツマンマシン 7

5 マルコフ連鎖モンテカルロ法（MCMC） 7

6 交換モンテカルロ法 7

7 MCMC を用いないボルツマンマシン 7

8 変分オートエンコーダー 7

9 階層変分オートエンコーダ 7

10 Appendix 7
- 10.1 微分に必要な線形代数の公式その 1 . . . 7
- 10.2 微分に必要な線形代数の公式その 2 . . . 8
- 10.3 微分に必要な線形代数の公式その 3 . . . 8

1　生成モデルでは何を行っているのか

生成モデル（拡散モデル）は、事前に与えられた教師データを元にモデルを学習することで、未知の入力に対して最適化された出力をアウトプットすることができる。例えば、いろいろな人間の顔データを学習させた生成モデルを用いると、「40 歳のおじさん」や「20 代のアイドル」といった入力を元にして、学習されたモデルで生成した顔画像を出力する。

生成モデルの学習というのは、「モデルの入力から欲しいデータを生み出す確率分布」をいかにして見つけるかということである。すなわち、任意のデータ $\vec{x}$ は、とある 1 つの確率分布 $p(\vec{x})$ に従って生み出されるという仮定のもと、自分のモデルをその $p(\vec{x})$ に限りなく近づけることが生成モデルにおける学習である。すなわち、あらゆるデータが従う確率分布 $p(\vec{x})$ そのものを手に入れることができれば、$p(\vec{x})$ の値を出力するような $\vec{x}$ を見つけることが可能であるだろう。これによって、$p(\vec{x})$ =「20 代のアイドル」を満たす画像 $\vec{x}$ を出力することができる。

学習を行う前に得られているもの

モデルの学習を行うためには、そのための学習データが必要である。n 個の学習データ $\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_n$ が手元にあるとする。このとき、

$$p_0(\vec{x}) = \frac{1}{n}\sum_{i=1}^{n} \delta(\vec{x} - \vec{x}_i) \tag{1.1}$$

を経験分布と呼ぶ。この経験分布は、n が十分に大きいとき、あらゆるデータを生み出す確率分布 $p(x)$ に非常に近いであろう：

$$p_0(\vec{x}) \approx p(\vec{x})\ 。 \tag{1.2}$$

なお、データを連続変数として扱う場合は $\delta(\vec{x} - \vec{x}_i)$ をデルタ関数、離散変数として扱う場合はクロネッカーのデルタとして扱う。

学習のために必要なもの

生成モデルにおいて学習させるための確率分布は自ら用意する必要がある。すなわち、パラメータ θ を用いて、

$$p(\vec{x}) \approx q_\theta(\vec{x}) \tag{1.3}$$

となるようなモデル $q_\theta(\vec{x})$ を見つけたい。モデル $q_\theta(\vec{x})$ の関数形は計算しやすい形を用いて、パラメータ θ を最適化することで $p(\vec{x})$ に近づけるというアプローチを取る。

2　ボルツマンマシン

ボルツマンマシンとは、モデル $q_\theta(\vec{x})$ が

$$q_\theta(\vec{x}) = \frac{1}{Z_\theta}\exp(-E_\theta(\vec{x}))\ 。 \tag{2.1}$$

と表されるものである。このような関数形を Gibbs-Boltzmann 分布（指数関数分布族）と呼ぶ。ここで、規格化定数 Z_θ は

$$Z_\theta = \sum_{\vec{x}} \exp(-E_\theta(\vec{x})) \tag{2.2}$$

である。このように定義された $q_\theta(\vec{x})$ を $p(\vec{x})$ に近づけることが目標である。しかし、$p(\vec{x})$ の形は誰も知らない。今手元にあるのは経験分布 $p_0(\vec{x})$ だけである。そのため、学習データの数 n が十分に多いとして、$q_\theta(\vec{x})$ を経験分布 $p_0(\vec{x})$ に近づけることを行う。

2.1 Kullback-Leibler 情報量

学習を行うためには、2 つの確率分布 $p(\vec{x})$, $q(\vec{x})$ の近さを定義する必要がある。この「近さ」は、データ $\vec{x}$ を入力した際の 2 つの確率分布間の「距離 $D(p \parallel q)$」を測ることで定義することができるだろう。このような関数 $D(p \parallel q)$ はどのように設定しても良いのだが、今回は、Kullback-Leibler divergence（KL 情報量）$D_{\mathrm{KL}}(p \parallel q)$ を導入する：

$$D_{\mathrm{KL}}(p(\vec{x}) \parallel q(\vec{x})) = \sum_{\vec{x}} p(\vec{x}) \ln \frac{p(\vec{x})}{q(\vec{x})} \text{。} \tag{2.3}$$

KL 情報量には

(i). $D(p \parallel q) \geq 0$（等号成立は $p = q$ のとき）
(ii). $D(p \parallel q) \neq D(q \parallel p)$（非対称性）

といった性質がある。特に、KL 情報量は 0 以上であるといった性質を良く用いる。この性質を 2 つの方法で証明してみる。

Gibbs の不等式を用いる方法

Gibbs の不等式とは、$x > 0$ において

$$x - 1 \geq \ln x \Longleftrightarrow \ln \frac{1}{x} \geq 1 - x \tag{2.4}$$

が成立する不等式である。

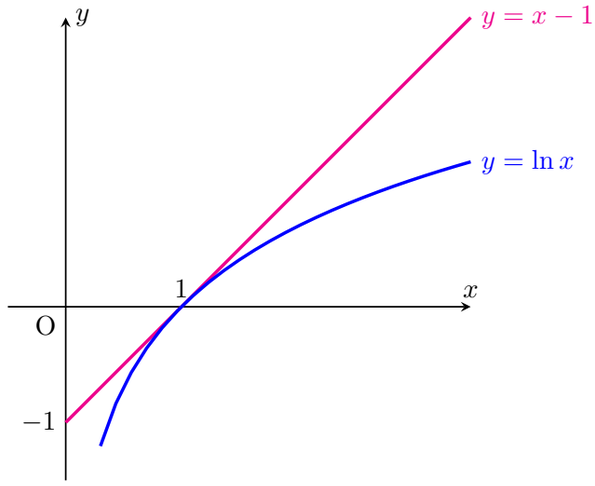

図 1: Gibbs の不等式の直感的な理解

Gibbs の不等式を用いると、

$$D_{\mathrm{KL}}(p \parallel q) = \sum_{\vec{x}} p(\vec{x}) \ln \frac{p(\vec{x})}{q(\vec{x})} \geq \sum_{\vec{x}} p(\vec{x}) \left(1 - \frac{q(\vec{x})}{p(\vec{x})}\right) = \sum_{\vec{x}} p(\vec{x}) - \sum_{\vec{x}} q(\vec{x}) = 0 \tag{2.5}$$

と示される。

Jensen の不等式を用いる方法

Jensen の不等式とは、上に凸な関数 $f(x)$ に対して、

$$f\left(\frac{a+b}{2}\right) \geq \frac{f(a)+f(b)}{2} \tag{2.6}$$

が成立する不等式である。

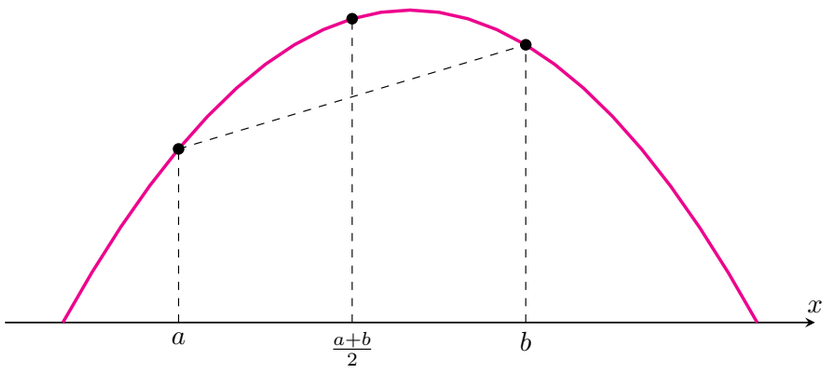

図 2: Jensen の不等式の直感的な理解

Jensen の不等式を用いると、

$$D_{\mathrm{KL}}(p \parallel q) = -\sum_{\vec{x}} p(\vec{x}) \ln \frac{q(\vec{x})}{p(\vec{x})} \geq -\ln\left(\sum_{\vec{x}} p(\vec{x}) \frac{q(\vec{x})}{p(\vec{x})}\right) = 0 \tag{2.7}$$

と示される。

2.2 例：ガウス分布の KL 情報量と学習

モデル $q_\theta(\vec{x})$ をガウス分布とおいて最適化を行う。人間が計算できるのはせいぜいガウス分布程度であるため、これ以上難しいモデルは考えない。まずは $\vec{x}$ が 1 次元のスカラーである場合を考える：

$$q_\theta(x) \propto \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)。 \tag{2.8}$$

規格化定数は

$$Z_\theta = \int_{-\infty}^{\infty} dx\, \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) = \sqrt{2\pi\sigma^2} \tag{2.9}$$

となる。したがって、

$$q_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \tag{2.10}$$

となる。このモデルの未知数は μ と σ^2 であるため、最適化すべきパラメータは $\theta = \{\mu, \sigma^2\}$ となる。次に、KL 情報量を考える。$q_\theta(x)$ を $p(x)$ に近づけることができれば良いのだが、あいにく $p(x)$ の具体形はわからない。そのため、経験分布 $p_0(x) = \sum_{i=1}^n \delta(x - x_i)$ にモデルを近づけることを考える。KL 情報量を計算すると、

$$\begin{aligned} D_{\mathrm{KL}}(p_0 \parallel q_\theta) &= \int dx\, p_0(x)(\ln p_0(x) - \ln q_\theta(x)) \underbrace{=}_{\theta\text{のみ}} - \int dx\, p_0(x)\left(\frac{1}{2\sigma^2}(x-\mu)^2 + \frac{1}{2}\ln\sigma^2\right) \\ &= \frac{1}{2\sigma^2}\frac{1}{n}\sum_{i=1}^n (x_i-\mu)^2 + \frac{1}{2}\ln\sigma^2 \end{aligned} \tag{2.11}$$

となる。KL 情報量を最小化（学習）すれば、経験分布に近いモデルが得られるため、μ と σ^2 で偏微分を行う。

$$\frac{\partial D_{\mathrm{KL}}}{\partial \mu} = -\frac{1}{\sigma^2}\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu) = 0 \qquad \Longleftrightarrow \mu = \frac{1}{n}\sum_{i=1}^{n} x_i \text{（データの平均）} \tag{2.12}$$

$$\frac{\partial D_{\mathrm{KL}}}{\partial \sigma^2} = -\frac{1}{2(\sigma^2)^2}\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 + \frac{1}{2\sigma^2} = 0 \qquad \Longleftrightarrow \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \text{（データの分散）} \tag{2.13}$$

この結果からわかることは、ガウス分布のモデルの平均 μ はデータの平均値に、モデルの分散 σ^2 はデータの分散にすれば、$p(x)$ に近いモデル $q_\theta(x)$ を生成できるということである。

2.3 最尤法（=KL 最小化）

最尤法とは、KL 情報量の最小化と同値なものであるが、良く使われるため紹介する。モデルの学習は、

$$\min_\theta \{D_{\mathrm{KL}}(p_0(\vec{x}) \parallel q_\theta(\vec{x}))\} = \max_\theta \left\{ \sum_{\vec{x}} p(\vec{x}) \ln q_\theta(\vec{x}) \right\} = \max_\theta \left\{ \frac{1}{n}\sum_{i=1}^{n} \ln q_\theta(\vec{x}_i) \right\} \tag{2.14}$$

と変形することができる。$\ln q_\theta(\vec{x}_i)$ を対数尤度関数と呼び、モデルの良さの指標として使われる。対数尤度を最大化することは、KL 情報量の最小化と同値なものである。

2.4 高次元化

先ほどはガウス分布のモデルの 1 次元バージョンの最適化を行った。一般の場合にどのように拡張されるかを見る。すなわち、

$$q_\theta(\vec{x}) \propto \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu})^\top \Sigma^{-1} (\vec{x} - \vec{\mu}) \right) \tag{2.15}$$

とする。規格化定数は

$$Z_\theta = \int d\vec{x}\, \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu})^\top \Sigma^{-1} (\vec{x} - \vec{\mu}) \right) \tag{2.16}$$

となる。このままでは規格化定数の計算ができないが、Σ^{-1} の対角化を行うことで計算を進めることができる。ある正方行列 P を用いて、

$$\begin{cases} P^{-1}\Sigma^{-1}P & = \Lambda = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_N \end{pmatrix} \\ P^{-1}(\vec{x} - \vec{\mu}) & = \vec{y} \end{cases} \tag{2.17}$$

と対角化できたとする。Σ は分散共分散行列と呼ばれる。このとき、

$$Z_\theta = \int d\vec{y}\, \exp\left(-\frac{1}{2}\vec{y}^\top \Lambda \vec{y} \right) = \prod_{k=1}^{N} \int d\vec{y}_k\, \exp\left(-\frac{1}{2}\lambda_k y_k^2 \right) = \prod_{k=1}^{N} \sqrt{\frac{2\pi}{\lambda_k}} \tag{2.18}$$

と計算できる。行列の determinant の性質

$$\prod_{k=1}^{N} \lambda_k = \det(\Lambda) = \det\left(\Sigma^{-1}\right) = \frac{1}{\det(\Sigma)} \tag{2.19}$$

を用いると、

$$Z_\theta = \sqrt{(2\pi)^N \det(\Sigma)} \tag{2.20}$$

のように、分散共分散行列 Σ を用いて計算できる。したがって、

$$q_\theta(\vec{x}) = \sqrt{\frac{1}{(2\pi)^N \det(\Sigma)}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^\top \Sigma^{-1}(\vec{x}-\vec{\mu})\right) \tag{2.21}$$

となる。KL 情報量は、

$$D_{\mathrm{KL}}(p_0 \parallel q_\theta) = \int d\vec{x}\, p_0(\vec{x})(\ln p_0(\vec{x}) - q_\theta(\vec{x})) \underbrace{=}_{\theta\text{のみ}} \frac{1}{2n}\sum_{i=1}^{n}(\vec{x}_i-\vec{\mu})^\top\Sigma^{-1}(\vec{x}_i-\vec{\mu}) + \frac{1}{2}\ln\det(\Sigma) \tag{2.22}$$

と計算できる。KL 情報量の最小化

$$\frac{\partial D_{\mathrm{KL}}}{\partial \vec{\mu}} = -\frac{1}{n}\sum_{i=1}^{n}\Sigma^{-1}(\vec{x}_i-\vec{\mu}) = 0 \qquad \Longleftrightarrow \vec{\mu} = \frac{1}{n}\sum_{i=1}^{n}\vec{x}_i \tag{2.23}$$

$$\frac{\partial D_{\mathrm{KL}}}{\partial \Sigma^{-1}} = \frac{1}{2n}\sum_{i=1}^{n}(\vec{x}_i-\vec{\mu})(\vec{x}_i-\vec{\mu})^\top - \frac{1}{2}\Sigma = 0 \qquad \Longleftrightarrow \Sigma = \frac{1}{n}\sum_{i=1}^{n}(\vec{x}_i-\vec{\mu})(\vec{x}_i-\vec{\mu})^\top \tag{2.24}$$

によって、パラメータの最適化を行うことができる。

2.5 よくあるボルツマンマシン

ボルツマンマシンのエネルギー関数はどのような形でも良いが、一般的にはエネルギー関数として Ising モデルを用いた

$$q_\theta(\vec{x}) = \frac{1}{Z_\theta}\exp\left(\sum_{k\neq l}J_{kl}x_kx_l + \sum_{k=1}^{N}h_kx_k\right), \quad \vec{x} = \{-1,1\}^N \tag{2.25}$$

$$Z_\theta = \sum_{\vec{x}}\exp\left(\sum_{k\neq l}J_{kl}x_kx_l + \sum_{k=1}^{N}h_kx_k\right) \tag{2.26}$$

が使われる。$\sum_{k\neq l}$ は異なる添字のペアについて 1 回ずつ足すことを意味しており、J_{kl} がランダムなモデルをスピングラスモデルと呼ぶ。このときの KL 情報量は

$$\begin{aligned} D_{\mathrm{KL}}(p_0 \parallel q_\theta) &= \sum_{\vec{x}} p_0(\vec{x})(\ln p_0(\vec{x}) - \ln q_\theta(\vec{x})) \\ &\underbrace{=}_{\theta\text{のみ}} -\sum_{\vec{x}} p_0(\vec{x})\left(\sum_{k\neq l}J_{kl}x_kx_l + \sum_{k=1}^{N}h_kx_k\right) + \ln Z_\theta \\ &= -\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k\neq l}J_{kl}x_k^{(i)}x_l^{(i)} + \sum_{k=1}^{N}h_kx_k^{(i)}\right) + \ln Z_\theta \end{aligned} \tag{2.27}$$

である。パラメータの学習は、

$$\frac{\partial D_{\mathrm{KL}}}{\partial h_k} = -\frac{1}{n}\sum_{i=1}^{n}x_k^{(i)} + \sum_{\vec{x}}x_k\frac{\exp\left(\sum_{k\neq l}J_{kl}x_kx_l + \sum_{k=1}^{N}h_kx_k\right)}{Z_\theta} = 0 \qquad \Longleftrightarrow \langle x_k\rangle_\theta = \frac{1}{n}\sum_{i=1}^{n}x_k^{(i)} \tag{2.28}$$

$$\frac{\partial D_{\mathrm{KL}}}{\partial J_{kl}} = -\frac{1}{n}\sum_{i=1}^{n}x_k^{(i)}x_l^{(i)} + \sum_{\vec{x}}x_kx_l\frac{\exp\left(\sum_{k\neq l}J_{kl}x_kx_l + \sum_{k=1}^{N}h_kx_k\right)}{Z_\theta} = 0 \quad \Longleftrightarrow \langle x_kx_l\rangle_\theta = \frac{1}{n}\sum_{i=1}^{n}x_k^{(i)}x_l^{(i)} \tag{2.29}$$

となる。したがって、モデルの平均値とデータの平均値を同じにするようにパラメータを決定すれば良いことがわかるが、それを実現する J_{kl}, h_k の値は不明なままである。

2.6　勾配法と代理関数

Ising モデルにおけるパラメータの最適化は勾配法によって行うことができる。すなわち、

$$\begin{cases} h'_k = h_k - \eta_k \left(\dfrac{1}{n} \displaystyle\sum_{i=1}^{n} x_k^{(i)} - \langle x_k \rangle_{\theta=\{h,J\}} \right) \\ J'_{kl} = J_{kl} - \eta_{kl} \left(\dfrac{1}{n} \displaystyle\sum_{i=1}^{n} x_k^{(i)} x_l^{(i)} - \langle x_k x_l \rangle_{\theta=\{h,J\}} \right) \end{cases} \tag{2.30}$$

のようにする。$\theta = \{h, J\} \to \theta' = \{h', J'\}$ へ更新する作業を繰り返すことで、パラメータの最適化を行うことができる。なお、η_k, η_{kl} は勾配法の学習率と呼ばれ、小さな値が設定される。パラメータがある値に収束するように学習率を適切に設定する。

3　隠れ変数を導入してリッチなモデルへ

4　制限ボルツマンマシン

5　マルコフ連鎖モンテカルロ法（MCMC）

6　交換モンテカルロ法

7　MCMC を用いないボルツマンマシン

8　変分オートエンコーダー

9　階層変分オートエンコーダ

10　Appendix

10.1　微分に必要な線形代数の公式その 1

$\vec{x}$ を n 次元ベクトル、A を n 次正方行列とする。このとき、$\vec{x}^\top A \vec{x}$ はスカラーであるため、

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\vec{x}} \left(\vec{x}^\top A \vec{x} \right) &= \frac{\mathrm{d}}{\mathrm{d}\vec{x}} \operatorname{tr}(\vec{x}^\top A \vec{x}) = \frac{\mathrm{d}}{\mathrm{d}\vec{x}} \operatorname{tr}\left(A \vec{x} \vec{x}^\top \right) = \frac{\mathrm{d}}{\mathrm{d}\vec{x}} \sum_{i,j=1}^{n} A_{ij} (\vec{x} \vec{x}^\top)_{ji} \\ &= \sum_{i=1}^{n} \boldsymbol{e}_i \frac{\mathrm{d}}{\mathrm{d}x_i} \left(\sum_{k,l=1}^{n} A_{kl} x_l x_k \right) = \sum_{i=1}^{n} \boldsymbol{e}_i \left(\sum_{k=1}^{n} A_{ki} x_k + \sum_{l=1}^{n} A_{il} x_l \right) = (A + A^\top) \vec{x} \end{aligned} \tag{10.1}$$

が成立する。なお、$\boldsymbol{e}_i$ を i 番目の基底ベクトルとして、$\mathrm{d}/\mathrm{d}\vec{x} = \sum_{i=1}^{n} \boldsymbol{e}_i \, \mathrm{d}/\mathrm{d}x_i$ であることを用いた。分散共分散行列は対称行列であるため、

$$\frac{\mathrm{d}}{\mathrm{d}\vec{x}} \left(\vec{x}^\top \Sigma^{-1} \vec{x} \right) = 2 \Sigma^{-1} \vec{x} \tag{10.2}$$

となる。

10.2　微分に必要な線形代数の公式その 2

$$\frac{\mathrm{d}}{\mathrm{d}A_{ij}}\left(\vec{x}^{\top}A\vec{x}\right)=\frac{\mathrm{d}}{\mathrm{d}A_{ij}}\operatorname{tr}\left(A\vec{x}\vec{x}^{\top}\right)=\frac{\mathrm{d}}{\mathrm{d}A_{ij}}\sum_{i,j=1}^{n}A_{ij}(\vec{x}\vec{x}^{\top})_{ji}=x_jx_i=(\vec{x}\vec{x}^{\top})^{\top}_{ij} \tag{10.3}$$

が成立するため、

$$\frac{\mathrm{d}}{\mathrm{d}A}\left(\vec{x}^{\top}A\vec{x}\right)=\vec{x}\vec{x}^{\top} \tag{10.4}$$

となる。

10.3　微分に必要な線形代数の公式その 3

determinant の余因子展開

$$\det(A)=\sum_{j=1}^{n}(-1)^{i+j}A_{ij}m_{ij}\quad(m_{ij}\text{は }A\text{ の }(i,j)\text{ 小行列式}) \tag{10.5}$$

を用いると、

$$\left(\frac{\mathrm{d}}{\mathrm{d}A}\det(A)\right)_{ij}=\frac{\mathrm{d}}{\mathrm{d}A_{ij}}\sum_{j=1}^{n}(-1)^{i+j}A_{ij}m_{ij}=(-1)^{i+j}m_{ij} \tag{10.6}$$

となる。さらに、余因子行列 $\tilde{A}$ を用いて $\det(A)=A\tilde{A}$ と表されるため、

$$\frac{\mathrm{d}}{\mathrm{d}A}\ln\det(A)=\frac{1}{\det(A)}\frac{\mathrm{d}}{\mathrm{d}A}\det(A)=\frac{1}{\det(A)}\tilde{A}^{\top}=(A^{-1})^{\top} \tag{10.7}$$

を得る。したがって、

$$\frac{\mathrm{d}}{\mathrm{d}\Sigma^{-1}}\ln\det(\Sigma)=-\Sigma^{\top}=-\Sigma \tag{10.8}$$

となる。