修士論文

Stiefel 空間上の
変分推論及び変分オートエンコーダ
Variational Inference and
Variational Auto-Encoder
on the Stiefel Space

東京大学大学院 情報理工学系研究科
電子情報学専攻
48-186446　三條 嵩明

指導教員　　喜連川 優 教授

2020 年 1 月 30 日 提出

本論文は東京大学大学院情報理工学系研究科に修士号授与の要件として提出した修士論文である.

要約

統計・機械学習において，観測データからその背後の潜在状態を推定する際には，通常ユークリッド空間上での定式化が行われる．しかし，データがユークリッド空間とは異なる空間に分布する場合，その本質的な構造を適切に捉えることは難しい．

本研究では，潜在的に正規直交性を持つようなデータを，その本質的な構造を損なわずに扱うことを目的とし，非ユークリッド空間である Stiefel 空間と呼ばれる空間における，変分推論及び変分オートエンコーダの学習手法を提案する．本提案手法を用いて変分推論と変分オートエンコーダの学習を行い，潜在的に正規直交性を持つようなデータに対する有効性を評価する．

目次

図目次

表目次

第 1 章

序論

1.1 背景

統計及び機械学習では，観測したデータからその背後にあるパラメタや潜在状態の事後分布を求めることで，予測や分類など，様々なタスクを解くことができる．この事後分布は，モデルが大規模で複雑になるほど，計算することが困難になる．変分推論は，そのような場合に真の事後分布を計算の容易な別の分布で近似的に表現する推論手法である．また，この変分推論をベースとした深層学習モデルである変分オートエンコーダ (Variational Auto-Encoder; VAE) は，教師無し生成モデルとして最も広く用いられる手法の一つであり，画像生成など様々な分野に応用されている [1, 2]．VAE は自己符号化器に変分推論の手法を取りこんだモデルであり，何らかの事前分布を仮定して，それに KL divergence の意味で近くなるような正則化を行いながら確率分布を推定する．

変分推論や VAE の学習の際には，計算を簡単にするために，データがユークリッド空間 $\mathbb{R}^m$ 上に分布することを仮定し，事前，事後分布としてそれらのユークリッド空間上の分布を用いることが一般的である．しかし，この仮定は $\mathbb{R}^m$ と同相でない標本空間上に分布するデータを学習する際には不適切である．

ユークリッド空間と同相でない標本空間の例として超球面が挙げられる．例えば，タンパク質構造の二面角や風向きといった方向データを扱う上では超球面の標本空間を用いた方が適切であることが古くから知られている [3, 4]．また近年の自然言語処理や画像処理の分野においても，cos 類似度を重視したい等の理由により特徴量ベクトルのノルムによる正規化が行われるような場合は，方向データとして扱われる方が適切で

ある．実際に，いくつかの機械学習タスクについて，ガウス分布の代わりに超球面上の確率分布である von Mises-Fisher 分布を事前分布として用いた VAE の方が安定して学習を行えることが報告されている．[5–7].

この超球面を一般化した空間として，Stiefel 空間と呼ばれる空間を考えることができる．Stiefel 空間は空間上の一点が k 個の正規直交基底の組に対応するような空間であり，この空間上の統計に関する研究が近年進められてきた．映像など，それぞれのデータ点が潜在的に正規直交性を持つデータについては，Stiefel 空間上の統計を取り入れることで機械学習手法の性能が向上することが報告されている [8].

変分推論及び VAE の学習に Stiefel 空間上の確率分布を用いることで，獲得される潜在状態に正規直交性を課すことができる．そのため，もし取り扱うデータが正規直交性を持つと分かっている場合，Stiefel 空間上の変分推論及び VAE を用いることでその本質的な構造を損なわずに確率的生成モデルを学習できると考えられる．

1.2　本研究の目的と貢献

本研究では，Stiefel 空間を潜在状態の空間として持つような変分推論及び VAE の学習手法の構成に取り組む．まず，Stiefel 空間上で最も基本的な分布である matrix Langevin 分布を用いた推論手法を提案する．次に，Stiefel 空間上で定義されるレトラクションとベクトル輸送という 2 つの演算を用いて巻き込み型正規分布を考案し，その分布を用いた推論手法を提案する．

それら 2 つの手法を，簡単な設定の下，人工データに対する変分推論によって評価した．実験の結果，巻き込み型正規分布を用いた手法は，matrix Langevin 分布のものよりも安定して高速に学習を行うことができることが確認された．更に Stiefel 空間上に潜在的に分布するようなデータを人工的に作成し，そのデータに対して巻き込み型正規分布を用いた VAE の学習を行い，ユークリッド空間上のガウス分布を用いた通常の VAE と比較した．その結果，ユークリッド空間上のガウス分布を用いた通常の VAE と比べての場合について高い性能を達成することを確認した．

1.3　本論文の構成

本論文の構成は以下の通りである．

第 2 章 微分幾何の基礎事項と多様体上の確率分布，Stiefel 空間に関して，本研究の前提知識を説明する．

第 3 章 多様体上の変分推論及び変分オートエンコーダについて，本研究に関連する手法について説明する．

第 4 章 Stiefel 空間上で定義される matrix Langevin 分布と呼ばれる分布を用いて変分推論及び変分オートエンコーダの学習を行う提案手法について説明する．

第 5 章 Stiefel 空間上に巻き込み型正規分布を定義し，変分推論及び変分オートエンコーダの学習を行う提案手法について説明する．

第 6 章 人工データを用意し，変分推論及び変分オートエンコーダの学習により，提案手法を評価する．

第 7 章 本研究のまとめと本研究で解明することのできなかった課題について説明する．

第 2 章

基礎知識

この章では，本研究を理解する上で必要となる基礎的事項について説明を行う．まず微分幾何について簡単な導入を行う．その後，Bayes 推論における事後確率分布の近似推論手法である変分推論について説明する．本章では簡単な説明にとどめた．証明等の詳細は参考文献 [9–18] を参照されたい．

2.1 微分幾何に関する基礎的事項

まず，多様体を扱う上で重要な微分幾何の用語について整理する．その後，多様体上で定義される演算と確率測度について導入を行う．

2.1.1 多様体論

局所的にユークリッド空間 $\mathbb{R}^n$ と見なせるような集合 M を多様体という．多様体やその上での接ベクトル，写像の微分，微分形式などの厳密な定義について以下で述べていく．本小節は主に [9–11] に基づいている．

定義 2.1 (位相). 集合 X の部分集合族 $\mathcal{O}$ が以下の条件を満たすとき，$\mathcal{O}$ を X の位相 (topology) と呼び，X と $\mathcal{O}$ の対 $(X, \mathcal{O})$ を位相空間 (topological space) と言う．

1. $X \in \mathcal{O}$ かつ $\phi \in \mathcal{O}$
2. $U_1, U_2, \cdots, U_k \in \mathcal{O}$ ならば $U_1 \cap U_2 \cap \cdots \cap U_k \in \mathcal{O}$
3. 任意の集合族 $\{U_\lambda\}_{\lambda \in \Lambda}$ について，$U_\lambda \in \mathcal{O} (\forall \lambda \in \Lambda)$ ならば $\bigcup_{\lambda \in \Lambda} U_\lambda \in \mathcal{O}$

X の部分集合 U が $\mathcal{O}$ に属する ($U \in \mathcal{O}$) とき，U を X の開集合という．位相 $\mathcal{O}$ を位相空間 X の開集合系と呼ぶことがある．

定義 2.2 (ハウスドルフ空間). $(X, \mathcal{O})$ を位相空間とする．X 上の任意の異なる 2 点 p, q に対して，p を含む開集合 U と q を含む開集合 V であって，$U \cap V = \emptyset$ となる U，V が存在するとする．このとき，$(X, \mathcal{O})$ をハウスドルフ空間 (Hausdorff space) という．

位相空間を用いて，連続写像を定義することができる．

定義 2.3 (連続写像). $f : X \to Y$ が連続写像であるとは，Y の任意の開集合 U について，その逆像 $f^{-1}(U) = \{p \in X \mid f(p) \in U\}$ が X の開集合になることである．

定義 2.4 (同相写像). $(X, \mathcal{O})$，$(Y, \mathcal{O}')$ を位相空間とする．写像 $f : X \to Y$ が以下の条件を満たすとき，f を同相写像 (homeomorphism) という．

1. $f : X \to Y$ は全単射である．
2. $f : X \to Y$ も $f^{-1} : Y \to X$ も，ともに連続写像である．

$f : X \to Y$ が同相写像であるとき，$f^{-1} : Y \to X$ も同相写像である．また，$f : X \to Y$ と $g : Y \to Z$ がともに同相写像であるとき，$g \circ f : X \to Z$ も同相写像である．

位相空間 X と Y の間に同相写像 $f : X \to Y$ が存在するとき，X と Y は互いに位相同型 (homeomorphic) であるといい，$X \cong Y$ と表す．

定義 2.5 (座標近傍). X を位相空間とし，U を X の開部分集合とする．U から，m 次元数空間 $\mathbb{R}^n$ 中のある開集合 U' への同相写像

$$\varphi : U \to U'$$

があるとする．このとき，U と φ の対 (U, φ) を m 次元座標近傍 (coordinate neighborhood) と呼び，φ を U 上の局所座標系 (local coordinate system) という．

定義 2.6 (局所座標). (U, φ) を位相空間 X 内の座標近傍とする．U 内の任意の点 p について，$\varphi(p)$ は U' の点である．したがって $\varphi(p)$ は $\mathbb{R}^m$ 内の点であるから，$\mathbb{R}^m$ の座

標を用いて，

$$\varphi(p) = (x_1, x_2, \cdots, x_m)$$

と書ける．$(x_1, x_2, \cdots, x_m)$ を，(U, φ) に関する p の局所座標 (local coordinates) という．

2 つの m 次元座標近傍 (U, φ) と (V, ψ) が交わっているとき，(U, φ) に関する局所座標 $(x_1, x_2, \cdots, x_m)$ と (V, ψ) に関する局所座標 $(y_1, y_2, \cdots, y_m)$ の間の関係を以下のように考えることができる．

定義 2.7 (座標変換). 同相写像 $\psi \circ \varphi^{-1} : \varphi(U \cap V) \to \psi(U \cap V)$ を，(U, φ) から (V, ψ) への座標変換 (coordinate transformation) と呼ぶ．

定義 2.8 (微分可能多様体). $r \geq 1$ を自然数もしくは ∞ とする．位相空間 M が以下の条件を満たすとき，M を m 次元 C^r 級微分可能多様体 (differential manifold of class C^r) という．

1. M はハウスドルフ空間である．
2. M は m 次元の座標近傍により被覆される．つまり，M の m 次元座標近傍からなる族 $\{(U_\lambda, \varphi_\lambda)\}_{\lambda \in \Lambda}$ があって，

   $$M = \bigcup_{\lambda \in \Lambda} U_\lambda$$

   が成り立つ．
3. $U_\alpha \cap U_\beta \neq \emptyset$ であるような任意の $\alpha, \beta \in \Lambda$ について，座標変換

   $$\varphi_\beta \circ \varphi_\alpha^{-1} : \varphi_\alpha(U_\alpha \cap U_\beta) \to \varphi_\beta(U_\alpha \cap U_\beta)$$

   は C^r 級写像である．

定義 2.8 中の条件 2. を満たす座標近傍の族 $\{(U_\lambda, \varphi_\lambda)\}_{\lambda \in \Lambda}$ を座標近傍系 (system of coordinate neighborhoods)，もしくはアトラス (atlas) という．また，条件 3. を満たす座標近傍系を C^r 級座標近傍系という．

以後簡単のために，C^r 級多様体といえば C^r 級微分可能多様体のことを指すものとする．M を m 次元 C^r 級多様体，N を n 次元 C^r 級多様体とする．

ユークリッド空間上の方向微分や接ベクトル，内積などと同様の概念を，多様体上で局所座標系の取り方によらない形で定義することができる．

定義 2.9 (方向微分). 多様体 M 上の点 p における方向微分 v とは，p の開近傍で定義された C^r 級関数 f に実数 $v(f)$ を対応させる操作であり，以下の性質を持つものをいう．

1. f と g が点 p の十分小さな開近傍上で一致すれば，$v(f) = v(g)$.
2. $v(af + bg) = av(f) + bv(g)$．ただし，$a, b \in \mathbb{R}$ であり，f，g は p の開近傍で定義された任意の C^r 級関数.
3. $v(fg) = v(f)g(p) + f(p)v(g)$.

例 2.10. 多様体 M 上において，p を含む座標近傍 $(U\,;\, x_1, x_2, \cdots x_m)$ を一つ固定する．p の周りで定義された C^r 級関数 f に対して，p における x_i 方向の偏微分係数

$$\frac{\partial f}{\partial x_i}(p) \in \mathbb{R}$$

を対応させる操作を $\left(\frac{\partial}{\partial x_i}\right)_p$ と書くとする．すなわち，

$$\left(\frac{\partial}{\partial x_i}\right)_p : f \mapsto \frac{\partial f}{\partial x_i}(p)$$

このとき，$\left(\frac{\partial}{\partial x_i}\right)_p$ は，p における方向微分の定義 2.9 を満たす．

このように多様体 M 上の方向微分を考えたとき，点 $p \in M$ における方向微分全体のなす集合を $D_p^r(M)$ とする．

定義 2.11 (接ベクトル空間). M を m 次元多様体とする．例 2.10 で導入した m 個のベクトル $\left(\frac{\partial}{\partial x_1}\right)_p, \left(\frac{\partial}{\partial x_2}\right)_p, \cdots, \left(\frac{\partial}{\partial x_m}\right)_p$ の張る $D_p^r(M)$ の部分ベクトル空間を，点 p における M の接ベクトル空間 (tangent vector space) と呼び，

$$T_pM$$

と表す．また，T_pM に属するベクトルを，点 p における M の接ベクトル (tangent vector) という．

定義 2.12 (速度ベクトル). $c : (-\epsilon, \epsilon) \to M$ を C^r 級曲線とし，$c(0) = p$ であるとする．

$$\left.\frac{dc}{dt}\right|_{t=0}(f) = \left.\frac{df(c(t))}{dt}\right|_{t=0}$$

で定義される写像 $\left.\frac{dc}{dt}\right|_{t=0} : C^r(M) \to \mathbb{R}$ を，曲線 c の $t = 0$ における速度ベクトルという．

命題 2.13. $\left.\frac{dc}{dt}\right|_{t=0}$ は T_pM の元である．すなわち，速度ベクトルは接ベクトルである．

次に，多様体 M から N への C^r 級写像 $f : M \to N$ の微分について説明する．

定義 2.14 (ヤコビ行列). 写像 $f : M \to N$ を C^r 級写像とする．点 p を含む M の座標近傍 $(U\,;\, x_1, \cdots, x_m)$ と，点 $q = f(p)$ を含む N の座標近傍 $(V\,;\, y_1, \cdots, y_n)$ を取る．$f(U) \subset V$ となるように U を十分小さく取るものとし，f を $(U\,;\, x_1, \cdots, x_m)$ と $(V\,;\, y_1, \cdots, y_n)$ について局所座標表示したものが，

$$\begin{aligned} y_1 &= f_1(x_1, \cdots, x_m) \\ &\vdots \\ y_n &= f_n(x_1, \cdots, x_m) \end{aligned}$$

と表されるとする．このとき，

$$(Jf)_p = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1}(p) & \dfrac{\partial f_1}{\partial x_2}(p) & \cdots & \dfrac{\partial f_1}{\partial x_m}(p) \\ \dfrac{\partial f_2}{\partial x_1}(p) & \dfrac{\partial f_2}{\partial x_2}(p) & \cdots & \dfrac{\partial f_2}{\partial x_m}(p) \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial f_n}{\partial x_1}(p) & \dfrac{\partial f_n}{\partial x_2}(p) & \cdots & \dfrac{\partial f_n}{\partial x_m}(p) \end{bmatrix}$$

で定義される n 行 m 列行列 $(Jf)_p$ をヤコビ行列 (Jacobian matrix) と呼ぶ．

命題 2.15. T_pM の任意の元 v に対し，点 p を通り，$\left.\frac{dc}{dt}\right|_{t=0} = v$ となるような C^r 級曲線

$$c : (-\epsilon, \epsilon) \to M, \qquad (c(0) = p)$$

が存在する．

命題 2.16. 点 p を通る C^r 級曲線 $c : (-\epsilon, \epsilon)$，$(c(0) = p)$ を考える．$\left.\frac{d(f\circ c)}{dt}\right|_{t=0}$ は c の $t = 0$ における局所的な振る舞いによって定まり，$\left.\frac{dc}{dt}\right|_{t=0}$ のみに依存する．

定義 2.17 (写像の微分). $f : M \to N$ を C^r 級写像とする．T_pM の任意の元 v に対し，$\left.\frac{dc}{dt}\right|_{t=0} = v$ となるような点 p を通る C^r 級曲線 $c : (-\epsilon, \epsilon)$ を取り，N 上の点 $q = f(p)$ における接ベクトル

$$\left.\frac{d(f \circ c(t))}{dt}\right|_{t=0} \in T_qN$$

を対応付ける写像を考える．この写像を

$$(df)_p : T_pM \to T_qN$$

と書き，点 p における $f : M \to N$ の微分 (differential) と呼ぶ．

命題 2.18. $(df)_p : T_pM \ni v \mapsto w \in T_qN$ は線型写像である．p，q の周りにそれぞれ座標近傍 $(U\ ;\ x_1, \cdots, x_m)$ と $(V\ ;\ y_1, \cdots, y_n)$ を取ると，任意の $v \in T_pM$，$w \in T_qN$ が，

$$v = \sum_{i=1}^{m} v_i \left(\frac{\partial}{\partial x_i}\right)_p$$
$$w = \sum_{j=1}^{n} w_i \left(\frac{\partial}{\partial y_i}\right)_q$$

と書ける．これにより，v と w がそれぞれベクトル $(v_1, \cdots, v_m)$ と $(w_1, \cdots, w_n)$ に対応付けられる．このとき，v と w は

$$\begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix} = (Jf)_p \begin{bmatrix} v_1 \\ \vdots \\ v_m \end{bmatrix}$$

の関係を持つ．

定義 2.19 (1 次形式). V を $\mathbb{R}$ 上の m 次元ベクトル空間とする．V 上の 1 次形式 (one-form) とは，V から $\mathbb{R}$ への写像

$$\omega : V \to \mathbb{R}$$

であって，任意のベクトル $X, Y \in V$ と任意の $a, b \in \mathbb{R}$ について，線型性

$$\omega(aX + bY) = a\omega(X) + b\omega(Y)$$

が成り立つようなものをいう．

V 上の 1 次形式全体のなす集合を V^* と書くとする．V^* は再び $\mathbb{R}$ 上の m 次元ベクトル空間になる．

定義 2.20 (双対ベクトル空間). V^* を，V の双対ベクトル空間 (dual vector space) という．

定義 2.21 (余接ベクトル空間). T_pM の双対ベクトル空間 T_pM^* のことを，多様体 M の点 p における余接ベクトル空間 (cotangent vector space) と呼び，T_p^*M と表記する．

定義 2.22 (k 次微分形式). M の各点 p に，p における T_pM 上の 1 次形式 $\omega_p \in T_p^*M$ を一つずつ対応させる対応 $\omega = \{\omega_p\}_{p \in M}$ を，M 上の 1 次微分形式 (differential 1-form) という．

定義 2.23 (k 次形式). V を $\mathbb{R}$ 上の m 次元ベクトル空間とする．V 上の k 次形式 (k-form) とは，V の k 個の直積から $\mathbb{R}$ への写像

$$\omega : V \times \cdots \times V \to \mathbb{R}$$

であって，$\omega(X_1, \cdots, X_k)$ が各 X_i について線型であるようなものを言う．

V 上の k 次形式全体のなす集合を $\bigotimes^k V^*$ と書くとする．V^* は再び $\mathbb{R}$ 上の m 次元ベクトル空間になる．

定義 2.24 (k 次テンソル場). M の各点 p に，p における T_pM 上の k 次形式 $\omega_p \in \bigotimes^k T_p^*M$ を一つずつ対応させる対応 $\omega = \{\omega_p\}_{p \in M}$ を，M 上の k 次テンソル場 (k-tensor field) という．

定義 2.25 (対称 k 次形式). V を $\mathbb{R}$ 上の m 次元ベクトル空間とする．V 上の k 次形式 ω が対称 k 次形式 (symmetric k-form) であるとは，$X_1, \cdots, X_k$ に置換を施しても $\omega(X_1, \cdots, X_k)$ の値が変わらないことである．

定義 2.26 (k 次対称テンソル場). C^∞ 級多様体 M 上のテンソル場 $\omega = \{\omega_p\}_{p \in M}$ が k 次対称テンソル場 (symmetric tensor field) であるとは，M の各点 p において，ω_p が T_pM 上の対称 k 次形式になっていることである．

2 次の対称テンソル場を用いて，多様体 M 上の点 p における接ベクトルの内積 $\omega : T_pM \times T_pM \to \mathbb{R}$ を定めることができる．

定義 2.27 (リーマン計量). C^∞ 級多様体 M 上の 2 次の対称テンソル場 ω が，M の各点 p において正定値，すなわち T_pM の任意の 0 でない接ベクトル v について，$\omega(v, v) > 0$ が成り立つとき，ω を M 上のリーマン計量 (Riemannian metric) という．

定義 2.28 (リーマン多様体). リーマン計量 g が一つ与えられた多様体 (M, g) のことを，リーマン多様体 (Riemannian manifold) という．

2.1.2　多様体上で定義される演算

本小節では，多様体上で数値計算を行う上で重要となる 2 つの演算について説明する．本小節は主に [9, 11–14] に基づいている．

1 つ目はレトラクションと呼ばれる演算である．この演算により，多様体上の点と接ベクトルを対応付けることができる．

定義 2.29 (レトラクション). M を m 次元多様体とする．写像 $R_p : T_pM \to M$ が以下の条件を満たすとき，R をレトラクション (retraction) という．

1. $R_p(0_p) = p$. ただし，0_p は T_pM の零元とする．
2. $T_{0_p}T_pM \simeq T_pM$ とみなしたとき，

$$(dR_p)_{0_p} = \mathrm{id}_{T_pM},$$

となる．ただし，id_{T_pM} は T_pM 上の恒等写像とする．

2 つ目はベクトル輸送と呼ばれる演算である．この演算により，異なる 2 地点の接ベクトル同士を対応付けることができる．

定義 2.30 (ベクトル輸送). M を m 次元多様体とし，$X \in T_pM$ とする．写像 $\mathcal{T}_X : T_pM \to T_{R_p(X)}M$ が以下の条件を満たすとき，$\mathcal{T}_X$ をベクトル輸送 (vector transport) という．

1. 任意の $Y \in T_pM$ について，$\mathcal{T}_{0_p}(Y) = Y$ が成り立つ．
2. 任意の $a, b \in \mathbb{R}$，$X, Y, Z \in T_pM$ について，$\mathcal{T}_X(aY + bZ) = a\mathcal{T}_X(Y) + b\mathcal{T}_X(Z)$ が成り立つ．

2.1.3　多様体上の確率測度

本小節では，確率分布を扱う上で必要となる多様体上で定義される測度について説明する．

M を $\mathbb{R}^n$ 中に埋め込まれた m 次元多様体とする．M 上の基準測度として，以下のハウスドルフ測度を定義できる [19–21].

定義 2.31 (ハウスドルフ測度). $A \subset \mathbb{R}^n$ を $\mathbb{R}^n$ 中の m 次元部分空間とする．このとき，

$$
\begin{aligned}
H^m(A) &= \lim_{\delta \to 0} \inf_{\substack{A \subset \bigcup_i S_i \\ \mathrm{diam}(S_i) < \delta}} \sum_i \alpha_m \left(\frac{\mathrm{diam}(S_i)}{2} \right)^m \\
where \quad & \mathrm{diam}(S_i) = \sup\{|x - y| \mid x, y \in S_i\} \\
& \alpha_m = \frac{\Gamma\left(\frac{1}{2}\right)^m}{\Gamma\left(\frac{m}{2} + 1\right)},
\end{aligned}
$$

で定義される $H^m(A)$ を m 次元ハウスドルフ測度 (Hausdorff measure) と呼ぶ．ただし，下限は可算被覆 $\{S_i\}_{i \in \mathbb{N}}$ の取り方を様々に変えて取るとする．

ハウスドルフ測度は，$\mathbb{R}^m$ に埋め込まれた m 次元多様体を半径 δ 以下の m 次元球で覆った際の m 次元球の体積の総和について，δ の無限小の極限を取ったものである．M から $\mathbb{R}$ への写像 g を M 上の m 次元ハウスドルフ測度 H^m を基準とする確率密度関数とする．また，f をユークリッド空間の開部分空間 $\mathcal{D} \subset \mathbb{R}^m$ から M への写像と

し，f による $\mathcal{D}$ の像を $\mathcal{I} = f(\mathcal{D}) \subseteq M$ とする．f が以下の制約を満たすとする．

1. M はほとんど至るところ f の像 $\mathcal{I}$ に含まれる．すなわち，$H^m(M \backslash \mathcal{I}) = 0$.
2. f は $\mathcal{D}$ 内で単射.
3. f は $\mathcal{D}$ 内で微分可能.

これらが成り立つとき，ハウスドルフ測度の変数変換公式を得ることができる．$x \in \mathcal{D}$ において $(T_f)_x = \left|\det\left((J_f)_x{}^T (J_f)_x\right)\right|^{\frac{1}{2}}$ という量 $(T_f)_x$ を定義する．このとき，$\mathcal{D}$ 上のルベーグ測度 L^m と M 上のハウスドルフ測度 H^m の間に以下の関係が成り立つ [19, 21–23].

定理 2.32. 任意のボレル集合 $A \subset \mathcal{D}$ について，

$$\int_A g(f(x))(T_f)_x L^m(dx) = \int_{f(A)} g(y) H^m(dy) \tag{2.1}$$

が成り立つ.

この定理は，M 上の確率変数 y が確率密度 g を持つとき，$y = f(x)$ として確率変数を x により表示すると，x は確率密度 $g(f(x))(T_f)_x$ を持つ，ということを意味している.

2.1.4　Stiefel 空間

k 個の m 次元正規直交基底の順序付き集合を k-frame (枠) といい，空間の一点が k-frame 一つに対応するような空間を Stiefel 空間という．この節ではまず Stiefel 空間について説明をする．その後，Stiefel 空間上で定義される演算としてレトラクションとベクトル輸送を導入し，Stiefel 空間上の確率分布として一様分布と，代表的な非一様分布である matrix Langevin 分布について説明する.

定義 2.33 (Stiefel 空間). Stiefel 空間 $\mathcal{V}_{m,k}$ は k 個の m 次元正規直交ベクトルの順序付き集合全体からなる空間であり，以下のように定義される.

$$\mathcal{V}_{m,k} = \left\{ X \in \mathbb{R}^{m \times k} \;\middle|\; X^T X = I_k \right\}. \tag{2.2}$$

ただし，$\mathbb{R}^{m\times k}$ は $m\times k$ 実行列全体からなる空間であり，I_k は $k\times k$ の単位行列とする．

$\mathcal{V}_{m,k}$ はコンパクトな $mk-k(k+1)/2$ 次元リーマン多様体であり，mk 次元ユークリッド空間 $\mathbb{R}^{mk}$ の部分多様体である．$k=1$ の場合に $\mathcal{V}_{m,k}$ は $(m-1)$ 超球面 $\mathbb{S}^{m-1}$ となり，$k=m$ の場合に $O(m)$ となる．ただし，$O(m)$ は m 次元直交群であり，$m\times m$ 実直交行列全体からなり，積が行列積として定義されるような群である．つまり，Stiefel 空間 $\mathcal{V}_{m,k}$ は m 次元正規直交縦ベクトルを k 個順に横に並べたものの全体からなる空間として考えることができて，更に特殊な場合として，$k=1$ のときに正規化された m 次元ベクトル全体からなる空間，$k=m$ のときに m 次元直交行列全体からなる空間として考えることができるような空間である．

点 $X\in\mathcal{V}_{m,k}$ における Stiefel 空間の接空間 $T_X\mathcal{V}_{m,k}$ は以下のように表される．

$$T_X\mathcal{V}_{m,k}=\left\{Z\in\mathbb{R}^{m\times k}\;\middle|\;X^TZ+Z^TX=0\right\} \tag{2.3}$$

$$=\left\{X\Omega+X_\perp K\;\middle|\;\Omega\in Skew(k),\,K\in\mathbb{R}^{(m-k)\times k}\right\}, \tag{2.4}$$

ただし，$X_\perp\in\mathbb{R}^{m\times(m-k)}$ を行列 X の列空間の直交補空間の基底を列として持つ行列とし，$Skew(k)$ は $k\times k$ の歪対象行列全体がなす空間とする．特に，上側 $k\times k$ のブロックが単位行列，残りが 0 となっている行列 $I_{m\times k}=\begin{bmatrix}I_k\\0\end{bmatrix}$ を原点としたとき，$I_{m\times k}$ における Stiefel 空間の接空間は，

$$T_{I_{m\times k}}\mathcal{V}_{m,k}=\left\{\begin{bmatrix}A\\B\end{bmatrix}\;\middle|\;A\in Skew(k),B\in\mathbb{R}^{(m-k)\times k}\right\}, \tag{2.5}$$

となる．

次に，Stiefel 空間上の演算について説明する．主に [12–14] に基づく．ケイリーによって提案されたケイリー変換は，歪対象行列 $W\in Skew(m)$ を特殊直交群 $SO(m)=\{S\in O(m)\mid\det S=1\}$ へと移す写像であり，以下のように定義される [24].

定義 2.34 (ケイリー変換). $Skew(m)$ から $SO(m)$ への写像

$$Cay(W)=(I_m-W)^{-1}(I_m+W),$$

をケイリー変換 (Cayley transform) という．

このケイリー変換を用いて Stiefel 空間上のレトラクションを定義できる．まず，任意の $Z \in T_X\mathcal{V}_{m,k}$ に対して，以下が成り立つような W_Z が存在する．

$$Z = W_Z X, \tag{2.6}$$

ただし，

$$W_Z = P_X Z X^T - X Z^T P_X \qquad \text{かつ} \qquad P_X = I_m - \frac{1}{2} X X^T. \tag{2.7}$$

この W_Z を用いて，X を始点とする Stiefel 空間上のレトラクションを以下のように定義することができる．

定義 2.35 (Stiefel 空間上のケイリー型レトラクション). $T_X\mathcal{V}_{m,k}$ から $\mathcal{V}_{m,k}$ への写像 R_X を，

$$R_X(tZ) = \left(I_m - \frac{t}{2} W_Z\right)^{-1} \left(I_m + \frac{t}{2} W_Z\right) X, \tag{2.8}$$

と定義する．

曲線 $R_X(tZ)$ は任意の t において $\mathcal{V}_{m,k}$ に含まれる．また，$R_X(0) = X$，$\left.\frac{dR_X(tZ)}{dt}\right|_{t=0} = W_Z X = Z$ を満たす．
また，ベクトル輸送も以下のように定義することができる．

定義 2.36 (Stiefel 空間上のケイリー型ベクトル輸送). $Y_X \in T_X\mathcal{V}_{m,k}$ から $Y_{R_X(Z)} \in T_{R_X(Z)}\mathcal{V}_{m,k}$ への写像 $\mathcal{T}_Z$ を，

$$\mathcal{T}_Z(Y_X) = \left(I_m - \frac{1}{2} W_Z\right)^{-1} \left(I_m + \frac{1}{2} W_Z\right) Y_X, \tag{2.9}$$

と定義する．

こうして定義した写像 $\mathcal{T}_Z$ は，$X \in \mathcal{V}_{m,k}$ から $R_X(Z) \in \mathcal{V}_{m,k}$ へのベクトル輸送となる．

次に，Stiefel 空間上の確率分布について導入を行う．$X \in \mathcal{V}_{m,k}$ として，$\mathcal{V}_{m,k}$ 上の微分形式 $\left(X^T dX\right) = \bigwedge_{i=1}^{k} \bigwedge_{j=i+1}^{m} x_j^T dx_i$ は Haar 測度を定める．これにより，$\mathcal{V}_{m,k}$

の表面積は以下のように計算できる．

$$\mathrm{vol}(\mathcal{V}_{m,k}) \coloneqq \int_{\mathcal{V}_{m,k}} \left(X^T dX\right) = \frac{2^k (\sqrt{\pi})^{mk}}{\Gamma_k(m/2)}. \tag{2.10}$$

ただし，$\Gamma_m(a)$ は多変量ガンマ関数 (multivariate gamma function) と呼ばれるものであり，以下のように定義される．

$$\begin{aligned}\Gamma_m(a) &:= \int_{S>0} \exp(\mathrm{tr}(-S))|S|^{a-(m+1)/2}(dS) \\ &= \pi^{m(m-1)/4} \prod_{i=1}^{m} \Gamma\left[a - \frac{1}{2}(i-1)\right], \\ &\quad with \qquad a > \frac{1}{2}(m-1).\end{aligned}$$

式 (2.10) より Stiefel 空間 $\mathcal{V}_{m,k}$ 上で一様な基準測度を以下のように定めることができる．

定義 2.37 (Stiefel 空間上の一様分布).

$$[dX] \coloneqq \frac{\left(X^T dX\right)}{\mathrm{vol}(\mathcal{V}_{m,k})}. \tag{2.11}$$

Matrix Langevin 分布は $\mathcal{V}_{m,k}$ 上の確率密分布として広く用いられており，[15] によって導入され，その後，[16, 17] による初期研究により，古典的状況下でのパラメタの最尤推定量について調べられた．また，漸近性などその他の性質について [18] に良くまとめられている．

定義 2.38 (Matrix Langevin 分布). Matrix Langevin 分布は $F \in \mathbb{R}^{m\times k}$ によってパラメタライズされ，基準測度 $[dX]$ に対する確率密度関数 $\mathcal{ML}(X\,;\,F)$ は以下によって与えられる．

$$\mathcal{ML}(X\,;\,F) = \frac{\exp\left(\mathrm{tr}\left(F^T X\right)\right)}{{}_0F_1\left(;\,\frac{1}{2}m\,;\,\frac{1}{4}F^T F\right)}. \tag{2.12}$$

ただし，${}_0F_1\left(;\,\frac{1}{2}m\,;\,\frac{1}{4}F^T F\right)$ は matrix Langevin 分布の正規化定数であり，行列引数超幾何関数と呼ばれる特殊関数である [16, 25].

この分布は多変量正規分布を $X^TX = I_k$ によって条件付けた条件付き確率分布として得られる．また，モーメント $\mathbb{E}[X]$ を定めたときに最大エントロピーを達成する分布として得ることもできる．$F = 0$ のときに一様分布と一致し，$k = 1$ のときに von Mises-Fisher 分布と一致する．

X の期待値はスコア関数の期待値が 0 になることを利用して以下のように求めることができる．

$$\begin{aligned}
&\mathbb{E}\left[\frac{\partial}{\partial F}\log\frac{\exp\bigl(\mathrm{tr}\bigl(F^TX\bigr)\bigr)}{{}_0F_1\bigl(;\,\frac{1}{2}m\,;\,\frac{1}{4}F^TF\bigr)}\right]\\
&=\mathbb{E}\left[\frac{\partial}{\partial F}\mathrm{tr}\bigl(F^TX\bigr)-\frac{\partial}{\partial F}\log{}_0F_1\left(;\,\frac{1}{2}m\,;\,\frac{1}{4}F^TF\right)\right]\\
&=\mathbb{E}\left[\frac{\partial}{\partial F}\mathrm{tr}\bigl(FX^T\bigr)-\frac{\partial}{\partial F}\log{}_0F_1\left(;\,\frac{1}{2}m\,;\,\frac{1}{4}F^TF\right)\right]\\
&=\mathbb{E}[X]-\frac{\partial}{\partial F}\log{}_0F_1\left(;\,\frac{1}{2}m\,;\,\frac{1}{4}F^TF\right)\\
&=0,\\
&\therefore\mathbb{E}[X]=\frac{\partial}{\partial F}\log{}_0F_1\left(;\,\frac{1}{2}m\,;\,\frac{1}{4}F^TF\right).
\end{aligned}\tag{2.13}$$

Matrix Langevin 分布の特性関数は以下の形となる．

$$\begin{aligned}
\Phi_X(T)&=\mathbb{E}\left[\exp\bigl(\mathrm{tr}\bigl(iT^TX\bigr)\bigr)\right]\\
&={}_0F_1\left(;\,\frac{1}{2}m\,;\,(F+iT)\right),\\
&\text{for } T\in\mathbb{R}^{m\times k}.
\end{aligned}\tag{2.14}$$

特性関数によっても X の期待値を得ることができる．

$$\mathbb{E}[X]=FR.\tag{2.15}$$

ただし，$R\in\mathbb{R}^{k\times k}$ であり，R の各成分は以下の形となる．

$$R_{ij}=2\left.\frac{\partial\log{}_0F_1\bigl(;\,\frac{1}{2}m\,;\,\frac{1}{4}G^TG\bigr)}{\partial G_{ij}}\right|_{G=F^TF}.\tag{2.16}$$

2.2　変分推論と変分オートエンコーダ

2.2.1　変分推論

変分推論とは，あるパラメタ z から観測値 x が $p(x \mid z)$ という分布に従って得られるようなモデルの下で，x が観測された条件の下での z の事後分布を近似的に求める手法である．この手法は事後分布 $p(z \mid x)$ を陽に書き表すことが困難な場合に用いられる．同じ事後分布の推論手法であるマルコフ連鎖モンテカルロ法 (Markov Chain Monte Carlo; MCMC) 等のサンプリングベースの手法と比べて，確率変数に適宜独立性を仮定することで計算を高速化できる，などの利点がある．

変分推論は，事後分布 $p(z \mid x)$ の推定を，あるパラメタ ψ で定められる近似事後分布 $q_\psi(z)$ のフィッティングによって行う．学習は以下のように近似事後分布と事後分布の間の KL divergence を最小化することで行う．

$$\min_{\psi} D_{KL}(q_\psi(z) \parallel p(z \mid x)). \tag{2.17}$$

この KL divergence は以下のように変形できる．

$$\begin{aligned}
&D_{KL}(q_\psi(z) \parallel p(z \mid x)) \\
&= \int q_\psi(z) \log \frac{q_\psi(z)}{p(z \mid x)} dz \\
&= \int q_\psi(z) \left(\log \frac{1}{p(x \mid z)} + \log \frac{q_\psi(z)}{p(z)} + \log p(x) \right) dz \\
&= -\mathbb{E}_{q_\psi(z)}[\log p(x \mid z)] + D_{KL}(q_\psi(z) \parallel p(z)) + \underbrace{\log p(x)}_{const.}.
\end{aligned} \tag{2.18}$$

よって，式 (2.18) 中に含まれる事後分布 $p(z \mid x)$ の正規化定数由来の項 $\log p(x)$ がパラメタ ψ に対して定数となるため，式 (2.17) と ELBO (式 (2.21)) の最大化が等価となる．

$$\begin{aligned}
&\min_{\psi} D_{KL}(q_\psi(z) \parallel p(z \mid x)) \\
\Longleftrightarrow &\max_{\psi} \mathbb{E}_{q_\psi(z)}[\log p(x \mid z)] - D_{KL}(q_\psi(z) \parallel p(z)).
\end{aligned} \tag{2.19}$$

2.2.2　変分オートエンコーダ

VAE では，生成モデルとして潜在状態からデータが生成されているようなモデルを考える．潜在状態を Z として，n 個の潜在状態 $\boldsymbol{Z}$ の分布を $P(\boldsymbol{Z})$ とする．また，n 個の潜在状態 $\boldsymbol{Z}$ それぞれから n 個の観測データ $\boldsymbol{x}$ が生成される確率を $p_\phi(\boldsymbol{x} \mid \boldsymbol{Z})$ として条件付確率で表す．VAE ではこの条件付確率は $\boldsymbol{Z}$ を入力とし，$\boldsymbol{x}$ を出力とする，パラメータ ϕ を持つニューラルネットとしてあらわされる．目的関数はデータ $\boldsymbol{x}$ が生成される対数尤度 (evidence 関数) $\log \int p_\phi(\boldsymbol{x} \mid \boldsymbol{Z}) p(\boldsymbol{Z}) d\boldsymbol{Z}$ であり，この関数を最大化するようにニューラルネットのパラメータ ϕ を学習する．しかし，$\boldsymbol{Z}$ について積分をして直接この関数を計算することは不可能であることが多く，代わりに以下のような evidence 関数の下限 (Evidence Lower BOund; ELBO) の最大化を行うことで，目的関数の最大化を行う．

$$\begin{aligned} &\log \int p_\phi(\boldsymbol{x} \mid \boldsymbol{Z}) p(\boldsymbol{Z}) d\boldsymbol{Z} \\ &\geq \mathbb{E}_{q(\boldsymbol{Z})}[\log p_\phi(\boldsymbol{x} \mid \boldsymbol{Z})] - D_{KL}(q(\boldsymbol{Z}) \parallel p(\boldsymbol{Z})). \end{aligned} \tag{2.20}$$

ここで，$q(Z)$ は潜在状態 Z の事後分布 $p_\phi(Z \mid x)$ の近似事後分布であり，実際に $q(Z) = p_\phi(Z \mid x)$ となるときに上記の不等式の等号が成立し，evidence 関数が最大化される．しかし，通常は正確に $q(Z) = p_\phi(Z \mid x)$ を求めることはせず，現実的な計算量で最適化できるように $q(Z)$ の関数クラスを制限し，$q_\psi(Z \mid x\,;\, \theta)$ というようにパラメータ ψ を持つニューラルネットで $q(Z)$ のパラメタ θ を出力することによって近似的に $q(Z)$ の推論をできるようにモデルを組み，そのようなアーキテクチャの下で ELBO の最大化を行う．結局，最終的な目的関数は以下のようになる．

$$\begin{aligned} \mathcal{L}(\phi, \psi) = &\underbrace{\mathbb{E}_{q_\psi(\boldsymbol{Z} \mid \boldsymbol{x}; \theta)}[\log p_\phi(\boldsymbol{x} \mid \boldsymbol{Z})]}_{\text{reconstruction error}} \\ &- \underbrace{D_{KL}(q_\psi(\boldsymbol{Z} \mid \boldsymbol{x}\,;\, \theta) \parallel p(\boldsymbol{Z}))}_{\text{KL divergence}}. \end{aligned} \tag{2.21}$$

この目的関数を最大化するようにエンコーダネットワークのパラメタ ψ とデコーダネットワークのパラメタ ϕ の最適化を行う．この目的関数は，推定事後分布 $q_\psi(Z \mid x\,;\, \theta)$ が事前分布 $p(Z)$ から離れ過ぎないように KL divergence によって正則化を行いながら，データ x を生成する尤度を reconstruction error の項によって最大化しようとし

ている式として理解できる．この目的関数によって VAE の学習を行い，生成モデル $p_\phi(x \mid Z)p(Z)$ を得ることができる．事前分布 $p(Z)$ や近似事後分布 $q_\psi(Z \mid x\,;\,\theta)$ の関数クラスには正規分布が広く用いられているが，この事前分布分布をデータに適した分布にすることにより，VAE の性能が向上することが報告されている．この事前分布と近似事後分布に Stiefel 空間上の確率分布を適用することを試みる．

第 3 章

関連研究

3.1 様々な空間や分布を用いた変分推論と変分オートエンコーダ

VAE の改善手法の多くは，取り扱うデータに適した事前確率分布と事後確率分布を潜在表現の分布として課すことによって行われる．代表的な手法としては normalizing flow が挙げられる [26]．この手法では，事後確率分布に変形を施すことで，学習可能な確率分布の表現力を高め，より複雑な事後分布を扱うことを可能にしている．しかしながら，この手法は依然としてユークリッド空間上のガウス分布を仮定している．

非ユークリッド空間上で確率モデルの学習を行う手法は，明示的に幾何学的構造を仮定するかどうかで大きく二分される．幾何学的構造を明示しない場合については，RSVGD と呼ばれる手法によって一般のリーマン多様体上で変分推論を行う手法が提案されている [27]．この手法では多数のパーティクルによって事後分布を近似するため，ノンパラメトリックに確率分布の近似を行うことができ，表現力が高い一方で，高次元での計算量は急峻に増大してしまう．

幾何学的構造を明示する場合については，超球面や双曲空間などの空間を標本空間に持つような確率モデルの学習手法が研究されている [6, 28]．それぞれ，規格化されたベクトルとして表現されるようなデータや階層構造を持つデータに対して，リンク予測などのタスクで低次元でも良い性能が達成されることが報告されている．

3.2 Stiefel 空間上での統計・機械学習モデル

Stiefel 空間と呼ばれる，空間の一点が k 個の正規直交基底の順序付き集合に対応する空間上で機械学習手法を考えることで，正規直交性を持つデータに対しては良い性

能が得られることが知られている．例えば，[8] では状態空間モデルの観測方程式中の行列に正規直交性を課し，Stiefel 空間上の確率分布を用いて行動認識タスクを解くことで，低次元でも良い性能を達成できることを示している．

以上の研究ではいくつかのタスクについて，変分推論及び VAE に対する非ユークリッド空間上の確率分布を考慮することの有効性と，Stiefel 空間上の機械学習の有効性とが示されている．これらの研究結果から，変分推論及び VAE の潜在空間を Stiefel 空間とすることで，正規直交性を持つようなデータに対して低次元な潜在空間でも高い性能を持つ確率的生成モデルを学習できることが予想されるが，知りうる限りそのような研究は未だ存在しない．

また，本研究では Stiefel 空間上で変分推論及び VAE の学習を安定して高速に行えるように，Stiefel 空間上の巻き込み型正規分布を考案した．しかし，このような分布の構成は完全に新しい訳ではなく，例えば [23, 29, 30] などの先行研究が存在する．しかし，[23, 30] では変分推論を行う上で必要である分布の具体的な正規化定数について考慮していない．また，[23] ではレトラクションによって分布を構成しているが，レトラクションの始点を原点に固定しているため，Stiefel 空間上の対蹠点の近傍を中心に持つ分布を考えた際に分布に歪みが生じてしまう．[29] では正規化定数について考慮しているものの，正規化定数を多数サンプルによるモンテカルロ近似によって求めているため，計算量が大きくなってしまう．更に，[29] では接空間上のルベーグ測度を基準とした確率密度関数を扱っているため，Stiefel 空間上の一様分布との KL ダイバージェンスを計算することが困難である．一方で，本研究において考案した分布は，正規化定数を考慮しており，ベクトル輸送によって原点における接ベクトルを Stiefel 空間上の任意の点の接ベクトルへと変換することにより，レトラクションの始点を Stiefel 空間上の任意の点に取ることが出来る．更に，確率密度関数は Stiefel 空間上の一様測度を基準としているため，一様分布との KL ダイバージェンスを計算することが容易である．

第 4 章

Matrix Langevin 分布を用いた変分推論手法

2.1.4 節で導入した Stiefel 空間上の一様分布や matrix Langevin 分布を用いることにより，潜在表現に正規直交性を明示的に課した VAE の学習手法を構成できる．まず 4.1 節では目的関数の具体的な表式を求める．次に 4.2 節では目的関数の最適化に必要となる勾配計算について議論する．最後に 4.2.1 節では勾配計算でネックとなる行列引数超幾何関数 ${}_0F_1$ の近似計算について説明する．

4.1 ELBO の導出

VAE の目的関数式 (2.21) について，事前分布 $p(Z)$ に $\mathcal{V}_{m,k}$ 上の無情報分布として一様分布 U を用い，近似事後分布 $q_\psi(Z \mid x\,;\,\theta)$ に matrix Langevin 分布を用いることで，Stiefel 空間上の VAE を学習できることが期待される．

目的関数式 (2.21) の第一項の reconstruction error については以下のように書くことができる．

$$\begin{aligned}&\mathbb{E}_{q_\psi(Z|x;\theta)}[\log p_\phi(x \mid Z)]\\ &= \mathbb{E}_{\mathcal{ML}(Z;F_\psi(x))}[\log p_\phi(x \mid Z)].\end{aligned} \tag{4.1}$$

ただし，$F_\psi(x)$ はパラメータ ψ を持つエンコーダニューラルネットに対し，あるデータ x を入力したときの出力とする．また，$p_\phi(x \mid Z)$ はパラメータ ϕ を持つデコーダニューラルネットに対し，matrix Langevin 分布からサンプリングした潜在変数 Z を入力して，エンコーダへの入力 x がどの程度復元されそうかを表す尤度である．

次に，目的関数式 (2.21) 中の第二項の KL divergence を求める．パラメータ F，G

を持つ 2 つの matrix Langevin 分布 f，g 間の KL divergence は以下の形となる．

$$\begin{aligned}
&D_{KL}(\mathcal{ML}(Z\,;\,F)\,\|\,g_{\mathcal{ML}}(Z\,;\,G)) \\
&= \mathrm{tr}\left((F-G)^T\frac{\partial}{\partial F}\log {}_0F_1\left(;\,\frac{1}{2}m\,;\,\frac{1}{4}F^TF\right)\right) \\
&\quad + \log\frac{{}_0F_1\left(;\,\frac{1}{2}m\,;\,\frac{1}{4}G^TG\right)}{{}_0F_1\left(;\,\frac{1}{2}m\,;\,\frac{1}{4}F^TF\right)}.
\end{aligned}$$

これは以下のように導出できる．

$$\begin{aligned}
&D_{KL}(\mathcal{ML}(Z\,;\,F)\,\|\,g_{\mathcal{ML}}(Z\,;\,G)) \\
&= \int_{\mathcal{V}_{m,k}} f(Z\,;\,F)\log\frac{f(Z\,;\,F)}{g(Z\,;\,G)}[dZ] \\
&= \int_{\mathcal{V}_{m,k}} f(Z\,;\,F)\big(\mathrm{tr}\big(F^TZ\big)-\mathrm{tr}\big(G^TZ\big)\big)[dZ] \\
&\quad + \int_{\mathcal{V}_{m,k}} f(Z\,;\,F)\log\frac{{}_0F_1\left(;\,\frac{1}{2}m\,;\,\frac{1}{4}G^TG\right)}{{}_0F_1\left(;\,\frac{1}{2}m\,;\,\frac{1}{4}F^TF\right)}[dZ] \\
&= \mathbb{E}_{f(Z;F)}\Big[\mathrm{tr}\Big((F-G)^TZ\Big)\Big] + \log\frac{{}_0F_1\left(;\,\frac{1}{2}m\,;\,\frac{1}{4}G^TG\right)}{{}_0F_1\left(;\,\frac{1}{2}m\,;\,\frac{1}{4}F^TF\right)} \\
&= \mathrm{tr}\Big((F-G)^T\mathbb{E}_{f(Z;F)}[Z]\Big) + \log\frac{{}_0F_1\left(;\,\frac{1}{2}m\,;\,\frac{1}{4}G^TG\right)}{{}_0F_1\left(;\,\frac{1}{2}m\,;\,\frac{1}{4}F^TF\right)} \\
&= \mathrm{tr}\left((F-G)^T\frac{\partial}{\partial F}\log {}_0F_1\left(;\,\frac{1}{2}m\,;\,\frac{1}{4}F^TF\right)\right) \\
&\quad + \log\frac{{}_0F_1\left(;\,\frac{1}{2}m\,;\,\frac{1}{4}G^TG\right)}{{}_0F_1\left(;\,\frac{1}{2}m\,;\,\frac{1}{4}F^TF\right)}.
\end{aligned}$$

ここで，$G=0$ とすれば分布 g が一様分布となるため，matrix Langevin 分布 $\mathcal{ML}$ と一様分布 U の間の KL divergence を求めることができる．これにより，式 (2.21) の第二項の KL divergence の項は以下の形となる．

$$\begin{aligned}
&D_{KL}(\mathcal{ML}(Z\,;\,F)\,\|\,U(Z)) \\
&= \mathrm{tr}\left(F^T\frac{\partial}{\partial F}\log {}_0F_1\left(;\,\frac{1}{2}m\,;\,\frac{1}{4}F^TF\right)\right) \\
&\quad - \log {}_0F_1\left(;\,\frac{1}{2}m\,;\,\frac{1}{4}F^TF\right).
\end{aligned} \tag{4.2}$$

4.2　ELBO の勾配計算

目的関数の最適化には勾配降下法を用いる．そのため，式 (4.1)，式 (4.2) の勾配を求める必要がある．ここで，matrix Langevin 分布からのサンプリング手法として，提案分布を Stiefel 空間上の一様分布 U とした棄却サンプリング法を用いるとする．このとき，式 (4.1) の勾配推定には VAE で広く用いられる期待値の勾配推定法である reparameterization trick [1] を用いることができない．そのため，代替として score function estimator と呼ばれる推定量を用いて期待値の勾配推定を行う．score function estimator による reconstruction error の勾配推定は以下のようになる．

$$\begin{aligned}
&\frac{\partial}{\partial F}\mathbb{E}_{\mathcal{ML}(Z;F_\psi(x))}[\log p_\phi(x \mid Z)] \\
&= \mathbb{E}_{\mathcal{ML}(Z;F_\psi(x))}\left[\log p_\phi(x \mid Z) \cdot \frac{\partial}{\partial F}\log \mathcal{ML}(Z\,;\,F_\psi(x))\right].
\end{aligned} \tag{4.3}$$

これは以下のように導出できる．

$$\begin{aligned}
&\frac{\partial}{\partial F}\mathbb{E}_{\mathcal{ML}(Z;F_\psi(x))}[\log p_\phi(x \mid Z)] \\
&= \frac{\partial}{\partial F}\int \mathcal{ML}(Z\,;\,F_\psi(x))\log p_\phi(x \mid Z)[dZ] \\
&= \int \frac{\partial}{\partial F}\mathcal{ML}(Z\,;\,F_\psi(x))\log p_\phi(x \mid Z)[dZ] \\
&= \int \mathcal{ML}(Z\,;\,F_\psi(x))\cdot\log p_\phi(x \mid Z)\cdot \\
&\qquad \cdot\frac{\partial}{\partial F}\log \mathcal{ML}(Z\,;\,F_\psi(x))[dZ] \\
&= \mathbb{E}_{\mathcal{ML}(Z;F_\psi(x))}\left[\log p_\phi(x \mid Z) \cdot \frac{\partial}{\partial F}\log \mathcal{ML}(Z\,;\,F_\psi(x))\right].
\end{aligned}$$

式 (4.2) の勾配の計算については，$\frac{\partial}{\partial F}\log {}_0F_1$ や $\frac{\partial^2}{\partial F^2}\log {}_0F_1$ の項の計算を除けば通常の自動微分によって計算することができる．

4.2.1　${}_0F_1$ の偏微分計算

式 (4.2) の勾配や式 (4.3) を求める際に，$\frac{\partial}{\partial F}\log {}_0F_1$ や $\frac{\partial^2}{\partial F^2}\log {}_0F_1$ の項の計算が問題となる．この項に含まれる ${}_0F_1$ は特殊関数となっており，値を陽に求めることがで

きない．そのため [31, 32] による鞍点近似によって計算する．

第 5 章

Stiefel 空間上の巻き込み型正規分布を用いた変分推論手法

第 4 章で提案した，matrix Langevin 分布を用いた場合の ELBO の勾配の計算には，reparameterization trick を適用できず，score function estimator と呼ばれる推定量を用いた．しかし，これは多くの場合に分散が大きくなってしまうことが知られている．そこで，本章では reparameterization trick を適用可能な Stiefel 空間上の確率正規分布を提案し，それを用いた ELBO とその勾配計算について説明する．

まず，5.1 節では Stiefel 空間上の巻き込み型正規分布と呼ぶ確率分布を提案する．次に，5.2 節でその確率分布を用いた際の目的関数 ELBO について説明する．最後に 5.3 節では reparameterization trick を用いた ELBO の勾配推定について説明する．

5.1 Stiefel 空間上の巻き込み型正規分布

5.1.1 分布の構成

分布の中心点 $M \in \mathcal{V}_{m,k}$ と分散ベクトル $\sigma^2 \in \mathbb{R}^{\dim \mathcal{V}_{m,k}}$ をパラメタとして持つ，Stiefel 空間上の巻き込み型正規分布 $\mathcal{SWN}\left(Z\,;\,M,\sigma^2\right)$ の構成について説明する．まず，$\mathcal{V}_{m,k}$ について，$k < m$ の場合を考える．サンプルの生成過程は以下のようになる．

1. 原点 $I_{m\times k}$ における接ベクトル $V_{I_{m\times k}} \in T_{I_{m\times k}}\mathcal{V}_{m,k}$ のサンプリング

まず，原点 $I_{I_{m\times k}}$ における接ベクトル $V_{I_{m\times k}} \in T_{I_{m\times k}}\mathcal{V}_{m,k}$ をサンプルするために，$V_{I_{m\times k}}$ を一意に表すベクトル $v \in \mathbb{R}^{\dim \mathcal{V}_{m,k}}$ について考える．原点におけ

る接ベクトルは式 (2.5) より，

$$V_{I_{m\times k}} = \begin{bmatrix} A \\ B \end{bmatrix} \quad where \quad A \in Skew(k),\, B \in \mathbb{R}^{(m-k)\times k},$$

として，$k \times k$ の歪対称行列 A と $(m-k) \times k$ の実行列 B により表される．よって，A の対角線を含まない下三角成分と B の成分を並べたベクトル $v = \left[\mathrm{tril}(A)^T, \mathrm{vec}(B)^T\right]^T \in \mathbb{R}^{\dim \mathcal{V}_{m,k}=mk-k(k+1)/2}$ により，$V_{I_{m\times k}}$ を一意に指定することができる．ただし，C を $m \times n$ 行列としたとき，$\mathrm{tril}(C)$ を，C の対角線を含まない下三角成分の各列を 1 列に並べる操作とし，$\mathrm{tril}(C) = [c_{2,1}, \cdots, c_{m,1}, c_{3,2}, \cdots, c_{m,2}, \cdots, c_{m,n-1}]^T \in \mathbb{R}^{mn-n(n+1)/2}$ とする．また，$\mathrm{vec}(C)$ を，C の各列を 1 列に並べる操作とし，$\mathrm{vec}(C) = [c_{1,1} \cdots, c_{m,1}, \cdots, c_{1,n}, \cdots, c_{m,n}]^T \in \mathbb{R}^{mn}$ とする．

任意の $\dim \mathcal{V}_{m,k}$-次元実ベクトル v が与えられたとき，それに対応する $V_{I_{m\times k}}$ を求める操作を，具体的な行列演算の形で表現することができる [23].

まず，$\Theta_1 = \begin{bmatrix} I_k \\ 0 \end{bmatrix} \in \mathbb{R}^{m\times k}$，$\Theta_2 = \begin{bmatrix} 0 \\ I_{m-k} \end{bmatrix} \in \mathbb{R}^{m\times m-k}$ とし，更に，特殊な行列 $\tilde{D}_m$ [33] を導入する．$\tilde{D}_m$ は与えられた $m(m-1)/2$ 次元ベクトルから，その要素を下三角成分に持つ歪対称行列を復元する $m^2 \times m(m-1)/2$ 行列である．C を $m \times m$ 歪対象行列としたとき，$\tilde{D}_m \,\mathrm{tril}\, C = \mathrm{vec}\, C$ となる．$E_{i,j}$ を (i,j)-成分が 1 でそれ以外の成分が 0 の $m \times m$ 行列とし，$\tilde{u}_{i,j}$ を $(j-1)m+i-j(j+1)/2$ 番目の要素が 1 でそれ以外の要素が 0 の $m(m-1)/2$-次元ベクトルとしたとき，$\tilde{D}_m$ は具体的に，

$$\tilde{D}_m = \sum_{i>j} (\mathrm{vec}(E_{i,j} - E_{j,i}))\tilde{u}_{i,j}^T, \tag{5.1}$$

と表される．

この Θ_1，Θ_2，$\tilde{D}_m$ を用いて，v に対応する $V_{I_{m\times k}}$ を，

$$\mathrm{vec}\, V_{I_{m\times k}} = \Xi v \tag{5.2}$$

$$where \quad \Xi = \left[(I_k \otimes \Theta_1)\tilde{D}_k \quad I_k \otimes \Theta_2\right], \tag{5.3}$$

として具体的に行列演算の形で求めることができる．

よって，原点における接ベクトル $V_{I_{m\times k}}$ をサンプリングするには，まずベクトル v を

$$v \sim \mathcal{N}\big(v\,;\,0,\sigma^2\big), \tag{5.4}$$

としてサンプリングし，次に式 (5.2) により，v を $V_{I_{m\times k}}$ に変換すればよい．

2. M **における接ベクトル** $V_M \in T_M\mathcal{V}_{m,k}$ **への** $V_{I_{m\times k}}$ **の変換**

次に，原点 $I_{m\times k}$ における接ベクトル $V_{I_{m\times k}}$ を，分布の中心 M における接ベクトル V_M に変換する．これを行うために，まず原点 $I_{m\times k}$ から見た M への方向 $X_{I_{m\times k}} \in T_{I_{m\times k}}\mathcal{V}_{m,k}$ を，式 (2.8) のレトラクションの逆写像 $X_{I_{m\times k}} = R^{-1}_{I_{m\times k}}(M)$ として求める．$X_{I_{m\times k}}$ と M を，

$$\begin{aligned} X_{I_{m\times k}} &= \begin{bmatrix} X_u \\ X_l \end{bmatrix} \quad where \quad X_u \in \mathbb{R}^{k\times k},\, X_l \in \mathbb{R}^{(m-k)\times k} \\ M &= \begin{bmatrix} M_u \\ M_l \end{bmatrix} \quad where \quad M_u \in \mathbb{R}^{k\times k},\, M_l \in \mathbb{R}^{(m-k)\times k}, \end{aligned}$$

というように，ブロック行列に分解する．このとき，X_u，X_l を，M_u，M_l から以下のように求めることができる [34].

$$\begin{aligned} F &= (I_k - M_u)(I_k + M_u)^{-1} \\ X_u &= F^T - F \\ X_l &= M_l(I_k + F). \end{aligned} \tag{5.5}$$

こうして得られた $X_{I_{m\times k}}$ を用いて，式 (2.9) で定義したベクトル輸送により，以下のように $V_{I_{m\times k}}$ を V_M に変換することができる．

$$\begin{aligned} V_M &= \mathcal{T}_{X_{I_{m\times k}}}\big(V_{I_{m\times k}}\big) \\ &= \left(I_m - \frac{1}{2}W_{X_{I_{m\times k}}}\right)^{-1}\left(I_m + \frac{1}{2}W_{X_{I_{m\times k}}}\right)V_{I_{m\times k}}. \end{aligned} \tag{5.6}$$

3. $\mathcal{V}_{m,k}$ **上のサンプル** Z **への** V_M **の変換**

最後に，式 (2.8) のレトラクションにより，M を始点として V_M の方向に Stiefel 空間に沿って進んだ点 Z へと V_M を変換することにより，M を中心とする分布に従う $\mathcal{V}_{m,k}$ 上のサンプル Z を得ることができる．

$$\begin{aligned} Z &= R_M(V_M) \\ &= \left(I_m - \frac{1}{2}W_{V_M}\right)^{-1}\left(I_m + \frac{1}{2}W_{V_M}\right)M. \end{aligned} \tag{5.7}$$

アルゴリズム **1** $\mathcal{SWN}\left(Z\,;\,M,\sigma^2\right)$ からのサンプリング

Input: パラメタ $M \in \mathcal{V}_{m,k}$，$\sigma^2 \in \mathbb{R}^{\dim \mathcal{V}_{m,k}}$

Output: サンプル $Z \in \mathcal{V}_{m,k}$

$v \sim \mathcal{N}\left(v\,;\,0,\sigma^2\right)$ をサンプリングする

式 (5.2) により，v を $V_{I_{m\times k}} \in T_{I_{m\times k}}\mathcal{V}_{m,k}$ に変換する

式 (5.5) により，$X_{I_{m\times k}} \in \mathcal{V}_{m,k}$ を計算する

式 (5.6) により，$V_{I_{m\times k}}$ を $V_M = \mathcal{T}_{X_{I_{m\times k}}}\left(V_{I_{m\times k}}\right) \in T_M\mathcal{V}_{m,k}$ に輸送する

式 (5.7) により，V_M を $Z = R_M(V_M) \in \mathcal{V}_{m,k}$ に変換する

これらの手続きによって構成される確率分布を，中心 M，分散 σ^2 を持つ Stiefel 空間上の巻き込み型正規分布と呼び，Stiefel 空間上の一様分布式 (2.11) を基準とした確率密度関数を $\mathcal{SWN}\left(Z\,;\,M,\sigma^2\right)$ と書くことにする．この分布からのサンプリング手順をまとめたものをアルゴリズム 1 に示す．

5.1.2　確率密度関数の計算

Stiefel 空間上の巻き込み型正規分布の密度関数 $\mathcal{SWN}\left(Z\,;\,M,\sigma^2\right)$ を求める．アルゴリズム 1 による，$M \in \mathcal{V}_{m,k}$ が与えられた下での $v \in \mathbb{R}^{\dim \mathcal{V}_{m,k}}$ から $Z \in \mathcal{V}_{m,k}$ への変換を f_M とし，$Z = f_M(v)$ とする．式 (2.1) より，Stiefel 空間の部分集合を

$A \subset \mathcal{V}_{m,k}$ とすると，A の規格化された測度は，

$$\begin{aligned}\int_A \mathcal{SWN}\big(Z\ ;\ M,\sigma^2\big)[dZ] &= \int_A \mathcal{SWN}\big(Z\ ;\ M,\sigma^2\big)\frac{H^{\dim \mathcal{V}_{m,k}}(dZ)}{\mathrm{vol}(\mathcal{V}_{m,k})} \\ &= \int_A \frac{\mathcal{SWN}\big(f_M(v)\ ;\ M,\sigma^2\big)(T_{f_M})_v}{\mathrm{vol}(\mathcal{V}_{m,k})} L^{\dim \mathcal{V}_{m,k}}(dv) \\ &= \int_A \mathcal{N}\big(v\ ;\ 0,\sigma^2\big) L^{\dim \mathcal{V}_{m,k}}(dv),\end{aligned}$$

となる．よって，

$$\mathcal{SWN}\big(Z\ ;\ M,\sigma^2\big) = \mathrm{vol}(\mathcal{V}_{m,k})\mathcal{N}\big(f_M^{-1}(Z)\ ;\ 0,\sigma^2\big)(T_{f_M})^{-1}_{f_M^{-1}(Z)}, \tag{5.8}$$

として確率密度関数 $\mathcal{SWN}\big(Z\ ;\ M,\sigma^2\big)$ が表される．これより，Z が与えられたとき，$v = f_M^{-1}(Z)$ と $(T_{f_M})_v$ が計算できれば，確率密度関数 $\mathcal{SWN}(Z\ ;\ M,\sigma^2)$ を求めることができる．

まず $f_M^{-1}(Z)$ を考える．式 (5.5) により，$M = R_{I_{m\times k}}\big(X_{I_{m\times k}}\big)$ となる $X_{I_{m\times k}} \in T_{I_{m\times k}}\mathcal{V}_{m,k}$ を求めることができる．こうして求めた $X_{I_{m\times k}}$ について，$m\times m$ 行列 Ω を $\Omega = \left(I_m - \frac{1}{2}W_{X_{I_{m\times k}}}\right)^{-1}\left(I_m + \frac{1}{2}W_{X_{I_{m\times k}}}\right)$ とすると，$M = \Omega I_{m\times k}$，$V_M = \Omega V_{I_{m\times k}}$ である．また，

$$\begin{aligned}\Omega^T\Omega &= \left(I_m + \frac{1}{2}W_{X_{I_{m\times k}}}\right)^T\left(I_m - \frac{1}{2}W_{X_{I_{m\times k}}}\right)^{-T} \\ &\qquad \left(I_m - \frac{1}{2}W_{X_{I_{m\times k}}}\right)^{-1}\left(I_m + \frac{1}{2}W_{X_{I_{m\times k}}}\right) \\ &= \left(I_m - \frac{1}{2}W_{X_{I_{m\times k}}}\right)\left(I_m + \frac{1}{2}W_{X_{I_{m\times k}}}\right)^{-1} \\ &\qquad \cdot \left(I_m - \frac{1}{2}W_{X_{I_{m\times k}}}\right)^{-1}\left(I_m + \frac{1}{2}W_{X_{I_{m\times k}}}\right) \\ &= \left(I_m - \frac{1}{2}W_{X_{I_{m\times k}}}\right)\left(I_m + \frac{1}{2}W_{X_{I_{m\times k}}}\right)^{-1} \\ &\qquad \cdot \left(I_m + \frac{1}{2}W_{X_{I_{m\times k}}}\right)\left(I_m - \frac{1}{2}W_{X_{I_{m\times k}}}\right)^{-1} \\ &= I_m,\end{aligned}$$

より，Ω は直交行列である．このとき，以下が成り立つ．

補題 5.1. W_{V_M} と $W_{V_{I_{m\times k}}}$ は，

$$W_{V_M} = \Omega W_{V_{I_{m\times k}}} \Omega^T, \tag{5.9}$$

という関係を持つ．

Proof. 式 (2.7) の W_V の定義より，以下が成り立つ．

$$\begin{aligned}
W_{V_M} &= P_M V_M M^T - M V_M^T P_M \\
&= \Big(I_m - \frac{1}{2}MM^T\Big)V_M M^T - M V_M^T \Big(I_m - \frac{1}{2}MM^T\Big) \\
&= V_M M^T - \frac{1}{2}MM^T V_M M^T - M V_M^T + \frac{1}{2}M V_M^T MM^T \\
&= \Omega V_{I_{m\times k}} I_{m\times k}^T \Omega^T - \frac{1}{2}\Omega I_{m\times k} I_{m\times k}^T \cancel{\Omega^T \Omega} V_{I_{m\times k}} I_{m\times k}^T \Omega^T \\
&\quad - \Omega I_{m\times k} V_{I_{m\times k}}^T \Omega^T + \frac{1}{2}\Omega I_{m\times k} V_{I_{m\times k}}^T \cancel{\Omega^T \Omega} I_{m\times k} I_{m\times k}^T \Omega^T,
\end{aligned}$$

ここで，$V_{I_{m\times k}} = \begin{bmatrix} A \\ B \end{bmatrix}$，$A \in Skew(k)$, $B \in \mathbb{R}^{(m-k)\times k}$ とすると，

$$\begin{aligned}
&= \Omega\Big(\begin{bmatrix} A & 0 \\ B & 0 \end{bmatrix} - \frac{1}{2}\begin{bmatrix} A & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} A^T & B^T \\ 0 & 0 \end{bmatrix} + \frac{1}{2}\begin{bmatrix} A^T & 0 \\ 0 & 0 \end{bmatrix}\Big)\Omega^T \\
&= \Omega \begin{bmatrix} A & -B^T \\ B & 0 \end{bmatrix} \Omega^T \\
&= \Omega\Big(P_{I_{m\times k}} V_{I_{m\times k}} I_{m\times k}^T - I_{m\times k} V_{I_{m\times k}}^T P_{I_{m\times k}}\Big)\Omega^T \\
&= \Omega W_{V_{I_{m\times k}}} \Omega^T.
\end{aligned}$$

□

以上を踏まえて，$v = f_M^{-1}(Z)$ に関する以下の補題が成り立つ．

補題 5.2. $Z = f_M(v)$ となる $v \in \mathbb{R}^{\dim \mathcal{V}_{m,k}}$ は，

$$\begin{aligned}
v &= f_M^{-1}(Z) \\
&= \begin{bmatrix} \mathrm{tril}\Big(\Big(\Psi_u + \big(\Psi_l \Phi_u^{-1}\big)^T \Phi_l\Big)\Phi_u^{-1}\Big) \\ \mathrm{vec}\big(\Psi_l \Phi_u^{-1}\big), \end{bmatrix},
\end{aligned} \tag{5.10}$$

として求められる．ただし，

$$\begin{bmatrix}\Phi_u\\ \Phi_l\end{bmatrix} = \Omega^T(Z+M) \qquad where \quad \Phi_u \in \mathbb{R}^{k\times k},\, \Phi_l \in \mathbb{R}^{(m-k)\times k} \tag{5.11}$$

$$\begin{bmatrix}\Psi_u\\ \Psi_l\end{bmatrix} = 2\Omega^T(Z-M) \qquad where \quad \Psi_u \in \mathbb{R}^{k\times k},\, \Psi_l \in \mathbb{R}^{(m-k)\times k}, \tag{5.12}$$

とする．

Proof.

$$\begin{aligned}
& Z = \left(I_m - \frac{1}{2}W_{V_M}\right)^{-1}\left(I_m + \frac{1}{2}W_{V_M}\right)M \\
\Leftrightarrow \quad & \left(I_m - \frac{1}{2}W_{V_M}\right)Z = \left(I_m + \frac{1}{2}W_{V_M}\right)M \\
\Leftrightarrow \quad & W_{V_M}(Z+M) = 2(Z-M) \\
\Leftrightarrow \quad & \Omega W_{V_{I_{m\times k}}}\Omega^T(Z+M) = 2(Z-M) \\
\Leftrightarrow \quad & W_{V_{I_{m\times k}}}\Omega^T(Z+M) = 2\Omega^T(Z-M) \\
\Leftrightarrow \quad & \begin{bmatrix}A & -B^T\\ B & 0\end{bmatrix}\Omega^T(Z+M) = 2\Omega^T(Z-M).
\end{aligned} \tag{5.13}$$

ここで，

$$\begin{bmatrix}\Phi_u\\ \Phi_l\end{bmatrix} = \Omega^T(Z+M) \quad where \quad \Phi_u \in \mathbb{R}^{k\times k},\, \Phi_l \in \mathbb{R}^{(m-k)\times k} \tag{5.14}$$

$$\begin{bmatrix}\Psi_u\\ \Psi_l\end{bmatrix} = 2\Omega^T(Z-M) \quad where \quad \Psi_u \in \mathbb{R}^{k\times k},\, \Psi_l \in \mathbb{R}^{(m-k)\times k}, \tag{5.15}$$

とすると，式 (5.13) より，

$$\begin{aligned}
& \begin{cases} A\Phi_u - B^T\Phi_l &= \Psi_u \\ B\Phi_u &= \Psi_l \end{cases} \\
\Leftrightarrow & \begin{cases} A &= \left(\Psi_u + \left(\Psi_l\Phi_u^{-1}\right)^T\Phi_l\right)\Phi_u^{-1} \\ B &= \Psi_l\Phi_u^{-1} \end{cases},
\end{aligned}$$

となる．よって，題意

$$
\begin{aligned}
v &= \begin{bmatrix} \operatorname{tril} A \\ \operatorname{vec} B \end{bmatrix} \\
&= \begin{bmatrix} \operatorname{tril}\Big(\Big(\Psi_u + \big(\Psi_l \Phi_u^{-1}\big)^T \Phi_l\Big)\Phi_u^{-1}\Big) \\ \operatorname{vec}\big(\Psi_l \Phi_u^{-1}\big) \end{bmatrix},
\end{aligned}
$$

が示された．　□

次に，$v \in \mathbb{R}^{\dim \mathcal{V}_{m,k}}$ における $\left(T_{f_M}\right)_v = \left|\det\Big(\left(J_{f_M}\right)_v{}^T \left(J_{f_M}\right)_x\Big)\right|^{\frac{1}{2}}$ を求める．まず，$\left(J_{f_M}\right)_v$ について以下が成り立つ．

補題 5.3. v における $f_M(v)$ のヤコビ行列は，

$$
\left(J_{f_M}\right)_v = \left(\left(\left(I_m - \frac{1}{2} W_{V_{I_{m\times k}}}\right)^{-1} I_{m\times k}\right)^T \otimes \left(\Omega\left(I_m - \frac{1}{2} W_{V_{I_{m\times k}}}\right)^{-1}\right)\right)\Gamma, \tag{5.16}
$$

となる．ただし，Γ は，$\operatorname{vec} W_{V_{I_{m\times k}}} = \Gamma \operatorname{vec} v$ となる行列とする [23]．具体的には，

$$
\Gamma = \left[(\Theta_1 \otimes \Theta_1)\tilde{D}_k \quad (I_{m^2} - K_{m,m})(\Theta_1 \otimes \Theta_2)\right], \tag{5.17}
$$

と表される．ここで，$K_{m,n}$ は $m \times n$ 行列の転置操作を表す commutation 行列と呼ばれる $mn \times mn$ 行列である [35]．C を $m \times n$ 行列としたとき，$K_{m,n} \operatorname{vec} C = \operatorname{vec} C^T$ となる．$H_{i,j}$ を (i,j)-成分が 1，それ以外の成分が 0 の $m \times n$ 行列としたとき，$K_{m,n}$ は具体的に，

$$
K_{m,n} = \sum_{i=1}^{m}\sum_{j=1}^{n}\big(H_{i,j} \otimes H_{i,j}^T\big). \tag{5.18}
$$

と表される．

Proof.

$$
\begin{aligned}
Z &= \left(I_m - \frac{1}{2}W_{V_M}\right)^{-1}\left(I_m + \frac{1}{2}W_{V_M}\right)M \\
&= \left(I_m - \frac{1}{2}\Omega W_{V_{I_{m\times k}}}\Omega^T\right)^{-1}\left(I_m + \frac{1}{2}\Omega W_{V_{I_{m\times k}}}\Omega^T\right)\Omega I_{m\times k} \\
&= \left(\Omega\left(I_m - \frac{1}{2}W_{V_{I_{m\times k}}}\right)\Omega^T\right)^{-1}\left(\Omega\left(I_m + \frac{1}{2}W_{V_{I_{m\times k}}}\right)\Omega^T\right)\Omega I_{m\times k} \\
&= \Omega^{-T}\left(I_m - \frac{1}{2}W_{V_{I_{m\times k}}}\right)^{-1}\cancel{\Omega^{-1}\Omega}\left(I_m + \frac{1}{2}W_{V_{I_{m\times k}}}\right)\cancel{\Omega^T\Omega} I_{m\times k} \\
&= \Omega\left(I_{m\times k} - \frac{1}{2}W_{V_{I_{m\times k}}}\right)^{-1}\left(I_{m\times k} + \frac{1}{2}W_{V_{I_{m\times k}}}\right)I_{m\times k}.
\end{aligned}
$$

ここで,

$$
d(C^{-1}) = -C^{-1}dCC^{-1},
$$

$$
\left(I_m - \frac{1}{2}W_{I_{I_{m\times k}}}\right)^{-1}\left(I_m + \frac{1}{2}W_{V_{I_{m\times k}}}\right) = 2\left(I_m - \frac{1}{2}W_{V_{I_{m\times k}}}\right)^{-1} - I_m,
$$

を用いると,

$$
\begin{aligned}
dZ = &\frac{1}{2}\Omega\left(I_m - \frac{1}{2}W_{V_{I_{m\times k}}}\right)^{-1}dW_{V_{I_{m\times k}}}\left(I_m - \frac{1}{2}W_{V_{I_{m\times k}}}\right)^{-1}\left(I_m + \frac{1}{2}W_{V_{I_{m\times k}}}\right)I_{m\times k} \\
&+ \frac{1}{2}\Omega\left(I_m - \frac{1}{2}W_{V_{I_{m\times k}}}\right)^{-1}dW_{V_{I_{m\times k}}}I_{m\times k} \\
= &\frac{1}{2}\Omega\left(I_m - \frac{1}{2}W_{V_{I_{m\times k}}}\right)^{-1}dW_{V_{I_{m\times k}}}\left(2\left(I_m - \frac{1}{2}W_{V_{I_{m\times k}}}\right)^{-1} - I_m\right)I_{m\times k} \\
&+ \frac{1}{2}\Omega\left(I_m - \frac{1}{2}W_{V_{I_{m\times k}}}\right)^{-1}dW_{V_{I_{m\times k}}}I_{m\times k} \\
= &\Omega\left(I_m - \frac{1}{2}W_{V_{I_{m\times k}}}\right)^{-1}dW_{V_{I_{m\times k}}}\left(I_m - \frac{1}{2}W_{V_{I_{m\times k}}}\right)^{-1}I_{m\times k},
\end{aligned}
$$

となる．これにより，以下が成り立つ．

$$d\operatorname{vec} Z = \left(\left(\left(I_m - \frac{1}{2}W_{V_{I_{m\times k}}}\right)^{-1} I_{m\times k}\right)^T \otimes \left(\Omega\left(I_m - \frac{1}{2}W_{V_{I_{m\times k}}}\right)^{-1}\right)\right) \operatorname{vec} W_{V_{I_{m\times k}}}$$
$$= \left(\left(\left(I_m - \frac{1}{2}W_{V_{I_{m\times k}}}\right)^{-1} I_{m\times k}\right)^T \otimes \left(\Omega\left(I_m - \frac{1}{2}W_{V_{I_{m\times k}}}\right)^{-1}\right)\right)\Gamma \operatorname{vec} v.$$

よって，v における $Z = f_M(v)$ のヤコビ行列は，

$$(J_{f_M})_v = \left(\left(\left(I_m - \frac{1}{2}W_{V_{I_{m\times k}}}\right)^{-1} I_{m\times k}\right)^T \otimes \left(\Omega\left(I_m - \frac{1}{2}W_{V_{I_{m\times k}}}\right)^{-1}\right)\right)\Gamma,$$

となる．□

よって，$(T_{f_M})_v$ は以下のように求められる．

$$(T_{f_M})_v = \left|\det\left((J_{f_M})_v{}^T (J_{f_M})_x\right)\right|^{\frac{1}{2}}$$
$$= \left|\det\left(\Gamma^T\left(\left(\left(I_m - \frac{1}{2}W_{V_{I_{m\times k}}}\right)^{-1} I_{m\times k} I_{m\times k}^T \left(I_m - \frac{1}{2}W_{V_{I_{m\times k}}}\right)^{-T}\right)\right.\right.$$
$$\left.\left.\otimes\left(\left(I_m - \frac{1}{2}W_{I_{m\times k}}\right)^{-T}\left(I_m - \frac{1}{2}W_{I_{m\times k}}\right)^{-1}\right)\right)\Gamma\right)\right|^{\frac{1}{2}}. \tag{5.19}$$

以上で求めた $v = f_M^{-1}(Z)$，$(T_{f_M})_v$ を用いて，確率密度 $\mathcal{SWN}\left(Z\,;\,M,\sigma^2\right)$ は，

$$\mathcal{SWN}\left(Z\,;\,M,\sigma^2\right) = \operatorname{vol}(\mathcal{V}_{m,k})\mathcal{N}\left(v\,;\,0,\sigma^2\right)(T_{f_M})_v^{-1}, \tag{5.20}$$

として求められる．確率密度 $\mathcal{SWN}\left(Z\,;\,M,\sigma^2\right)$ の計算手順をアルゴリズム 2 に示す．

これまでは $k < m$ の場合について考えてきた．次に，$k = m$ の場合について考える．第 4 章で紹介した matrix Langevin 分布と異なり，Stiefel 空間上の巻き込み型正規分布は，このままでは $k = m$ の場合に用いることができない．これは，直交行列がなす空間が，行列式が $+1$ となる直交行列空間 $SO_+(m)$ と，行列式が -1 となる直交行列空間 $SO_-(m)$ とからなる 2 つの連結成分に分かれており，連続写像で互いに移り合うことができないためである．$SO_+(m)$ 内の始点 M_+ から連続写像であるレトラク

アルゴリズム **2** $\mathcal{SWN}\big(Z\,;\,M,\sigma^2\big)$ の確率密度計算

Input: パラメタ $M \in \mathcal{V}_{m,k}$，$\sigma^2 \in \mathbb{R}^{\dim \mathcal{V}_{m,k}}$，サンプル $Z \in \mathcal{V}_{m,k}$

Output: Z の確率密度 $\mathcal{SWN}\big(Z\,;\,M,\sigma^2\big)$

式 (5.10) によって $v = f_M^{-1}(Z)$ を計算

式 (5.19) によって $\left(J_{f_M}\right)_v$ を計算

式 (5.20) によって $\mathcal{SWN}\big(Z\,;\,M,\sigma^2\big)$ を計算

ションによって移された点は $SO_+(m)$ 内にしか到達できず，また，同様に $SO_-(m)$ 内の始点 M_- からもレトラクションによって $SO_-(m)$ 内にしか到達できない．このように，ある一点を中心として持つ Stiefel 空間上の巻き込み型正規分布は直交行列空間全体を網羅できない．

このとき，$M_+ \in SO_+(m)$，$M_- \in SO_-(m)$ をそれぞれ中心としてアルゴリズム 1 によって構成した $SO_s(m)$，$s \in \{+,-\}$ 上の巻き込み型正規分布を考え，その確率密度関数を $\mathcal{SOWN}\big(Z\,;\,M_s,\sigma_s^2\big)$ と書くことにする．この密度関数は，$\mathrm{vol}(SO_s(m)) = \mathrm{vol}(\mathcal{V}_{m,m})/2$ であることを用いれば，以下のように表される．

$$\mathcal{SOWN}\big(Z\,;\,M_s,\sigma_s^2\big) = \frac{\mathrm{vol}(\mathcal{V}_{m,m})}{2}\mathcal{N}\big(v\,;\,0,\sigma_s^2\big)\left(T_{f_{M_s}}\right)_v^{-1}. \tag{5.21}$$

よって，$k = m$ となるときは，Stiefel 空間上の巻き込み型正規分布として，以下のような $SO_s(m)$ 上の巻き込み型正規分布の混合分布を用いることにする．

$$\mathcal{SWN}(Z\,;\,\pi_\pm,M_\pm,\sigma_\pm) = \sum_{s\in\{+,-\}} \pi_s \mathcal{SOWN}\big(Z\,;\,M_s,\sigma_s^2\big) \tag{5.22}$$

$$where \quad \sum_{s\in\{+,-\}} \pi_s = 1. \tag{5.23}$$

ただし，パラメタを $\theta = (\pi_\pm, M_\pm, \sigma_\pm)$ とし，$M_\pm$ は $\det M_\pm = \pm 1$ となるような直交行列とする．

5.1.3　確率密度関数の計算の効率化

$k = m$ の場合には，確率密度関数の中のヤコビ行列の行列式 $\left(T_{f_M}\right)_v$ を以下の補題のように変形し，$m \times m$ 行列の行列式の形で計算を効率化することが出来る．

補題 5.4. $k = m$ のとき，

$$(T_{f_M})_v = 2^{1/4m(m-1)} \left| \det\left(I_m + \frac{1}{2} W_{V_{I_m}} \right) \right|^{-(m-1)}, \tag{5.24}$$

が成り立つ．

Proof. まず，任意の $n \times n$ 行列 A について，

$$\det\bigl(\Gamma^T (A \otimes A)\Gamma\bigr) = 2^{1/2n(n-1)} (\det A)^{n-1} \tag{5.25}$$

の関係式が成り立つ ([33] の補題 4.4)．

また，式 (5.19) について，

$$\left(I_m - \frac{1}{2} W_{V_{I_m}} \right)^{-1} \left(I_m - \frac{1}{2} W_{V_{I_m}} \right)^{-T} = \left(I_m - \frac{1}{2} W_{I_m} \right)^{-T} \left(I_m - \frac{1}{2} W_{I_m} \right)^{-1}, \tag{5.26}$$

が成り立つ．実際，

$$\begin{aligned}
\left(I_m - \frac{1}{2} W_{V_{I_m}} \right)^{-1} \left(I_m - \frac{1}{2} W_{V_{I_m}} \right)^{-T} &= \left(\left(I_m - \frac{1}{2} W_{V_{I_m}} \right)^T \left(I_m - \frac{1}{2} W_{V_{I_m}} \right) \right)^{-1} \\
&= \left(\left(I_m + \frac{1}{2} W_{V_{I_m}} \right) \left(I_m - \frac{1}{2} W_{V_{I_m}} \right) \right)^{-1} \\
&= \left(\left(I_m - \frac{1}{2} W_{V_{I_m}} \right) \left(I_m + \frac{1}{2} W_{V_{I_m}} \right) \right)^{-1} \\
&= \left(I_m - \frac{1}{2} W_{I_m} \right)^{-T} \left(I_m - \frac{1}{2} W_{I_m} \right)^{-1},
\end{aligned}$$

となる．

式 (5.25) より，式 (5.19) が $\left|\det\bigl(\Gamma^T (A \otimes A)\Gamma\bigr)\right|^{1/2}$ の形で書けることが分かる．こ

れに式 (5.26) を適用すると，式 (5.19) を以下のように変形することが出来る．

$$
\begin{aligned}
\text{式 (5.19)} &= \left|\det\left(\Gamma^T\left(\left(I_m - \frac{1}{2}W_{V_{I_m}}\right)^{-1}\left(I_m - \frac{1}{2}W_{V_{I_m}}\right)^{-T}\right.\right.\right. \\
&\qquad \left.\left.\left. \otimes\left(I_m - \frac{1}{2}W_{V_{I_m}}\right)^{-1}\left(I_m - \frac{1}{2}W_{V_{I_m}}\right)^{-T}\right)\Gamma\right)\right|^{1/2} \\
&= \left|2^{1/2m(m-1)}\left(\det\left(\left(I_m - \frac{1}{2}W_{V_{I_m}}\right)^{-1}\left(I_m - \frac{1}{2}W_{V_{I_m}}\right)^{-T}\right)\right)^{m-1}\right|^{1/2} \\
&= 2^{1/4m(m-1)}\left|\det\left(I_m + \frac{1}{2}W_{V_{I_m}}\right)\right|^{-(m-1)}.
\end{aligned}
$$

□

5.2　ELBO の導出

$Z \in \mathcal{V}_{m,k}$ の事前分布として，$p(Z)$ を $\mathcal{V}_{m,k}$ 上の一様分布とする．このとき，$p(Z) = 1$ となり，$\log p(Z) = 0$ となるため，データ X を生成する対数尤度の下限は，

$$
\log p(X) \geq \int_{\mathcal{V}_{m,k}} \mathcal{SWN}(Z\,;\,\theta)(\log p(X \mid Z) - \log \mathcal{SWN}(Z\,;\,\theta))[dZ], \tag{5.27}
$$

となる．$k < m$ のときは，$\theta = \left(M, \sigma^2\right)$ を用い，$\mathcal{SWN}\left(Z\,;\,M, \sigma^2\right)$ とすれば，ELBO は式 (5.27) の右辺となる．$k = m$ の場合，$\theta = (\pi_\pm, M_\pm, \sigma_\pm)$ とし，$\mathcal{SWN}(Z\,;\,\theta)$ を式 (5.22) とすると，式 (5.27) について以下が成り立つ．

補題 5.5. X が生成される尤度について，以下の不等式が成り立つ．

$$
\begin{aligned}
\log p(X) \geq &\sum_{s\in\{+,-\}} \pi_s \mathbb{E}_{Z\sim\mathcal{SOWN}(Z;M_s,\sigma_s^2)}\left[\log(X \mid Z) - \log \mathcal{SOWN}\left(Z\,;\,M_s, \sigma_s^2\right)\right] \\
&- \sum_{s\in\{+,-\}} \pi_s \log \pi_s.
\end{aligned} \tag{5.28}
$$

Proof. 直交行列がなす空間は，$SO_+(m)$ と $SO_-(m)$ からなる 2 つの連結成分からなり，連続写像で互いに移り合うことができない．このため，$SO_+(m)$ 内の始点 M_+ か

ら連続写像であるレトラクションによって移された点は $SO_+(m)$ 内にしか到達できず，分布 $\mathcal{SOWN}\big(Z\,;\,M_+,\sigma_+^2\big)$ から $SO_-(m)$ 内の点 W_- が得られる確率は 0 になる．$\mathcal{SOWN}\big(Z\,;\,M_-,\sigma_-^2\big)$ についても同様に，$SO_+(m)$ 内の点 W_+ が得られる確率は 0 となる．よって，$s\in\{+,-\}$ として，以下が成り立つ．

$$\mathcal{SOWN}\big(Z\,;\,M_s,\sigma_s^2\big)=\begin{cases}\mathcal{SOWN}\big(Z\,;\,M_s,\sigma_s^2\big) & (Z\in SO_s(m))\\ 0 & (otherwise)\end{cases}.$$

これにより，式 (5.27) が次のように変形できる．

$$
\begin{aligned}
\text{式 (5.27)} &= \int_{O(m)} \left(\sum_{s\in\{+,-\}} \pi_s \mathcal{SOWN}\big(Z\,;\, M_s, \sigma_s^2\big) \right) \\
&\qquad \cdot \left(\log p(X \mid Z) - \log \sum_{s\in\{+,-\}} \pi_s \mathcal{SOWN}\big(Z\,;\, M_s, \sigma_s^2\big) \right) [dZ] \\
&= \int_{SO_+(m)} \pi_+ \mathcal{SOWN}\big(Z\,;\, M_+, \sigma_+^2\big) \\
&\qquad \cdot \big(\log p(X \mid Z) - \log\big(\pi_+ \mathcal{SOWN}\big(Z\,;\, M_+, \sigma_+^2\big)\big)\big)[dZ] \\
&\quad + \int_{SO_-(m)} \pi_- \mathcal{SOWN}\big(Z\,;\, M_-, \sigma_-^2\big) \\
&\qquad \cdot \big(\log p(X \mid Z) - \log\big(\pi_- \mathcal{SOWN}\big(Z\,;\, M_-, \sigma_-^2\big)\big)\big)[dZ] \\
&= \pi_+ \int_{SO_+(m)} \mathcal{SOWN}\big(Z\,;\, M_+, \sigma_+^2\big) \\
&\qquad \cdot \big(\log p(X \mid Z) - \log \mathcal{SOWN}\big(Z\,;\, M_+, \sigma_+^2\big)\big)[dZ] \\
&\quad + \pi_- \int_{SO_-(m)} \mathcal{SOWN}\big(Z\,;\, M_-, \sigma_-^2\big) \\
&\qquad \cdot \big(\log p(X \mid Z) - \log \mathcal{SOWN}\big(Z\,;\, M_-, \sigma_-^2\big)\big)[dZ] \\
&\quad - \pi_+ \log \pi_+ \underbrace{\int_{SO_+(m)} \mathcal{SOWN}\big(Z\,;\, M_+, \sigma_+^2\big)[dZ]}_{=1} \\
&\quad - \pi_- \log \pi_- \underbrace{\int_{SO_-(m)} \mathcal{SOWN}\big(Z\,;\, M_-, \sigma_-^2\big)[dZ]}_{=1} \\
&= \sum_{s\in\{+,-\}} \pi_s \mathbb{E}_{Z\sim\mathcal{SOWN}(Z;M_s,\sigma_s^2)} \big[\log(X \mid Z) - \log \mathcal{SOWN}\big(Z\,;\, M_s, \sigma_s^2\big)\big] \\
&\quad - \sum_{s\in\{+,-\}} \pi_s \log \pi_s.
\end{aligned}
$$

□

よって，$k = m$ の場合の ELBO は，式 (5.28) の右辺とすればよい．

5.3　ELBO の勾配計算

最後に，ELBO の勾配計算について説明する．本章で提案した Stiefel 空間上の巻き込み型正規分布は，reparameterization trick [1] と呼ばれる分散の小さな期待値勾配推定手法を適用可能である．アルゴリズム 1 における $v \in \mathbb{R}^{\dim \mathcal{V}_{m,k}}$ は，$v = \sigma \cdot \epsilon$，$\epsilon \sim \mathcal{N}(\epsilon\,;\, 0, 1)$ というように，σ と，ELBO の最適化変数に依存しない確率分布 $\mathcal{N}(\epsilon\,;\, 0, 1)$ に従う確率変数 ϵ との関数として表すことができる．よって，$h\big((M, \sigma^2), \epsilon\big) \coloneqq f_M(\sigma \cdot \epsilon)$ として関数 h を定義すれば，$Z = h\big(\theta = \big(M, \sigma^2\big), \epsilon\big)$ というように，確率変数 Z を最適化変数 θ と最適化変数に依存しない確率変数 ϵ の関数として表すことが出来る．

このとき，Z の関数 $g(Z)$ を Z について期待値を取り，その期待値を θ により微分する操作は，$g(h(\theta, \epsilon))$ を θ について微分して ϵ について期待値を取る操作に以下のように変形できる．

$$\nabla_\theta \mathbb{E}_{Z \sim p(Z;\theta)}[g(Z)] = \mathbb{E}_{\epsilon \sim \mathcal{N}(\epsilon;0,1)}[\nabla_\theta g(h(\theta, \epsilon))] \tag{5.29}$$

よって，この式の右辺をモンテカルロ近似することにより，期待値の勾配の推定量を得ることができる．この手法を reparameterization trick と呼ぶ．

ここで $g(Z)$ を式 (5.27) とすれば，その勾配を reparameterization trick によって推定することが出来る．

第 6 章

人工データを用いた評価

本章では，第 4 章，第 5 章で述べた手法について，人工データを用いたシンプルな設定で変分推論と変分オートエンコーダの実験を行い，それらの振る舞いを調べる．

6.1 変分推論

まず，提案手法を用いて，最小構成の実験設定で変分推論を行い，正規直交性を持つ潜在状態の事後確率を近似的に求められることを確認する．

6.1.1 タスク

実験のモデルを図 1 に示す．まず，ある正規直交行列 $Z\big(\in \mathbb{R}^{m\times k}; Z^TZ = I_2\big)$ の各成分に正規分布に従うノイズが加わり，観測値列として N 個の行列 $\big\{X_t \in \mathbb{R}^{m\times k}\big\}_{t=1}^N$ が得られる状況を考える．

$$\{X_t\}_{t=1}^N \overset{\text{i.i.d.}}{\sim} \mathcal{N}\big(X \mid Z, \sigma^2\big). \tag{6.1}$$

この状況下では Z の事後分布 $p\Big(Z \Big| \{X_t\}_{t=1}^N\Big)$ を陽に求めることができる．Z の事前分布として $\mathcal{V}_{m,k}$ 上の一様分布 式 (2.11) を仮定する．このとき，事後分布は以下のようになる．

$$p\Big(Z \Big| \{X_t\}_{t=1}^N\Big) = \mathcal{ML}\left(Z\,;\ \frac{1}{\sigma^2}\sum_{t=1}^N X_t\right). \tag{6.2}$$

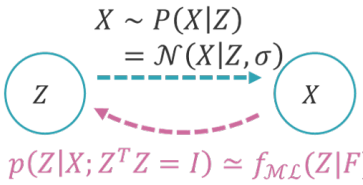


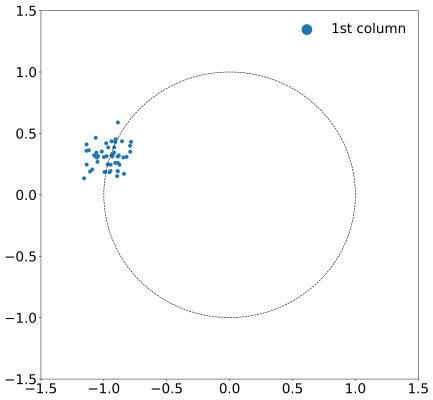
図 1　変分推論タスクの概略

これは以下のように求めることができる．

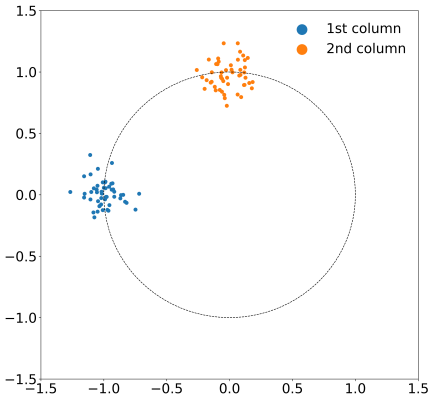
$$
\begin{aligned}
& p\left(Z \;\middle|\; \{X_t\}_{t=1}^N \,;\, Z^T Z = I_k\right) \\
& \propto p\left(\{X_t\}_{t=1}^N \;\middle|\; Z \,;\, Z^T Z = I_k\right) p(Z) \\
& = \mathcal{N}\left(\{X_t\}_{t=1}^N \;\middle|\; Z, \sigma^2 \,;\, Z^T Z = I_k\right) U(Z) \\
& \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{t=1}^N \sum_{i,j} (X_{t,i,j} - Z_{i,j})^2 \right) \Bigg|_{Z^T Z = I_k} \\
& = \exp\left(-\frac{1}{2\sigma^2} \sum_{t=1}^N \sum_{i,j} (X_{t,i,j}^2 + Z_{i,j}^2 - 2X_{t,i,j} Z_{i,j}) \right) \Bigg|_{Z^T Z = I_k} \\
& = \exp\left(-\frac{1}{2\sigma^2} \sum_{t=1}^N (\mathrm{tr}(\underbrace{X_t^T X_t + Z^T Z}_{const.} - 2X_t^T Z)) \right) \Bigg|_{Z^T Z = I_k} \\
& \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{t=1}^N \left(\mathrm{tr}\left(-2X_t^T Z\right)\right) \right) \Bigg|_{Z^T Z = I_k} \\
& = \exp\left(\mathrm{tr}\left(\left(\frac{1}{\sigma^2} \sum_{t=1}^N X_t \right)^T Z \right) \right) \Bigg|_{Z^T Z = I_k} .
\end{aligned}
\tag{6.3}
$$

これが $\mathcal{ML}$ 分布の表式 (式 (2.12)) と正規化定数を除いて一致していることから，式 (6.3) はパラメタ $F = \frac{1}{\sigma^2} \sum_{t=1}^N X_t$ を持つ $\mathcal{ML}$ 分布となることが分かり，事後分布 $p\left(Z \;\middle|\; \{X_t\}_{t=1}^N \,;\, Z^T Z = I_k\right)$ を式 (6.2) の形で表せられることが示された．

次に，行列 X を観測した下での Z の従う事後分布 $p(Z \mid X)$ を変分推論により近似する．Z の事前分布として $\mathcal{V}_{m,k}$ 上の一様分布 式 (2.11) を仮定する．近似分布 $q(Z \,;\, \theta)$



(a) $m = 2$，$k = 1$ の場合．青い点は行列 X_i の第一列を表す．



(b) $m = 2$，$k = 2$ の場合．青い点とオレンジ色の点はそれぞれ行列 X_i の第一列と第二列を表す．

図 2　変分推論の実験に用いたサンプル行列 $\{X_i\}_{i=1}^N$ の例

として matrix Langevin 分布 $\mathcal{ML}(Z\,;\,F)$, $(\theta = F)$ と Stiefel 空間上の巻き込み型正規分布 $\mathcal{SWN}\big(Z\,;\,M, \sigma^2\big)$, $(\theta = (M, \sigma^2))$ を用い，第 4 章，第 5 章 で説明した手法により，分布のパラメタ θ を更新しながら式 (2.19) でも示したように ELBO の最大化を行う．

$$\max_{\theta}\Big\{\mathbb{E}_{q(Z;\theta)}\big[\log \mathcal{N}\big(X \;\big|\; Z, \sigma^2\big)\big] - D_{KL}(q(Z\,;\,\theta) \,\|\, p(Z))\Big\}.$$

この最適化によって得られた近似事後分布 $q(Z\,;\,\theta^*)$ が式 (6.2) を良く近似できているかどうかによって，ELBO の最適化が正常に行えていかどうかを確認できる．

6.1.2　実験設定

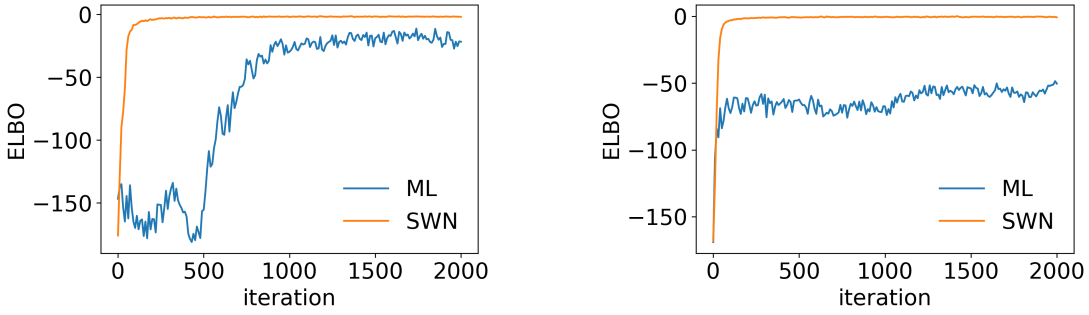
次元 m，k について，$m = 2$，$k \in \{1, 2\}$ の組み合わせで実験を行った．観測データを次のような設定の下で生成した．分布の中心となる $Z \in \mathcal{V}_{m,k}$ をランダムに生成し，ノイズの偏差を $\sigma = 0.1$ として，観測点を $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\big(X \;\big|\; Z, \sigma^2\big)$ によりサンプリングした．また，観測点数を $N = 50$ とした．生成された行列データ $\{X\}_{t=1}^N$ の例として，

表 1　変分推論に用いたハイパーパラメータ

最適化手法	Adam [36]
初期学習率	0.1
バッチサイズ	50
イテレーション数	1000

表 2　変分推論の実験結果

	$k=1$			$k=2$		
	ELBO	Recon. Loss	KL Loss	ELBO	Recon. Loss	KL Loss
$\mathcal{ML}$	-11.3	10.4	0.83	-48.5	45.0	3.49
$\mathcal{SWN}$	-1.40	-1.76	3.16	0.154	-2.92	2.77



(a) $k=1$ の場合　　(b) $k=2$ の場合

図 3　各イテレーション毎の ELBO の推移

$m=2$，$k\in\{1,2\}$ の場合について図 2 に示す．

6.1.3　実験結果

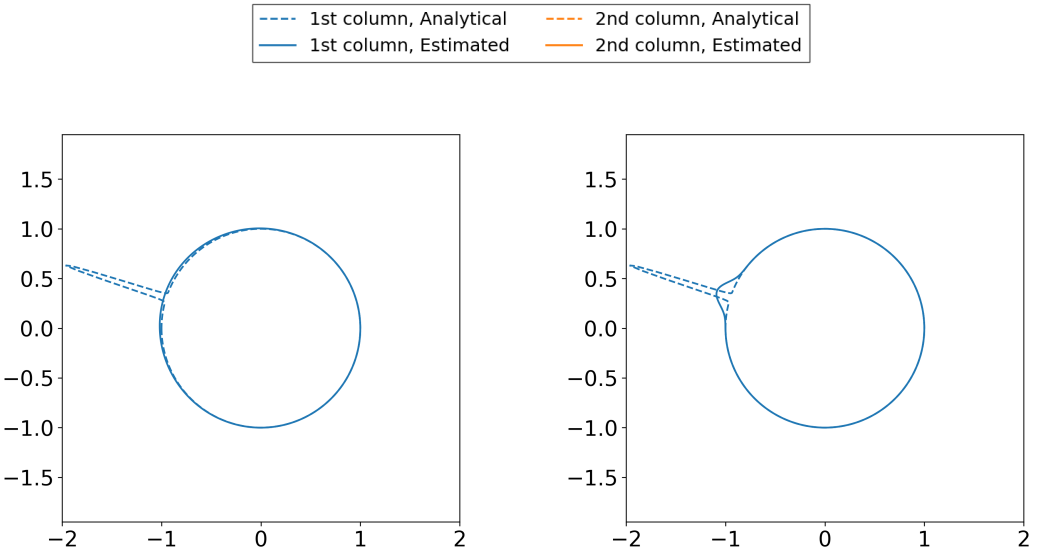
変分推論の実験結果について述べる．まず，勾配降下法を 1000 回イテレーションして得られた中で得られた ELBO の最大値と，その ELBO を達成したとき点における，

ELBO 式中の再構成誤差 (Recon. Loss)，KL 誤差 (KL Loss) の 2 項を表 2 に示す．ELBO と再構成損失については，Stiefel 空間上の巻き込み型正規分布を用いた方が，matrix Langevin 分布を用いたものよりも良い結果となった．また，$k = 1$ の場合の KL 誤差が小さな値となった．これは，推定された近似事後分布が Stiefel 空間上の一様分布に近いことを表している．

訓練ときにおける ELBO の推移を図 3 に示す．この図 3 から，matrix Langevin 分布を用いたものは，不安定で収束も遅いのに対し，Stiefel 空間上の巻き込み型正規分布を用いたものは，安定して素早く学習が安定し，最終的に収束した先の値も matrix Langevin 分布のものより良い値となっていることが分かる．

解析的に求めた事後分布式 (6.2) と ELBO の最適化式 (2.19) によって求めた事後分布について，行列 Z の各列の確率密度関数をプロットしたものを図 4 に示す．行列 Z は正規直交性を持つため，その各列は正規化されたベクトルとなって円周上に分布する．そこで，Z の各列がある方向を取る確率を，その方向の半径で表現した．図 4 からも，Stiefel 空間上の巻き込み型正規分布 $\mathcal{SWN}$ を用いたものの方が，matrix Langevin 分布 $\mathcal{ML}$ を用いたものよりも，真の分布 (式 (6.2)) をよく近似していることが見て取れる．

学習の 1 イテレーションあたりにかかった計算とき間を図 5 に示す．この図 5 から，Stiefel 空間上の巻き込み型正規分布を用いた手法は matrix Langevin 分布を用いた手法に比べて 100 倍以上高速であることが分かる．



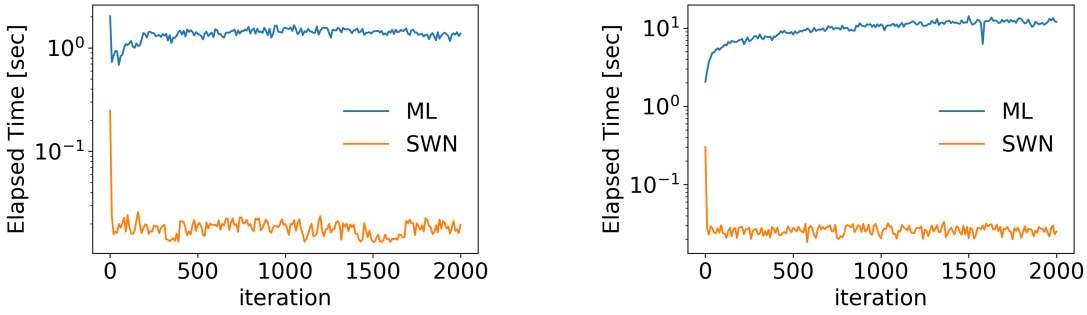
(a) $\mathcal{ML}(Z\,;\,F)$　　(b) $\mathcal{SWN}\big(Z\,;\,M,\sigma^2\big)$

$m=2,\,k=1$ の場合

(c) $\mathcal{ML}(Z\,;\,F)$　　(d) $\mathcal{SWN}\big(Z\,;\,M,\sigma^2\big)$

$m=2,\,k=2$ の場合

図 4　Z の事後確率密度関数のプロット．実線は変分推論 (式 (2.19)) によって近似的に推定された Z の事後分布を，点線は式 (6.2) による解析的な Z の事後分布を表す．



(a) $k = 1$ の場合　　(b) $k = 2$ の場合

図 5　1 イテレーションあたりの計算とき間の推移

6.1.4　考察

本実験では，第 4 章の matrix Langevin 分布と第 5 章の Stiefel 空間上の巻き込み型正規分布を用いた手法のどちらにおいても，事後分布を近似的に推定できることを確認した．また，Stiefel 空間上の巻き込み型正規分布を用いた手法は，学習の安定性と高速性の点で matrix Langevin 分布を用いたものよりも優れていることが示された．

学習の安定性に寄与している要因については，期待値の勾配推定量の違いが考えられる．Matrix Langevin 分布を用いた変分推論手法では Score function estimator と呼ばれる推定量によって期待値の勾配を推定しているが，この推定量は分散が大きく，従来のユークリッド空間上の変分オートエンコーダの研究でも学習が安定的に行えないことが述べられている [37]．Stiefel 空間上の巻き込み型正規分布ではより分散の小さい期待値勾配推定手法である Reparameterization trick を使用しているため，一貫した方向へ勾配降下をすることができ，結果として学習の安定性が得られている，と予想される．

また，学習の高速性については，潜在状態のサンプリング手法の違いが考えられる．Matrix Langevin 分布を用いた手法では，潜在表現のサンプリングに採択-棄却法 [18, 38] を用いている．このサンプリング手法は，一様分布から多数の標本を得た後，目的の分布との確率比によってそれらの標本の一部を採択し，その他を捨てるアルゴリズムとなっている．そのため，捨てられる標本の数だけ計算に無駄が生じ，目的

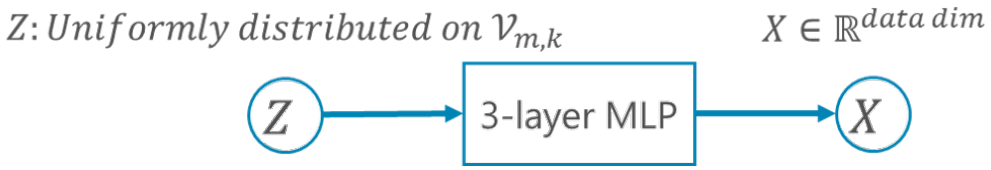


図 6　変分オートエンコーダで学習するデータセットの生成過程

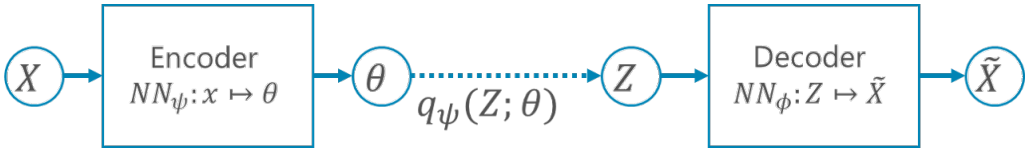


図 7　変分オートエンコーダモデルの概略

の分布に従うサンプルを一定個数得るために必要なとき間が長くなる．目的の分布が鋭くなり，一様分布から離れるほど，捨てられる標本の割合は増加し，サンプリングとき間は長くなる．一方で，Stiefel 空間上の巻き込み型正規分布の方は，ユークリッド空間上の正規分布に従う確率変数を変数変換するだけなので，無駄になる標本が一切なく，効率の良いアルゴリズムとなっている．この点が学習の高速性に寄与している，と予想される．

6.2　変分オートエンコーダ

次に，変分オートエンコーダによる人工データの学習を行う．潜在的に Stiefel 空間に分布するようなデータを作り，Stiefel 空間上の巻き込み型正規分布を用いた変分オートエンコーダとユークリッド空間上の正規分布を用いた通常の変分オートエンコーダを比較する．

6.2.1　実験設定

まず，用いる人工データの構成を図 6 に示す．このように，まず潜在状態 Z を $\mathcal{V}_{m,k}$ から一様にサンプリングした後，それを 3 層パーセプトロンに通し，高次元空間へと非線形変換を施すことにより，潜在的に Stiefel 空間 $\mathcal{V}_{m,k}$ に分布するデータを生成する．

表 3　変分オートエンコーダの学習に用いたハイパーパラメータ

エンコーダの中間層の数	2 層
エンコーダの中間層の次元	128
デコーダの中間層の数	1 層
デコーダの中間層の次元	$m \times k$
最適化手法	Adam [36]
初期学習率	$1.0 \cdot 10^{-3}$
バッチサイズ	100
エポック数	1000
勾配クリッピング	10

次に，モデルについて説明する．モデルのエンコーダが出力する近似分布 $q_\psi(Z\,;\,\theta)$ のパラメタ θ については，ユークリッド空間上のガウス分布を用いた変分オートエンコーダについては $\theta = \left(\mu \in \mathbb{R}^{\dim \mathcal{V}_{m,k}}, \sigma \in \mathbb{R}^{\dim \mathcal{V}_{m,k}}\right)$ とする．また，Stiefel 空間上の巻き込み型正規分布を用いた変分オートエンコーダについては，$\theta = \left(M \in \mathcal{V}_{m,k}, \sigma \in \mathbb{R}^{\dim \mathcal{V}_{m,k}}\right)$ とする．ここで $M \in \mathcal{V}_{m,k}$ は，特異値分解を用いて，エンコーダの出力 $output \in \mathbb{R}^{m \times k}$ を $\mathcal{V}_{m,k}$ 上の最近傍点に射影することによって次のようにして得る．

$$U\Sigma V^T = output, \tag{6.4}$$

$$M = UV^T. \tag{6.5}$$

また，変分オートエンコーダの学習に用いた各種パラメタを表 3 に示す．

6.2.2　結果

変分オートエンコーダの学習結果を表 4，表 5 に示す．表 4 は $m = 5$ として，$k \in \{1, 2, 3, 4\}$ として k の値を様々に取った結果である．表 5 は，1000 エポックを計算し終わった後，目的関数である ELBO が最大になるとき点における対数尤度 (LL)，ELBO，再構成誤差 (Recon. Loss)，KL ダイバージェンスによる正則化項 (KL Loss) を示す．一番重要な指標は対数尤度であり，大きければ大きいほど良い．表 4 を見ると，$m = 5$，$k = 1, 2$ の場合については Stiefel 空間上の巻き込み型正規分布を用いた

表 4　$m = 5$ とし，k の値を変動させた場合の変分オートエンコーダの実験結果

	$k=1$		$k=2$		$k=3$		$k=4$	
	$\mathcal{SWN}$	$\mathcal{N}$	$\mathcal{SWN}$	$\mathcal{N}$	$\mathcal{SWN}$	$\mathcal{N}$	$\mathcal{SWN}$	$\mathcal{N}$
LL	-17.2	-56.7	-55.5	$-2.56 \cdot 10^2$	$-4.06 \cdot 10^2$	$-2.03 \cdot 10^2$	$-8.35 \cdot 10^2$	$-6.81 \cdot 10^2$
ELBO	-18.5	-58.5	-62.2	$-2.64 \cdot 10^2$	$-4.29 \cdot 10^2$	$-2.18 \cdot 10^2$	$-8.84 \cdot 10^2$	$-7.07 \cdot 10^2$
Recon. Loss	3.81	50.5	30.1	$2.47 \cdot 10^2$	$3.84 \cdot 10^2$	$1.82 \cdot 10^2$	$8.03 \cdot 10^2$	$6.64 \cdot 10^2$
KL Loss	14.7	7.99	32.1	17.8	45.1	35.9	80.5	43.0

表 5　$k = 4$ とし，m の値を変動させた場合の変分オートエンコーダの実験結果

	$m=10$		$m=20$	
	$\mathcal{SWN}$	$\mathcal{N}$	$\mathcal{SWN}$	$\mathcal{N}$
LL	$-3.47 \cdot 10^3$	$-4.47 \cdot 10^3$	$-3.65 \cdot 10^4$	$-6.61 \cdot 10^4$
ELBO	$-3.82 \cdot 10^3$	$-4.59 \cdot 10^3$	$-3.66 \cdot 10^4$	$-6.67 \cdot 10^4$
Recon. Loss	$3.34 \cdot 10^3$	$4.49 \cdot 10^3$	$3.58 \cdot 10^4$	$6.65 \cdot 10^4$
KL Loss	$4.78 \cdot 10^2$	95.8	$7.92 \cdot 10^2$	$2.57 \cdot 10^2$

変分オートエンコーダがユークリッド空間上のガウス分布を用いたものよりも高い性能を示している．一方で，$m = 5$，$k = 3, 4$ を見ると，性能差は逆転し，ユークリッド空間上の変分オートエンコーダの方が性能が高くなる．次元が上がるにつれてユークリッド空間上のガウス分布の変分オートエンコーダが優勢になるのかを調べるために，今度は $k = 4$ を固定し，m の方を変えて性能を調べてみる．その結果が表 5 である．これを見ると，$m = 10, 20$ となると再び Stiefel 空間上の変分オートエンコーダが優勢になっていることが分かる．これにより，次元の多寡よりも $k \ll m$ であるかどうかが性能に対して重要である，と予想される．

6.2.3　考察

Stiefel 空間上の巻き込み型正規分布を用いた変分オートエンコーダが，ガウス分布のものと比べて，$k \ll m$ であるときに性能が高く，$k \approx m$ のときに性能が低くなる理

由はまだ解明できていない．憶測の域を出ないが，$\mathcal{V}_{m,k}$ の k が大きくなるにつれ，各変数間の制約が多くなり，それが最適化問題を難しくしているのではないか，と考えている．

第 7 章

結論

本研究では，Stiefel 空間を潜在状態の空間として持つようなデータに対して，そうした構造を保ちながら変分推論及び VAE の学習を行う手法の開発について取り組んだ．matrix Langevin 分布と Stiefel 空間上の巻き込み型正規分布を用いた変分推論手法を提案した．簡単な設定の下で人工データに対して変分推論を行い，巻き込み型正規分布を用いた変分推論手法は，matrix Langevin 分布を用いた手法と比べて，安定性と速度の両面で優れていることを実験的に確認した．また，人工データに対して VAE を学習することにより，ガウス分布を用いた通常の VAE と比べて，データが $k \ll m$ となるような Stiefel 空間 $\mathcal{V}_{m,k}$ に潜在的に分布する場合に優れた性能を発揮することを確認した．

未解決の課題としては，第一に実世界データでの実験が挙げられる．本研究では，シンプルな設定で人工データの学習を行うにとどまった．実世界のデータへの応用先として，言語横断写像の学習 [39, 40] など，直交性を課されるようなタスクが考えられる．本手法をそれらのタスクに適用することで，不確かさや情報量といったものを考慮した学習ができるようになると考えられる．第二の課題として，本研究で行った VAE の実験結果について，Stiefel 空間 $\mathcal{V}_{m,k}$ 上の巻き込み型正規分布を用いた VAE は，なぜ $k \approx m$ となる場合に通常の VAE よりも性能が悪くなってしまうのかの解明が挙げられる．その要因を理論的に解析し，原因を明確化することで，変分推論手法のさらなる改善が期待される．また，本手法を適用するにはデータが潜在的に Stiefel 空間上に分布していることを知っている必要がある．しかし，実際の状況では何らかのデータが与えられた際に，そのデータがどのような構造を本質的に持っているのかは明らかでない．このような状況で，データやタスクに応じて適切な空間や分布を選択する規

準を開発することが重要だと考えられる.

謝辞

本研究に取り組むにあたり，本当に多くの方々にご指導ご支援を賜りました．研究テーマが非ユークリッド空間上での統計・機械学習と，当研究室とは関連の薄い分野であるにも関わらず，このように無事修論を書き上げることができたのは，ひとえにこの研究室に関わる様々な人々のお力添えと，のびのびと自由に研究でき，どんな分野でも興味を持って受け入れてもらえる研究室の雰囲気のおかげです．

豊田正史教授には，分野外であるにもかかわらず，論文の添削や発表の仕方，研究に詰まった際のアドバイスなど，様々なご指導を賜りました．研究のディスカッションなどにおいて，どんな突飛な提案をしても，決して否定せず，真剣に吟味してくださり，おかげさまで 2 年間楽しく研究に取り組むことができました．また，精神的に辛いときや失敗してしまったときなども，常に落ち着いて対応してくださり，幾度となく救われました．

現在は New York 大学へ赴任されている小宮山純平助教と，NTT の岩田具治さん，石畠正和さんには，機械学習分野全般について様々な話題を聞かせていただき，知見を広げることができました．機械学習の研究として，どのような実験をし，評価していけばいいのかなどの，メタ的なノウハウを指導いただきありがとうございました．また，企業とアカデミアを両方とも経験されている立場からの助言は，進路を考える上で大変参考になりました．

吉永直樹准教授には分かり易い発表の仕方や研究者の心得など，様々なご指導を賜りました．また，住んでいる場所が近かったのもあり，最寄り駅〜学校間の美味しい料理やビールを教わり，また，ときには御馳走していただき，ありがとうございました．

喜連川優教授には素晴らしい研究環境を用意していただき，おかげさまで何不自由なく研究に取り組むことができました．また，喜連川教授の本質を的確につくご指摘には，何度も感銘を受けました．

合田和生准教授，早水悠登助教には研究室に所属してから最初の半年間，データベースアルゴリズムについてご指導いただき，見識を深めることができました．特に最初に勧められて読んだ乱択データ構造についての一連の論文などでは，統計的推定量の

漸近性について洗練された証明がなされており，その美しさに衝撃を受けました．

秘書・経理である，周佐亜樹さん，池田鈴子さん，井崎葉子さんには，事務手続きについて手厚くサポートしていただき，おかげさまで雑務に煩わされることなく快適に研究を進めることができました．また，研究に疲れた際に雑談の相手になってくださり，気分転換になりました．

研究室の先輩方である，石渡祥之佑さん，佐藤文一さん，金洪善さん， 佐藤翔悦さん，澤田頌子さん，赤崎智さん，吉岡弘隆さん，保田和彦さん，清水洸希さん，根石将人さん，張翔さん，佐久間仁さん，遠田哲史さん，羅博明さんには，最も身近な先人として，発表スライドのチェックや研究方針のディスカッション，研究以外の相談事など，あらゆる面でお世話になりました．

また，研究室の後輩である，王子晗君，左天池君，磯川弘基君，塚田涼太郎君，大前拓巳君，詹浩森君，馬唯焜君，袁月皓君，中村夏子君とは研究の話をしていてとても面白かったです．僕なんかよりとてもしっかりしているので，教えることよりも教わることの方が多く，非常に勉強になりました．

同期の蔦侑磨君，福田展和君，杉山普君，大葉大輔君，別所祐太朗君，土屋潤一郎君には，研究分野が異なるながらもディスカッションし合い，切磋琢磨できたことを嬉しく思います．他分野の最先端の研究について知ることができました．

最後に，修士 2 年間を支えてくれた家族に多大な感謝を捧げます．帰省するたびに暖かく出迎えてくれて，心の拠り所となりました．

2020 年 1 月 30 日

参考文献

[1] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations*, 2014.

[2] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *Proceedings of the 31st International Conference on Machine Learning*, pp. 1278–1286, 2014.

[3] K V Mardia. Statistics of Directional Data. *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 37, No. 3, pp. 349–393, 1975.

[4] N I Fisher, T Lewis, and B J J Embleton. Statistical analysis of spherical data. *Cambridge: University Press, 1987*, 1987.

[5] Md. Abul Hasnat, Julien Bohné, Jonathan Milgram, St é phaneGentric, Liming Chen. von Mises-Fisher Mixture Model-based Deep learning: Application to Face Verification. 2017.

[6] Tim R Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M Tomczak. Hyperspherical Variational Auto-Encoders. In *34th Conference on Uncertainty in Artificial Intelligence*, 2018.

[7] Jiacheng Xu and Greg Durrett. Spherical Latent Spaces for Stable Variational Autoencoders. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4503–4513, 2018.

[8] Pavan Turaga, Ashok Veeraraghavan, and Rama Chellappa. Statistical analysis on Stiefel and Grassmann manifolds with applications in computer vision.

In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[9] 宏今野. 微分幾何学. 東京大学出版会, 2013.

[10] 幸夫松本. 多様体の基礎. 東京大学出版会, 1988.

[11] 青季西川. 幾何学的変分問題. 岩波書店, 2006.

[12] Alan Edelman, Tomas A Arias, and Steven T Smith. The Geometry of Algorithms with Orthogonality Constraints. *SIAM Journal on Matrix Analysis and Applications*, Vol. 20, No. 2, pp. 303–353, 1998.

[13] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, 2008.

[14] Xiaojing Zhu. A Riemannian conjugate gradient method for optimization on the Stiefel manifold. *Computational Optimization and Applications*, Vol. 67, No. 1, pp. 73–110, 5 2017.

[15] Thomas D. Downs. Orientation Statistics. *Biometrika*, Vol. 59, No. 3, pp. 665–676, 1972.

[16] C G Khatri and K V Mardia. The Von Mises-Fisher Matrix Distribution in Orientation Statistics. *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 39, No. 1, pp. 95–106, 1977.

[17] P.E. Jupp and K.V. Mardia. Maximum likelihood estimators for the matrix von Mises-Fisher and Bingham distributions. *The Annals of Statistics*, Vol. 7, No. 3, pp. 599–606, 1979.

[18] Yasuko Chikuse. *Statistics on Special Manifolds*, Vol. 174. Springer Science & Business Media, 2003.

[19] Herbert Federer. *Geometric Measure Theory*. Springer-Verlag Berlin Heidelberg, 1969.

[20] Simon Byrne and Mark Girolami. Geodesic Monte Carlo on embedded manifolds. *Scandinavian Journal of Statistics*, Vol. 40, No. 4, pp. 825–845, 2013.

[21] Persi Diaconis, Susan Holmes, and Mehrdad Shahshahani. Sampling from a Manifold. *Advances in Modern Statistical Theory and Applications: A Festschrift in honor of Morris L. Eaton*, Vol. 10, pp. 102–125, 2013.

[22] Tim Traynor. Change of Variable for Hausdorff Measure (from the begginning). In *Workwhop on Measure Theory and Real Analysis*, pp. 327–347, 1993.

[23] Michael Jauch, Peter D. Hoff, and David B. Dunson. Random orthogonal matrices and the Cayley transform. *Bernoulli*, Vol. 26, No. 2, pp. 1560–1586, 5 2020.

[24] Arthur Cayley. Sur quelques Propriétés des Déterminants Gauches. *Journal für die reine und angewandte Mathematik*, Vol. 32, pp. 119–123, 1846.

[25] Alan T. James. Distributions of Matrix Variates and Latent Roots Derived from Normal Samples. *The Annals of Mathematical Statistics*, Vol. 35, No. 2, pp. 475–501, 1964.

[26] Danilo Jimenez Rezende and Shakir Mohamed. Variational Inference with Normalizing Flows. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1530–1538, 2015.

[27] Chang Liu and Jun Zhu. Riemannian Stein Variational Gradient Descent for Bayesian Inference. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[28] Ivan Ovinnikov. Poincaré Wasserstein Autoencoder. In *Third workshop on Bayesian Deep Learning (NeurIPS 2018)*, 2018.

[29] Pavan Turaga, Ashok Veeraraghavan, Anuj Srivastava, and Rama Chellappa. Statistical computations on grassmann and stiefel manifolds for image and video-based recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 33, No. 11, pp. 2273–2286, 2011.

[30] Rudrasis Chakraborty and Baba C Vemuri. Statistics on the Stiefel Manifold: Theory and Applications. *The Annals of Statistics*, Vol. 47, No. 1, pp. 415–438, 2019.

[31] Alfred Kume and Andrew T A Wood. Saddlepoint approximations for the Bingham and Fisher - Bingham normalising constants. *Biometrika*, Vol. 92, No. 2, pp. 465–476, 2005.

[32] Alfred Kume, S P Preston, and Andrew T A Wood. Saddlepoint approximations for the normalizing constant of Fisher-Bingham distributions on products of spheres and Stiefel manifolds. *Biometrika*, Vol. 100, No. 4, pp. 971–984, 2013.

[33] Heinz Neudecker. On Jacobians of Transformations with Skew-Symmetric, Strictly (Lower) Triangular or Diagonal Matrix Arguments. *Linear and Multilinear Algebra*, Vol. 14, No. 3, pp. 271–295, 11 1983.

[34] Ron Shepard, Scott R. Brozell, and Gergely Gidofalvi. The Representation and Parametrization of Orthogonal Matrices. *The Journal of Physical Chemistry A*, Vol. 119, No. 28, pp. 7924–7939, 7 2015.

[35] Jan R. Magnus and Heinz Neudecker. The Commutation Matrix: Some Properties and Applications. *The Annals of Statistics*, Vol. 7, No. 2, pp. 381–394, 3 1979.

[36] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. International Conference on Learning

Representations, ICLR, 2015.

[37] Ming Xu, Matias Quiroz, Robert Kohn, and Scott A. Sisson. Variance reduction properties of the reparameterization trick. *Proceedings of Machine Learning Research*, Vol. 89, pp. 2711–2720, 2019.

[38] David J C Mackay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[39] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2289–2294. Association for Computational Linguistics (ACL), 2016.

[40] Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, Herv é Jégou. Word Translation without Parallel Data. In *International Conference on Learning Representations*, 2018.

発表文献

査読無し国内会議・ワークショップ

1. 三條 嵩明，小宮山 純平，豊田 正史，喜連川 優，
   "Stiefel 空間上の変分オートエンコーダ"，
   第 33 回 人工知能学会全国大会 (JSAI 2019)，札幌，2019.

2. 三條 嵩明，岩田 具治，石畠 正和，小宮山 純平，豊田 正史，喜連川 優，
   "Stiefel 空間上の Reparameterization 可能な分布を用いた変分オートエンコーダ"，
   第 22 回 情報論的学習理論ワークショップ (IBIS 2019)，名古屋，2019