２．【研究計画】適宜概念図を用いるなどして、わかりやすく記入してください。なお、本項目は１頁に収めてください。様式の変更・追加は不可。

(1) 研究の位置づけ

特別研究員として取り組む研究の位置づけについて、当該分野の状況や課題等の背景、並びに本研究計画の着想に至った経緯も含めて記入してください。

着想経緯：データサイエンスの発展にはベイズ法の発展が重要

申請者は学部時代にデータサイエンティストとしてのインターンを通じ，ベイズ統計の手法が解析結果を推定値の**一点**だけでなく**誤差の広がりを持って視覚化できる**ことに魅力を感じた（図 1 参照）．

しかし，現状のベイズ法は万人にとって使いやすいものであると言えない．後述の HMC アルゴリズムに習熟していない場合は，最適な設定を見つけるまで何度も試行錯誤を必要とすることがある．

そのことから，ベイズ法が敬遠されたり，安易な近似手法が選好されたりしがちな現状には苦い思いを感じていた．

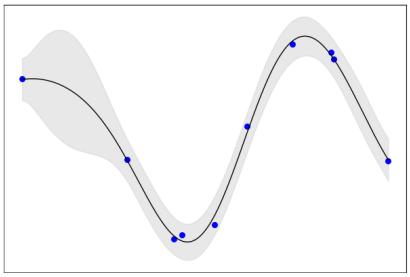


図 1: ベイズ法（ガウス過程回帰）の結果．推定値（実線）だけでなく 95% 信用区間（灰色）も自然に得られる．データが少ない領域では予測が不確実である（＝モデルに自信がない）ことがわかる．

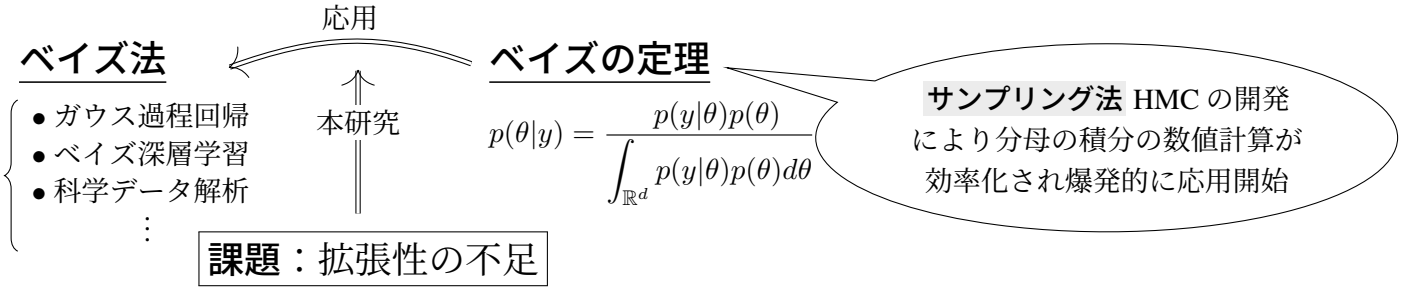
当該分野の状況：ベイズ法の発展にはサンプリング法の発展が重要

ベイズ法とは，機械学習や統計学において，ベイズの定理に基づく手法群の総称である．ベイズ法は最適である（＝どの他手法よりも悪くない）ことが示されている上に，事前情報を取り入れた柔軟なモデリングが可能であるため大変魅力的であるが，実際の計算が困難であることが応用範囲を狭めてきた．

ほとんどすべてのベイズ法は，確率測度 $P \in \mathcal{P}(\mathbb{R}^d)$ に関する積分の数値計算の問題に帰着する．決定論的な数値計算法は高々 $d \leq 3$ までの場合にしか現実的な時間内で実行できない．そのため，収束レートは落ちるが，P に従う乱数を生成（**サンプリング**という）してモンテカルロ推定により積分を近似する：

$$\frac{1}{N}\sum_{n=1}^{N} f(X_n) \xrightarrow[\text{a.s.}]{N\to\infty} \int_{\mathbb{R}^d} f(x)\,P(dx), \qquad X_n \overset{\text{i.i.d.}}{\sim} P, \quad f \in \mathcal{L}^1(P).$$

広汎な確率分布 P に使える**汎用サンプリング法**であるハミルトニアンモンテカルロ法（HMC）が開発 [1] されてから，ベイズ法は爆発的に応用され始めた．しかし問題は，この **HMC を 40 年近くも改良できないまま現在でも用いている**点にある．前述の通り **HMC は万人に使いやすい訳ではない**ため，ベイズ法の普及と応用を妨げる要因となってしまっている．



当該分野の課題：サンプリング法の発展には拡張性（スケーラビリティ）が重要

現状，$d \gg 1$ の高次元の場合でも広い範囲の分布 $P \in \mathcal{P}(\mathbb{R}^d)$ に汎用的に使えるサンプリング法が得られていない（右表参照）．

モデル次元 d	HMC	本研究
$10^0 \sim 10^4$	○	○
$10^4 \sim$	×	◎

現代のモデルは自然言語処理・画像解析・経済学・疫学をはじめとし，ほとんどの領域で $d = 10^n$ $(n \geq 4)$ などの高次元になることも多い．その場合はデータも巨大で複雑であるのが常である．

このような大規模モデル・大規模データの場面では，有効なサンプリング法が欠如しているために，ベイズ法が採用されにくい状況にある．そこで本研究では，**拡張性を持ったサンプリング法**の創出を目指す．

【研究計画】（続き） 適宜概念図を用いるなどして、わかりやすく記入してください。なお、各事項の字数制限はありませんが、全体で２頁に収めてください。様式の変更・追加は不可。

(2) 研究目的・内容等

① 特別研究員として取り組む研究計画における研究目的、研究方法、研究内容について記入してください。
② どのような計画で、何を、どこまで明らかにしようとするのか、特別研究員奨励費の応募区分（下記（※）参照）に応じて、具体的に記入してください。
③ 研究の特色・独創的な点（先行研究等との比較、本研究の完成時に予想されるインパクト、将来の見通し等）にも触れて記入してください。
④ 研究計画が所属研究室としての研究活動の一部と位置づけられる場合は申請者が担当する部分を明らかにしてください。
⑤ 研究計画の期間中に受入研究機関と異なる研究機関（外国の研究機関等を含む。）において研究に従事することも計画している場合は、具体的に記入してください。
（※）特別研究員奨励費の研究期間が 3 年の場合の応募総額は（A 区分）が 240 万円以下、（B 区分）が 240 万円超 450 万円以下（DC1 のみ）。2 年の場合は（A 区分）が 160 万円以下、（B 区分）が 160 万円超 300 万円以下。1 年の場合は（A 区分）が 80 万円以下、（B 区分）が 80 万円超 150 万円以下。（B 区分については研究計画上必要な場合のみ記入）

研究目標：次の２点を兼ね備えた拡張性に優れたサンプリング法の開発と実装を目指す

(i) **モデル拡張性**：高次元空間上の分布 P でも精度と速度が落ちないこと
(ii) **データ拡張性**：大規模なデータでも計算時間が爆発せず，現実的な時間内で実行可能であること

研究の特色とインパクト：サンプリングに注目することにより，モデル中心のパラダイムに資する

本研究では具体的なモデル $\{P_i\} \subset \mathcal{P}(\mathbb{R}^d)$ に特化した手法よりもむしろ，HMC に代わることができるような**モデルに依存せず普遍的に使える汎用サンプリング法の構成をめざす**のが特徴である．

種々のモデルに汎用的に使える手法の発展は，ベイズ法が広く応用されるために必要不可欠である．

ベイズ統計の悲願は，具体的な推論エンジンの設計にとらわれず，モデルの設計に集中出来る枠組みの達成にある．

	ベイズ法	非ベイズ法
アイデア	モデルベース	推論ベース
推論エンジン	**サンプリング法**（共通）	推論手法ごとに構成
長所	推論手法が統一なので モデリングに集中可能	専用に作っているので アルゴリズムが高速

そこで，本研究は開発したサンプリング法を Python パッケージの形で一般公開することも目指す．本研究により大規模モデル・大規模データでも効率よく推論できる手法が非専門家でも簡単に利用できるようになれば，ベイズ法を複雑な実社会の現象に応用することが産学両面で促進されると期待されるためである．

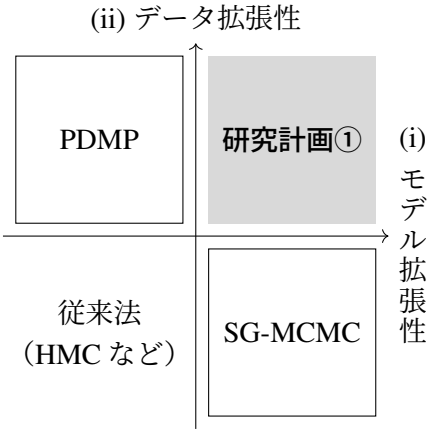
本研究は上述の研究目標に向けた研究計画 ①, ②, ③ の３部構成からなり，各アプローチに１年ずつかける．

研究計画①：２つの有望な方向の検討と限界の明瞭化

現状，**拡張性**を持つ次世代のサンプリング法として，次の２つが候補に挙がっている：

(1) **区分確定的マルコフ過程（PDMP）**
ランダムな時刻にランダムに変化する以外は確定的な動きをする連続時間マルコフ過程を用いて空間を探索する手法群

(2) **確率的勾配マルコフ連鎖モンテカルロ法（SG-MCMC）**
データの一部のみから計算した対数尤度の勾配の推定量を用いてエルゴード的なマルコフ連鎖を構成する MCMC 手法群



いずれの手法も **２つの拡張性** (i), (ii) のうち片方のみを満たさないため，**互いの弱点を補い合うことで新たな手法が創出できないかを探究する**．

PDMP はバイアスを導入しない効率的なサブサンプリングが可能で **(ii) データ拡張性** の条件を満たし，高次元での収束速度も期待されている [2] が，適用可能なモデルが限られており **(i) モデル拡張性** が課題である．一方で，SG-MCMC は勾配さえ推定可能であれば使える手法であり適用可能なモデルは広いが，高次元空間での収束レートが速くなく，**(ii) データ拡張性** が未だ達成されていない．

そこで，勾配の情報のみを用いた PDMP アルゴリズムを考案することで，**２つの拡張性**を同時に満たすアルゴリズムの創出を目指す．

(研究目的・内容等の続き)

本研究の独自性：連続時間極限という観点

研究計画①で取り上げた２つの手法（PDMP と SG-MCMC）は他の手法と比べて特殊なものである（そのために１年目で取り組む）．それはいずれも**連続時間の確率過程**をベースとした手法であるということである．

離散時間手法（従来法）	連続時間手法（拡張性あり）
メトロポリス法	PDMP
HMC	SG-MCMC
ギブス法	（以上の２つのみ）
粒子フィルター	
⋮	

そもそも，汎用的なサンプリング法（従って本研究で扱う手法）は，全てマルコフ連鎖モンテカルロ法と呼ばれる手法に属する．これは分布 P から直接サンプリングするのではなく，P に収束するマルコフ過程を構成し，その時間発展を追うことでサンプリングする，というものである．当該マルコフ過程がエルゴード性を持つならば，十分時間が経ったあとは P からの独立同分布なサンプルとみなせるというカラクリである．

計算機上で実装するため，最終的には離散化が必要である．そのため，基本的には最初から離散時間のマルコフ過程を構成するアルゴリズムが多いが，PDMP と SG-MCMC は**連続時間のマルコフ過程**を考え，これを最後に適当な時間幅で区切ってサンプルとする手法であり，この点が成功を収めてきた（右上図参照）．

中でも特に, PDMP は元々，離散時間ベースの手法の１つであるメトロポリス法を，連続時間極限 $\Delta t \to 0$ を考えることで，より効率的な空間の探索を可能にし，高い収束レートを達成した手法である [3].

研究計画②：連続時間極限をキーワードに既存のアルゴリズムを改良する

PDMP の成功を模倣し，既存のアルゴリズムの連続時間極限を考えることで，研究計画①で考えたサンプリング法から更なる拡張性の達成を目指す．

元の手法	連続時間極限	得たもの
メトロポリス法	PDMP	(ii) データ拡張性
粒子フィルター	（**研究計画②**）	（**研究計画②**）

特に, 粒子フィルターと呼ばれる手法に注目して，その連続時間極限を特定し，新たなアルゴリズムとして定式化し，**どのような拡張性を持つか検証する**．

粒子フィルターについては，先行研究 [4] が連続時間極限から得られるジャンプ測度から，アルゴリズムの最適なイベント時刻（方向転換，リサンプリングなど）への示唆を得ることに成功している．だが，その数学的な議論はほとんど単一の書籍 [5] にある定理を適用するのみで，極限として得られた過程の分析が出来ておらず，どのようなアルゴリズムが得られるのかが不透明である．

先行研究の受難は，書籍 [5] に丸投げされたジャンプあり確率過程への収束理論が，従来本分野では用いられなかった高度な数学を必要とするためである．この収束理論を専門とする数学者の多くは金融データの統計解析を主な応用先としている．サンプリング法などの計算統計の分野でもジャンプ付き連続時間確率過程が有用であり, 同様の数学が応用可能であるとは, まだ多くの人の知るところでない．申請者は日本の金融データの統計解析を牽引する研究グループの出身であり，**必要な数学を備えつつ，重要な応用先との稀有な交差点に居る**．

そこで，**研究計画②ではこの結果の改良と数学的定式化から着手し，極限として得られた過程をサンプリングに用いる「新たな PDMP」を検討する**．

研究計画③：連続時間極限をキーワードに提案アルゴリズムを比較する

研究計画①は PDMP と SG-MCMC の比較，研究計画②は PDMP と SG-MCMC 以外の連続時間手法の創出であった．研究計画③では，②で創出した手法を，①の比較の俎上に載せ，更なる検討と比較を重ねる．

在外研究の計画：確率過程論と機械学習数理との交流

本研究では海外研究グループとの交流が重要である．実際，研究計画②の一部は，関連する業績 1 の発表の際に Omar Chehab 氏との**ディスカッションを通じて着想**された．彼の所属する研究室が先行研究 [4] の出所である．彼らはその後，PDMP だけでなく機械学習数理にも重点を移したが，同様に連続時間極限をキーワードとしている．そこで，研究計画②に取り組む段階で CREST-ENSAE での滞在研究を計画している．

参考文献 (1) Duane, S., et al. 1987, Physics Letters B, 195, 216 (2) Fearnhead, P., et al. 2018, Statistical Science, 33, 386 (3) Peters, E. A. J. F. & de With, G. 2012, Physical Review E, 85 (4) Chopin, N., et al. 2022, The Annals of Statistics, 50, 3197 (5) Ethier, S. N. & Kurtz, T. G. 1986, Markov Processes

３．人権の保護及び法令等の遵守への対応 本項目は１頁に収めてください。様式の変更・追加は不可。

本欄には、「２.研究計画」を遂行するにあたって、相手方の同意・協力を必要とする研究、個人情報の取り扱いの配慮を必要とする研究、生命倫理・安全対策に対する取組を必要とする研究や安全保障貿易管理を必要とする研究など指針・法令等（国際共同研究を行う国・地域の指針・法令等を含む）に基づく手続が必要な研究が含まれている場合、講じる対策と措置を記入してください。

例えば、個人情報を伴うアンケート調査・インタビュー調査、行動調査（個人履歴・映像を含む）、国内外の文化遺産の調査等、提供を受けた試料の使用、侵襲性を伴う研究、ヒト遺伝子解析研究、遺伝子組換え実験、動物実験、機微技術に関わる研究など、研究機関内外の情報委員会や倫理委員会等における承認手続が必要となる調査・研究・実験などが対象となりますので手続の状況も具体的に記入してください。

なお、該当しない場合には、その旨記入してください。

本研究の遂行上最も関連するものは個人情報の取り扱いについてであるが, アンケート調査・インタビュー調査や行動調査等を実行することも極めて考えにくく，該当しないと言って良いと思われる.

４．【研究遂行力の自己分析】各事項の字数制限はありませんが、全体で２頁に収めてください。様式の変更・追加は不可。
本申請書記載の研究計画を含め、当該分野における⑴「研究に関する自身の強み」及び⑵「今後研究者として更なる発展のため必要と考えている要素」のそれぞれについて、これまで携わった研究活動における経験などを踏まえ、具体的に記入してください。

(1) 研究に関する自身の強み

先述の【研究計画】を遂行するために重要な資質は，以下の通りである：

(a) 統計計算：新たなサンプリング法を創出するためのアルゴリズム設計力・計算機科学の深い見識

(b) 数学：サンプリング法を解析する数学（確率過程の収束理論と Malliavin 解析）の深い素養

(c) 社会実装：提案手法を万人が利用できるパッケージへ実装する技術力と主体性

(d) コミュニケーション力：多様な研究者と交流して問題意識を共有し，共に問題解決へと向かう力

この４分野の全てにおいて高い適性を持つ者は，相当に限られるものと思われる.

以下，(a) 統計計算，(b) 数学，(c) 社会実装 そして (d) コミュニケーション力 を順に分析する．

(a) 統計計算について

アルゴリズムと計算機への深い理解

申請者は学部時代にコンピュータサークルで活動した経験があり，アルゴリズムと計算機科学に関する深い知識がある．実際，学部１年生時点で，世界最大の計算機セキュリティーコンテストである SECCON 決勝大会への出場経験をもつ．

コンピュータセキュリティーというと統計計算とは無関係に思えるが，「計算機に何ができるか」を表面的な理解にとどまらず「低いレイヤーではどうやって動いているのか」の深い知識を楽しく学ぶ大変良い題材となった．実際，申請者はリサーチ・アシスタントとして統計計算パッケージ Yuima の開発に従事しているが，その際全く違う２つのプログラム間でデータを受け渡すことが必要になった．これを文字列の受け渡しで行うという場当たり的な対処から，**C++というより低レイヤーの言語の知識を用いて，根本的な解決に導くことができた**．現在，一般公開への調整の最中である．

応用・実践への主体的関与

上述の事例は，通常の数学科生活では絶対得られなかったはずの他分野の知識でも，研究に関連する内容を申請者が主体的に獲得できる証左でもある．

また申請者は，現場での**実践の経験**も重要だと考え，学部時代に自ら経営コンサルティング会社にて**データサイエンティストとしてのインターンを志願**し，実際に１年に渡り継続的に関わった．

このように，申請者は理論を実践に還元するだけでなく，実践の場での問題を理論的な興味に還元し，双方向にシナジーを生み出すことができる．今後とも，「現場で必要とされている理論は何か」と「この理論がどのように現場に資するか」の複眼的思考を持ちながら理論研究を進めていきたい．

(b) 数学について

確固たる数学力

申請者は東大数学科において吉田朋広教授に指導教員をお願いし，４年次の１年をかけて Nualart & Nualart (2018) *Introduction to Malliavin Calculus* を読み，毎週 90 分を指導教員の前で発表した．初めは失敗ばかりであったが必死に準備するうちに，何も見ずに定義・定理・例・証明を説明する力が身についた．現在も，**このセミナーの経験が大きな数学的基礎体力となっている**.

確率過程の収束理論への深い造形

吉田教授の下で学びたかった理由は，金融データ解析の分野において林-吉田推定量に名前を残すなど，確率過程の収束理論を用いて大きな業績を上げていることを知っていたからである．申請者は当時から金融ではなく統計計算の分野に興味があったが，その理論の有用さを直感し，同様の数学は必ず応用先があるだろうという信念から，吉田教授に指導をお願いしたのであった．

実際，本研究で扱うサンプリング法の分野において，ここ数年**連続時間確率過程**が突如キーワードとして浮かび上がってきた．ここまでの奇跡は予想していなかったが，分野に貢献しながら自身の強みをさらに伸ばす格好の機会だと考え，連続時間極限を用いたサンプリング法をテーマに掲げた．

(研究遂行力の自己分析の続き)

(c) 社会実装について

実装力：中規模開発と継続運用の経験

データサイエンティストとしても主体的に活動し，公開ライブラリ OSS を有効活用して Twitter から特定の単語を含むツイートを自動収集し[1]形態素分析を通じて分析するという**複数のパイプラインを持ったツールを一人で開発した**.

加えて，このパイプラインはチームの大きな役に立ち，大規模言語モデルを活用した感情分析など，会社の他チームによる種々のタスクに応用された．すなわち，申請者は自分でプログラムを開発するだけでなく，それを実際に他人に使ってもらう力も持つ.

問題解決能力：データから統計分析を通じ経営提言まで

データサイエンティストとしてのインターン中には，実際のものづくり企業に対して，データ分析の結果をレポートの形にまとめ，事業改善の提言を行った．その過程で，経営学に関してはメンタリングを受けながら，**制作工程から収集されたデータの分析から事業改善の提言までを一人で実行**した.

(d) コミュニケーション力

高い分野横断的コミュニケーション力

東京大学を卒業後も先端科学技術研究センター連携研究員という立場で継続的な共同研究や意見交換を継続している．実際，AI の信頼性から知的財産に関する問題まで，10 を超える（英語）シンポジウムと研究会を主催し，登壇者とのディスカッションや運営に貢献した.

その結果，統計と機械学習をはじめとして**極めて広い分野の研究者がサンプリング法の発展状況に強い興味を抱いている**ことがわかり，自身の研究に対する深いモチベーションを得ている.

国際コミュニケーション能力

申請者は**日本語だけでなく中国語も母語レベルに話せ，また英語も TOEFL iBT で 100 点の語学力を持つ**．そのため，国際学会によっては参加者の大半と先方の母語で会話をすることが出来る.

実際，現状までの研究活動も英語が中心であり，本研究も一部英語でのポスター発表（業績 1）でのディスカッションから着想を得たものである．国内で「サンプリング法」または「ベイズ法」を専門とする研究者は少なく，この 2 つにまたがる研究となるとさらに少ないため，申請者がすでに現段階から国際交流を交えながら研究を進めていることは重要なことである.

業績 1 Shiba, H. *A Recent Development of Particle Filter*. MLSS 2024, Mar. OIST, Okinawa. ポスター発表（査読なし）.　本発表では研究計画②の粒子フィルターの極限の特定を試みた.

業績 2 Shiba, H. 「新時代の MCMC を迎えるために」. 統数研オープンハウス, 2024 May. ポスター発表（査読なし，採録予定）.　本発表では研究計画①の PDMP の性質と未解決問題をサーベイした.

(2) 今後研究者として更なる発展のため必要と考えている要素

成果の少なさ → 落ち着いて深い数理的素養の獲得を目指す

申請者はまだ研究成果が少ない状態である．これは，申請者の研究分野が，手法開発だけでなく，これを数学的に理解することで統一的な枠組みを提供しようとするものであるためである．研究対象こそ応用志向であれど，必要であれば数学的結果も創出することが必要である．そのためには，高度な数学（確率過程・確率解析・関数解析・偏微分方程式）への熟練が必要であり，時間はかかるかもしれないが腰を落ち着けた習得が肝心だと考えている.

高度な数学的概念もわかりやすく説明する力 → 幅広い背景の研究者と研究交流をする

本研究分野は物理学とも深い関わりを持つ．実際，最も代表的なサンプリング法である MCMC は物理シミュレーションのために開発されたものである．そこで申請者は物理学・機械学習など幅広い分野の学会に出席した．その結果，自分の研究成果を他の分野の研究者に説明するためには，自身の分野に詳しくなるだけでなく，他分野の問題意識と専門用語を深く学ぶ必要もあると悟った.

[1]2023 年 2 月当初は可能であったが，Twitter のサービスが X に変更後はスクレイピングは不可能になり，現在はスクリプトとして公開・継続運用はされていない.

５．【目指す研究者像等】 各事項の字数制限はありませんが、全体で１頁に収めてください。様式の変更・追加は不可

日本学術振興会特別研究員制度は、我が国の学術研究の将来を担う創造性に富んだ研究者の養成・確保に資することを目的としています。この目的に鑑み、(1)「目指す研究者像」、(2)「目指す研究者像に向けて特別研究員の採用期間中に行う研究活動の位置づけ」を記入してください。

(1) 目指す研究者像 ※目指す研究者像に向けて身に付けるべき資質も含め記入してください。

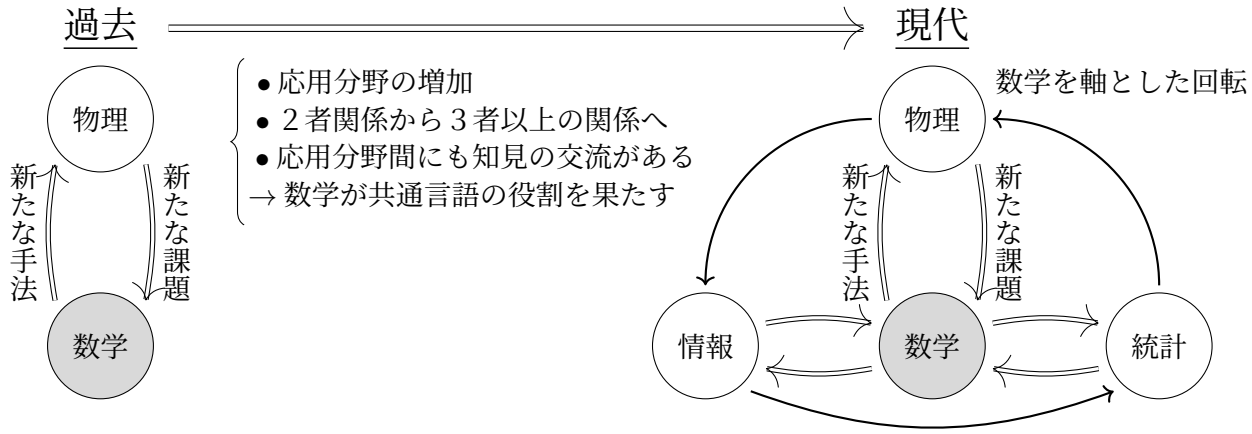
目指す研究者像：数学に軸足を置いた応用ができる「真の意味の数学者」

申請者は**数学に軸足を置いた応用ができる数学者**になりたい．これについては，日本の数学者岡潔が「数学の応用には，真の意味の数学者をじかに使うのが最も簡単で，最も先鋭で，しかも適用範囲が比較にならないほど広い」という言葉を残している．

申請者は，この岡潔のいう「**真の意味の数学者**」に当たる資質を身につけた応用数学者となることで，広い範囲の分野に貢献する仕事ができる研究者を目指している．

身に付けるべき資質：(a) 相互理解を促進する (b) 共通言語を提供できる 力

申請者は，現代の応用数学では，**数学を軸とした「回転」**が起こっていると考える．



そこで申請者は，「**真の意味の数学者**」として必要な資質は次の２つだと考え，己に課している：

(a) 相互理解：各分野でどのような相互理解が必要とされているかを見抜く洞察力

(b) 共通言語：相互理解の場として共通言語を与え，その理解を深める数学力

(2) 上記の「目指す研究者像」に向けて、特別研究員の採用期間中に行う研究活動の位置づけ

本研究は２つの資質 (a) 相互理解の促進 と (b) 共通言語の提供 を鍛えるための格好のテーマとなっている．

(a) 本研究は各応用分野の相互理解のために極めて肝心なテーマである

申請者は数学を軸としつつも，種々の応用分野の学習を欠かさなかった．その中で，現在の指導教員を通じて，統計計算とサンプリング法という，**極めて多くの分野が交差する魅力的な研究テーマ**に出会うことができた．現代のサンプリング法は統計や機械学習で必要不可欠な技術であるだけでなく，もともと物理学でシミュレーションのために開発されたものであり，現在では物性科学一般で広く使われる手法になっている．まさに多くの応用分野のハブとしての役割を果たしていると言っても過言でないテーマである．

(b) 本研究は各応用分野に強力な共通言語を提供する

サンプリング法においてマルコフ過程が必要不可欠な役割を果たす．マルコフ過程を分析する際は $\mathcal{P}(\mathbb{R}^d)$ 上の（決定論的な）力学系として捉える見方が極めて自然になる．有限次元の幾何学には多くの理論の蓄積があるが，無限次元空間内の凸集合である $\mathcal{P}(\mathbb{R}^d)$ 上の力学系にはまだ理解が及ばない部分が多い．近年，機械学習や最適輸送の分野でも $\mathcal{P}(\mathbb{R}^d)$ の解明が重要になっており，サンプリング法との関係性の解明が切望されている．つまり，$\mathcal{P}(\mathbb{R}^d)$ は**共通理解の場を提供する言語として更なる理解が待たれている**．