統計物理学の視点で読み解く 機械学習の理論

ほの

2025 年 9 月 19 日

概要

近年、目覚ましい発展を遂げている**機械学習**と、物理学の一分野である**統計物理学**。これら二つの分野は、一見すると全く異なる学問体系に思えるかもしれません。しかし、その根底には驚くほど深く、美しい理論的なつながりが存在します。多数の要素が相互作用する系が全体としてどのような振る舞いを示すか、という根源的な問いは、両分野に共通するテーマです。

本書は、機械学習の応用事例を紹介する多くの入門書とは一線を画します。その目的は、統計物理学という**羅針盤**を用いて、現代機械学習の理論体系がどのように構築されてきたのか、その理論的な源流を解き明かすことにあります。なぜボルツマンマシンは物理学のイジングモデルと同じ形をしているのか。なぜ深層学習の階層構造は、物理学のくりこみ群とこれほど似ているのか。本書は、こうした問いに答えることを目指します。

執筆にあたり、本書では数学的な厳密性のみを追求するのではなく、物理的な直観やアナロジーを大切にしました。**なぜ**そのような理論が生まれたのか、そのモチベーションを理解することで、読者の皆さんが機械学習の数式やアルゴリズムの背後にある豊かな構造を掴む手助けとなることを願っています。

本書を読み終えたとき、これまでブラックボックスに見えていたかもしれない機械学習モデルが、統計物理学の言葉で語りかける、秩序とゆらぎの美しいシステムとして見えてくるでしょう。

目次

第 I 部

序論 ー二つの分野の共鳴ー

1　統計物理と機械学習の歴史的邂逅

—物理学者がシリンダーの中の気体分子の振る舞いを追い、情報科学者がコンピュータに猫の画像を認識させる—この二つの営みは、全く異なる世界に属しているように見えるかもしれません。片や、自然界の法則を探求する物理学。もう片や、データから知的な振る舞いを構築する情報科学。しかし、もしこの二つの分野が、同じ山の頂を異なるルートから目指している登山隊だとしたらどうでしょうか。

本書でこれから解き明かしていく物語の核心は、まさにそこにあります。すなわち、**機械学習**と**統計物理学**は、その根底において驚くほど深く共鳴しあっている、という事実です。

この歴史的な邂逅が明確な形で現れたのは、1980 年代のことでした。当時、人工知能の研究は一つの停滞期にありましたが、物理学者のジョン・ホップフィールドが発表した一つの論文が、新たな息吹を吹き込みます。彼が提案した**ホップフィールド・ネットワーク**は、相互に結合した多数の単純なニューロン（神経細胞）のモデルが、全体として「記憶」や「計算」といった創発的な機能を持つことを示しました。

このモデルの画期的な点は、その動作原理が物理学における磁性体の理論、特に**スピングラス**の理論と深く結びついていたことです。スピングラスとは、原子間の相互作用が複雑に入り混じり、「フラストレーション」（どの向きを向けば安定するのか決められない状態）を抱えた特殊な磁石のことです。このフラストレーションが、無数の安定状態を持つ複雑な**エネルギー地形（ランドスケープ）**を生み出します。ホップフィールドは、このエネルギー地形の谷底を「記憶されたパターン」と見なすことで、物理学の安定状態を求めるプロセスを、連想記憶という計算問題に見事に結びつけたのです。

この流れは、ジェフリー・ヒントンらによる**ボルツマンマシン**の提案によって、さらに決定的となります。その名の通り、ボルツマンマシンは統計物理学の根幹をなす**ボルツマン分布** $P(\boldsymbol{x}) \propto \exp(-E(\boldsymbol{x})/T)$ をその動作原理に直接取り入れました。ここに至り、ネットワークの状態の「エネルギー」を定義し、その確率分布を考えるという統計物理学の思考法そのものが、機械がデータを「学習」するためのアルゴリズムとして再定式化されたのです。

このように、機械学習と統計物理学の出会いは、単なる偶然の産物や表面的なアナロジーではありません。それは、**「多数の単純な要素の相互作用から、いかにして複雑で知的な全体的振る舞いが生まれるか」**という、両分野が共有する根源的な問いに対する、必然的な邂逅だったのです。本章では、この歴史的な出会いを出発点として、二つの分野が織りなす豊かな理論的世界への扉を開いていきます。

1.1　共有された根源的な問い

前節で述べた歴史的な邂逅は、単なる偶然ではありませんでした。それは、統計物理学と機械学習が、その核心に全く同じ構造の問いを抱えていたからこそ、起こるべくして起きた必然でした。その問いとは、次のように要約できます。

根源的な問い

無数の単純な構成要素が、単純なルールに従って相互作用するとき、そこからいかにして複雑で、秩序だった「全体としての振る舞い」が生まれるのか？

この問いの構造を、両分野の具体的な対象に当てはめて考えてみましょう。

統計物理学の場合

統計物理学が対象とするのは、例えばコップ一杯の水です。この中には、10^{23} 個という天文学的な数の水分子（H_2O）が存在します。個々の水分子は、隣接する分子と比較的単純な電磁気的な力（水素結合など）で相互作用しているに過ぎません。これが**ミクロなルール**です。

しかし、この単純なミクロなルールから、私たちは「水」が示す驚くほど多様な**マクロな振る舞い**を知っています。温度を下げれば、分子が綺麗に整列した固体の「氷」になり、温度を上げれば、分子が自由に飛び回る気体の「水蒸気」になります。氷が水に溶ける「融解」や、水が水蒸気になる「蒸発」は、**相転移**と呼ばれる劇的な全体的変化です。

物理学者の興味は、個々の水分子の運動を一つ一つ追跡することにはありません。そうではなく、温度や圧力といったマクロなパラメータを変化させたときに、系全体がどのように応答するのか、その統計的な性質を理解することにあります。

機械学習の場合

一方、現代の機械学習、特に深層学習が対象とするのは、何百万、何千万というパラメータ（重み）を持つ巨大なニューラルネットワークです。個々の人工ニューロンは、物理の構成要素よりもさらに単純です。入力信号に重みを掛けて足し合わせ、活性化関数という単純な非線形関数を適用するだけ。これがここでの**ミクロなルール**です。

しかし、この単純な計算ユニットを膨大な数だけ組み合わせ、適切にその結合強度（重み）を調整することで、ネットワーク全体は驚くべき**マクロな知能**を獲得します。例えば、あるネットワークは写真に写った動物が「猫」であると識別し、別のネットワークは自然言語を流暢に翻訳します。

機械学習の研究者の興味もまた、個々のニューロンの活動を詳細に監視することにはありません。彼らの目的は、学習データという環境に適応させることで、ネットワーク全体が賢明な判断を下す「良い安定状態」へと自律的にたどり着くように、その学習プロセスを設計することです。

結論：同じ構造、異なる舞台

このように、舞台は「自然」と「情報」で異なりますが、両者が解き明かそうとしている問題の構造は全く同じです。**ミクロな相互作用**から**マクロな創発現象**への橋渡しを理解すること。この共通の課題意識こそが、統計物理学の洗練された理論的道具立てが、機械学習という新たな分野で強力な武器となる土壌だったのです。

1.2 原点：スピングラス理論とホップフィールドネットワーク

前節で提示した「根源的な問い」への橋渡しは、1982 年、物理学者ジョン・ホップフィールドによって具体的な形で示されました。彼が提案した**ホップフィールドネットワーク**は、その後の機械学習と物理学の歴史的な共鳴の、まさに「原点」と言える存在です。

ホップフィールドネットワーク：記憶を「地形の谷」として捉える

ホップフィールドネットワークは、相互に結合した多数の人工ニューロンから構成される、一種の**連想記憶モデル**です。その動作は非常に直観的です。

- **記憶の貯蔵**：ネットワークにいくつかのパターン（例えば、白黒の画像）を「記憶」させます。これは、ニューロン間の結合の強さ（シナプス荷重 W_{ij}）を調整することで行われます。
- **記憶の想起**：不完全な、あるいはノイズの乗ったパターンをネットワークに入力します。すると、各ニューロンが状態を次々と更新していき、ネットワーク全体が最終的にもとの完璧な記憶パターンへと収束（想起）します。

この想起プロセスの鍵となるのが、ネットワーク全体の状態に対して定義される**エネルギー関数** E です。

$$E = -\frac{1}{2}\sum_{i \neq j} W_{ij} S_i S_j$$

ここで、S_i はニューロン i の状態（例えば $+1$ か -1）、W_{ij} はニューロン i と j の結合の強さを表します。ネットワークは、この**エネルギー** E **が低くなるように**、自律的に状態を変化させていきます。

重要なのは、あらかじめ記憶させたパターンが、このエネルギー関数の作る「地形」の**安定な谷底（局所的極小点）**に対応するように設計されている点です。ノイズの多い入力から出発することは、いわば山の斜面にボールを置くようなものです。ボールが自然に転がり落ちて最も近い谷底に落ち着くように、ネットワークの状態もエネルギーの坂道を下り、最も近い記憶パターンへと引き込まれていくのです。

スピングラス理論との出会い

ホップフィールドが物理学者であったことは、ここで決定的な意味を持ちます。彼は、自らが定義したネットワークのエネルギー関数が、物性物理学で長年研究されてきた**スピングラス**と呼

ばれる物質のエネルギーモデル（ハミルトニアン）と、数学的に完全に同一であることに気づきました。

スピングラスとは？

スピングラスとは、磁性原子（スピン）の向きの間に働く力が、場所によって「同じ向きを向け（強磁性的）」と「逆の向きを向け（反強磁性的）」という指令がランダムに混在している特殊な磁性体です。全ての指令を同時に満たすことができないため、スピンは**フラストレーション**と呼ばれる状態に陥り、極低温に冷やしても単純な秩序状態になれず、極めて複雑なエネルギー地形を持つ多数の安定状態（準安定状態）に行き着きます。

この対応関係は、衝撃的でした。

ホップフィールドネットワーク	$\longleftrightarrow$	スピングラス
ニューロンの状態 S_i	$\longleftrightarrow$	スピンの向き σ_i
シナプス荷重 W_{ij}	$\longleftrightarrow$	スピン間の相互作用 J_{ij}
エネルギー関数 E	$\longleftrightarrow$	ハミルトニアン $\mathcal{H}$
記憶されたパターン	$\longleftrightarrow$	準安定状態（エネルギーの谷）
想起のプロセス	$\longleftrightarrow$	物理的な緩和プロセス

この発見により、単なるアナロジーではない、**数学的な同型性**が明らかになったのです。それまで物理学者たちがスピングラスを解析するために発展させてきた難解な数学的手法（レプリカ法など）が、ニューラルネットワークの性質を解明するために直接使えるようになりました。例えば、ネットワークが記憶できるパターンの上限（記憶容量）はどれくらいか、といった問いに、物理学の言葉で理論的に答える道が拓かれたのです。

このホップフィールドネットワークとスピングラス理論の出会いこそ、物理学の思考法が機械学習の理論構築に大きく貢献する、壮大な物語の幕開けでした。

1.3 本書の羅針盤

これまでの節で見てきたように、統計物理学と機械学習の間のつながりは、単なる歴史的な逸話や表面的な類似点に留まるものではありません。それは、両分野が同じ構造を持つ問題に取り組んできたがゆえの、深く本質的な結びつきです。一見すると無秩序で、様々な手法が乱立しているように見える機械学習の理論の世界にも、物理学の視点を通すことで見えてくる、確かな秩序と道筋が存在します。

本書の目的は、その道筋を照らし出すことです。そのために、私たちは統計物理学の洗練された概念体系を**「羅針盤」**として手に取ります。この羅針盤は、私たちが機械学習の広大な理論の海を航海する際に、現在地を確かめ、進むべき方向を示してくれます。

本書の旅路は、以下のように進んでいきます。

1. まず、両分野の**「共通言語」**を学びます。機械学習における確率的な「学習」や「推論」と

いった概念が、いかにして統計物理学の「分配関数」や「自由エネルギー」といった中心的な量と結びつくのかを明らかにします。（第 II 部）

2. 次に、共通の課題、すなわち「複雑すぎて厳密には解けない」問題に、両分野がどのように挑んできたかを見ます。物理学の**「平均場近似」**が、機械学習の世界では**「変分推論」**という名で呼ばれる強力な武器となっていることを理解します。（第 III 部）
3. そして、現代機械学習の中核である深層学習の謎に迫ります。ニューラルネットワークの階層構造を物理学の**「くりこみ群」**という考え方で読み解き、また拡散モデルのような最新の生成モデルが**「非平衡物理学」**の過程として見事に記述できることを学びます。（第 IV 部）
4. 最後に、学習というプロセスそのものを物理現象として捉えます。学習の進展に伴い、ネットワークの性質が劇的に変化する様を、物質の**「相転移」**として理解する視点を獲得します。（第 V 部）

この羅針盤が指し示すのは、個々のアルゴリズムの操作方法だけではありません。それらの理論が**なぜ**そのような形をしているのか、その背後にある統一的な構造への深い洞察です。

第 II 部

共通言語としての確率分布とエネルギー

2　統計物理学の視点 ー分配関数と自由エネルギーー

この部では、機械学習と統計物理学の理論的なつながりを具体的に見ていきます。まずはその出発点として、物理学者が複雑な系をどのように記述するのか、その基本的な「言語」を学びましょう。中心となるのは、**分配関数**と**自由エネルギー**という二つの重要な概念です。

2.1　最大エントロピー原理とカノニカル分布

物理学的な対象（例えば、箱の中の気体）を考えます。この系が取りうる状態（気体分子の位置や運動量など）は天文学的な数にのぼり、その全てを把握することは不可能です。我々が知ることができるのは、系のエネルギーの平均値 $\langle E \rangle$ のような、ごく一部のマクロな情報だけです。

では、この限られた情報だけを頼りに、系が特定の状態 $\boldsymbol{x}$ をとる確率 $P(\boldsymbol{x})$ を推定するには、どのような分布を考えれば最も公平で、客観的と言えるでしょうか。ここで極めて強力な指導原理となるのが、**最大エントロピー原理**です。

最大エントロピー原理

私たちが持つ知識（情報）を制約条件として満たす確率分布の中で、情報エントロピー（シャノンエントロピー）S を最大にするものが、最も偏りのない（unbiased）確率分布である。

$$S = -\sum_{\boldsymbol{x}} P(\boldsymbol{x}) \ln P(\boldsymbol{x})$$

情報エントロピーは、確率分布の「不確かさ」の度合いを表す量です。エントロピーを最大化するとは、すなわち**「知っていること（制約条件）以外は、何も知らない（最大限、不確かである）」**という立場をとることを意味します。何か特定の知識がない限り、すべての可能性を平等に扱うという、科学的に誠実な態度と言えるでしょう。

さて、この原理を物理的な系に適用してみましょう。ここでの制約条件は、系のエネルギーの平均値が特定の値 $\langle E \rangle$ になる、ということです。

$$\sum_{\boldsymbol{x}} P(\boldsymbol{x}) E(\boldsymbol{x}) = \langle E \rangle$$

この制約の下でエントロピーを最大化するような確率分布 $P(\boldsymbol{x})$ を数学的に求めると（詳細は割愛しますが、ラグランジュの未定乗数法を用います）、その解は常に以下の指数関数の形をとります。

カノニカル分布（ボルツマン分布）

エネルギーの平均値が固定された系において、最大エントロピー原理から導かれる、状態 $\boldsymbol{x}$ の出現確率は以下の式で与えられる。

$$P(\boldsymbol{x}) = \frac{1}{Z}\exp(-\beta E(\boldsymbol{x}))$$

ここで、$E(\boldsymbol{x})$ は状態 $\boldsymbol{x}$ のエネルギー、β は平均エネルギーを定めるパラメータ（物理学では逆温度 $1/k_BT$ に対応）、Z は確率の総和を 1 にするための規格化定数である。

このカノニカル分布は、統計物理学における最も重要な結果の一つです。これは、私たちが平均エネルギーというマクロな量しか知らないとき、系がエネルギーの低い状態を指数関数的にとりやすい、という非常に自然な結論を導き出します。

重要なのは、この分布が物理学の特別な法則から導かれたのではなく、**情報が不完全な場合の最も合理的な推論の帰結**であるという点です。この考え方は、後に見るように、データからモデルを構築する機械学習の分野においても、全く同じ形で現れることになります。

2.2　分配関数と自由エネルギーの物理的意味

前の節で導出したカノニカル分布には、Z という規格化定数が現れました。

$$P(\boldsymbol{x}) = \frac{1}{Z}\exp(-\beta E(\boldsymbol{x})), \quad \text{ただし} \quad Z = \sum_{\boldsymbol{x}} \exp(-\beta E(\boldsymbol{x}))$$

この Z は、単に確率の総和を 1 にするための数学的な要請から導入されましたが、実はそれ自体が系の熱力学的な情報をすべて内包する、極めて重要な物理量です。この Z を**分配関数 (partition function)** と呼びます。

分配関数：系の全情報を集約した関数

分配関数がなぜ重要かというと、それを用いることで、系の様々なマクロな物理量を計算できるからです。つまり、ミクロな状態（エネルギー $E(\boldsymbol{x})$）の情報から、マクロな量（平均エネルギーなど）を導出する「橋渡し」の役割を果たします。

例えば、系のエネルギーの平均値 $\langle E \rangle$ は、分配関数 Z の対数を β で偏微分することで簡単に計算できます。

計算：平均エネルギー $\langle E \rangle$ の導出

平均エネルギーの定義は以下の通りです。

$$\langle E \rangle = \sum_{\boldsymbol{x}} E(\boldsymbol{x})P(\boldsymbol{x}) = \frac{1}{Z}\sum_{\boldsymbol{x}} E(\boldsymbol{x})\exp(-\beta E(\boldsymbol{x}))$$

一方、分配関数 Z を β で偏微分すると、

$$\frac{\partial Z}{\partial \beta} = \frac{\partial}{\partial \beta} \sum_{\boldsymbol{x}} \exp(-\beta E(\boldsymbol{x})) = \sum_{\boldsymbol{x}} (-E(\boldsymbol{x})) \exp(-\beta E(\boldsymbol{x}))$$

となります。この二つの式を比較すると、以下の関係が導かれます。

$$\langle E \rangle = -\frac{1}{Z}\frac{\partial Z}{\partial \beta} = -\frac{\partial (\ln Z)}{\partial \beta}$$

このように、一旦**分配関数 Z という一つの量を計算してしまえば、そこから微分などの操作によって様々なマクロな量が導出できる**のです。分配関数は、系の統計的な性質をすべて集約した「生成関数」としての役割を持っています。

自由エネルギー：エネルギーとエントロピーのバランス

分配関数 Z は非常に便利ですが、その値は天文学的に大きくなることがあり、対数をとった $\ln Z$ の方が扱いやすい場合が多くあります。物理学では、この $\ln Z$ と密接に関連した**ヘルムホルツの自由エネルギー F** という量を定義します。

$$F = -\frac{1}{\beta} \ln Z \quad (\text{物理学では } F = -k_B T \ln Z \text{ と書くことが多い})$$

自由エネルギー F の物理的な意味は、システムの安定性を決定する指標である、という点にあります。一定の温度・体積に置かれた系は、**自由エネルギー F が最も小さくなる状態**を安定な平衡状態として選択します。

なぜなら、自由エネルギーは、系の内部エネルギー $\langle E \rangle$ とエントロピー S の間のバランスを表しているからです。具体的には、$F = \langle E \rangle - TS$ という関係が成り立ちます（T は温度）。

自由エネルギーの最小化原理

系は、以下の二つの相反する要求のバランスをとることで、安定な状態（F が最小の状態）に落ち着きます。

- **エネルギー $\langle E \rangle$ をできるだけ低くしたい**：これは、より秩序だった安定な状態へ向かう傾向を表します。
- **エントロピー S をできるだけ高くしたい**：これは、より乱雑で、可能な状態の数が多い状態へ向かう傾向を表します。

温度 T は、どちらの要求を優先するかの「重み」の役割を果たします。低温ではエネルギー最小化が優先され、高温ではエントロピー最大化が優先されます。

まとめると、分配関数 Z は系の情報を数学的に集約したものであり、自由エネルギー F はその情報から系の安定性を物理的に議論するための量と言えます。この二つの概念は、統計物理学の視点から機械学習を理解する上で、繰り返し登場する重要なキーワードとなります。

3　機械学習の視点 ーベイズ推論と生成モデルー

前のセクションでは、統計物理学の中心的な概念であるカノニカル分布、分配関数、そして自由エネルギーについて学びました。これらは、情報が不完全な中で物理系を記述するための強力な理論的枠組みでした。

このセクションでは、視点を大きく転換し、機械学習の世界に入っていきます。そして、一見すると全く異なるこの分野の根底に、実は統計物理学と全く同じ数学的構造と課題意識が存在することを見ていきます。特に、**ベイズ推論**と**生成モデル**という二つの重要な考え方を通して、その対応関係を明らかにしていきます。

3.1　学習とは確率分布の推定である

機械学習における「学習」とは、一体何をしているのでしょうか。最も単純なイメージは、データにフィットするような「関数を見つける」ことかもしれません。例えば、入力 x（身長）から出力 y（体重）を予測する問題を考えたとき、データ点 (x_i, y_i) をうまく説明するような直線 $y = ax + b$ のパラメータ a, b を見つける、といった具合です。

しかし、この考え方には限界があります。現実のデータは、常にばらつきやノイズを含んでいます。同じ身長の人でも、体重は様々です。決定的な関数を一つ見つけるよりも、より強力で柔軟なアプローチは、「学習」を**データが生成される確率分布を推定する**ことだと捉え直すことです。

機械学習の確率的視点

機械学習の目的は、観測されたデータ（訓練データ）を手がかりに、その背後にある**確率分布を推定**することである。

この視点に立つと、様々な機械学習のタスクを統一的に理解することができます。

教師あり学習の場合

身長 x から体重 y を予測する問題は、「x が与えられたときの y の条件付き確率分布 $P(y|x)$ を推定する」問題として再定式化されます。

例えば、$P(y|x)$ が平均 $ax + b$、分散 σ^2 の正規分布に従うと仮定してみましょう。

$$P(y|x; a, b, \sigma^2) = \mathcal{N}(y|\mu = ax + b, \sigma^2)$$

このとき「学習」とは、手元にある訓練データ $\{(x_i, y_i)\}$ が、この確率分布から生成されたものである可能性（尤度、ゆうど）が最も高くなるように、パラメータ a, b, σ^2 を調整するプロセスになります。これにより、単一の予測値だけでなく、予測の不確かさ（分散 σ^2）も同時に得ることができます。

教師なし学習の場合

入力と出力のペアがない、データそのものの特徴を捉えたい場合もあります。例えば、たくさんの猫の画像データを集めてきたとします。このタスクは、「猫の画像」の確率分布 $P(\text{画像})$ を推定する問題と考えることができます。

この確率分布 $P(\text{画像})$ を学習することができれば、私たちはその分布からサンプリングすることで、**本物そっくりの新しい猫の画像を生成する**ことが可能になります。このようなモデルは**生成モデル (generative model)** と呼ばれ、近年の機械学習において非常に重要な役割を担っています。

物理学との接点

これまでの議論で見てきた統計物理学の世界観と、機械学習の確率的な世界観。これら二つの話は、ここで本書の根幹をなす一つのアイデアによって、強力に結びつきます。ここが、本書における最も重要な視点の転換点です。

本書の中心的アナロジー：学習と物理系の同一視

私たちが機械学習を通じて推定したい、データの背後にある未知の確率分布 $P(\text{データ})$ を、**物理的な系が従うカノニカル分布 $P(\boldsymbol{x}) = \frac{1}{Z}\exp(-E(\boldsymbol{x}))$ と同一のものと見なす。**

この大胆なアナロジーを受け入れると、機械学習における抽象的な活動が、途端に物理的なイメージを伴った具体的なプロセスとして見えてきます。この視点の転換は、「モデル設計」と「学習」という二つの中心的な活動の捉え方を、根底から変革する力を持っています。

1. **モデル設計から、エネルギー関数の設計へ**
   「機械学習モデルを設計する」という作業は、もはや抽象的な数式やネットワーク構造の構築だけを意味しません。それは、**系の物理的な性質を決定づける「エネルギー関数 $E(\boldsymbol{x})$」を設計する**という、より具体的で物理的な作業へと翻訳されます。私たちの目標は、観測データとしてあり得そうな状態には低いエネルギーを、あり得そうにない状態には高いエネルギーを与えるような、適切なエネルギーの地形（ランドスケープ）をデザインすることになるのです。
2. **学習から、エネルギー地形の彫刻へ**
   そして、「学習」というアルゴリズミックなプロセスは、**観測された訓練データという「現実」に合うように、エネルギー関数の形を動的に調整していく物理的なプロセス**として再解釈されます。それはまるで、手元のデータという名の彫刻刀を使い、エネルギーという名の素材を削っていく作業です。データ点が存在する場所のエネルギーを低く（谷を深く掘り）、存在しない場所のエネルギーを高く（山を築く）していく。このエネルギー地形を「彫刻」していくプロセスこそが、学習なのです。

この「物理系との同一視」という視点こそが、統計物理学の分野で培われてきた膨大な理論的

資産を、機械学習という新しい問題領域を解き明かすための、強力なツールキットとして解放する鍵となります。

3.2 最大エントロピー原理は機械学習でも自然な仮定か？

前の節で、私たちは統計物理学におけるカノニカル分布が、平均エネルギーという制約の下でエントロピーを最大化する、最も「偏りのない」分布として導かれることを見ました。物理的な平衡状態を記述する上で、これは非常に自然な要請です。

しかし、これを機械学習の文脈に持ち込む際には、一度立ち止まって考える必要があります。私たちの目標は、手元にある有限個の訓練データ $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ から、その背後にある真のデータ生成分布 $P_{\text{true}}(\boldsymbol{x})$ を推定することです。このとき、「制約条件以外は全て未知とみなす」という最大エントロピーの立場は、果たして正当化されるのでしょうか。

この問いに対する答えは、私たちがモデルに何を「制約」として課すか、という解釈に懸かっています。

制約条件としての「データの特徴量」

機械学習の文脈では、物理学における「系の平均エネルギーを固定する」という制約を、より一般的に**「モデルが生成するデータの統計的特徴量の期待値を、訓練データのそれと一致させる」**という制約に置き換えて解釈します。

例えば、画像データを扱うモデルを考えましょう。私たちは、モデルに以下のような性質を学習してほしいと考えるかもしれません。

- 「生成される画像の平均的な明るさが、訓練データの画像の平均的な明るさと一致してほしい」
- 「特定のフィルター（例えば、縦縞を検出するフィルター）をかけたときの応答の平均値が、訓練データとモデルで一致してほしい」

これらの一つ一つが、モデルに対する制約条件となります。ある特徴量を記述する関数を $f_k(\boldsymbol{x})$ とすると、制約条件は数学的に次のように書けます。

$$\langle f_k(\boldsymbol{x}) \rangle_{\text{model}} = \langle f_k(\boldsymbol{x}) \rangle_{\text{data}}$$

ここで、$\langle \cdot \rangle_{\text{model}}$ はモデルの分布 $P(\boldsymbol{x}|\theta)$ での期待値、$\langle \cdot \rangle_{\text{data}}$ は訓練データから計算される経験的な平均値を意味します。

機械学習における最大エントロピー原理の解釈

「データから抽出した特定の特徴量の期待値を再現する」という制約条件の下で、それ以外の余計な仮定（未知の相関など）を一切置かない、最も不確実な（＝エントロピーが最大の）確率分布を構築する。

この立場は、**オッカムの剃刀**、すなわち「必要以上に複雑な仮定を置くべきではない」という

科学の基本原則に合致します。データが示している統計的な証拠は忠実に再現するが、データが直接語っていないことについては、何も知らないという謙虚な態度をとる。これが、機械学習において最大エントロピー原理に基づくモデル（エネルギーベースモデルなど）が強力である理由の一つです。

仮定の限界と注意点

一方で、この仮定が万能でないことも認識しておく必要があります。

1. **有限なデータの問題**：私たちが計算できるのは、あくまで有限な訓練データにおける特徴量の平均値 $\langle f_k(\boldsymbol{x}) \rangle_{\text{data}}$ です。これは、真の分布 $P_{\text{true}}(\boldsymbol{x})$ における期待値の推定値に過ぎません。もし訓練データが偏っていたり、サンプル数が少なかったりすれば、私たちの課す制約自体が真実からずれている可能性があります。
2. **モデルの表現力の問題**：最大エントロピー原理は、与えられた制約を満たす中で「最も自然な」分布を選び出しますが、その分布がそもそも私たちのモデル（例えば、特定の構造を持つニューラルネットワークで定義されたエネルギー関数）で表現可能であるという保証はありません。どの特徴量 $f_k(\boldsymbol{x})$ を制約として選ぶか、という選択自体が、モデルに対する強力なバイアス（inductive bias）となります。

結論として、最大エントロピー原理を機械学習に適用することは、「真の分布を完璧に当てる魔法」ではありません。そうではなく、**「手元のデータという限られた証拠と、モデルの表現力という制約の中で、最も合理的で偏りのない推論を行うための一貫した枠組み」**と理解するのが適切です。この枠組みの強力さとその限界を理解した上で、私たちは次のエネルギーベースモデルの議論に進みます。

3.3　エネルギーベースモデルとボルツマンマシン

「学習とは確率分布の推定である」という考え方と、統計物理学のカノニカル分布を結びつける具体的な枠組みが、**エネルギーベースモデル (Energy-Based Model, EBM)** です。

エネルギーベースモデルの考え方

エネルギーベースモデルのアイデアは非常に直接的です。学習したいデータのあらゆる配置（状態）$\boldsymbol{x}$ に対して、**エネルギー** $E(\boldsymbol{x})$ というスカラー値を割り当てます。そして、状態 $\boldsymbol{x}$ が出現する確率 $P(\boldsymbol{x})$ を、統計物理学のカノニカル分布と全く同じ形で定義します。

エネルギーベースモデル

パラメータ θ を持つエネルギー関数 $E_\theta(\boldsymbol{x})$ を用いて、確率分布を以下のように定義するモデルをエネルギーベースモデルと呼ぶ。

$$P(\boldsymbol{x}|\theta) = \frac{\exp(-E_\theta(\boldsymbol{x}))}{Z(\theta)}$$

ここで、$Z(\theta) = \sum_{\boldsymbol{x}'} \exp(-E_\theta(\boldsymbol{x}'))$ は分配関数であり、確率の総和を 1 に保つための規格化定数である。

このモデルの直観的な意味は次の通りです。

- 訓練データに似ている「ありそうな」状態 $\boldsymbol{x}$ には、**低いエネルギー**を割り当てる。
- 訓練データとは似ていない「ありえない」状態 $\boldsymbol{x}$ には、**高いエネルギー**を割り当てる。

学習の目標は、訓練データが生成される確率（尤度）が高くなるように、エネルギー関数の形を決めるパラメータ θ を調整することです。エネルギー関数 $E_\theta(\boldsymbol{x})$ の具体的な形としては、ニューラルネットワークなどが用いられます。

しかし、このモデルには統計物理学と同様の、大きな計算上の困難が伴います。それは**分配関数 $Z(\theta)$ の計算**です。状態 $\boldsymbol{x}$ がとりうる全ての配置について和を取る必要があり、$\boldsymbol{x}$ の次元が高くなると、この計算は事実上不可能になります。

ボルツマンマシン：物理学から生まれたモデル

エネルギーベースモデルの考え方を体現した、歴史的にも重要なモデルが**ボルツマンマシン**です。これは、ホップフィールドネットワークを確率的な生成モデルへと発展させたもので、その構造は物理学のイジングモデルと密接に関係しています。

ボルツマンマシンは、$\{0, 1\}$ または $\{-1, 1\}$ の二値をとるニューロン（ユニット）から構成されるネットワークです。ユニットは、データに対応する**可視ユニット $\boldsymbol{v}$** と、データには直接現れない内部的な**隠れユニット $\boldsymbol{h}$** に分かれています。

そして、ネットワーク全体の状態 $(\boldsymbol{v}, \boldsymbol{h})$ のエネルギーは、以下のように定義されます。

$$E(\boldsymbol{v}, \boldsymbol{h}) = -\sum_{i,j} W_{ij} v_i h_j - \sum_{i,k} U_{ik} v_i v_k - \sum_{j,l} J_{jl} h_j h_l - \sum_i a_i v_i - \sum_j b_j h_j$$

ここで、W, U, J はユニット間の結合重み、a, b は各ユニットのバイアスです。これらがモデルの学習すべきパラメータ θ にあたります。このエネルギーの形式は、相互作用のあるスピン系のハミルトニアン（物理的なエネルギー）と全く同じ形をしています。

ボルツマンマシンが学習したいのは、データが観測される確率分布 $P(\boldsymbol{v})$ です。これは、隠れユニット $\boldsymbol{h}$ のあらゆる状態について確率を足し合わせる（周辺化する）ことで得られます。

$$P(\boldsymbol{v}) = \sum_{\boldsymbol{h}} P(\boldsymbol{v}, \boldsymbol{h}) = \frac{1}{Z} \sum_{\boldsymbol{h}} \exp(-E(\boldsymbol{v}, \boldsymbol{h}))$$

ここでもやはり、分配関数 Z の計算と、隠れユニットについての総和計算が困難であるという問題に直面します。この計算困難性こそが、統計物理学とエネルギーベースモデルが共有する中心的な課題であり、後に様々な近似学習法が開発される動機となりました。

3.4　自由エネルギーと尤度の対応

これまでの議論で、私たちは物理学と機械学習の「言語」が、確率分布という共通の土台の上に成り立っていることを見てきました。この節では、両分野における「第一原理」とも言える指導原理が、実は数学的に同じものであることを示します。これは、本書の核心をなすアナロジーです。

まず、機械学習の立場から「良いモデル」とは何かを考えてみましょう。私たちの目標は、手元にある観測データ $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ をうまく説明できるような、モデル（とそのパラメータ θ）を見つけることです。

この「うまく説明できる」という曖昧な要求を、確率の言葉で定量的に表現する、極めて自然で強力な考え方が**最尤（さいゆう）推定の原理**です。

最尤推定の原理 (Principle of Maximum Likelihood)

様々なモデル（パラメータ θ）の中で、私たちが実際に観測したデータ D が生成される確率 $P(D|\theta)$ を最大にするモデルこそが、最も「もっともらしい」良いモデルであると考える。

この確率 $P(D|\theta)$ は、θ を変数とみなしたとき、**尤度関数 (Likelihood Function)** と呼ばれます。つまり、最尤推定とは、尤度を最大化するパラメータ θ を探すことに他なりません。これは、データに最もよく適合するモデルを見つけるための、統計的推論における最も基本的なアプローチの一つです。

この考え方は、ベイズ推論の文脈でさらに正当化されます。ベイズの定理によれば、データ D を観測した後のパラメータ θ の事後分布 $P(\theta|D)$ は、尤度 $P(D|\theta)$ と事前分布 $P(\theta)$ の積に比例します。

$$P(\theta|D) \propto P(D|\theta)P(\theta)$$

もし、パラメータについて特定の事前知識がない場合、事前分布 $P(\theta)$ を平坦（すべての θ が同程度に確からしい）と仮定できます。その場合、事後分布 $P(\theta|D)$ を最大化すること（MAP 推定）は、尤度 $P(D|\theta)$ を最大化すること（最尤推定）と等価になります。

以上の考察から、機械学習における中心的な原理を次のようにまとめることができます。

機械学習における指導原理

モデルの学習とは、観測データ D に対するモデルの尤度 $P(D|\theta)$ **を最大化する**ようなパラメータ θ を見つけ出すプロセスである。

さて、これで準備が整いました。これまでの議論で、私たちは二つの異なる分野から、それぞれ中心的な原理を学びました。

- **統計物理学の原理**：一定温度の系は、**ヘルムホルツの自由エネルギー** F **を最小化する**ことで、熱平衡状態に達する。
- **機械学習（ベイズ推論）の原理**：モデルの良さを評価し、学習を進めるためには、観測デー

タ D がそのモデルから生成される確率、すなわち**尤度 $P(D|\theta)$ を最大化する**ことが望ましい。

驚くべきことに、これら二つの原理は、一見すると無関係に見えますが、数学的には完全に等価なものです。この対応関係を理解することは、物理学の視点から機械学習を読み解く上で最も重要なステップの一つです。

尤度の再確認

まず、尤度 $P(D|\theta)$ について考えます。これは、モデルのパラメータ θ が与えられたときに、データ D が観測される確率でした。特に、ボルツマンマシンのように隠れ変数 $\boldsymbol{h}$ を持つモデルの場合、尤度は隠れ変数がとりうる全ての状態について、同時確率 $P(D, \boldsymbol{h}|\theta)$ を足し合わせる（周辺化する）ことで計算されます。

$$P(D|\theta) = \sum_{\boldsymbol{h}} P(D, \boldsymbol{h}|\theta)$$

ここで、同時確率 $P(D, \boldsymbol{h}|\theta)$ がエネルギーベースの形式で書けるとします。

$$P(D, \boldsymbol{h}|\theta) = \frac{\exp(-E_\theta(D, \boldsymbol{h}))}{Z(\theta)}$$

これを代入すると、尤度は以下のようになります。

$$P(D|\theta) = \frac{1}{Z(\theta)} \sum_{\boldsymbol{h}} \exp(-E_\theta(D, \boldsymbol{h}))$$

物理学とのアナロジー

ここで、右辺の和の部分 $\sum_{\boldsymbol{h}} \exp(-E_\theta(D, \boldsymbol{h}))$ に注目してください。これは、観測データ D を「固定された外部パラメータ」とみなした場合の、隠れ変数 $\boldsymbol{h}$ に関する**分配関数**の形と全く同じです。

このアナロジーに基づき、観測データ D に依存する「自由エネルギー」を定義してみましょう。物理学における $F = -(1/\beta) \ln Z$ の定義に倣い、簡単のため $\beta = 1$ とすると、次のようになります。

変分自由エネルギー

観測データ D とモデルパラメータ θ が与えられたときの変分自由エネルギー $F(D, \theta)$ を以下のように定義する。

$$F(D, \theta) = -\ln \left(\sum_{\boldsymbol{h}} \exp(-E_\theta(D, \boldsymbol{h})) \right)$$

この定義を用いると、尤度 $P(D|\theta)$ は、変分自由エネルギー $F(D, \theta)$ を使って非常にコンパクトに書き直せます。

$$P(D|\theta) = \frac{\exp(-F(D, \theta))}{Z(\theta)}$$

この式の対数をとると、以下の重要な関係が得られます。

$$\ln P(D|\theta) = -F(D,\theta) - \ln Z(\theta)$$

通常、分配関数 $Z(\theta)$ はデータ D には依存しないため、データに対するモデルの適合度を考える上では、$F(D,\theta)$ の部分が本質的です。

自由エネルギーと尤度の対応関係

尤度 $\ln P(D|\theta)$ を最大化するという機械学習の目標は、**変分自由エネルギー $F(D,\theta)$ を最小化する**という目標と（ほぼ）等価である。

機械学習	$\longleftrightarrow$	統計物理学
尤度の最大化	$\longleftrightarrow$	自由エネルギーの最小化
$\ln P(D\|\theta)$	$\longleftrightarrow$	$-F$

この見事な対応関係は、本書の根幹をなすアイデアです。これにより、**データに基づいてモデルを学習させるという情報科学的なプロセスを、外部環境（データ）に適応して安定な状態（平衡状態）に達するという物理的なプロセスとして解釈する**道が拓かれます。

この視点に立つことで、なぜ物理学における近似計算手法（例えば、次の部で学ぶ平均場近似）が、機械学習の近似推論アルゴリズム（変分推論）として有効に機能するのか、その理由が明らかになるのです。

第 III 部

近似手法のアナロジー

4　平均場近似から変分推論へ

第 II 部で、私たちは統計物理学と機械学習が「確率分布」と「エネルギー」という共通言語を持つこと、そして「自由エネルギーの最小化」と「尤度の最大化」という共通の指導原理を持つことを見てきました。しかし、この美しい対応関係には、共通の困難が伴います。それは、**分配関数 Z の計算が極めて困難である**という問題です。

分配関数は、系がとりうる全ての状態についての和 $(\sum_{\boldsymbol{x}})$ を含みます。変数が多く、それらが複雑に相互作用している場合、この和を厳密に計算することは、スーパーコンピュータを使っても不可能な「組み合わせ爆発」を引き起こすのです。

この計算不可能性の壁を乗り越えるために、両分野では様々な**近似手法**が開発されてきました。この部では、その中でも特に重要で、美しい対応関係を持つアプローチ、すなわち物理学の**平均場近似**と機械学習の**変分推論**のアナロジーを解き明かしていきます。

4.1　物理からのアプローチ：平均場近似

計算を困難にしている根源は、変数間の**相互作用**です。例えばイジングモデルにおいて、あるスピン S_i の振る舞いは、隣接するスピン S_j の状態に依存します。その S_j の振る舞いは、さらにその隣の S_k に依存し…という具合に、全ての変数が複雑に絡み合っています。この「多対多」の関係が、問題を難しくしているのです。

平均場近似 (Mean-Field Approximation, MFA) は、この困難を乗り越えるための、大胆かつ強力なアイデアです。

平均場近似の核心的アイデア

ある一つの構成要素（例：スピン S_i）に注目したとき、その周囲の全ての要素からの複雑で揺れ動く相互作用を、**それらの平均的な効果を表す一つの「有効な場（平均場）」**で置き換えてしまう。

この近似によって、元の複雑な多体問題は、**独立な一体問題の集まり**へと劇的に単純化されます。各スピンは、もはや他の個々のスピンの状態を気にする必要はなく、自分にかかる「平均的な場」の向きだけを考えて振る舞えばよくなります。

イジングモデルへの適用

具体的に、イジングモデルのエネルギー $E = -J\sum_{\langle i,j\rangle} S_iS_j$ で考えてみましょう。スピン S_i が感じるエネルギーの部分は、$E_i = -S_i\left(J\sum_{j\in N(i)} S_j\right)$ と書けます。括弧の中の $h_i = J\sum_{j\in N(i)} S_j$ は、隣接スピン S_j の状態によって揺れ動く「場」です。

平均場近似では、この揺れ動く S_j を、その統計的な平均値である磁化 $m = \langle S_j \rangle$ で置き換えます。

$$h_i = J \sum_{j \in N(i)} S_j \quad \xrightarrow{\text{MFA}} \quad h_{\text{MF}} = J \sum_{j \in N(i)} m$$

すると、スピン S_i が感じるエネルギーは $E_i \approx -S_i h_{\text{MF}}$ となり、あたかもスピン S_i が外部から一定の磁場 h_{MF} を受けているかのような、非常に単純な問題に変わります。

もちろん、この近似は「鶏が先か卵が先か」という問題をはらんでいます。平均の場 h_{MF} を決めるためには平均の磁化 m が必要ですが、その m は平均の場の中でスピンがどう振る舞うかによって決まります。このため、両者が矛盾しないように、すなわち**セルフコンシステント（自己無撞着）**になるように m の値を決定する必要があります。

利点と限界

平均場近似の最大の利点は、その**計算の容易さ**です。相互作用がなくなり、全ての変数が独立と見なせるため、困難だった分配関数の計算が、各変数の簡単な和の積へと分解され、解析的に計算可能になるのです。

一方で、これはあくまで近似です。変数間の**相関や揺らぎを無視する**という大胆な仮定を置いているため、特に相転移点近傍など、相関が重要になる現象を正確に記述することはできません。

しかし、この「複雑な確率分布を、より単純な（独立な変数の積で書ける）確率分布で近似する」という発想は非常に強力です。次の節では、この平均場近似の考え方が、機械学習の文脈で「変分推論」として、どのように一般化され、定式化されるかを見ていきます。

4.2 機械学習からのアプローチ：変分推論

前節で見た平均場近似の根底には、「複雑で扱いにくい確率分布を、独立な確率分布の積という、単純で扱いやすい分布で近似する」という思想がありました。機械学習の分野では、このアイデアをより一般化し、数学的に洗練させた**変分推論 (Variational Inference, VI)** というフレームワークが発展しました。

目標：事後分布の近似

機械学習、特にベイズ推論における中心的な課題は、観測データ $\boldsymbol{X}$ が与えられたときの、隠れ変数やパラメータ $\boldsymbol{Z}$ の**事後分布** $P(\boldsymbol{Z}|\boldsymbol{X})$ を求めることです。この事後分布には、データから学習されたモデルに関する全ての情報が凝縮されています。

しかし、ベイズの定理 $P(\boldsymbol{Z}|\boldsymbol{X}) = P(\boldsymbol{X}, \boldsymbol{Z})/P(\boldsymbol{X})$ を思い出すと、分母 $P(\boldsymbol{X}) = \int P(\boldsymbol{X}, \boldsymbol{Z}) d\boldsymbol{Z}$ の計算が困難であるため、事後分布 $P(\boldsymbol{Z}|\boldsymbol{X})$ を直接計算することは、ほとんどの場合不可能です。

そこで変分推論では、この真の事後分布 $P(\boldsymbol{Z}|\boldsymbol{X})$ を、より簡単な形を持つ**近似分布** $Q(\boldsymbol{Z})$ で置き換えることを考えます。

変分推論の戦略

真の事後分布 $P(\boldsymbol{Z}|\boldsymbol{X})$ は複雑すぎて計算できない。そこで、扱いやすい単純な確率分布の族（例えば、全変数が独立な分布族）の中から、真の事後分布に**最も近い**ものを探し出し、それを代理として用いる。

ここでの問題は、「近さ」をどのように測るかです。その尺度として、情報理論における **KL ダイバージェンス (Kullback-Leibler Divergence)** を用います。KL ダイバージェンス $D_{KL}(Q||P)$ は、二つの確率分布 Q と P がどれだけ異なっているかを表す量で、常に非負の値 ($D_{KL} \geq 0$) をとり、$Q = P$ のときにのみ 0 となります。私たちの目標は、この $D_{KL}(Q(\boldsymbol{Z})||P(\boldsymbol{Z}|\boldsymbol{X}))$ を最小化するような $Q(\boldsymbol{Z})$ を見つけることです。

KL ダイバージェンスの定義

二つの確率分布 $Q(\boldsymbol{Z})$ と $P(\boldsymbol{Z})$ の間の KL ダイバージェンスは、以下のように定義される。

$$D_{KL}(Q||P) = \int Q(\boldsymbol{Z}) \ln \frac{Q(\boldsymbol{Z})}{P(\boldsymbol{Z})} d\boldsymbol{Z}$$

これは、$Q(\boldsymbol{Z})$ という分布の下での、対数確率比 $\ln(Q(\boldsymbol{Z})/P(\boldsymbol{Z}))$ の期待値 $\mathbb{E}_{\boldsymbol{Z}\sim Q}\left[\ln \frac{Q(\boldsymbol{Z})}{P(\boldsymbol{Z})}\right]$ とも解釈できる。

KL ダイバージェンスは、単なる数学的な定義に留まらず、明確な情報理論的な意味を持っています。それは、「真の分布が P であるときに、それを不正確な分布 Q を用いて表現（符号化）した場合に、平均的にどれだけの情報損失が生じるか」を表す量と解釈できます。

より直観的には、KL ダイバージェンス $D_{KL}(Q||P)$ の最小化は、Q と P の間の以下の不一致に対してペナルティを課す、と考えることができます。

- $Q(\boldsymbol{Z})$ が大きいのに $P(\boldsymbol{Z}|\boldsymbol{X})$ が小さい領域：近似分布 Q が、「真の事後分布 P がほとんど確率を与えていない領域」に確率を割り当ててしまうと、$\ln(Q/P)$ が大きくなり、ペナルティが増大します。これにより、Q は P が低い確率を持つ領域を避けるようになります。
- $Q(\boldsymbol{Z})$ が小さいのに $P(\boldsymbol{Z}|\boldsymbol{X})$ が大きい領域：これは、Q が真の事後分布の重要な部分（モード）を捉え損ねている状況に対応しますが、Q 自身が小さい値をとるため、ペナルティへの寄与は比較的小さくなります。

この非対称な性質のため、KL ダイバージェンスは数学的な意味での「距離」ではありません（$D_{KL}(Q||P) \neq D_{KL}(P||Q)$）。そして、変分推論で用いられる $D_{KL}(Q||P)$ の最小化は、近似分布 Q が真の分布 P のいずれかのモード（確率が高い山）を一つ選び、そこに集中するような性質を持つことが知られています。

ELBO：計算可能な目的関数

しかし、KL ダイバージェンスの定義には、計算不可能であるはずの事後分布 $P(\boldsymbol{Z}|\boldsymbol{X})$ が含まれており、このままでは最小化できません。ところが、この定義式にベイズの定理を適用するこ

とで、問題を計算可能な目的関数へと見事に変換できます。

ELBO の導出

まず、KL ダイバージェンスの定義から出発します。

$$
\begin{aligned}
D_{KL}(Q||P) &= \int Q(\boldsymbol{Z}) \ln \frac{Q(\boldsymbol{Z})}{P(\boldsymbol{Z}|\boldsymbol{X})} d\boldsymbol{Z} \\
&= \int Q(\boldsymbol{Z}) \ln \frac{Q(\boldsymbol{Z})}{P(\boldsymbol{X},\boldsymbol{Z})/P(\boldsymbol{X})} d\boldsymbol{Z} \\
&= \int Q(\boldsymbol{Z}) \left(\ln Q(\boldsymbol{Z}) - \ln P(\boldsymbol{X},\boldsymbol{Z}) + \ln P(\boldsymbol{X})\right) d\boldsymbol{Z} \\
&= \int Q(\boldsymbol{Z}) \ln Q(\boldsymbol{Z}) d\boldsymbol{Z} - \int Q(\boldsymbol{Z}) \ln P(\boldsymbol{X},\boldsymbol{Z}) d\boldsymbol{Z} + \int Q(\boldsymbol{Z}) \ln P(\boldsymbol{X}) d\boldsymbol{Z} \\
&= \int Q(\boldsymbol{Z}) \ln Q(\boldsymbol{Z}) d\boldsymbol{Z} - \int Q(\boldsymbol{Z}) \ln P(\boldsymbol{X},\boldsymbol{Z}) d\boldsymbol{Z} + \ln P(\boldsymbol{X}) \underbrace{\int Q(\boldsymbol{Z}) d\boldsymbol{Z}}_{=1} \\
&= -\left(\int Q(\boldsymbol{Z}) \ln \frac{P(\boldsymbol{X},\boldsymbol{Z})}{Q(\boldsymbol{Z})} d\boldsymbol{Z} \right) + \ln P(\boldsymbol{X}) \\
&= -\mathcal{L}(Q) + \ln P(\boldsymbol{X})
\end{aligned}
$$

この結果 $D_{KL}(Q||P) = -\mathcal{L}(Q) + \ln P(\boldsymbol{X})$ を並べ替えることで、最終的に以下の重要な関係式が得られます。

$$\ln P(\boldsymbol{X}) = \mathcal{L}(Q) + D_{KL}(Q(\boldsymbol{Z})||P(\boldsymbol{Z}|\boldsymbol{X}))$$

この等式が、変分推論の核心です。

ELBO の最大化と KL ダイバージェンスの最小化

$\ln P(\boldsymbol{X})$ は Q に依らない定数であり、かつ $D_{KL} \geq 0$ なので、**ELBO $\mathcal{L}(Q)$ を最大化すること**は、**KL ダイバージェンス $D_{KL}(Q||P)$ を最小化すること**と完全に等価である。

決定的に重要なのは、**ELBO の式には、扱いにくい事後分布 $P(\boldsymbol{Z}|\boldsymbol{X})$ が現れない**という点です。同時分布 $P(\boldsymbol{X},\boldsymbol{Z})$（モデルを定義すれば計算できる）と近似分布 $Q(\boldsymbol{Z})$（私たちが設計する）だけで計算できるため、ELBO は最適化可能な目的関数となります。

こうして、計算不可能な分布との「近さ」を測るという直接的なアプローチから、計算可能な ELBO を最大化するという間接的な最適化問題へと、問題を変換することに成功したのです。

平均場近似との関係：なぜ因子分解が「平均場」なのか

変分推論は一般的な枠組みですが、その真価は近似分布 $Q(\boldsymbol{Z})$ に具体的な構造を仮定したときに発揮されます。ここで、物理学の平均場近似の思想に倣い、全ての変数が互いに独立である、と

いう大胆な仮定を置きます。これを**平均場近似族**と呼びます。

$$Q(\boldsymbol{Z}) = \prod_{i=1}^{M} q_i(Z_i)$$

■計算上の恩恵　この「因子分解」の仮定を置くことの恩恵は絶大です。元々は M 個の変数が絡み合った高次元の分布 $Q(\boldsymbol{Z})$ を最適化するという困難な問題だったものが、**M 個の一次元の分布 $q_i(Z_i)$ をそれぞれ個別に最適化する**という、はるかに簡単な問題へと分解されるからです。これにより、計算が劇的に扱いやすくなります。

■自己無撞着場（セルフコンシステント・フィールド）の導出　では、この仮定の下で ELBO を最大化すると、具体的にどのような解が得られるのでしょうか。ELBO をある一つの分布 $q_j(Z_j)$ について最大化するよう計算を進めると、最適な $q_j^*(Z_j)$ は以下の関係式を満たすことが分かります。

変分推論における最適解

$$\ln q_j^*(Z_j) = \mathbb{E}_{\boldsymbol{Z}_{\neg j} \sim \prod_{i \neq j} q_i} \left[\ln P(\boldsymbol{X}, \boldsymbol{Z})\right] + \text{const.}$$

したがって、　$q_j^*(Z_j) \propto \exp\left(\mathbb{E}_{\boldsymbol{Z}_{\neg j}} \left[\ln P(\boldsymbol{X}, \boldsymbol{Z})\right]\right)$

ここで、$\mathbb{E}_{\boldsymbol{Z}_{\neg j}}$ は、Z_j を除く全ての変数 $\boldsymbol{Z}_{\neg j}$ についての期待値をとる操作を表します。

この式こそが、平均場近似と変分推論を結びつける鍵です。この式が意味するところを読み解きましょう。

変分推論と平均場の対応

ある一つの変数 Z_j の最適な分布 $q_j^*(Z_j)$ を決定するためには、他の全ての変数 $\boldsymbol{Z}_{\neg j}$ がモデルに与える影響を計算する必要がある。
しかし、その影響を計算する際、他の変数の詳細な状態を追うのではなく、それらの変数が現在従っている分布 $q_{i \neq j}$ に基づく**平均的な影響**のみを考慮に入れる。
この「他の変数からの平均的な影響」こそが、物理学で言うところの**平均場（有効な場）**に他ならない。

この更新式は、q_j^* が他の全ての $q_{i \neq j}$ に依存することを示しています。そのため、全ての分布を一度に解くことはできず、次のような反復的なアルゴリズムで最適解を探索します。

1. まず、各 $q_i(Z_i)$ を適当に初期化する。
2. $q_1(Z_1)$ を、他の $q_2, \ldots, q_M$ を固定した上で更新する。
3. $q_2(Z_2)$ を、他の $q_1, q_3, \ldots, q_M$ を固定した上で更新する。
4. ... これを全ての q_i について繰り返し、全体の ELBO が収束するまで続ける。

この、各変数が「他の変数たちが作る平均場」に適応し、その結果として平均場自体も変化し、最

終的に全体が矛盾のない解に落ち着く、という反復的なプロセスは、物理学で**自己無撞着な（セルフコンシステントな）**解を求める手続きと完全に一致しています。

このように、**平均場近似とは、変分推論の枠組みにおいて、事後分布を完全に因子分解できる分布で近似すること**であり、物理学者の直観から生まれた近似手法が、機械学習の分野でより厳密かつ一般的に定式化され、強力な推論アルゴリズムとして確立されているのです。

4.3 理論の統一：変分自由エネルギーと KL ダイバージェンス

ここまで、私たちは二つの異なるアプローチを見てきました。

- **物理学の平均場近似**：相互作用のある複雑な系を、独立な粒子が有効な「場」を感じる単純な系で近似する。その目標は、近似的な系の**変分自由エネルギーを最小化する**ことで、真の系の状態に最も近づけることでした。
- **機械学習の変分推論**：計算不可能な真の事後分布 P を、扱いやすい単純な分布 Q で近似する。その目標は、Q と P の間の **KL ダイバージェンスを最小化する**ことでした。

これら二つの目標は、異なる分野で、異なる動機から生まれたように見えます。しかし、数学の光を当てると、両者が完全に同じものであることが明らかになります。

変分自由エネルギーの定義

まず、物理学の文脈で現れる「変分自由エネルギー」を、より一般的に定義し直しましょう。真の系の確率分布が $P(\boldsymbol{x}) \propto \exp(-\beta E(\boldsymbol{x}))$ で与えられるとします。これに対する任意の近似分布を $Q(\boldsymbol{x})$ としたとき、変分自由エネルギー $F[Q]$ は以下のように定義されます。

$$F[Q] \equiv \underbrace{\mathbb{E}_{\boldsymbol{x}\sim Q}\left[E(\boldsymbol{x})\right]}_{\text{エネルギーの期待値}} - \underbrace{TS(Q)}_{\text{温度×エントロピー}}$$

ここで、$S(Q) = -k_B \mathbb{E}_{\boldsymbol{x}\sim Q}[\ln Q(\boldsymbol{x})]$ は近似分布 Q のエントロピーです。この $F[Q]$ は、もし系が確率分布 Q に従うとしたら、真のエネルギー関数 $E(\boldsymbol{x})$ の下でどれくらいの自由エネルギーを持つことになるか、を表す量です。物理学の要請は、この $F[Q]$ を最小化するような近似分布 Q を見つけることです。

KL ダイバージェンスとの数学的関係

それでは、この変分自由エネルギー $F[Q]$ と、変分推論の目標であった KL ダイバージェンス $D_{KL}(Q||P)$ の関係を導出してみましょう。

変分自由エネルギーと KL ダイバージェンスの関係式の導出

KL ダイバージェンスの定義から出発します。

$$
\begin{aligned}
D_{KL}(Q||P) &= \int Q(\boldsymbol{x}) \ln \frac{Q(\boldsymbol{x})}{P(\boldsymbol{x})} d\boldsymbol{x} \\
&= \int Q(\boldsymbol{x}) \ln \frac{Q(\boldsymbol{x})}{\exp(-\beta E(\boldsymbol{x}))/Z} d\boldsymbol{x} \\
&= \int Q(\boldsymbol{x}) \left(\ln Q(\boldsymbol{x}) + \beta E(\boldsymbol{x}) - \ln Z\right) d\boldsymbol{x} \\
&= \int Q(\boldsymbol{x}) \ln Q(\boldsymbol{x}) d\boldsymbol{x} + \beta \int Q(\boldsymbol{x}) E(\boldsymbol{x}) d\boldsymbol{x} - \ln Z \int Q(\boldsymbol{x}) d\boldsymbol{x} \\
&= \mathbb{E}_Q[\ln Q(\boldsymbol{x})] + \beta \mathbb{E}_Q[E(\boldsymbol{x})] - \ln Z \\
&= -\beta T S(Q) + \beta \mathbb{E}_Q[E(\boldsymbol{x})] - \ln Z \\
&= \beta \left(\mathbb{E}_Q[E(\boldsymbol{x})] - T S(Q)\right) + \beta (T \ln Z) \\
&= \beta \left(F[Q] - F_{\text{true}}\right)
\end{aligned}
$$

この導出から、以下の極めて重要な関係が明らかになりました。

理論の統一原理

近似分布 Q と真の分布 P の間の KL ダイバージェンスは、変分自由エネルギー $F[Q]$ と真の自由エネルギー F_{true} の差に、逆温度 β を掛けたものに等しい。

$$D_{KL}(Q||P) = \beta \left(F[Q] - F_{\text{true}}\right)$$

真の自由エネルギー F_{true} は定数なので、**KL ダイバージェンスを最小化することと、変分自由エネルギーを最小化することは**、完全に等価な目標である。

この関係式こそが、物理学の平均場近似と機械学習の変分推論を一つの理論として統一する、美しい架け橋です。私たちが前節で ELBO の最大化として定式化した問題は、物理学者の言葉で言えば、変分自由エネルギーの最小化に他ならなかったのです。

こうして、計算困難な問題に対する現実的な近似解を求めるという共通の目的のために、異なる分野で発展した二つのアプローチが、同じ数学的構造の上に成り立っていることが示されました。この深い結びつきは、両分野の知見が互いに交換可能であることを意味し、より強力な手法を開発するための豊かな土壌となっているのです。

5　サンプリング手法の融合

前のセクションでは、複雑な確率分布を、解析的に扱いやすい単純な分布で近似する変分推論の枠組みを見てきました。これは、分布全体の形状を捉えようとする決定論的なアプローチでした。

しかし、問題によっては、分布の全体の形を陽に知る必要はなく、その分布に従うサンプル（標本）をいくつか生成できれば十分な場合があります。例えば、物理的な観測量（エネルギーの平均値など）を計算したい場合、その量の期待値を求められれば良いわけです。

このセクションでは、このような要請に応えるための、もう一つの強力な近似手法の系統である**サンプリング法**、特に**マルコフ連鎖モンテカルロ法 (Markov Chain Monte Carlo, MCMC)** について解説します。

5.1 物理シミュレーションにおけるマルコフ連鎖モンテカルロ法

物理学者が知りたいのは、多くの場合、ある物理量 A の統計的な平均値（期待値）$\langle A \rangle$ です。これは、カノニカル分布 $P(\boldsymbol{x})$ を用いて次のように定義されます。

$$\langle A \rangle = \sum_{\boldsymbol{x}} A(\boldsymbol{x}) P(\boldsymbol{x})$$

しかし、ここでもやはり状態空間が巨大であるため、全ての状態 $\boldsymbol{x}$ について和を取ることは不可能です。

モンテカルロ法という発想

この問題を解決する素朴かつ強力なアイデアが**モンテカルロ法**です。もし、我々が何らかの方法で、目標とする確率分布 $P(\boldsymbol{x})$ に従うサンプル状態の集合 $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ を手に入れることができたとします。そうすれば、期待値は単純なサンプルの平均値で近似できます。

$$\langle A \rangle \approx \frac{1}{N} \sum_{i=1}^{N} A(\boldsymbol{x}_i)$$

大数の法則により、N が大きければ大きいほど、この近似は正確になります。

問題は、**どのようにして複雑なカノニカル分布 $P(\boldsymbol{x}) \propto \exp(-\beta E(\boldsymbol{x}))$ からサンプルを生成するか**、という点です。特に、規格化定数である分配関数 Z が未知であるため、標準的なサンプリング手法は使えません。

マルコフ連鎖によるサンプリング

この難問を解決するのが、**マルコフ連鎖**を利用するアイデアです。マルコフ連鎖とは、未来の状態が現在の状態のみに依存して決まるような、状態の確率的な連鎖のことです。

MCMC の核心は、**最終的に訪れる状態の分布（定常分布）が、目標とする分布 $P(\boldsymbol{x})$ と一致するような、都合の良いマルコフ連鎖を設計する**ことにあります。そのようなマルコフ連鎖を構築できれば、あとは適当な初期状態から連鎖を長時間動かし続けるだけで、得られるサンプルは実質的に $P(\boldsymbol{x})$ から生成されたものと見なせるようになります。

マルコフ連鎖の「なぜ」: 定常分布と詳細釣り合い

この一連の操作がうまくいく仕組みを、もう少し詳しく見ていきましょう。アナロジーとして、たくさんの部屋がある家の中を、一人の人が特定のルールに従って移動する様子を想像してみてください。

- **状態 $\boldsymbol{x}$**：家の中の各部屋。

- **目標分布** $P(\boldsymbol{x})$：各部屋の「居心地の良さ」や「人気度」。私たちはこの人気度分布を知りたいと思っています。
- **マルコフ連鎖**：人が部屋から部屋へ移動するプロセス。
- **サンプラー**：部屋を移動している人。その人の居場所が「サンプル」となります。

人が移動を延々と続けたとき、長時間経過した後では、その人が各部屋に存在する確率（滞在時間の割合）は、ある一定の分布に落ち着くと考えられます。この、マルコフ連鎖が十分に長い時間の後に収束する確率分布を**定常分布** $\pi(\boldsymbol{x})$ と呼びます。

MCMC の目標は、この最終的な滞在時間の割合 $\pi(\boldsymbol{x})$ が、部屋の人気度 $P(\boldsymbol{x})$ とぴったり一致するように、人の移動ルールを設計することです。では、どのような移動ルールにすれば、そのような都合の良いことが起きるのでしょうか。その答えが、**詳細釣り合い条件 (Detailed Balance Condition)** です。

詳細釣り合い条件

任意の二つの状態（部屋）$\boldsymbol{x}$ と $\boldsymbol{x}'$ の間を考えたとき、定常状態においては、「$\boldsymbol{x}$ から $\boldsymbol{x}'$ への遷移の流れ」と「$\boldsymbol{x}'$ から $\boldsymbol{x}$ への遷移の流れ」が等しくなる。
数式で書くと、遷移確率を $T(\boldsymbol{x}'|\boldsymbol{x})$ として、以下の式が成り立つ。

$$\pi(\boldsymbol{x})T(\boldsymbol{x}'|\boldsymbol{x}) = \pi(\boldsymbol{x}')T(\boldsymbol{x}|\boldsymbol{x}')$$

この条件は、系が平衡状態にあるとき、ミクロなレベルで全ての遷移がバランスしている、という物理学の考え方に基づいています。そして、ここが最も重要な点です。

MCMC の基本原理

もし、私たちが設計する移動ルール（遷移確率 T）が、目標分布 P に対して詳細釣り合い条件を満たすように、すなわち

$$P(\boldsymbol{x})T(\boldsymbol{x}'|\boldsymbol{x}) = P(\boldsymbol{x}')T(\boldsymbol{x}|\boldsymbol{x}')$$

となるように作られていれば、そのマルコフ連鎖の定常分布 $\pi(\boldsymbol{x})$ は、**数学的に目標分布 $P(\boldsymbol{x})$ と一致することが保証される。**

つまり、私たちは「このルールに従って動き続ければ、最終的に滞在時間の割合が部屋の人気度と一致する」ような、魔法の移動ルールを設計すればよいのです。

サンプリングがうまくいくのは、偶然ではありません。詳細釣り合いという条件を満たすマルコフ連鎖を注意深く設計することで、サンプラーを長時間歩かせれば、その足跡の密度が、必然的に目標の確率分布の形を浮かび上がらせるのです。

次に紹介するメトロポリス・ヘイスティングス法は、この詳細釣り合い条件を巧みに満たす「移動ルール」を具体的に与える、天才的なアルゴリズムの一例です。

メトロポリス・ヘイスティングス法

詳細釣り合い条件を満たすマルコフ連鎖を構築するための、具体的で天才的な「レシピ」が、**メトロポリス・ヘイスティングス法**です。このアルゴリズムは、遷移のプロセスを「候補の提案」と「採択か棄却か」という二段階に分解することで、詳細釣り合いを巧みに満たします。

まず、状態 $\boldsymbol{x}$ から別の状態 $\boldsymbol{x}'$ への遷移確率 $T(\boldsymbol{x}'|\boldsymbol{x})$ を、以下のように二つの部分の積で表現します。

$$T(\boldsymbol{x}'|\boldsymbol{x}) = g(\boldsymbol{x}'|\boldsymbol{x})A(\boldsymbol{x}'|\boldsymbol{x}) \quad (\text{for } \boldsymbol{x} \neq \boldsymbol{x}')$$

- $g(\boldsymbol{x}'|\boldsymbol{x})$：**提案分布 (Proposal Distribution)**。現在の状態 $\boldsymbol{x}$ から、次の候補として $\boldsymbol{x}'$ を「提案」する確率。これは私たちが自由に設計できます（例：「現在の状態からランダムに一つのスピンを反転させる」など）。
- $A(\boldsymbol{x}'|\boldsymbol{x})$：**採択確率 (Acceptance Probability)**。提案された候補 $\boldsymbol{x}'$ を、次の状態として「採択」する確率。この採択確率をうまく設計することが、アルゴリズムの核心です。

■詳細釣り合い条件からの採択確率の導出　私たちの目標は、この遷移確率 T が、目標分布 P に対して詳細釣り合い条件 $P(\boldsymbol{x})T(\boldsymbol{x}'|\boldsymbol{x}) = P(\boldsymbol{x}')T(\boldsymbol{x}|\boldsymbol{x}')$ を満たすようにすることです。上記の分解した式を代入すると、

$$P(\boldsymbol{x})g(\boldsymbol{x}'|\boldsymbol{x})A(\boldsymbol{x}'|\boldsymbol{x}) = P(\boldsymbol{x}')g(\boldsymbol{x}|\boldsymbol{x}')A(\boldsymbol{x}|\boldsymbol{x}')$$

この式が成り立つように、採択確率 A を設計すればよいのです。式を整理すると、採択確率の比は以下のようになります。

$$\frac{A(\boldsymbol{x}'|\boldsymbol{x})}{A(\boldsymbol{x}|\boldsymbol{x}')} = \frac{P(\boldsymbol{x}')g(\boldsymbol{x}|\boldsymbol{x}')}{P(\boldsymbol{x})g(\boldsymbol{x}'|\boldsymbol{x})}$$

この関係を満たすような採択確率の決め方はいくつかありますが、メトロポリス・ヘイスティングスが採用したのが、以下のシンプルで強力な形式です。

$$A(\boldsymbol{x}'|\boldsymbol{x}) = \min\left(1, \frac{P(\boldsymbol{x}')g(\boldsymbol{x}|\boldsymbol{x}')}{P(\boldsymbol{x})g(\boldsymbol{x}'|\boldsymbol{x})}\right)$$

この形式が詳細釣り合いを満たすことは、簡単な計算で確認できます。

■物理シミュレーションへの応用　さて、この一般的な形式を、物理学のカノニカル分布 $P(\boldsymbol{x}) \propto \exp(-\beta E(\boldsymbol{x}))$ に適用してみましょう。さらに、最も簡単な場合として、提案分布が対称である、すなわち $g(\boldsymbol{x}'|\boldsymbol{x}) = g(\boldsymbol{x}|\boldsymbol{x}')$ であると仮定します（例えば「ランダムに選んだスピンを反転する」という提案は対称です）。すると、採択確率は劇的に単純化されます。

物理系における採択確率の導出

採択確率の比の部分を計算します。

$$\frac{P(\boldsymbol{x}')g(\boldsymbol{x}|\boldsymbol{x}')}{P(\boldsymbol{x})g(\boldsymbol{x}'|\boldsymbol{x})} = \frac{P(\boldsymbol{x}')}{P(\boldsymbol{x})} = \frac{\exp(-\beta E(\boldsymbol{x}'))/Z}{\exp(-\beta E(\boldsymbol{x}))/Z} = \exp(-\beta(E(\boldsymbol{x}') - E(\boldsymbol{x}))) = e^{-\beta\Delta E}$$

ここで、未知の分配関数 Z が綺麗に打ち消し合うことが、この手法の極めて強力な点です。したがって、採択確率は、

$$A(\boldsymbol{x}'|\boldsymbol{x}) = \min\left(1, e^{-\beta\Delta E}\right)$$

となります。

この数学的な結果が、何を意味しているかを読み解くと、以前に提示した直観的なアルゴリズムそのものになります。

メトロポリス法のアルゴリズム

時刻 t で系が状態 $\boldsymbol{x}_t$ にあるとする。

1. **候補の提案**：対称な提案分布 $g(\boldsymbol{x}'|\boldsymbol{x}_t)$ に従い、候補状態 $\boldsymbol{x}'$ を生成する。
2. **採択確率の計算**：エネルギー変化 $\Delta E = E(\boldsymbol{x}') - E(\boldsymbol{x}_t)$ を計算し、採択確率 $A = \min(1, e^{-\beta\Delta E})$ を求める。
   - もし $\Delta E \leq 0$ なら、$e^{-\beta\Delta E} \geq 1$ なので $A = 1$。必ず採択する。
   - もし $\Delta E > 0$ なら、$e^{-\beta\Delta E} < 1$ なので $A = e^{-\beta\Delta E}$。確率 A で採択する。
3. **状態の更新**：確率 A の試行が成功すれば $\boldsymbol{x}_{t+1} = \boldsymbol{x}'$ とし、失敗すれば $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t$ とする。
4. このプロセスを繰り返す。

このように、一見すると直観的なヒューリスティクスのようにも見えるメトロポリス法の採択ルールは、実はマルコフ連鎖が目標分布に正しく収束するための数学的な要請である**詳細釣り合い条件**を、最小限の要素で満たすように注意深く設計された、見事な結果なのです。このアルゴリズムの発見が、計算物理学の発展に大きく貢献しました。そしてこの強力なサンプリングの考え方は、機械学習における確率的推論にも、そのまま応用されるのです。

5.2　機械学習におけるサンプリングの役割

前節で見たように、MCMC は物理学において、分配関数を知ることなくボルツマン分布からサンプルを生成し、物理量の期待値を計算するための強力な手法でした。この「未知の規格化定数を持つ確率分布からサンプリングする」という機能は、機械学習におけるベイズ推論が直面する問題と、驚くほど合致しています。

ベイズ推論における計算困難性

思い出してみましょう。ベイズ推論の中心的な目的は、データ D を観測した後の、モデルのパラメータ θ に関する**事後分布** $P(\theta|D)$ を求めることでした。そして、この事後分布の計算を困難にしていたのが、分母に現れる $P(D)$ でした。

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

$P(D)$ はパラメータ θ の全空間にわたる積分を含むため、多くの場合、計算不可能です。

しかし、MCMC の観点から見ると、事後分布は以下のように捉えることができます。

$$P(\theta|D) \propto \underbrace{P(D|\theta)}_{\text{尤度}} \times \underbrace{P(\theta)}_{\text{事前分布}}$$

これは、規格化されていない確率密度であれば、右辺の尤度と事前分布の積を計算することで、いつでも評価できることを意味します。これはまさに、MCMC が得意とする状況です。物理学における「エネルギー関数 $E(\boldsymbol{x})$」の役割を、機械学習では「負の対数事後確率 $-\ln(P(D|\theta)P(\theta))$」が果たすことになります。

MCMC による事後分布からのサンプリング

そこで、物理シミュレーションで用いたメトロポリス・ヘイスティングス法を、ベイズ推論の問題にそのまま適用することができます。

ベイズ推論のための MCMC アルゴリズム

現在のパラメータが θ_t であるとする。

1. **候補の提案**：現在のパラメータ θ_t の近くから、新しい候補パラメータ θ' をランダムに提案する（例：θ_t を中心とするガウス分布からサンプリング）。
2. **採択確率の計算**：以下の比を計算する。（簡単のため対称な提案分布を仮定）

$$r = \frac{P(\theta'|D)}{P(\theta_t|D)} = \frac{P(D|\theta')P(\theta')}{P(D|\theta_t)P(\theta_t)}$$

   ここでも、計算不能な $P(D)$ は打ち消し合って消えることに注意してください。
3. **採択か棄却か**：確率 $A = \min(1, r)$ で候補 θ' を採択する。採択されれば $\theta_{t+1} = \theta'$ とし、棄却されれば $\theta_{t+1} = \theta_t$ とする。
4. このプロセスを何千、何万回と繰り返す。

このプロセスを十分に長い時間実行した後、得られるパラメータのサンプル列 $\{\theta_N, \theta_{N+1}, \ldots, \theta_M\}$ は、真の事後分布 $P(\theta|D)$ からのサンプルと見なすことができます。

サンプルの活用法

一度、事後分布からのサンプルを手に入れてしまえば、私たちは様々な統計的推論を行うことができます。

- **パラメータの期待値**：サンプルの平均値を計算することで、パラメータの事後期待値を近似できます。$\mathbb{E}[\theta|D] \approx \frac{1}{M-N}\sum_{t=N}^{M}\theta_t$
- **信頼区間**：サンプルの分布を見ることで、パラメータが 95% の確率でどの範囲に収まるか、といった信頼区間を評価できます。
- **事後分布の可視化**：サンプルのヒストグラムを描くことで、事後分布の形状そのものを直接見ることができます。

- **予測分布**：新しいデータ $\boldsymbol{x}_{\text{new}}$ に対する予測を行うには、得られた各サンプル θ_t を用いて予測分布 $P(\boldsymbol{x}_{\text{new}}|\theta_t)$ を計算し、それらを平均します。

■MCMC と変分推論の比較　この部で学んだ二つの近似手法は、相補的な関係にあります。

- **変分推論 (VI)**：最適化問題として解を求めます。計算が高速で、大規模なデータにも適用しやすいですが、近似分布の形状（例：平均場近似）に解が制約されるという限界があります。
- **マルコフ連鎖モンテカルロ法 (MCMC)**：サンプリングによって解を求めます。計算コストが高い傾向にありますが、十分な時間があれば真の事後分布を（サンプルとして）正確に表現できるという理論的な保証があります。

どちらの手法も、計算不可能性の壁を乗り越え、複雑な確率モデルからの推論を可能にするための、現代の機械学習と統計学において不可欠なツールキットなのです。

第 IV 部

深層学習の物理的解釈

6　深層学習とくりこみ群

これまでの部で、私たちは統計物理学の視点を用いて、機械学習モデルの背後にある理論的構造を解き明かしてきました。特に、エネルギー関数や確率分布を設計し、平均場近似やサンプリングといった手法で推論を行うという考え方は、ボルツマンマシンのようなモデルを理解する上で強力な武器となりました。

しかし、現代の AI の発展を牽引する**深層学習 (Deep Learning)** は、これらのモデルとは一線を画す、ある重要な特徴を持っています。この部では、深層学習のその特徴、すなわち「深さ（階層性）」に焦点を当て、それを物理学の**くりこみ群 (Renormalization Group)** という強力な概念的枠組みを通して解釈していきます。

まずは、くりこみ群の議論に入る前に、私たちがこれまで扱ってきた「機械学習」と、これから扱う「深層学習」の本質的な違いは何か、そしてアプローチがどう異なるのかを明確にしておきましょう。

機械学習と深層学習の本質的な違い：特徴量の自動獲得

これまでの機械学習アプローチでは、多くの場合、人間がデータに関する専門知識を駆使して、モデルが学習しやすいように**特徴量 (features)** を設計する必要がありました。例えば、画像から猫を認識するモデルを作る場合、画像の中から「耳の形」「ひげの存在」「目の色」といった情報を抽出し、それをモデルの入力とする、といった具合です。このプロセスは**特徴量エンジニアリング**と呼ばれ、モデルの性能を大きく左右する、職人芸的な側面を持っていました。

これに対し、深層学習、特に多層のニューラルネットワーク（DNN）がもたらした革命は、この特徴量エンジニアリングを**自動化**した点にあります。

古典的機械学習と深層学習の決定的違い

- **古典的機械学習**：人間が設計した**特徴量**を入力とし、その特徴量と出力の関係性を学習する。
- **深層学習**：生のデータ（例：画像のピクセル値そのもの）を入力とし、データの本質を捉えるための**特徴量そのものを、階層的に学習する**（これを**表現学習**と呼ぶ）。

深層学習モデルでは、入力に近い層はエッジや色のグラデーションのような単純で局所的な特徴を捉え、中間層はそれらを組み合わせて目や鼻、質感といった、より複雑な部品（パーツ）を認識し、出力に近い層ではさらにそれらを組み合わせて「猫の顔」全体のような、大域的で抽象的な概念を捉えるようになります。

このように、深層学習のアプローチは、人間が特徴量を設計する代わりに、特徴量を自動で発見するための**階層的な構造**を設計することに重点が置かれます。この「深さ」こそが、深層学習が

複雑な現実世界のデータを、これほどまでうまく扱える理由の核心なのです。

この「階層的な特徴抽出」という深層学習の振る舞いは、物理学のある考え方と驚くほどよく似ています。物理学には、ミクロな世界の詳細な記述から出発し、不要な情報（細かすぎる自由度）を段階的に消去していくことで、マクロなスケールでの本質的な振る舞いを抜き出す、という強力な手法が存在します。それが、次に解説する「くりこみ群」の思想です。

※注意点

ただし、この深層学習とくりこみ群のアナロジーは、あくまで概念的な類似点を探るための「レンズ」であり、両者が全く同じものだと考えるべきではありません。このアナロジーの面白さと限界を正しく理解し、本書の議論をより深く味わっていただくために、両者の重要な違いを先に確認しておきましょう。

第一に、**目的が異なります**。物理学のくりこみ群は、いわば「神の視点」で、どんな現象にも共通する普遍的な法則を探究しようとします。情報の取捨選択ルールは、物理法則によって予め決まっています。一方、深層学習はもっと現実的で、特定の「お題」（タスク）、例えば「猫と犬を見分ける」を解くことだけを目指します。どの情報が重要で、どれが不要かは、そのお題をうまく解けるかどうかで決まり、データから学習していきます。

第二に、**「表現学習」の歴史**です。データから自動で特徴を見つけ出すという考え方自体は、実は深層学習が生まれる前からありました。深層学習の凄みは、そのアイデアを、何百万ものパラメータを持つ巨大なモデルと膨大な計算能力と組み合わせることで、これまでとは比較にならないほど複雑で豊かな表現を獲得できるようにした点にあります。

第三に、**階層性の実現方法**です。情報を階層的にまとめていく「まとめ方」も、モデルの種類（アーキテクチャ）によって様々です。画像認識で活躍する CNN は、隣接するピクセルから徐々に視野を広げていくように、まさに「ズームアウト」するような方法で情報を集約します。一方、文章を扱う Transformer のようなモデルは、文中の遠く離れた単語同士の関係性を一気に見るような、よりグローバルな方法で情報をまとめます。

このように多くの違いがあることを念頭に置いた上で、本書では、くりこみ群の持つ**「多くの情報の中から、本質的に重要なもの（relevant）と、些末で無視すべきもの（irrelevant）を巧みに選り分ける」**という視点を借ります。この視点を通して、深層学習の「深さ」が、複雑なデータの中からいかにして「有効な情報」だけを抽出していくのか、そのメカニズムを解き明かしていきます。

この視点に立ち、まずはアナロジーの出発点となる、物理学における「くりこみ群」の思想そのものを見ていきましょう。

6.1　くりこみ群の思想：ミクロからマクロへ

くりこみ群 (Renormalization Group, RG) は、もともと場の量子論で現れる発散の問題を解決するために開発され、その後、統計物理学における相転移現象の理解に革命をもたらした、物理学における最も深遠な概念の一つです。

その数学的な詳細は非常に難解ですが、根底に流れる思想は、あるシステムの「見方」をスケールに応じて変えていく、という極めて強力なものです。

問題意識：スケールがすべてを支配する

物理学者が相転移（例えば、水が沸騰して水蒸気になる現象）を理解しようとするとき、特異な困難に直面します。相転移が起こる「臨界点」においては、一つの水分子の小さな揺らぎが、隣の分子、さらにその隣の分子へと伝播し、巨視的なスケールにまで及ぶ相関が生まれるのです。

このような、ミクロな詳細からマクロな現象まで、あらゆるスケールが複雑に絡み合った状態を記述することは、平均場近似のような「平均」で物事を捉えるアプローチでは全く歯が立ちません。スケールを横断して、系の本質を捉える新しい視点が必要でした。くりこみ群は、まさにこの課題への答えでした。

くりこみ群の操作：「粗視化」と「スケール変換」

くりこみ群の思想を、イジングモデルを例に、具体的な操作として見てみましょう。

くりこみ群の基本操作

くりこみ群とは、系のミクロな自由度を段階的に消去（積分消去）し、より大きなスケールでの有効な理論を導出する手続きである。これは主に二つのステップから成る。

■1. **粗視化 (Coarse-graining)**　まず、元のスピン格子を、例えば 2×2 のような小さなブロックに分割します。そして、各ブロックの中のミクロなスピン（この例では 4 つ）の情報を、一つの**ブロックスピン**という新しい変数に集約します。ブロックスピンの状態は、例えばブロック内のスピンの多数決で決めます（上が多ければ「上」、下が多ければ「下」）。

この「粗視化」のステップが核心です。私たちは、**ブロック内部の細かなスピンの揺らぎに関する情報を意図的に捨て去り**、ブロック全体の平均的な振る舞いのみを次のスケールの記述として残すのです。これは、カメラのズームを引いて、解像度を落として世界を見る操作に似ています。

■2. **スケール変換 (Rescaling)**　粗視化によって、格子のサイズは元の $1/2$ になりました。この新しいブロックスピンの格子を、元の格子と同じサイズに見えるように、空間の物差し自体を拡大します。これにより、変換後の系を、変換前の系と同じ土俵で比較できるようになります。

この「粗視化＋スケール変換」という一連の操作を、くりこみ変換と呼びます。この変換によって、元の系のハミルトニアン（エネルギー関数）は、ブロックスピンの間の新しい有効な相互作用を記述する、新しいハミルトニアンへと書き換えられます。

くりこみフローと固定点

くりこみ変換を繰り返し適用していくと、系のパラメータ（イジングモデルでは相互作用の強さ J など）は、変換のたびに変化していきます。このパラメータの変化の軌跡を、**くりこみフロー**と呼びます。

このフローを追いかけていくと、最終的に「固定点」と呼ばれる、それ以上くりこみ変換を施してもパラメータが変化しない点に行き着くことがあります。この固定点は、系がスケール変換に対して不変である、すなわち**自己相似性**を持つことを意味しており、相転移が起こる臨界点と密接に関係しています。

くりこみ群の思想とは、ある一つのスケールだけで系を理解しようとするのではなく、**スケールを変えたときに、系の「記述」そのものがどのように変化していくか**を追いかけることに本質があります。それは、ミクロな世界の無数の自由度の中から、マクロな現象を支配する本質的な情報だけを、段階的に「蒸留」していくプロセスなのです。

この「情報の蒸留」と「階層的な記述の変換」という考え方が、まさに深層学習における階層的な特徴抽出と見事なアナロジーを成していることを、次の節で見ていきます。

6.2　深層ネットワークの階層構造とのアナロジー

前節で解説した、くりこみ群（RG）の思想、すなわち「ミクロな詳細を段階的に捨て、マクロな本質を抽出する」という考え方は、深層ニューラルネットワーク（DNN）、特に画像認識で成功を収めた**畳み込みニューラルネットワーク (Convolutional Neural Network, CNN)** の動作原理と、驚くほど美しいアナロジーを成します。

このアナロジーの核心は、**ネットワークの層を深く進んでいくデータの流れを、物理学におけるくりこみ群フローとして解釈する**、という視点にあります。

両者の対応関係を、具体的に見ていきましょう。

概念の対応関係

RG と DNN のアナロジー対応表

くりこみ群 (RG)		深層ニューラルネットワーク (DNN)
ミクロな自由度（例：スピン）	$\longleftrightarrow$	入力データ（例：画像のピクセル）
粗視化の操作	$\longleftrightarrow$	層の変換（特に Convolution + Pooling）
スケール（物理的な長さ）	$\longleftrightarrow$	抽象度のレベル（ネットワークの深さ）
有効な相互作用	$\longleftrightarrow$	タスクに有効な特徴量
くりこみフローの固定点	$\longleftrightarrow$	最終的な分類・認識結果

この対応関係を、CNN を例にさらに詳しく見ていきます。

■粗視化としての Convolution と Pooling　CNN における中心的な操作は、**畳み込み（Convolution）とプーリング（Pooling）**です。

- **畳み込み**：入力画像に対して、小さなフィルター（カーネル）を適用し、局所的な特徴（エッジ、模様など）を抽出します。これは、RG においてスピンをブロックにまとめる「ブロッキング」の操作に似ています。ある局所領域の情報を、一つの特徴マップの値へと集約しています。

- **プーリング**：畳み込みで得られた特徴マップの解像度を落とす操作です。例えば、マックスプーリングでは、2×2 の領域から最大値のみを取り出し、他の 3 つの値は捨て去ります。これは、RG における**粗視化**の思想そのものです。すなわち、**特徴の正確な位置というミクロな情報は捨て去り、その特徴が存在するかどうかというマクロな情報のみを保持する**操作です。

この「畳み込み＋プーリング」という一連の変換を層の数だけ繰り返すことで、ネットワークは RG のくりこみ変換を何度も適用するように、情報の「蒸留」を段階的に行っていきます。

■スケールと抽象度　ネットワークの浅い層では、フィルターは小さな受容野しか持たず、エッジや特定の色といった、物理的なスケールの小さい、局所的で単純な特徴にしか反応しません。

しかし、層が深くなるにつれて、後の層のニューロンは、前の層のニューロンの出力を受け取ります。プーリングによって解像度が落ちているため、後の層のニューロンは、元の画像のより広い領域（大きな受容野）の情報を間接的に見ることになります。これにより、深い層では、単純な特徴を組み合わせた、より複雑で、スケールの大きい、抽象的な特徴（目、鼻、車輪など）に反応できるようになるのです。

これは、RG がミクロなスケールの相互作用から出発し、くりこみ変換を繰り返すことで、より大きなスケールでの有効な物理法則を明らかにしていくプロセスと、まさしく対応しています。

■タスクによる特徴の選別　もちろん、導入で述べたように決定的な違いもあります。RG における粗視化のルール（多数決など）は人間が予め決めますが、DNN における変換ルール（フィルターの重みなど）は、**タスクの損失関数を最小化するように、データから自動で学習されます。**

これは、RG の言葉で言えば、どの自由度が「関連があり (relevant)」、どれが「無関係 (irrelevant)」かを、タスクが決定していると解釈できます。猫を認識するタスクでは、「ひげ」や「三角の耳」に関連する特徴量はくりこみフローの中で増幅され（relevant な自由度）、背景の模様のような特徴量は抑制されていきます（irrelevant な自由度）。

このように、深層学習の階層構造は、くりこみ群という物理学のレンズを通して見ることで、単なる多層の関数近似器ではなく、**データの本質的な構造を、スケールを横断しながら自動的に発見していく、情報の粗視化プロセス**として、より深く理解することができるのです。

6.3　スケール不変性と特徴抽出

深層学習、特に画像認識モデルがなぜ優れているのかを理解する鍵は、その**特徴抽出**の仕方にあります。優れた特徴とは、対象の本質を捉え、些末な変化に対しては影響を受けない**不変性 (Invariance)** を持つべきです。

機械学習における不変性の重要性

例えば、ある画像認識モデルの仕事が「猫」を見つけることだとします。その猫が写真の中で大きく写っていても小さくても（**スケール不変性**）、画面の右端にいても中央にいても（**並進不変性**）、私たちはそれを同じ「猫」として認識できます。

モデルが人間のように頑健な認識能力を持つためには、その内部で学習される「猫らしさ」を表現する特徴量が、このような見かけ上の変化に対して鈍感でなければなりません。入力データのピクセルレベルの詳細情報から、こうした不変な概念的特徴をいかにして抽出するかが、モデルの性能を決定づけるのです。

物理学からの視点：スケール不変性と固定点

実は、この「不変性」という概念は、物理学、特に**くりこみ群 (RG)** の理論において中心的な役割を果たします。RG の文脈では、スケール不変性とは、系をどのような倍率で見ても（スケール変換しても）、その統計的な性質が変わらない状態を指します。

このような特殊な状態は、物質が相転移を起こす臨界点で見られます。RG の言葉を使うと、スケール不変な系は「くりこみフローの**固定点**」に対応します。くりこみ変換という情報の粗視化プロセスを何度も繰り返しても、もはや系の本質的な記述が変化しなくなった究極の状態です。この固定点に至る過程で、個々の物質のミクロな詳細（物理学者が「無関係 (irrelevant)」と呼ぶ情報）は洗い流され、マクロな振る舞いを支配する普遍的な性質だけが残ります。

RG 固定点と不変特徴量の対応

物理学において、くりこみ群フローがミクロな詳細を洗い流して**スケール不変な固定点**へと収束していくプロセスは、深層学習において、ネットワークが入力データの些末な変化（スケール、位置など）を吸収し、**変換に対して不変な特徴表現**を抽出していくプロセスに深く対応している。

CNN はいかにして不変な特徴に到達するか

CNN の階層構造は、まさにこの RG フローのように振る舞い、不変な特徴量を効率的に学習します。

- **畳み込みとプーリングの反復**：畳み込み層が局所的な特徴を検出し、プー ling 層がその特徴の正確な位置情報を「忘れる」ことで、局所的な並進不変性が生まれます。この操作が層を重ねるごとに繰り返されることで、より大域的で、より抽象的で、そしてより変換に対して不変な特徴が段階的に構築されていきます。
- **情報の蒸留**：入力画像のピクセル値という膨大な「ミクロな自由度」から出発し、各層はタスク（例：「猫」の分類）にとって「無関係 (irrelevant)」な情報を捨て、「関連のある (relevant)」な情報だけを次の層に渡します。最終的に、ネットワークの最深部では、入力画像の様々な変化に対して不変な、「猫」という概念そのものを表す、安定した特徴表現（RG の言葉で言う「固定点」）が得られるのです。

このアナロジーは、深層学習の「深さ」が持つ意味を、物理学の視点から照らし出してくれます。それは、データに内在する普遍的で本質的な構造を、見かけの些末な変化から分離して抽出するための、極めて合理的で強力なプロセスなのです。

7　拡散モデルと非平衡統計物理学

これまでの章で見てきたボルツマンマシンや、くりこみ群とアナロジーを結んだ深層学習モデルは、ある意味で「静的」な、あるいは「平衡状態」にあるシステムを扱っていると見なすことができます。エネルギー関数を定義し、その最も安定な状態を探したり、スケール不変な固定点へと収束させたり、といった具合です。

しかし、近年、画像生成などの分野で絶大な性能を発揮している**拡散モデル (Diffusion Models)** は、これらとは全く異なる物理的なアナロジーに基づいています。それは、系が時間とともに変化していく「動的」なプロセス、特に**非平衡統計物理学**の描く世界観です。

このセクションでは、データ生成というプロセスを、秩序だった状態が次第に乱雑になっていき（拡散過程）、そしてその時間を巻き戻すかのようにして秩序を回復する（逆拡散過程）という、非平衡な物理過程として捉え直します。

7.1　非平衡過程としてのデータ生成

拡散モデルの根底にあるアイデアは、直観的で美しく、そして二つの対照的なプロセスから成り立っています。

順過程：秩序から無秩序へ（情報の破壊）

まず、コップの中の澄んだ水に、一滴のインクを垂らす様子を想像してみてください。最初、インクは小さくまとまった、非常に秩序だった状態（情報量が多い状態）にあります。これが、私たちの手元にあるデータ、例えば一枚の猫の画像に対応します。

時間が経つにつれて、インクの粒子はブラウン運動によってランダムに動き回り、次第に水全体へと広がっていきます。この過程が**拡散**です。最終的に、インクは水全体に均一に混ざり合い、もとの「一滴のインク」という秩序だった構造は完全に失われ、最大限に乱雑な状態（情報量が少ない、ただの色のついた水）になります。

拡散モデルにおける**順過程 (Forward Process)** は、まさにこのインクの拡散と同じです。

1. 完璧なデータ（例：猫の画像 $\boldsymbol{x}_0$）から出発します。
2. このデータに、微小なノイズ（ガウスノイズ）を少しだけ加えます。これが時間ステップ 1 の状態 $\boldsymbol{x}_1$ です。
3. さらに $\boldsymbol{x}_1$ にノイズを加え、$\boldsymbol{x}_2$ を作ります。
4. この操作を何百、何千回と繰り返していきます（$\boldsymbol{x}_0 \to \boldsymbol{x}_1 \to \cdots \to \boldsymbol{x}_T$）。

最終的に、十分な時間 T が経過した後、元の猫の画像の構造は完全に破壊され、後に残るのはただのランダムなノイズ（正規分布に従うピクセルの集まり）だけになります。この順過程は、人間が設計した固定のプロセスであり、どんなデータも最終的には純粋なノイズへと変換します。

逆過程：無秩序から秩序へ（情報の生成）

ここからが生成モデルとしての本領発揮です。順過程が「秩序を破壊する」プロセスだったのに対し、私たちが本当にやりたいのは、その逆、すなわち**「無秩序から秩序を創造する」**ことです。

これは、均一に混ざったインク水を、再び元の一滴のインクに戻すようなものです。物理学の熱力学第二法則に逆らうかのような、時間を巻き戻すプロセスです。これが**逆過程 (Reverse Process)** です。

もちろん、このような奇跡は自然には起こりません。しかし、もし私たちが「神の手」を持っていて、水中の全てのインク粒子が、次の瞬間にどちらへ動けば元の位置に近づくかを正確に知っており、その方向にわずかに後押ししてやることができたとしたらどうでしょうか。

拡散モデルでは、この「神の手」の役割を**ニューラルネットワーク**が担います。

1. 純粋なランダムノイズ $\boldsymbol{x}_T$ から出発します。
2. 現在の状態 $\boldsymbol{x}_t$（ノイズまみれの画像）と現在の時刻 t をニューラルネットワークに入力します。
3. ネットワークは、状態 $\boldsymbol{x}_t$ に含まれているノイズの成分を予測するように学習します。
4. 予測されたノイズを、現在の状態 $\boldsymbol{x}_t$ からわずかに引き算することで、一つ前の、よりクリーンな状態 $\boldsymbol{x}_{t-1}$ を推定します。
5. この「ノイズ除去」のプロセスを T 回繰り返すことで（$\boldsymbol{x}_T \to \boldsymbol{x}_{T-1} \to \cdots \to \boldsymbol{x}_0$）、最終的に純粋なノイズから、本物そっくりのデータ（猫の画像）を生成するのです。

拡散モデルと非平衡物理学

拡散モデルは、データ生成を時間発展する物理過程として捉える。

- **順過程**：データにノイズを加え、平衡状態（完全な無秩序）へと向かう、物理的な**拡散過程**。
- **逆過程**：ノイズから出発し、学習されたニューラルネットワークの助けを借りて時間を巻き戻し、非平衡な経路を辿って秩序を生成する、**逆時間発展過程**。

この動的な視点は、静的なエネルギー地形を考えるこれまでのモデルとは全く異なり、物理学におけるランジュバン方程式やフォッカー・プランク方程式といった、非平衡過程を記述する理論と深く結びついています。

7.2　フォッカー・プランク方程式と順過程

前節では、拡散モデルの順過程を、データに少しずつノイズを加えていく離散的な時間ステップのプロセスとして紹介しました。

$$\boldsymbol{x}_t = \sqrt{1-\beta_t}\boldsymbol{x}_{t-1} + \sqrt{\beta_t}\boldsymbol{\epsilon}_{t-1}$$

このプロセスは、物理学者が連続的な時間で記述する、より根源的な物理過程の「デジタルシミュレーション」と見なすことができます。この連続時間での記述こそが、非平衡統計物理学との強力な接点となります。

連続時間での記述：ランジュバン方程式

時間ステップを無限に小さくする極限を考えると、上記のプロセスは**確率的微分方程式 (Stochastic Differential Equation, SDE)** で記述されます。物理学では、このような方程式は**ランジュバン方程式**として知られています。これは、個々の粒子（例えば、水中のインク粒子）の運動を記述する「粒子視点」の方程式です。

順過程に対応するランジュバン方程式は、次のような構造を持ちます。

$$d\boldsymbol{x} = \underbrace{-\frac{1}{2}\beta(t)\boldsymbol{x}dt}_{\text{ドリフト項}} + \underbrace{\sqrt{\beta(t)}d\boldsymbol{W}_t}_{\text{拡散項}}$$

この式は、データ点 $\boldsymbol{x}$ の微小時間 dt での変化 $d\boldsymbol{x}$ が、二つの力の合算で決まることを意味しています。

- **ドリフト項（流れの項）**：系を決定論的に原点 $(\boldsymbol{x} = 0)$ へと引き寄せる力です。$-\boldsymbol{x}$ に比例することから、原点から遠いほど強く中心に引き戻されることがわかります。
- **拡散項（ノイズの項）**：$d\boldsymbol{W}_t$ はウィーナー過程と呼ばれるランダムな力（ブラウン運動）を表します。これが、粒子をランダムに揺さぶり、拡散させる原因となります。

つまり、順過程とは、全てのデータ点を原点に引き寄せながら、同時にランダムなノイズでかき混ぜていくプロセスなのです。

確率分布のダイナミクス：フォッカー・プランク方程式

ランジュバン方程式が一個の粒子の軌跡を追いかけるのに対し、**フォッカー・プランク方程式**は、粒子全体の集団、すなわち**確率分布 $p(\boldsymbol{x}, t)$ そのものが時間とともにどう変化していくか**を記述します。これは、インクの「雲」全体の濃度分布がどう広がっていくかを記述する「大局的視点」の方程式です。

上記のランジュバン方程式に対応するフォッカー・プランク方程式は、以下のように書かれます。

$$\frac{\partial p(\boldsymbol{x}, t)}{\partial t} = -\nabla \cdot \left(-\frac{1}{2}\beta(t)\boldsymbol{x}p(\boldsymbol{x}, t) \right) + \nabla^2 \left(\frac{1}{2}\beta(t)p(\boldsymbol{x}, t) \right)$$

この方程式もまた、ドリフトと拡散という二つの効果で、確率分布の時間発展を記述しています。

- **ドリフト項**：確率分布の「山」全体を、原点に向かって移動させる効果を持ちます。
- **拡散項**：確率分布の「山」を、時間とともになだらかに広げ、平坦にしていく効果を持ちます。

順過程の物理的描像

拡散モデルの順過程は、非平衡物理学の言語で二通りに記述できる。

- **ランジュバン方程式**：一個のデータ点 $\boldsymbol{x}$ が、ノイズを受けながら原点へと向かう**ミクロな軌跡**を記述する。
- **フォッカー・プランク方程式**：データ全体の確率分布 $p(\boldsymbol{x})$ が、その形を崩しながら原点を中心とするガウス分布へと変化していく**マクロなダイナミクス**を記述する。

この物理的描像は、拡散モデルの順過程が、単なる思いつきの操作ではなく、**任意の初期分布を、最終的に単純なガウス分布という名の「熱平衡状態」へと緩和させる、物理的に自然なプロセス**であることを示しています。

この明確な物理的基盤があるからこそ、次のステップである「時間の巻き戻し」、すなわち逆過程を理論的に考察することが可能になるのです。

7.3 時間反転対称性と逆過程

順過程が、物理的に自然な「拡散」として、データをノイズへと変換するプロセスであることを見てきました。これは、エントロピーが増大していく、不可逆な過程のように思えます。インクが混ざった水を、自然に元の一滴に戻すことはできません。

しかし、物理学のミクロな法則（例えば、粒子一つ一つの運動方程式）は、多くの場合、時間反転に対して対称です。つまり、時間を逆再生しても、物理法則としては成立します。このミクロなレベルでの可逆性と、マクロなレベルでの不可逆性のギャップを埋め、拡散過程の「時間を巻き戻す」レシピを数学的に与えるのが、非平衡物理学における時間反転 SDE（確率的微分方程式）の理論です。

拡散過程の時間反転

驚くべきことに、前節で見た順過程の SDE

$$d\boldsymbol{x} = f(\boldsymbol{x}, t)dt + g(t)d\boldsymbol{W}_t$$

に対して、時間を T から 0 に向かって逆再生する、対応する**逆過程の SDE** が存在することが知られています。その方程式は、以下のような形をしています。

$$d\boldsymbol{x} = \left[-f(\boldsymbol{x}, t) + g(t)^2 \nabla_{\boldsymbol{x}} \ln p(\boldsymbol{x}, t)\right] dt + g(t)d\bar{\boldsymbol{W}}_t$$

ここで、$d\bar{\boldsymbol{W}}_t$ は逆向き時間のウィーナー過程です。この方程式を順過程のものと比較すると、二つの重要な点がわかります。

1. ドリフト項 $-f(\boldsymbol{x}, t)$ は、元のドリフトの向きを単純に反転させたものです。
2. しかし、それだけではなく、$\boldsymbol{g(t)^2 \nabla_x \ln p(x, t)}$ という追加のドリフト項が現れます。

スコア関数：時間を巻き戻すための「羅針盤」

この追加項こそが、時間を巻き戻すための鍵です。特に、その中心的な要素である $\nabla_{\boldsymbol{x}} \ln p(\boldsymbol{x}, t)$ は、**スコア関数 (Score Function)** と呼ばれます。

スコア関数

スコア関数 $\nabla_{\boldsymbol{x}} \ln p(\boldsymbol{x}, t)$ は、ある時刻 t における確率分布 $p(\boldsymbol{x}, t)$ の対数を、状態 $\boldsymbol{x}$ に関して勾配をとったものである。これは、その地点 $\boldsymbol{x}$ において、**確率が最も急激に増加する方向**を指し示すベクトル場となる。

逆過程の SDE が意味するところは、極めて直観的です。拡散を巻き戻すためには、元のドリフトを反転させるだけでなく、各ステップで「確率がより高い方向」へと向かうような、追加の「押し」が必要なのです。スコア関数は、この「押し」の方向と強さを教えてくれる、まさに**時間を巻き戻すための羅針盤**の役割を果たします。

ニューラルネットワークによるスコア関数の学習

しかし、ここでもまた、計算不可能性の壁が立ちはだかります。スコア関数 $\nabla_{\boldsymbol{x}} \ln p(\boldsymbol{x}, t)$ は、時間とともに変化していく未知の確率分布 $p(\boldsymbol{x}, t)$ に依存するため、直接計算することはできません。

ここで、深層学習がその真価を発揮します。私たちは、この**未知のスコア関数を、ニューラルネットワーク $s_\theta(\boldsymbol{x}, t)$ を用いて近似する**のです。

$$s_\theta(\boldsymbol{x}, t) \approx \nabla_{\boldsymbol{x}} \ln p(\boldsymbol{x}, t)$$

ニューラルネットワークは、様々な時刻 t とノイズまみれのデータ $\boldsymbol{x}_t$ を入力とし、その時刻におけるスコア（ノイズが加えられた方向）を予測するように訓練されます。

拡散モデルにおける生成の物理的意味

拡散モデルによるデータ生成とは、物理的に正当な**時間反転した拡散過程（逆過程 SDE）**のシミュレーションである。
そのシミュレーションには、本来であれば未知である「スコア関数」が必要となるが、その役割を**データから学習したニューラルネットワーク**が担う。

こうして、純粋なノイズ $\boldsymbol{x}_T$ から出発し、ニューラルネットワークという羅針盤に導かれながら、逆過程 SDE を数値的に解いていくことで、私たちは時間を遡り、最終的に秩序だったデータ $\boldsymbol{x}_0$ へとたどり着くことができるのです。

これは、生成モデルの学習という問題を、非平衡物理過程における時間反転に必要な「力場」を学習する問題へと見事に読み替えた、物理学と機械学習の幸福な融合と言えるでしょう。

第 V 部

機械学習の最前線 ー学習ダイナミクスと相転移ー

8　学習過程の物理学

これまでの部で、私たちは主に「学習が完了したモデル」を物理的な系と見なして、その性質を分析してきました。統計物理学の平衡状態や、非平衡過程の時間発展といったアナロジーは、モデルの構造や性能を理解するための強力なレンズとなりました。

しかし、現代の深層学習が提起する最も興味深く、そして難解な問いの多くは、学習の「結果」だけでなく、学習の「過程」そのものに潜んでいます。何百万、何億というパラメータを持つ巨大なニューラルネットワークが、なぜ訓練データに完璧にフィットしつつも、未知のデータに対して高い汎化性能を保てるのか。この「過学習の謎」は、モデルの最終的な状態だけを見ていては完全には理解できません。

この部では、視点を大きく転換し、**学習プロセスそのものを、時間発展する物理的なダイナミクスとして捉え直します。**

学習という名の高次元空間探査

この新しい視点では、以下の対応関係を考えます。

- **物理的な系** $\longleftrightarrow$ **ニューラルネットワーク**
- **系の状態** $\longleftrightarrow$ **ネットワークの全パラメータ（重み）** θ
- **エネルギー地形** $\longleftrightarrow$ **損失関数が作るランドスケープ** $L(\theta)$
- **時間発展の法則** $\longleftrightarrow$ **学習アルゴリズム（例：勾配降下法）**

学習とは、パラメータという名の粒子が、損失という名のポテンシャル地形の上を、学習アルゴリズムという運動法則に従って転がり落ちていく、壮大な高次元空間の探査プロセスなのです。

この描像に立つことで、私たちは物理学、特に複雑系や統計力学の言語を用いて、学習中に起こる様々な現象を分析することができます。例えば、学習の進展に伴って、ネットワークの性質が突如として劇的に変化する現象が観測されることがあります。これは、物理学における**相転移 (Phase Transition)**、すなわち水が氷になるような、系のマクロな性質の質的な変化と非常によく似ています。

この部で探求する問い

- なぜ巨大なネットワークは、悪い局所解に陥ることなく、汎化性能の高い解を見つけられるのか？（損失ランドスケープの幾何学）
- 学習の途中で見られる急激な性能の変化は、「相転移」として記述できないか？

- 過学習は、ある種の臨界現象として理解できないか？

この部は、現在進行形で世界中の研究者が探求している、まさに「最前線」の領域です。確立された理論だけでなく、刺激的なアイデアや未解決の問題も多く含みます。物理学の最も強力な概念の一つである「相転移」を武器に、深層学習の謎に満ちたダイナミクスの世界へと足を踏み入れていきましょう。

8.1　損失関数のランドスケープという考え方

学習プロセスを物理系として捉えるとき、その最も基本的な舞台設定が**損失関数のランドスケープ (Loss Landscape)** という考え方です。これは、学習のダイナミクスを直観的に理解するための、非常に強力なメタファーとなります。

学習とは地形を転がり落ちる旅

まず、このランドスケープがどのようなものかを定義しましょう。

- **場所（座標）**：モデルが持つ全てのパラメータ $\theta = (w_1, w_2, \ldots, w_D)$ の値の組。パラメータが 100 万個あれば、ここは 100 万次元の空間となります。
- **標高**：その「場所」（パラメータの組）における、**損失関数の値** $L(\theta)$ 。訓練データに対するモデルの性能が悪いほど、「標高」は高くなります。

この定義に従うと、学習とは、この広大な高次元地形の上に置かれたボール（現在のパラメータ θ）が、重力（勾配降下法）に引かれて、最も標高の低い谷底（損失の最小値）を目指して転がり落ちていく旅路として描くことができます。

このランドスケープの「地形」がどのような形をしているかは、学習の難易度を決定的に左右します。もし地形が、なめらかで一つの谷底（大域的最小解）しかないお椀のような形（凸関数）をしていれば、学習は非常に簡単です。どこからボールを転がし始めても、必ず同じ最低地点にたどり着きます。

高次元ランドスケープの意外な性質

古典的な機械学習の文脈では、ニューラルネットワークの損失ランドスケープは、極めて複雑で厄介な地形だと考えられてきました。性能の悪い「谷底」（**悪い局所解**）が無数に存在し、ボールがいったんそこに囚われると、二度と抜け出せずに学習が停滞してしまう、と恐れられていたのです。

しかし、深層学習に関する近年の研究は、この古典的なイメージが、特に現代の巨大な（過剰パラメータな）ネットワークにおいては、必ずしも正しくないことを示唆しています。高次元のランドスケープは、私たちの 3 次元的な直観が及ばない、驚くべき性質を持っているのです。

1. **悪い局所解は稀である**：理論的・実験的研究により、巨大なネットワークの損失ランドスケープでは、ほとんどの局所解が、大域的最小解とほぼ同じくらい良い性能（低い損失値）

を持つことが分かってきました。つまり、「質の悪い谷底」は実は非常に稀で、学習が局所解に「ハマった」としても、それはほとんどの場合、性能の良い解なのです。

2. **鞍点（あんてん）の優勢**：局所解よりもはるかに多く存在するのが、**鞍点（Saddle Point）**です。鞍点とは、ある方向から見れば谷底（極小）ですが、別の方向から見れば山の尾根（極大）になっているような場所です。通常の勾配降下法は、このような平坦な領域で学習の速度が著しく低下することがありますが、近年の最適化アルゴリズム（Adam など）は、こうした鞍点を効率的に脱出するメカニズムを備えています。
3. **解の多様性**：物理学の知見によれば、高次元空間では、孤立した「点」として存在する解よりも、解が広大な「部分空間」を形成する方が一般的です。これは、性能の良い解（谷底）が、針で突いたような一点ではなく、広くて平らな盆地のように存在していることを意味します。この「平坦な最小解」の存在が、深層学習モデルの高い汎化性能と密接に関わっていると考えられています。

ランドスケープに対する視点の変化

かつては学習を妨げる「罠」に満ちていると恐れられた損失ランドスケープは、現代の深層学習の文脈では、**高次元性のおかげで、むしろ探索しやすい、性質の良い地形をしている**と考えられるようになっている。

このように、学習のダイナミクスを理解するためには、まずその舞台となる損失ランドスケープの幾何学的な性質を理解することが不可欠です。次の節では、この地形の上で実際に学習が進む中で、ネットワーク全体がどのようにその性質を変化させていくか、「相転移」という視点から見ていきます。

8.2　学習と相転移現象

損失ランドスケープという静的な地形の上を、パラメータが転がり落ちていく。この描像は直観的ですが、学習中に起こる最も興味深い現象のいくつかを説明するには、もう一歩踏み込む必要があります。それは、パラメータ集団の「全体としての振る舞い」が、学習の進行に伴ってどのように変化するか、という動的な視点です。

ここで、物理学における最も劇的で美しい概念の一つである、**相転移 (Phase Transition)** のアナロジーが力を発揮します。

物理学における相転移

相転移とは、ある制御パラメータ（例えば温度）を連続的に変化させたとき、系のマクロな性質が、ある点を境に**不連続かつ劇的に**変化する現象です。最も身近な例は、水が氷になる「凝固」です。

温度を 1 ℃から 0.1 ℃へと滑らかに下げても、水は液体のままで、その性質は連続的に変化するだけです。しかし、0 ℃という臨界点を跨いだ瞬間、分子が自由に動き回る無秩序な液体状態から、分子が結晶格子に固定された、極めて秩序だった固体（氷）へと、系の状態は質的に全く異な

るものに変化します。 ☒

このような相転移は、個々の分子の振る舞いだけを見ていては理解できません。それは、無数の分子間の相互作用から生まれる、**協力的な集団現象（創発現象）**なのです。

8.2.1 ：

深層学習における「相転移」では、パラメータを連続的に更新していく深層学習のプロセスに、これと似たような劇的な変化は存在するのでしょうか。近年の研究は、その答えが「イエス」であることを強く示唆しています。

学習ダイナミクスにおける相転移

ニューラルネットワークの学習過程において、訓練の反復回数やデータセットのサイズといった制御パラメータを連続的に変化させると、モデルの汎化性能や内部表現といったマクロな性質が、ある点を境に**急激に、質的に変化する**現象が観測されている。

具体的には、以下のような現象が「学習における相転移」の候補として活発に研究されています。

- **急激な汎化（”Grokking”）**：モデルが訓練データに対する正解率を 100% にした後も、未知のデータに対する正解率はランダムなまま、という状態が長く続くことがあります。しかし、さらに学習を続けると、ある時点で突如として、未知のデータに対してもほぼ完璧な正解率を示すようになるのです。これは、モデルが単なる「暗記」のフェーズから、本質を理解した「汎化」のフェーズへと、相転移したかのように見えます。
- **ダブル・ディセント (Double Descent)**：モデルのパラメータ数を増やしていくと、性能は向上しますが、ある点を超えると過学習によって一旦悪化します。これは古典的な統計学の常識です。しかし、さらにパラメータ数を増やしていくと、驚くべきことに、再び性能が向上し始めるという現象です。これは、モデルが「過学習」の相から、別の新しい「良質な補間」の相へと移行することを示唆しています。

これらの現象は、損失ランドスケープの特定の谷を見つけるという静的な描像だけでは説明が困難です。そうではなく、学習のダイナミクスそのものが、ある臨界点を超えると、パラメータ全体の集団的な振る舞いをがらりと変え、それまで到達できなかったような、性質の異なる解（例えば、暗記解から汎化解へ）へとアクセスを可能にする、と考える方が自然です。

この「学習と相転移」というアナロジーは、深層学習の謎を解き明かすための、まさに最前線の研究テーマです。それは、学習というプロセスを、単なる最適化ではなく、無数のパラメータが相互作用し、自己組織化していく、豊かで複雑な物理現象として捉える、新しい扉を開くものなのです。

8.3　スピングラス理論と汎化能力

本書の序盤、ホップフィールドネットワークとの関連で登場した**スピングラス理論**を、ここで再び取り上げます。かつては連想記憶モデルのアナロジーとして登場しましたが、ここでは深層学習の損失ランドスケープそのものを、スピングラスのエネルギー地形として捉え直します。この視点は、なぜ巨大なネットワークが高い**汎化能力**を持つのか、という深層学習の根源的な謎に光を当てる可能性を秘めています。

汎化と損失ランドスケープの形状

経験的に、深層学習において高い汎化性能を示すモデルは、損失ランドスケープの**幅が広く、平坦な最小解 (wide, flat minima)** に対応することが知られています。一方で、訓練データにだけ過剰に適合（過学習）してしまったモデルは、**幅が狭く、鋭い最小解 (sharp minima)** に対応する傾向があります。

この違いは、直観的に次のように理解できます。

- **鋭い最小解**：パラメータを少しでも動かすと損失が急激に増加する、非常に敏感な解です。これは、訓練データに含まれるノイズの一 つ一つにまで過剰に反応し、無理やり損失を下げた「暗記型」の解に対応します。テストデータのように、少しでも分布がずれると、性能が大きく劣化してしまいます。
- **平坦な最小解**：パラメータをある程度動かしても損失があまり変わらない、頑健（ロバスト）な解です。これは、データの細かいノイズを無視し、本質的で単純な構造を捉えた「汎化型」の解に対応します。このような解は、未知のデータに対しても安定した性能を発揮します。

したがって、「なぜ深層学習は汎化するのか？」という問いは、「なぜ学習アルゴリズムは、無数にある最小解の中から、都合よく平坦な最小解を見つけ出すことができるのか？」という、損失ランドスケープの幾何学に関する問いへと言い換えることができます。

スピングラス理論からの答え

驚くべきことに、この「平坦な最小解の優位性」は、まさにスピングラス理論が予測する高次元エネルギー地形の性質そのものなのです。

スピングラスのエネルギー地形は、単純な谷底の集まりではありません。その低エネルギー状態は、**互いによく似た多数の状態が集まって、広大な「盆地」や「台地」のような領域を形成する**という、階層的な構造を持つことが知られています。鋭く孤立した谷底よりも、広大で平坦な領域の方が、状態空間の中で圧倒的に大きな「体積」を占めているのです。

スピングラス理論と汎化のアナロジー

深層学習の損失ランドスケープを、スピングラスのエネルギー地形と同一視する。

- 汎化性能の高い**平坦な最小解**の領域は、スピングラスにおける**広大でエントロピーの高い低エネルギー状態のクラスター**に対応する。
- 学習アルゴリズム（特に SGD）は、その確率的な揺らぎによって、鋭い谷底を通り抜け、より体積の大きい（見つけやすい）平坦な領域へと引き寄せられる性質を持つ可能性がある。

このアナロジーが正しければ、深層学習の成功は、単にアルゴリズムが優れているから、というだけではありません。**高次元パラメータ空間が持つ、物理学の法則に根差した普遍的な幾何学的性質そのものが、モデルを自然と汎化しやすい解へと導いている**のかもしれないのです。

物理学者がスピングラスの複雑な解空間を分析するために開発したレプリカ法などの高度な数学的ツールは、現在、深層学習の汎化の謎を解き明かすための新しい武器として期待されています。

統計物理学、特に無秩序系の理論は、単なるアナロジーを超えて、機械学習の最も根源的な問いに答えるための、新しい数学的言語を提供する可能性を秘めているのです。この広大なフロンティアの探求は、まだ始まったばかりです。

第 VI 部
結論

9　物理学の言葉で語る機械学習

本書の旅は、一見すると遠く離れた二つの分野、物理学と機械学習の出会いから始まりました。私たちは、統計物理学の洗練された言語体系を羅針盤として、現代機械学習の理論の海を航海してきました。この旅路を、ここで振り返ってみたいと思います。

まず私たちは、両分野の根底に「エネルギー」と「確率分布」という**共通言語**が存在することを明らかにしました。物理学における自由エネルギーの最小化原理が、ベイズ推論における尤度の最大化という、機械学習の指導原理と数学的に等価であることを見たとき、両者の結びつきが単なる表面的なアナロジーではないことを確信しました。

次に、この理論的枠組みが直面する「計算不可能性」という共通の壁に対し、両分野がいかにして酷似した解決策を見出してきたかを探りました。**平均場近似と変分推論**という決定論的なアプローチ、そして**マルコフ連鎖モンテカルロ法**という確率的なアプローチは、異なる名前で呼ばれながらも、その思想において見事な対応関係を示していました。

そして、現代 AI の中核である深層学習へと視点を移し、その「深さ」の意味を物理学の概念で読み解きました。**くりこみ群**のアナロジーは、深層ネットワークの階層的な特徴抽出を、ミクロな詳細を捨てマクロな本質に至る情報の「蒸留」プロセスとして描き出しました。また、**拡散モデ**

ルと非平衡統計物理学の対応は、データ生成という創造的なプロセスを、秩序と無秩序の間を行き来する、動的な時間発展過程として捉える視点を与えてくれました。

最後に私たちは、学習の「結果」だけでなく「過程」そのものに目を向け、**学習ダイナミクス**を物理現象として考察しました。高次元の**損失ランドスケープ**の幾何学、そして学習中に観測される劇的な変化を**相転移**として捉える視点は、なぜ深層学習がこれほどまでに成功しているのか、その根源的な謎に迫るための、まさに最前線の研究領域です。

アナロジーを超えて

本書を通して、私たちは何度も「アナロジー」という言葉を使ってきました。しかし、この旅を終えた今、その結びつきは単なるアナロジーやメタファー以上の、より深く、構造的なものであると感じていただけたのではないでしょうか。

物理学とは、元来、無数の構成要素からなる複雑な系が、全体としてどのような法則に従うのかを探求する学問です。そのために物理学者が発展させてきた数学的な言語や概念的枠組みは、本質的に「集団現象」を記述するためのものです。

一方で、現代の機械学習モデル、特に深層学習は、何百万、何億というパラメータ（自由度）が相互作用し、データという外部環境に適応することで、知性という名の「創発現象」を引き起こす、人類が作り出した最も複雑な人工システムです。

そう考えれば、物理学の言葉で機械学習を語ることは、決して突飛なことではありません。むしろ、このような巨大で複雑なシステムを理解するために、物理学の言語を用いるのは、極めて自然で、必然的なアプローチであるとさえ言えるでしょう。

この視点は、私たちに機械学習を理解するための新しい「直観」を与えてくれます。そして、その直観は、今後の AI 研究における新たなフロンティアを指し示しています。物理学における対称性や保存則の概念を応用した、よりデータ効率が良く、信頼性の高いモデルの構築。あるいは、因果推論や解釈可能性といった難問に、物理的な実在論の視点からアプローチすること。

情報と知性の物理学は、まだ始まったばかりです。本書が、読者の皆様にとって、この刺激的なフロンティアを探求するための一助となれば、それに勝る喜びはありません。

9.1　理論体系の総括

本書を通じて、私たちは物理学と機械学習という二つの広大な大陸の間に、いくつもの橋を架けてきました。この最終節では、その全体像を改めて俯瞰し、本書が提示してきた「物理学の言葉で語る機械学習」という理論体系を総括します。

この理論体系の核心は、二つの分野における概念の対応関係、いわば「翻訳辞書」を構築することにありました。その対応関係は、以下の表のようにまとめられます。

物理学の概念	機械学習の概念	本書の対応する部
I. 平衡系の統計力学 $\longleftrightarrow$ 確率的モデリング		
エネルギー関数 $E(\boldsymbol{x})$	(負の) 対数確率 $-\ln P(\boldsymbol{x})$	_
分配関数 Z	尤度 $P(D)$	第 II 部
自由エネルギー F	(負の) 対数尤度 $-\ln P(D)$	(共通言語)
II. 近似計算手法 $\longleftrightarrow$ 近似推論アルゴリズム		
平均場近似	変分推論	第 III 部
マルコフ連鎖モンテカルロ法	MCMC サンプリング	(近似手法)
III. 階層・スケールとダイナミクス $\longleftrightarrow$ 深層学習と生成モデル		
くりこみ群	階層的特徴学習 (CNN など)	_
スケール不変性（固定点）	変換不変な特徴表現	第 IV 部
非平衡過程（拡散）	拡散モデル (生成過程)	(深層学習)
時間反転対称性	スコアマッチングによる逆過程	_
IV. 複雑系・無秩序系の物理学 $\longleftrightarrow$ 学習のダイナミクス		
エネルギー地形	損失ランドスケープ	_
相転移	急激な汎化・ダブルディセント	第 V 部
スピングラス理論	汎化能力の理論	(最前線)

この対応表は、本書が辿ってきた道のりを要約しています。

まず第一に、私たちはボルツマンマシンのようなモデルを「平衡状態にある物理系」と見なすことで、両分野の**静的な構造**が共通の言語で記述できることを見ました (I)。

次に、分配関数や事後分布の計算という共通の困難に対し、物理学と機械学習が、それぞれ**決定論的近似（平均場・変分推論）**と**確率論的近似（MCMC）**という、同じ発想の計算手法を独立に発展させてきたことを確認しました (II)。

さらに現代的な深層学習へと進み、その**動的なプロセスや構造**に物理学のアナロジーを適用しました。くりこみ群はネットワークの**階層性**を、非平衡物理学は拡散モデルの**生成プロセス**を、そして複雑系の物理学は**学習ダイナミクス**そのものを理解するための、強力な視座を提供してくれました (III, IV)。

この理論体系が示すのは、単なる興味深い類似点の寄せ集めではありません。それは、**「多数の自由度が相互作用する複雑な系を、限られた情報からいかにして理解し、モデル化するか」**という、両分野に共通する根源的な課題に対する、驚くほど収斂した知的営為の姿です。

物理学が自然という究極の複雑系を理解するために築き上げてきた概念的道具立てが、知性というもう一つの複雑な現象を人工的に構築しようとする機械学習の分野で、再びその力を発揮している。この事実は、私たちに、分野の垣根を超えた知識の普遍性について、深く考えさせてくれるのではないでしょうか。

9.2　今後の展望：因果推論と解釈可能性への視点

本書で展開してきた物理学と機械学習の理論体系は、二つの分野の間にいかに深く豊かな関係が築かれているかを示してきました。しかし、現代の人工知能が直面している課題は、単にデータへの適合精度を高めることだけではありません。最後に、この物理的な視点が、AI の未来における二つの大きな挑戦、すなわち**解釈可能性**と**因果推論**に対して、どのような展望を与えうるかを考えてみましょう。

解釈可能性への視座：有効な自由度は何か？

深層学習モデルは、しばしば「ブラックボックス」と批判されます。なぜモデルがそのような判断を下したのか、その内部メカニズムを人間が理解するのは極めて困難です。

この問題に対して、本書で議論した物理学の概念は、新しい解釈の糸口を提供します。

- **くりこみ群と有効自由度**：くりこみ群の思想は、無数のミクロな自由度の中から、マクロな現象を支配する本質的な「有効自由度」を抽出するプロセスでした。同様に、ニューラルネットワークの階層構造を解析し、その決定を支配している少数の「有効な特徴量」や「概念的な自由度」を特定できるかもしれません。これは、複雑なモデルの振る舞いを、人間が理解できる低次元の記述へと圧縮するアプローチです。
- **相転移と秩序パラメータ**：相転移は、「秩序パラメータ」というマクロな量によって特徴づけられます。モデルが何かを「認識」する状態を、ある種の相転移と見なすならば、その認識に対応する「秩序パラメータ」がネットワークの内部に存在するはずです。この秩序パラメータを特定する試みは、モデルが何を学習したのかを理解する上で、本質的な洞察を与える可能性があります。

因果推論への視座：相関の先へ

現代の機械学習モデルの多くは、データに潜む**相関関係**を見つけ出すことには長けていますが、**因果関係**を理解することはできません。「鶏の鳴き声」と「日の出」が強く相関していても、鶏が太陽を昇らせるわけではないことを、モデルは知りません。科学や医療、政策決定など、世界に「介入」する必要がある分野では、この限界は致命的です。

物理学は、本質的に**因果の学問**です。その法則は、ある原因（力）がどのような結果（運動）をもたらすか、という時間発展の形で記述されます。

- **対称性と保存則**：物理学における対称性の要請は、系が従うべき普遍的な因果構造を規定します。このような物理的な構造（既知の因果関係）をモデルのアーキテクチャに組み込むアプローチ（物理情報ニューラルネットワークなど）は、単なる相関を超えた推論を可能にする上で重要です。
- **動的な世界観**：拡散モデルで見たような、時間の矢を明確に意識した非平衡物理学の視点は、原因から結果へと流れる因果の構造と親和性が高いです。静的な平衡状態を仮定する

モデルから、動的な生成プロセスを記述するモデルへの移行は、因果推論への道を開くかもしれません。

宇宙の法則を探求する物理学の営みと、知性の法則を構築しようとする機械学習の営み。この二つの壮大な冒険は、今、まさに交差しようとしています。本書で探検してきた共通の言語は、未来の科学と技術の姿を垣間見せてくれる、一つの窓であったのかもしれません。この知のフロンティアにおける探求は、まだ始まったばかりです。

参考文献

1. **第 I 部 序論 ―二つの分野の共鳴―**
   - [1]：1.2 節「スピングラス理論とホップフィールドネットワーク」に対応。エネルギー最小化と連想記憶の対応を原論文で提示。
   - [2]：1.2 節の「記憶容量」言及に対応。ホップフィールドの容量解析。
   - [3]：1 章末〜3.3 節の**ボルツマンマシン**登場の歴史的流れに対応。
   - [4], [5]：1.2 節のスピングラス背景（レプリカ法・フラストレーション）に対応。
2. **第 II 部 共通言語としての確率分布とエネルギー**
   - [6]：2.1 節「最大エントロピー原理」に対応する古典的出典。
   - [7], [8]：2.1 節の情報エントロピーと推論の基礎。最大エントロピーの解釈に対応。
   - [9], [10]：2.2 節「分配関数と自由エネルギー」の物理的意味。
   - [11]：3.3 節「エネルギーベースモデル（EBM）」の総説。
   - [12]：3.4 節「自由エネルギーと尤度の対応」および ELBO 視点の基礎。
   - [13]：3.4 節の「自由エネルギー最小化と EM」の対応関係に直接対応。
3. **第 III 部 近似手法のアナロジー**
   - [14]：4 章の平均場近似・ベイズ推論・グラフ的モデルの橋渡し（Bethe/MF）。
   - [15], [16]：4.2 節「変分推論」の体系的まとめ。
   - [17]：4.2 節の**変分推論の実践**（ブラックボックス VI など）。
   - [18], [19], [20], [21]：5 章（MCMC）の古典から HMC まで。
4. **第 IV 部 深層学習の物理的解釈**
   - [22], [23]：6.1 節「くりこみ群の思想」の原典。
   - [24], [25]：6.2 節「深層と RG のアナロジー」に対応（RBM と RG の写像，階層表現の物理的理解）。
   - [26], [27], [28, 29]：7 章「拡散モデルと非平衡統計物理学」に対応（順過程・逆過程・スコアの定式化）。
   - [30], [31]：7.2〜7.3 節のフォッカー・プランク方程式と時間反転拡散に対応。
5. **第 V 部 機械学習の最前線 ―学習ダイナミクスと相転移―**
   - [32], [33]：8 章「学習過程の物理学」の総論（統計力学的学習理論）。
   - [34]：8.1 節の損失ランドスケープとスピングラス近似の接続に対応。

- [35], [36], [37], [38]：8.1〜8.3 節の「フラット最小・汎化・局所エントロピー」の関係に対応。
- [39]：8.2 節の深層学習に対する統計力学的視座の俯瞰。

6. **第 VI 部 結論（因果・解釈可能性）**
   - [40]：9.2 節の因果推論の基礎。
   - [41], [42]：9.2 節「解釈可能性」への導入。

参考文献

[1] J. J. Hopfield, “Neural networks and physical systems with emergent collective computational abilities,” *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982.

[2] D. J. Amit, H. Gutfreund, and H. Sompolinsky, “Storing infinite numbers of patterns in a spin-glass model of neural networks,” *Physical Review Letters*, 55(14):1530–1533, 1985.

[3] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, “A learning algorithm for Boltzmann machines,” *Cognitive Science*, 9(1):147–169, 1985.

[4] M. Mézard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond*. World Scientific, 1987.

[5] H. Nishimori, *Statistical Physics of Spin Glasses and Information Processing*. Oxford University Press, 2001.

[6] E. T. Jaynes, “Information theory and statistical mechanics,” *Physical Review*, 106(4):620–630, 1957; and 108(2):171–190, 1957.

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Wiley, 2006.

[8] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[9] H. B. Callen, *Thermodynamics and an Introduction to Thermostatistics*, 2nd ed. Wiley, 1985.

[10] R. K. Pathria and P. D. Beale, *Statistical Mechanics*, 3rd ed. Elsevier, 2011.

[11] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang, “A tutorial on energy-based learning,” in *Predicting Structured Data*, 2006 (tutorial manuscript).

[12] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[13] R. M. Neal and G. E. Hinton, “A view of the EM algorithm that justifies incremental, sparse, and other variants,” in *Learning in Graphical Models*, pp. 355–368, Kluwer, 1998.

[14] M. Mézard and A. Montanari, *Information, Physics, and Computation*. Oxford University Press, 2009.

[15] M. J. Wainwright and M. I. Jordan, “Graphical models, exponential families, and variational inference,” *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008.

[16] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Machine Learning*, 37:183–233, 1999.

[17] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[18] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equations of state calculations by fast computing machines," *Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[19] W. K. Hastings, "Monte Carlo sampling methods using Markov chains and their applications," *Biometrika*, 57(1):97–109, 1970.

[20] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

[21] R. M. Neal, "MCMC using Hamiltonian dynamics," in *Handbook of Markov Chain Monte Carlo*, pp. 113–162, CRC Press, 2011.

[22] L. P. Kadanoff, "Scaling laws for Ising models near T_c," *Physics*, 2:263–272, 1966.

[23] K. G. Wilson, "Renormalization group and critical phenomena. I. Renormalization group and the Kadanoff scaling picture," *Physical Review B*, 4(9):3174–3183, 1971.

[24] P. Mehta and D. J. Schwab, "An exact mapping between the variational renormalization group and deep learning," arXiv:1410.3831, 2014.

[25] H. W. Lin, M. Tegmark, and D. Rolnick, "Why does deep and cheap learning work so well?," *Journal of Statistical Physics*, 168:1223–1247, 2017.

[26] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *ICML*, 2015.

[27] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *NeurIPS*, 2020.

[28] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *NeurIPS*, 2019 (conf. year 2019; journalized later); see also [29].

[29] Y. Song, J. Sohl-Dickstein, D. P. Kingma, et al., "Score-based generative modeling through stochastic differential equations," *ICLR*, 2021.

[30] H. Risken, *The Fokker–Planck Equation*, 2nd ed. Springer, 1996.

[31] B. D. O. Anderson, "Reverse-time diffusion equation models," *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

[32] A. Engel and C. Van den Broeck, *Statistical Mechanics of Learning*. Cambridge University Press, 2001.

[33] D. Saad (ed.), *On-line Learning in Neural Networks*. Cambridge University Press, 1998; and related chapters in *Advances in Neural Information Processing Systems* (various years) for statistical mechanics of learning.

[34] A. Choromanska, M. Henaff, M. Mathieu, G. Arous, and Y. LeCun, "The loss surfaces

of multilayer networks," in *AISTATS*, 2015.

[35] S. Hochreiter and J. Schmidhuber, "Flat minima," *Neural Computation*, 9(1):1–42, 1997.

[36] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, "On large-batch training for deep learning: Generalization gap and sharp minima," in *ICLR*, 2017.

[37] P. Chaudhari, A. Choromanska, S. Soatto, et al., "Entropy-SGD: Biasing gradient descent into wide valleys," in *JSTSP*, 11(4): 592–604, 2017.

[38] C. Baldassi, C. Borgs, J. Chayes, A. Ingrosso, C. Lucibello, L. Saglietti, and R. Zecchina, "Unreasonable effectiveness of learning neural networks: From accessible states and robust ensembles to basic algorithmic schemes," *PNAS*, 113(48):E7655–E7662, 2016.

[39] Y. Bahri, J. Kadmon, J. Pennington, S. S. Schoenholz, J. Sohl-Dickstein, and S. Ganguli, "Statistical mechanics of deep learning," *Annual Review of Condensed Matter Physics*, 11:501–528, 2020.

[40] J. Pearl, *Causality*, 2nd ed. Cambridge University Press, 2009.

[41] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, 1:206–215, 2019.

[42] C. Molnar, *Interpretable Machine Learning*, 2nd ed. (online book), 2022.