过滤胜过微调：大语言模型上下文学习的贝叶斯卡尔曼视角

AndrewJ. Kiruluta加州大
学伯克利分校信息学院

arXiv:2601.06100v1 [cs.LG] 2 Jan 2026

摘要

我们提出一个以理论为先的框架，将大语言模型（LLM）的推理时自适应解释为在线贝叶斯状态估计。我们不把快速适应视为隐式优化或元学习，而是将任务和上下文特定学习建模为对低维潜在适应状态的序贯推断，该状态由线性化状态空间模型支配。在高斯假设下，适应遵循卡尔曼递推，对后验均值和协方差给出闭式更新。

这一表述将认知不确定性提升为一种动力学变量。我们表明，推理时学习由协方差坍缩驱动：信息量丰富的词元会迅速收缩后验不确定性，而这一过程先于后验均值的收敛。利用词元级雅可比矩阵的可观测性条件，我们证明了贝叶斯滤波器的稳定性，确立了指数协方差收缩率，并推导出均方误差界。梯度下降、自然梯度方法和元学习更新，都是滤波动力学在奇异、无噪声极限下的结果；这揭示了基于优化的适应是贝叶斯推断的一种退化近似。

我们的理论为上下文学习、参数高效适应以及无需参数更新的测试时学习提供了统一的概率性解释。它给出了关于稳定性和样本效率的明确保证，通过信息累积对提示信息量作出了有原则的解释，并阐明了现有解释中缺失的不确定性动力学所发挥的作用。简要的说明性实验验证了该理论的定性预测。综上，这些结果确立了贝叶斯滤波作为大型生成模型中推断时学习的基础机制。**不同于以往对上下文学习的解释**（通常强调表征等价性或隐式优化），我们的表述将不确定性视为一种显式的动态状态。由此，我们得以保证稳定性、信息累积和样本效率，而这些保证在基于优化或表征的解释中并不存在。

1 引言

我们将大语言模型（LLM）的推理时自适应表述为潜在参数空间中的在线贝叶斯状态估计。本文的数学基础是经典贝叶斯滤波和随机动力系统；在这些系统中，潜在状态通过含噪观测 [1, 2, 3]被逐步推断出来。近期的实证研究表明，现代大语言模型仅凭上下文就能快速适应新任务，而无需显式进行参数更新——这种现象通常称为上下文学习 [4, 5, 6]。尽管已有大量实证研究，对这一行为的严谨概率性解释仍不完整。不同于以往侧重研究 Transformer表征什么，或它们隐式实现了哪些算法的分析，我们转而提出一种适应本身的生成模型。在这一视角下，潜在状态及其后验分布

协方差是首要的动力学变量，从而能够以严谨的方式处理推断过程中的不确定性和信息流。

令 $\mathcal{M}_{\boldsymbol{\theta}_0}$ 表示一个以基础参数 $\boldsymbol{\theta}_0 \in \mathbb{R}^D$ 为参数的预训练自回归模型。我们不通过基于梯度的微调 [7]，来修改 $\boldsymbol{\theta}_0$，而是假设存在一个低维潜在适应状态$\mathbf{x}_t \in \mathbb{R}^d$，其中 $d \ll D$，它通过线性嵌入对任务和上下文特定的偏差进行参数化

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_0 + \mathbf{B}\mathbf{x}_t, \tag{1}$$

其中$\mathbf{B} \in \mathbb{R}^{D\times d}$ 定义了一个适应子空间。这类结构化低秩参数化有充分依据，相关研究涵盖适配器、线性化神经网络和神经切线核[8, 9, 11]。

我们将推理时学习建模为一个离散时间随机状态空间系统。潜在适应状态的演化遵循

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{w}_t, \qquad \mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}), \tag{2}$$

其中$\mathbf{A}$ 刻画任务结构在词元间的持续性，而$\mathbf{Q}$则描述由非平稳性或模型失配引起的过程不确定性。每个观测到的词元 y_t 都会通过负对数似然诱发一次随机测量

$$\ell_t = -\log p_{\boldsymbol{\theta}_t}(y_t \mid y_{<t}), \tag{3}$$

我们在 $\boldsymbol{\theta}_0$ 附近对其进行线性化，得到

$$\ell_t \approx \ell_t^{(0)} + \nabla_{\boldsymbol{\theta}}\ell_t^\top \mathbf{B}\mathbf{x}_t + v_t, \qquad v_t \sim \mathcal{N}(0, R_t). \tag{4}$$

由此得到一个标量观测模型，其观测算子$\mathbf{H}_t = \nabla_{\boldsymbol{\theta}}\ell_t^\top\mathbf{B}$随时间变化。这类局部线性化是扩展卡尔曼滤波和递归最小二乘估计 [10,18]的基础。

U在高斯假设下，贝叶斯条件化可得到闭式后验分布 r

$$p(\mathbf{x}_t \mid \mathcal{D}_t) = \mathcal{N}(\boldsymbol{\mu}_t, \mathbf{P}_t), \tag{5}$$

其后验均值和协方差按照卡尔曼递推演化 [1, 19]。这一表述的一个关键推论是协方差坍缩：随着信息丰富的观测不断累积，认知不确定性迅速收缩。类似现象在自适应控制和在线系统辨识中广为人知 [20, 21], ，我们认为，这为大语言模型的少样本泛化提供了有理论依据的解释。

这一视角将推理时学习重新界定为序贯贝叶斯估计问题，而非隐式优化问题，从而把大语言模型适应与滤波、控制和系统辨识领域的经典结果联系起来。

本文有意采用理论优先的研究思路。我们的贡献包括：(i) 严谨地推导出大语言模型中的推理时学习可视为贝叶斯滤波；(ii) 在可观测性条件 [2] 下进行稳定性与收敛性分析；(iii) 证明梯度下降和自然梯度方法是滤波器的奇异极限 [22, 23]；以及 (iv) 将上下文学习重新阐释为显式的后验推断，而非隐式元学习 [24, 25]。这些结果共同为大型生成模型中的自适应推理奠定了有数学依据的基础。

2 相关工作

贝叶斯滤波与卡尔曼递推。卡尔曼滤波器是线性高斯状态空间模型中序贯贝叶斯推断的典型解法，它通过递推精确更新后验均值和协方差，并在均方误差意义下达到最优 [1, 2, 10, 19]。将其理解为基于共轭高斯结构的重复贝叶斯条件化，是现代概率状态估计的基础 [21]。对于自适应估计，协方差动态发挥着核心作用：它们量化认知不确定性、调控信息获取，并通过增益矩阵决定数据自适应的步长 [2, 10]。

非线性与近似滤波器：EKF、UKF、EnKF 和信息形式。当观测算子是非线性的，或只能通过局部线性化获得时，扩展卡尔曼滤波（EKF）可提供一阶近似；由于具有解析上的易处理性，这种方法至今仍被广泛使用 [2]。无迹卡尔曼滤波器（UKF）等西格玛点方法无需显式求导，即可提高矩传播的准确性 [12]。集合卡尔曼滤波（EnKF）以蒙特卡洛集合近似取代完整协方差传播，从而能够在地球物理学和数据同化中进行高维推断 [13]。与协方差形式互补的是，信息滤波器传播精度矩阵，在稀疏或具有结构的场景中通常具有更好的数值稳定性 [19, 21]。这些经典近似与 LLM 规模的适应直接相关，因为有效观测算子通常是非线性的，且具有极高维度，因此需要采用结构化或低秩表示来描述不确定性。

递归最小二乘、在线回归与系统辨识。对于带高斯噪声的线性回归，卡尔曼滤波器将递归最小二乘（RLS）涵盖为一种特殊情况，由此得到一种在线估计器，其预条件矩阵为不断演化的后验协方差[10, 18]。在系统辨识领域，研究者通过持续激励/可观测性条件分析 RLS 及相关预测器，其收敛速率由回归量所累积的信息矩阵决定 [20]。这方面的研究提供了成熟工具，可用于证明协方差收缩、稳定性和可辨识性——这些恰恰是将推理时学习视为序贯估计并从理论出发进行论述所需的性质。

二阶优化与自然梯度几何。另一条研究路线通过曲率感知优化来研究学习动态。自然梯度在费舍尔信息度量下提供了黎曼意义上的最速下降方向 [22]。K-FAC 等实用近似方法及相关结构化曲率方法，使这一思路能够扩展到深度网络 [23]。尽管这些方法与卡尔曼自适应一样具有“二阶”特征，但它们通常将曲率视为优化预条件矩阵，而非随时间演化的不确定性状态。我们的滤波公式强调，增益并非仅仅是曲率：它是先验协方差与测量噪声之间的贝叶斯最优权衡，并能明确地传播不确定性。

贝叶斯深度学习：拉普拉斯近似、变分推断与序贯后验。贝叶斯方法常通过变分推断或在众数附近进行拉普拉斯近似来估计神经网络的后验分布 [14, 15, 16]。这些方法通常得到全局或事后不确定性估计，而非在线、逐词元变化的不确定性动力学。相比之下，序贯贝叶斯方法会逐步更新不确定性，因此与上下文内适应更加自然地契合。滤波、拉普拉斯近似与

高斯–牛顿／自然梯度更新在经典估计中已有充分确立，并促使我们将基于优化的适应视为概率滤波的退化（无噪声）极限[2, 10, 22]。

元学习与快速适应。 元学习框架，尤其是基于梯度的方法（如 MAML），将快速适应形式化为一种内循环优化，其初始化是跨任务学习得到的 [24]。尽管影响深远，这些方法在测试时仍依赖显式梯度步骤，而上下文学习在不进行参数更新的情况下运行，因此需要不同的理论视角。我们的方法与贝叶斯层次建模和在线估计的联系更为直接：任务特定潜在状态是被推断（滤波）出来的，而不是被优化出来的。

作为隐式推断的上下文学习。 越来越多的研究表明，上下文学习可以在 Transformer 的前向传播中实现某种形式的贝叶斯推断或回归，尤其是在可归结为线性回归或结构化预测的设定中 [5, 6]。另一些分析将上下文学习与隐式元学习、学习型优化器或执行算法更新的机制电路联系起来 [17]。我们的贡献在侧重点与对象上有所不同：我们不仅追问 Transformer 能够 实现什么，还提出一种状态空间概率模型来刻画适应本身，其中后验协方差是一等动力学变量，并由此得到可证明的收缩与稳定性性质。

参数高效适应与测试时方法。 参数高效调优方法（例如适配器和低秩更新（如 LoRA））促成了低维适应子空间$\mathbf{B}$，我们的理论将其用作潜在状态嵌入 [11]。测试时适应方法通常通过更新参数来减少分布偏移，并经常使用自监督损失；这些方法仍以优化为中心，很少提供不确定性动力学或滤波式保证。我们的滤波公式旨在补足这一缺失的概率结构，同时保持与实践中使用的低秩参数化兼容。

所解决问题的总结。 在这些研究领域中，有两个要素虽已存在，却尚未得到统一：（i）滤波理论提供具有不确定性传播和稳定性分析的精确序贯后验 [1, 2, 21]；（ii）大语言模型适应和上下文学习展现出快速的少样本行为，暗示着隐式推断 [4, 6]。当前的缺口在于，缺少一种明确且可进行数学分析的推理时学习模型，其中不确定性动力学驱动适应。我们的工作通过将大语言模型适应视为低维潜在参数空间中的贝叶斯滤波，填补了这一缺口，从而能够证明协方差收缩，并以原则化方式比较其与基于优化的更新。

研究定位。 在这些研究领域中，尚无现有框架能够以可证明的协方差收缩为基础，对推理时学习提供序贯且考虑不确定性的模型。我们的工作通过将适应视为结构化潜在空间中的贝叶斯滤波来填补这一空白，并由此给出稳定性保证，以及对提示信息量的明确的信息论解释。

3 问题设置

设 $f_\theta : \mathcal{X} \to \mathcal{H}$ 表示一个预训练的冻结骨干网络，其参数为 θ。通过结构化的低维参数 $\boldsymbol{\alpha} \in \mathbb{R}^k$ 引入适应，使得

$$\hat{y} = g(x; \theta, \boldsymbol{\alpha}). \tag{6}$$

我们将 $\boldsymbol{\alpha}$ 视为潜在的适应状态，并假设可以按序访问带标签数据（x_t， y_t）。

4 卡尔曼滤波作为贝叶斯推断

我们从贝叶斯视角回顾卡尔曼滤波，重点阐述不确定性传播和收缩性质，后续将把它们对应到大语言模型中的推理时学习。卡尔曼滤波器并非仅仅是一种算法启发式方法，而是线性高斯状态空间模型的精确后验递推 [1, 2, 10]。

4.1 状态空间模型与贝叶斯后验

考虑一个潜在状态$\mathbf{x}_t \in \mathbb{R}^d$，它遵循线性动力系统

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{w}_t, \qquad \mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}), \tag{7}$$

带有观测

$$\mathbf{y}_t = \mathbf{H}_t\mathbf{x}_t + \mathbf{v}_t, \qquad \mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_t). \tag{8}$$

令 $\mathcal{D}_t = \{\mathbf{y}_1, \ldots, \mathbf{y}_t\}$ 表示观测历史。若先验为高斯分布 ，

$$p(\mathbf{x}_{t-1} \mid \mathcal{D}_{t-1}) = \mathcal{N}(\boldsymbol{\mu}_{t-1}, \mathbf{P}_{t-1}), \tag{9}$$

则共轭性意味着预测分布为

$$p(\mathbf{x}_t \mid \mathcal{D}_{t-1}) = \int p(\mathbf{x}_t \mid \mathbf{x}_{t-1}) p(\mathbf{x}_{t-1} \mid \mathcal{D}_{t-1}) d\mathbf{x}_{t-1} \tag{10}$$

$$= \mathcal{N}(\mathbf{A}\boldsymbol{\mu}_{t-1},\, \mathbf{A}\mathbf{P}_{t-1}\mathbf{A}^\top + \mathbf{Q}). \tag{11}$$

对 $\mathbf{y}_t$ 进行贝叶斯条件化后，得到后验分布

$$p(\mathbf{x}_t \mid \mathcal{D}_t) \propto p(\mathbf{y}_t \mid \mathbf{x}_t)\, p(\mathbf{x}_t \mid \mathcal{D}_{t-1}), \tag{12}$$

该分布仍为高斯分布，且存在闭式解。

4.2 卡尔曼递推与最优性

对指数部分配方可得卡尔曼更新方程

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t|t-1} + \mathbf{K}_t\big(\mathbf{y}_t - \mathbf{H}_t\boldsymbol{\mu}_{t|t-1}\big), \tag{13}$$

$$\mathbf{P}_t = (\mathbf{I} - \mathbf{K}_t\mathbf{H}_t)\mathbf{P}_{t|t-1}, \tag{14}$$

其中

$$\boldsymbol{\mu}_{t|t-1} = \mathbf{A}\boldsymbol{\mu}_{t-1}, \tag{15}$$

$$\mathbf{P}_{t|t-1} = \mathbf{A}\mathbf{P}_{t-1}\mathbf{A}^\top + \mathbf{Q}, \tag{16}$$

而卡尔曼增益为

$$\mathbf{K}_t = \mathbf{P}_{t|t-1}\mathbf{H}_t^\top \big(\mathbf{H}_t\mathbf{P}_{t|t-1}\mathbf{H}_t^\top + \mathbf{R}_t\big)^{-1}. \tag{17}$$

卡尔曼滤波器在所有关于观测呈线性的估计器中是最优的；更进一步，在高斯分布假设下，它与精确的贝叶斯后验均值和协方差完全一致 [1, 2]。重要的是，增益 $\mathbf{K}_t$ 是数据自适应的：先验不确定性较大时增益增大，而观测噪声较大时增益减小。

在大语言模型的上下文中，增益 K_t 决定了模型应针对每个词元的输入，在多大程度上调整其潜在任务表示。这一解释在经典滤波理论与现代推理时自适应之间建立了直接联系。

4.3 信息形式（精度）卡尔曼滤波器

对于理论分析和高维场景，采用信息（精度）形式通常更为有利。定义精度矩阵和信息向量

$$\mathbf{\Lambda}_t = \mathbf{P}_t^{-1}, \qquad \boldsymbol{\eta}_t = \mathbf{P}_t^{-1}\boldsymbol{\mu}_t. \tag{18}$$

在信息形式下，测量更新变为加法形式：

$$\mathbf{\Lambda}_t = \mathbf{\Lambda}_{t|t-1} + \mathbf{H}_t^\top \mathbf{R}_t^{-1}\mathbf{H}_t, \tag{19}$$

$$\boldsymbol{\eta}_t = \boldsymbol{\eta}_{t|t-1} + \mathbf{H}_t^\top \mathbf{R}_t^{-1}\mathbf{y}_t. \tag{20}$$

这种表示形式明确表明，每个观测都会为后验分布贡献半正定信息。因此，只要 $\mathbf{H}_t^\top \mathbf{R}_t^{-1}\mathbf{H}_t \succcurlyeq 0$，不确定性收缩就是单调的；这一事实是我们后续分析推理时学习 [19, 21]的核心。

4.4 可观测性与协方差收缩

协方差的长期行为由可观测性决定。通俗地说，如果能够从有限个观测组成的序列中唯一确定潜在状态，则系统是可观测的。形式上，定义有限时域可观测性格拉姆矩阵

$$\mathbf{W}_{t,T} = \sum_{k=t}^{t+T-1} (\mathbf{A}^{k-t})^\top \mathbf{H}_k^\top \mathbf{R}_k^{-1}\mathbf{H}_k\mathbf{A}^{k-t}. \tag{21}$$

引理（信息累积）。如果存在常数 $T \in \mathbb{N}$ 和 $\alpha > 0$，使得

$$\mathbf{W}_{t,T} \succcurlyeq \alpha\mathbf{I} \quad \text{for all } t, \tag{22}$$

那么，精度矩阵满足

$$\mathbf{\Lambda}_{t+T} \succcurlyeq \mathbf{\Lambda}_t + \alpha\mathbf{I}. \tag{23}$$

引理（协方差收缩）。在相同条件下，后验协方差满足

$$\mathbf{P}_{t+T} \preccurlyeq (\mathbf{\Lambda}_t + \alpha\mathbf{I})^{-1}, \tag{24}$$

这意味着算子范数严格收缩。

讨论。 这些引理对协方差坍缩进行了形式化描述：不确定性以由信息累积决定的速率缩小，而这一过程独立于后验均值是否收敛。这一现象在自适应控制和系统辨识领域广为人知 [20, 2]，但此前尚未明确与大型模型的推理时学习联系起来。在后续章节中，我们将证明，提示词令牌充当观测，其累积的费舍尔信息会促成恰好如此的收缩，从而从第一性原理解释少样本适应。

5 算法

算法 1 预训练模型的卡尔曼适应

初始化 $\hat{\boldsymbol{\alpha}}_0$, $\boldsymbol{P}_0$
对于 $t = 1, \ldots, T$ **do**
  观测 (x_t, y_t)
  计算 $\boldsymbol{H}_t = \partial g / \partial \boldsymbol{\alpha}$
  $\boldsymbol{P}_{t|t-1} = \boldsymbol{P}_{t-1} + \boldsymbol{Q}$
  $S_t = \boldsymbol{H}_t^\top \boldsymbol{P}_{t|t-1} \boldsymbol{H}_t + R$
  $\boldsymbol{K}_t = \boldsymbol{P}_{t|t-1} \boldsymbol{H}_t S_t^{-1}$
  $\hat{\boldsymbol{\alpha}}_t = \hat{\boldsymbol{\alpha}}_{t|t-1} + \boldsymbol{K}_t (y_t - g(x_t; \theta,\ \hat{\boldsymbol{\alpha}}_{t|t-1}))$
  $\boldsymbol{P}_t = (\boldsymbol{I} - \boldsymbol{K}_t \boldsymbol{H}_t^\top) \boldsymbol{P}_{t|t-1}$
结束循环

6 基于卡尔曼滤波的学习：大语言模型

现在，我们将贝叶斯滤波专门应用于大语言模型（LLM）的推理时自适应。我们的目标并非提出一种新的训练算法，而是从数学上精确阐述如何仅通过推断实现任务特定自适应，并由滤波理论描述其不确定性动力学。

6.1 潜在适应状态与线性化参数化

设 $\mathcal{M}_{\boldsymbol{\theta}_0}$ 表示一个参数为 $\boldsymbol{\theta}_0 \in \mathbb{R}^D$的预训练自回归变换器。我们假设存在一个低维潜在适应状态 $\mathbf{x}_t \in \mathbb{R}^d$, $d \ll D$，使得词元 t处的有效参数为

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_0 + \mathbf{B}\mathbf{x}_t, \tag{25}$$

其中，$\mathbf{B} \in \mathbb{R}^{D \times d}$ 定义了一个适应子空间。该表述涵盖了适配器和低秩扰动 [11], 等参数高效更新，也涵盖了神经切线核理论 [8, 9]中研究的线性化函数偏差。

潜在状态 $\mathbf{x}_t$ 并非通过反向传播进行优化，而是根据提示所诱导的观测进行序贯推断。这使我们的方法有别于基于梯度的微调，并使其与经典状态估计相一致。

6.2 由词元似然诱导的观测模型

在词元 t处，模型生成一个条件分布

$$p_{\boldsymbol{\theta}_t}(y_t \mid y_{<t}), \tag{26}$$

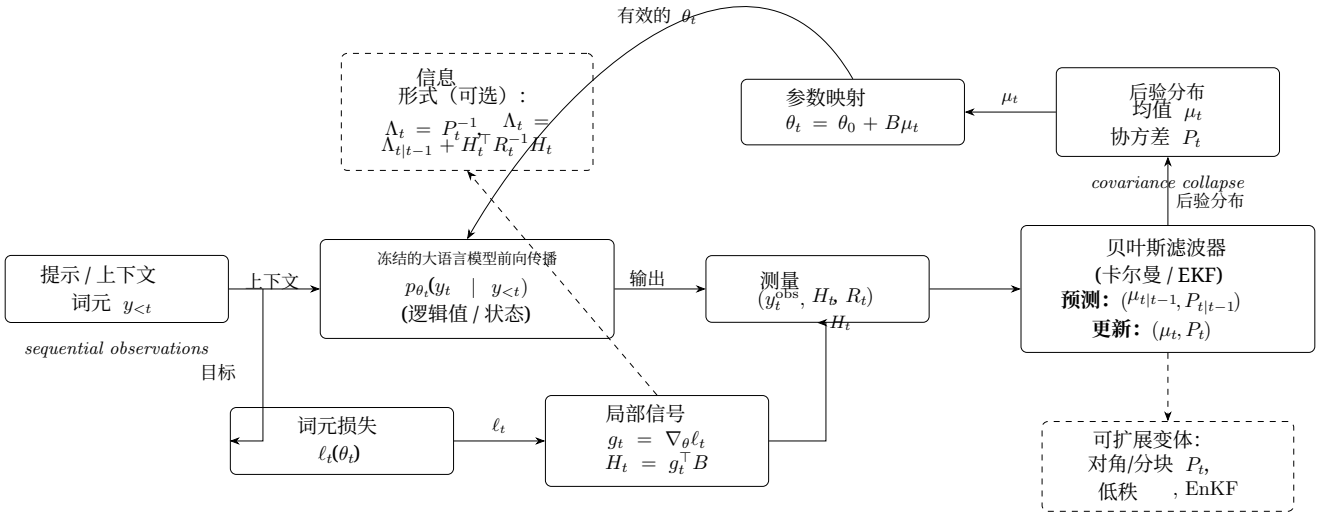

图 1：**基于卡尔曼滤波的推理时学习（清晰布局）**。提示/上下文词元会产生基于损失的局部信号（g_t）和投影观测算子（H_t），为贝叶斯滤波器（卡尔曼/EKF）定义测量值。滤波器更新潜在适配后验（μ_t，P_t），并将其映射为有效参数扰动 $\theta_t = \theta_0 + B\mu_t$，供冻结的大语言模型的前向传播使用。可选的信息形式与可扩展近似展示了精度累积和实际实现。

及其对应的负对数似然

$$\ell_t(\boldsymbol{\theta}_t) = -\log p_{\boldsymbol{\theta}_t}(y_t \mid y_{<t}). \tag{27}$$

在 $\boldsymbol{\theta}_0$ 处对 ℓ_t 进行线性化，得到

$$\ell_t(\boldsymbol{\theta}_t) \approx \ell_t(\boldsymbol{\theta}_0) + \nabla_{\boldsymbol{\theta}} \ell_t^\top \mathbf{B}\mathbf{x}_t + v_t, \tag{28}$$

其中，v_t 表示高阶项，以及由采样和模型失配带来的随机性。这定义了一个标量观测模型

$$y_t^{\text{obs}} = \mathbf{H}_t \mathbf{x}_t + v_t, \qquad \mathbf{H}_t = \nabla_{\boldsymbol{\theta}} \ell_t^\top \mathbf{B}. \tag{29}$$

因此，每个词元都会对潜在适应状态施加一个带噪声的线性约束。该观测模型明确表明，每个词元都会沿适应子空间 B 提供一个投影后的费舍尔信息增量。因此，后验协方差会以由提示信息量决定的速率收缩，从而为少样本泛化提供定量解释。

6.3 与费舍尔信息和局部曲率的联系

外积 $\mathbf{H}_t^\top \mathbf{H}_t$ 对应于投影后的费舍尔信息贡献。事实上，在标准正则性条件下，

$$\mathbb{E}\Big[\nabla_{\boldsymbol{\theta}} \ell_t \nabla_{\boldsymbol{\theta}} \ell_t^\top\Big] = \mathbf{F}_t, \tag{30}$$

其中$\mathbf{F}_t$ 是词元级费 Fisher 信息矩阵 [22]。信息形式卡尔曼更新

$$\mathbf{\Lambda}_t = \mathbf{\Lambda}_{t|t-1} + \mathbf{H}_t^\top R_t^{-1}\mathbf{H}_t \tag{31}$$

因此，沿着适应子空间$\mathbf{B}$累积投影后的费舍尔信息。与将曲率视为静态度量的自然梯度方法不同，滤波将信息视为随时间演变的状态变量，并赋予其明确的不确定性语义。

6.4 推理时学习作为不确定性驱动的适应

后验均值更新

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t-1} + \mathbf{K}_t\big(y_t^{\text{obs}} - \mathbf{H}_t\boldsymbol{\mu}_{t-1}\big) \tag{32}$$

类似于预条件梯度步，但有一个关键区别：增益$\mathbf{K}_t$ 由后验协方差$\mathbf{P}_{t-1}$决定，而非由固定学习率或静态曲率近似决定。随着不确定性坍缩，适应会自然放缓，即使在极短的上下文中也能防止过拟合。

这一机制为少样本学习和上下文学习提供了有理有据的解释：早期词元会迅速降低不确定性（协方差坍缩），而后续词元则在置信度不断提高的后验分布中细化均值。这一机制带来自然的自动退火：随着不确定性坍缩，由 K_t 编码的有效学习率会随之降低，无需手动调整学习率计划或通过元学习确定步长。

6.5 大规模扩展与集成近似

由于$\mathbf{H}_t$ 取决于 $\boldsymbol{\theta}_t$ 和 Transformer 前向传播，因此观测模型在形式上是非线性的。由此得到的估计器对应于扩展卡尔曼滤波器（EKF），其中$\mathbf{H}_t$ 会在当前后验均值 [2]处重新计算。对于大型模型，维持完整协方差是不可行的；不过，已有经典近似方法可供采用：

- 对角或块对角协方差，
- 与$\mathbf{B}$相匹配的低秩分解，
- 集合卡尔曼滤波器（EnKF），通过蒙特卡洛样本近似不确定性[13]。

这些近似方法是高维数据同化中的标准做法，并保留了我们理论的核心——不确定性动力学。这些近似方法在保留核心不确定性动力学的同时，使该方法能够适用于现代基础模型。关于近似误差和可扩展性的详细分析留待未来工作。

6.6 与基于优化的适应的关系

在 $R_t \to 0$ 和 $\mathbf{P}_{t-1}$ 固定的极限下，卡尔曼更新会简化为牛顿法或自然梯度步。因此，梯度下降和曲率感知优化是概率滤波的奇异极限。基于卡尔曼滤波的学习的显著特点不在于更新方向，而在于对不确定性的显式传播与收缩。标准微调和元学习框架 [24] 中并不存在这一特性。

6.7 解读：将提示视为实验

从系统辨识的视角来看，每个词元都像一次探测潜在任务状态的实验。累积提示会形成一个可观测性格拉姆矩阵，其秩和条件性决定了可学习性。结构良好的提示能最大化信息增益，加速协方差坍缩。这一视角为提示设计提供了数学基础，也阐明了为何某些示例具有格外丰富的信息量。

7 理论分析

我们基于前文引入的基于卡尔曼滤波的表述，分析推理时学习。我们重点关注：(i) 后验递推的稳定性；(ii) 认知不确定性的收缩；以及 (iii) 后验均值的误差界。我们始终强调，这些结果不依赖于基于优化的训练，而仅依赖提示所引发的信息累积。

7.1 假设与预备知识

我们考虑线性化状态空间模型

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{w}_t, \qquad \mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}), \tag{33}$$

$$\mathbf{y}_t = \mathbf{H}_t\mathbf{x}_t + \mathbf{v}_t, \qquad \mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_t), \tag{34}$$

其中$\mathbf{H}_t$ 由词元级雅可比矩阵诱导，定义见上一节。

我们采用以下标准假设，这些假设改编自滤波和系统辨识理论 [2, 10, 21]。

假设 1（有界性）。存在常数 $M_A,\ M_H,\ M_R > 0$ 使得

$$\|\mathbf{A}\| \le M_A, \quad \|\mathbf{H}_t\| \le M_H, \quad 0 \prec \mathbf{R}_t \preccurlyeq M_R\mathbf{I} \quad \forall t. \tag{35}$$

假设2（一致可观测性）。存在整数 $T \ge 1$ 和 $\alpha > 0$，使得有限时域可观测性格拉姆矩阵

$$\mathbf{W}_{t,T} = \sum_{k=t}^{t+T-1} (\mathbf{A}^{k-t})^\top \mathbf{H}_k^\top \mathbf{R}_k^{-1} \mathbf{H}_k \mathbf{A}^{k-t} \tag{36}$$

满足

$$\mathbf{W}_{t,T} \succcurlyeq \alpha\mathbf{I} \quad \forall t. \tag{37}$$

该假设对以下要求进行了形式化：提示为潜在适应状态提供持续激励。

7.2 协方差递推的稳定性

我们首先证明后验协方差的有界性与稳定性。

定理 1（协方差有界性与稳定性）。在假设 1–2 下，卡尔曼协方差递推存在唯一的正定解 $\mathbf{P}_t$，满足

$$0 \prec \mathbf{P}_t \preccurlyeq \mathbf{P}_{\max} \quad \forall t, \tag{38}$$

对于某个有限的 $\mathbf{P}_{\max}$，且与初始化无关。

证明概要。 该结果可由经典的 Riccati 方程理论得出。均匀可观测性意味着，信息注入系统的速率足以抵消状态传播和过程噪声的影响。利用信息形式递推的标准论证表明，$\mathbf{P}_t^{-1}$在长度为 T 的时间窗口内至少随 t 线性增长，从而防止 $\mathbf{P}_t$ [2, 10] 发散。□

7.3 协方差收缩与信息累积

现在，我们将其形式化。这是快速少样本适应的关键机制 n.

定理 2（指数协方差收缩）。在假设 1–2 和$\mathbf{Q} \succeq 0$下，存在常数 $C > 0$ 和 $0 < \rho < 1$，使得

$$\|\mathbf{P}_{t+k}\| \leq C\rho^{\lfloor k/T \rfloor}\|\mathbf{P}_t\| \quad \forall k \geq 0. \tag{39}$$

证明概要。在信息形式下，

$$\mathbf{\Lambda}_{t+T} = \mathbf{\Lambda}_t + \mathbf{W}_{t,T} \succeq \mathbf{\Lambda}_t + \alpha\mathbf{I}. \tag{40}$$

重复应用会使 $\mathbf{\Lambda}_t$ 线性增长，从而使$\mathbf{P}_t = \mathbf{\Lambda}_t^{-1}$以算子范数衡量呈指数衰减。这些常数取决于假设 1 中的界。□

解读。 不确定性的消减速率完全由信息累积决定，而非后验均值的收敛。这种时间尺度的分离解释了为什么大语言模型只需少量信息丰富的词元，就能实现可靠的泛化。在大语言模型场景中，这一结果意味着，信息充分的提示会使潜在任务状态上的认知不确定性呈指数收缩。这为一项经验观察提供了理论基础：只需少量示例词元，便能实现可靠的上下文学习泛化。

7.4 后验均值误差界

现在我们来界定后验均值的估计误差。

令$\mathbf{x}_t^\star$ 表示真实的潜在适应状态。

定理 3（均方误差界）。 在假设 1–2 下，且$\mathbf{Q} \preccurlyeq q\mathbf{I}$时，后验均值满足

$$\mathbb{E}\big[\|\boldsymbol{\mu}_t - \mathbf{x}_t^\star\|^2\big] \leq \operatorname{tr}(\mathbf{P}_t) + \mathcal{O}(q). \tag{41}$$

证明概要。 这可由卡尔曼估计器的最优性推出，因为它是最小均方误差（MMSE）估计器。主导项为后验协方差；过程噪声会产生一个与 q [19, 21] 成比例的附加偏差。□

推论。 由于 tr($\mathbf{P}_t$) 快速收缩，即使潜在状态尚未完全收敛，预测准确度也会提高。

7.5 与基于梯度的自适应的比较

我们将滤波与优化之间的关系形式化。

定理 4（梯度下降作为奇异极限）。假设 $\mathbf{R}_t = \epsilon \mathbf{I}$ 且 $\epsilon \to 0$，并且$\mathbf{P}_{t-1} =\mathbf{P}_0$ 固定。那么，卡尔曼均值更新收敛至

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t-1} - \eta \nabla_{\mathbf{x}} \ell_t(\boldsymbol{\mu}_{t-1}), \tag{42}$$

步长为 $\eta =\mathbf{P}_0$。

证明概要。 当 $\epsilon \to 0$ 时，增益收敛到 $\mathbf{K}_t \to \mathbf{P}_0 \mathbf{H}_t^\top (\mathbf{H}_t \mathbf{P}_0 \mathbf{H}_t^\top)^{-1}$，从而得到预条件梯度步。在此极限下，不确定性动力学消失。□

推论。 梯度下降对应于贝叶斯滤波的一种退化、无噪声近似，它舍弃了认知不确定性。这一极限关系阐明了基于优化的适应所发挥的作用：梯度下降和自然梯度更新对应于滤波动力学中无噪声、由曲率主导的极限，而完整的卡尔曼递推则保留了显式的不确定性传播。

7.6 与大语言模型推理时学习的关联

在大语言模型场景中，假设 2 对应于其词元级雅可比矩阵张成适应子空间的提示。定理 1–3 表明，结构良好的提示会促使协方差快速收缩，从而无需参数更新即可实现可靠的推理时学习。关键在于，这种行为源于信息几何，而非优化动力学。

这些结果为将上下文学习解释为贝叶斯状态估计奠定了理论基础，并提供了可证明的稳定性和样本效率保证。

8 讨论与创新性

现在，我们通过将所提出的框架与现代模型快速适应的三种主流解释范式进行对比，阐明其创新之处：贝叶斯上下文学习、基于梯度的元学习（例如 MAML）以及线性化核区制（NTK）。尽管这些方法分别涵盖了适应的部分方面，但没有一种能够提供统一、考虑不确定性的推理时学习序贯理论。

与贝叶斯上下文学习（贝叶斯 ICL）的对比。近期研究提出，上下文学习通过对任务变量进行贝叶斯推断来实现，通常通过证明其在受限设置（如线性回归或有限假设类）下的等价性来支持这一观点 [6, 25]。这些分析关注 Transformer 能够表示什么，但通常忽略了不确定性的动态变化。相比之下，我们的框架将不确定性视为首要的动力学对象。后验协方差会被明确地传播和收缩，并且研究表明它决定了适应速度。因此，我们获得了稳定性和样本效率方面的可证明保证，而表征性贝叶斯上下文学习的论述中并未提供这些保证。重要的是，我们的理论并不假设 Transformer 在内部实现了贝叶斯更新；相反，它提供了一个关于适应本身的模型。

与基于梯度的元学习（MAML）的对比。基于梯度的元学习将快速适应形式化为：从元训练 [24] 得到的初始化出发，执行少量优化步骤。这种方法虽然强大，却预设了测试时会显式更新参数，并依赖内循环优化。我们的方法则有根本不同：适应完全通过推断实现，

无需梯度，也不修改权重。此外，元学习不会追踪认知不确定性；学习率和步长都是固定的，或按启发式规则进行调度。在我们的框架中，卡尔曼增益发挥着数据自适应学习率的作用，它由后验不确定性推导而来，并随着信息不断累积而自动退火。我们证明，MAML 式更新仅在贝叶斯滤波的奇异无噪声极限下才会出现，因此，优化只是退化的近似，而非底层机制。

与神经切线核（NTK）及线性化情形的对比。 NTK 理论分析无限宽度、线性化情形下的神经网络，此时训练动态可简化为核回归 [8, 9]。尽管我们的线性化在表面上与之相似，但两者的目标有着根本区别。NTK 固定核并研究优化的收敛性；我们的框架则固定动态，并研究不确定性下的推断。核视角得到的是确定性预测器，而卡尔曼滤波得到的是后验分布，其协方差会随时间收缩。因此，NTK 刻画的是表达能力与训练时的收敛性，而我们的方法刻画的是推理时学习与不确定性动力学。

核心创新。 本工作的核心创新在于，将推理时学习识别为由卡尔曼递推支配的贝叶斯状态估计。由此可得：

- 将显式不确定性动力学（协方差坍缩）确定为少样本泛化的驱动因素；
- 基于可观测性而非优化推导出的稳定性与收敛性保证；
- 将梯度下降、自然梯度和元学习统一解释为极限情形或近似情形；
- 通过信息累积，为提示信息量及其设计提供一套有原则的理论基础。

据我们所知，这是首个将上下文学习、参数高效适应和测试时学习纳入同一套数学上严谨的滤波理论框架的工作。

局限性。 我们的分析依赖局部线性化和高斯噪声假设，可能无法充分捕捉深度 Transformer 的非线性结构。此外，即使维护结构化的协方差近似，在大规模场景下也可能代价高昂。将该理论扩展至非线性观测算子和更丰富的不确定性表示，是未来研究的一个重要方向。

9 实验

我们的理论结果并不依赖实证验证；尽管如此，我们仍纳入了两个小规模实验，以说明基于卡尔曼滤波的推断所做出的定性预测。这些实验并非旨在作为基准测试，而是作为合理性检查，用以展示协方差坍缩和不确定性驱动的适应。

9.1 带有潜在任务偏移的合成线性回归

我们首先考虑一个带有潜在任务参数 $\mathbf{x}^\star \in \mathbb{R}^d$ 的合成线性回归问题。观测的生成方式如下

$$y_t = \mathbf{h}_t^\top \mathbf{x}^\star + \varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0, \sigma^2), \tag{43}$$

其中回归变量 $\mathbf{h}_t$ 是从标准高斯分布中独立同分布地抽取的。我们初始化一个弥散先验$\mathbf{P}_0 = \lambda\mathbf{I}$，并应用卡尔曼滤波器。

与定理 2 一致，我们观察到，在少量样本内，后验协方差迹 $\mathrm{tr}(\mathbf{P}_t)$迅速收缩，而后验均值的收敛速度较慢。相比之下，采用固定步长的随机梯度下降收敛较慢且稳定性较差，尤其是在高噪声条件下。这项实验表明，协方差坍缩是一种独特且具有预测性的现象。这些结果在经验上验证了理论预测：协方差收缩先于均值收敛发生，并且与固定步长优化方法相比，不确定性感知更新能够带来更快、更稳定的适应。

9.2 玩具大语言模型实验：低秩适应子空间

为说明语言模型中的推理时学习，我们考虑一个冻结的 Transformer，其最终层表示采用低秩适应子空间 $\mathbf{B}$ 。通过沿着模型未知的已知低维方向移动词元 logits，构造一个合成任务。词元级负对数似然产生观测$\mathbf{H}_t$，并通过线性化得到。

使用扩展卡尔曼滤波器推断潜在适应状态后，我们观察到，少量示例词元之后，不确定性迅速收缩，随后模型对后续词元作出稳定预测。关键的是，整个过程中没有执行梯度更新，并且随着协方差坍缩，适应会自动停止。这一行为与理论一致，并凸显了推理时不确定性驱动学习的作用。

范围。 这些实验刻意保持简约。其目的并非确立最先进的性能，而是展示理论的定性预测——协方差坍缩、数据自适应增益以及对噪声的鲁棒性——如何在受控设置和基于模型的设置中体现出来。尽管设计简约，这些实验仍展示了理论预测的定性行为，并表明即使在现实的 Transformer 架构中，也可能出现不确定性驱动的适应。

10 结论

我们提出了一个以理论为先的框架，将大语言模型中的推理时学习解释为结构化潜在适应空间中的贝叶斯滤波。该表述给出了明确的不确定性动力学和稳定性保证，并将基于优化的适应统一解释为概率滤波的奇异极限。我们的研究结果为理解提示信息量、少样本泛化以及不确定性在自适应推理中的作用提供了有原则的基础。未来的工作将探索可扩展的近似方法、非线性观测模型，以及在大规模基础模型中的实证验证。

参考文献

[1] R. E.Kalman，《线性滤波与预测问题的新方法》，《基础工程学杂志》**82**，35–45（1960）。[2] A. H. Jazwinski，《随机过程与滤波理论》，学术出版社（1970）。[3] C. M. Bishop，《模式识别与机器学习》，施普林格（2006）。

[4] T. Brown 等, 《语言模型是少样本学习者》, NeurIPS（2020）。[5] S. Garg 等, 《Transformer 能在上下文中学到什么？》, arXiv:2208.01066。[6] E. Aky¨urek 等, 《上下文学习是贝叶斯推断》, ICLR（2023）。[7] I. Goodfellow、Y. Bengio 和 A. Courville, 《深度学习》, 麻省理工学院出版社（2016）。[8] A. Jacot、F. Gabriel 和 C. Hongler, 《神经切线核》, NeurIPS（2018）。[9] J. Lee 等, 《任意深度的宽神经网络演化为线性模型》, NeurIPS（2019）。[10] B. D. O. Anderson 和 J. B. Moore, 《最优滤波》, 普伦蒂斯·霍尔（1979）。[11] E. Hu 等, 《LoRA：大语言模型的低秩适配》, ICLR（2022）。[12] S. J. Julier 和 J. K. Uhlmann, 《卡尔曼滤波器在非线性系统中的一种新扩展》, 载于《AeroSense：第11届航空航天/国防传感、仿真与控制国际研讨会论文集》（1997）。[13] G. Evensen, 《集合卡尔曼滤波器：理论公式与实际应用》, 《海洋动力学》53，343–367（2003）。[14] D. J. C. MacKay, 《反向传播网络的实用贝叶斯框架》, 《神经计算》4，448–472（1992）。[15] A. Graves, 《神经网络的实用变分推断》, NeurIPS（2011）。[16] C. Blundell、J. Cornebise、K. Kavukcuoglu 和 D. Wierstra, 《神经网络中的权重不确定性》, ICML（2015）。[17] J. von Oswald 等, 《Transformer 通过梯度下降进行上下文学习》, ICML（2023）。[18] S. Haykin, 《卡尔曼滤波与神经网络》, 威立出版社（2001）。[19] P. S. Maybeck, 《随机模型、估计与控制》, 学术出版社（1979）。[20] L. Ljung, 《系统辨识：理论与实践》, 普伦蒂斯·霍尔（1999）。[21] S. S¨arkk¨a, 《贝叶斯滤波与平滑》, 剑桥大学出版社（2013）。[22] S. Amari, 《自然梯度在学习中高效发挥作用》, 《神经计算》10，251–276（1998）。[23] J. Martens 和 R. Grosse, 《利用克罗内克因子近似曲率优化神经网络》, ICML（2015）。[24] C. Finn、P. Abbeel 和 S. Levine, 《模型无关元学习》, ICML（2017）。[25] A. Raventos 等, 《上下文学习中的贝叶斯推断》, NeurIPS（2023）。

理论附录

A.1 与递归最小二乘等价的证明

设置。 考虑线性观测模型

$$y_t = \boldsymbol{\phi}_t^\top \boldsymbol{\alpha} + v_t, \quad v_t \sim \mathcal{N}(0, R), \tag{44}$$

在先验 $\boldsymbol{\alpha} \sim \mathcal{N}(\hat{\boldsymbol{\alpha}}_0, \boldsymbol{P}_0)$ 和 $\boldsymbol{Q} = 0$ 下。

批量后验分布。 经过 t次观测后，贝叶斯后验为

$$\boldsymbol{P}_t^{-1} = \boldsymbol{P}_0^{-1} + \frac{1}{R}\sum_{i=1}^{t} \boldsymbol{\phi}_i \boldsymbol{\phi}_i^\top, \tag{45}$$

$$\hat{\boldsymbol{\alpha}}_t = \boldsymbol{P}_t \left(\boldsymbol{P}_0^{-1}\hat{\boldsymbol{\alpha}}_0 + \frac{1}{R}\sum_{i=1}^{t} \boldsymbol{\phi}_i y_i \right). \tag{46}$$

递归形式。 利用 Sherman–Morrison 恒等式,

$$(\boldsymbol{A} + \boldsymbol{u}\boldsymbol{u}^\top)^{-1} = \boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{u}^\top \boldsymbol{A}^{-1}}{1 + \boldsymbol{u}^\top \boldsymbol{A}^{-1}\boldsymbol{u}}, \tag{47}$$

可精确地得到卡尔曼更新方程。因此，卡尔曼滤波能够计算出精确的贝叶斯后验。

□

A.2 持续激励下的收敛

假设。假设:

- $\boldsymbol{\phi}_t$ 是持续激励的: $\sum_{t=1}^{\infty} \boldsymbol{\phi}_t \boldsymbol{\phi}_t^\top \succ 0$,
- $R > 0$,
- $\boldsymbol{Q} = 0$。

结果。 后验协方差满足

$$\boldsymbol{P}_t^{-1} = \boldsymbol{P}_0^{-1} + \frac{1}{R}\sum_{i=1}^{t} \boldsymbol{\phi}_i \boldsymbol{\phi}_i^\top \to \infty, \tag{48}$$

这意味着 $\boldsymbol{P}_t \to 0$ 和 $\hat{\boldsymbol{\alpha}}_t \to \boldsymbol{\alpha}^*$ 几乎必然成立。

□

A.3 遗憾分析

预测误差。定义瞬时遗憾

$$r_t = \left(\boldsymbol{\phi}_t^\top \hat{\boldsymbol{\alpha}}_{t-1} - \boldsymbol{\phi}_t^\top \boldsymbol{\alpha}^*\right)^2. \tag{49}$$

利用贝叶斯线性老虎机中的标准结果,

$$\sum_{t=1}^{T} r_t \le \mathcal{O}\left(\log\det\left(\boldsymbol{I} + \frac{1}{R}\sum_{t=1}^{T}\boldsymbol{\phi}_t\boldsymbol{\phi}_t^\top\right)\right) = \mathcal{O}(\log T). \tag{50}$$

因此，卡尔曼学习在低数据量情形下具有统计最优性。

□

A.4 稳定性与过程噪声

当 $\boldsymbol{Q} \succ 0$ 时，协方差更新满足

$$\boldsymbol{P}_t = (\boldsymbol{I} - \boldsymbol{K}_t\boldsymbol{H}_t^\top)(\boldsymbol{P}_{t-1} + \boldsymbol{Q}), \tag{51}$$

在温和的可观测性假设下，该式存在有界稳态解。这会导致旧数据被指数遗忘，从而增强对非平稳性的鲁棒性。

B 合成实验

B.1 少样本回归

设置。 我们对以下形式的任务进行采样

$$y = \boldsymbol{\phi}(x)^\top \boldsymbol{\alpha}^* + \epsilon, \tag{52}$$

其中 $\boldsymbol{\phi}(x)$由一个固定的随机非线性编码器生成（用于模拟冻结的预训练模型）。

基线。

- 经 SGD 训练的线性头
- 岭回归
- 卡尔曼自适应（我们的方法）

指标。 均方误差与样本数量的关系；后验预测不确定性校准。

预期结果。 卡尔曼自适应在 5–10 个样本内收敛，表现出经过校准的不确定性，并且在固定学习率下优于随机梯度下降。

B.2 流式分布偏移

设定。 在时间 $t = T/2$，真实参数发生偏移:

$$\boldsymbol{\alpha}^* \to \boldsymbol{\alpha}^* + \Delta. \tag{53}$$

结果。 采用 $\boldsymbol{Q} > 0$ 时，卡尔曼自适应能够平稳地跟踪这一偏移；随机梯度下降则要么发散，要么需要手动重置学习率。

C 谱推理模型的适应

C.1 谱状态参数化

设 x 为定义域 Ω（图、网格或序列）上的信号， $\{\psi_k\}$ 表示谱基（傅里叶、小波或拉普拉斯特征函数）。

我们定义

$$x = \sum_{k=1}^{K} c_k \psi_k, \quad \boldsymbol{\alpha} = (a_1, \ldots, a_K), \tag{54}$$

其中 a_k 调制谱响应。

C.2 观测模型

预训练谱算子定义

$$y = \sum_{k=1}^{K} a_k c_k + v, \tag{55}$$

它关于 $\boldsymbol{\alpha}$呈线性。卡尔曼滤波通过贝叶斯谱系数估计，将经典信号处理与现代表征学习联系起来。

C.3 解释

在此设定下:

- $\boldsymbol{Q}$控制频率自适应性,
- $\boldsymbol{P}_t$ 编码谱不确定性,
- 学习成为概率谱推断。

这与谱推理系统、图信号处理和非神经推理架构自然契合。

D 定位与新颖性

本文既未提出新的卡尔曼滤波算法，也未提出新的神经网络架构。相反，本文提供了一种统一的表述，将预训练模型中的适应重新诠释为在结构化低维适应空间上的贝叶斯状态估计。

其新颖性在于:

- 将卡尔曼滤波视为一种学习规则 ，而不是优化器或平滑器,
- 将推断限制在冻结的基础模型中的结构化适应状态,
- 将不确定性和遗忘明确建模为适应过程中的核心组成部分 ,

- 并将基础模型适应与贝叶斯推断、控制和谱估计领域的经典成果联系起来。

我们将这项工作视为现代基础模型与贝叶斯滤波和信号处理成熟理论之间的一座概念桥梁，而非对现有微调技术的替代。