语义率失真与后验设计：计算约束、多模态与策略性推理

Emrah Akyol，*IEEE* 高级会员

arXiv:2602.03949v1 [cs.IT] 3 Feb 2026

摘要

我们研究受率和计算约束的策略性高斯语义压缩，其中编码器与解码器分别优化不同的二次目标函数。潜在高斯状态生成依赖于任务的语义变量，而解码器通过最小均方误差估计进行最优响应，从而将编码器的问题归结为信息率约束下的后验协方差设计。我们刻画了直接、远程和全信息机制下的策略性率失真函数，推导出语义注水和受率约束的高斯劝说解，并证明了在目标不一致时的高斯最优性。我们进一步表明，架构计算限制会形成隐式速率约束；随着模型深度和推理时计算的增加，语义准确率呈指数级提升，而多模态观测则消除了远程编码固有的几何均值惩罚。这些结果为数据与能源高效人工智能奠定了信息论基础，并为将现代多模态语言模型视为资源约束下的后验设计机制提供了原则性阐释。

索引术语

语义通信、策略通信、率失真、贝叶斯劝说、高斯信号传输、多模态学习、缩放定律

本研究由美国国家科学基金会通过CCF/CIF（CAREER）项目#2048042资助。Emrah Akyol就职于宾厄姆顿大学电气与计算机工程系，美国纽约州宾厄姆顿13902（电子邮箱：eakyol@binghamton.edu）。

I. 引言

当前的人工智能时代由缩放定律所定义：增加数据量、模型规模和计算量，一直是提升性能并获得涌现能力的可靠途径。然而，这一轨迹面临硬性约束。数据可用性是有限的，训练与推理算力代价高昂，能耗已成为负责任地部署大规模人工智能系统的核心瓶颈。这些现实要求有原则的统计与信息论工具，以刻画能量、数据、计算与性能之间的基本权衡。本文通过发展资源约束下语义推断的率失真（RD）与信息设计理论，贡献了这样一套工具。

实现数据与能量高效学习的一个关键障碍在于，现代人工智能系统很少以重构原始观测为目标。相反，它们是依据分类准确率、决策质量或语义正确性等目标函数进行训练和评估的。在这些场景中，相关信息并非观测本身，而是一个依赖于任务的潜在量。此外，人工智能系统常常在目标不一致的多智能体环境中运行：用户、平台和自动化决策模块可能对同一信息的不同方面赋予不同价值。这些事实从两个方面对经典信息论提出了挑战。首先，标准率失真理论假设编码器与解码器共享同一失真度量，而语义和面向任务的系统自然会引出不同的目标函数。其次，经典贝叶斯劝说与信息设计刻画了发送方如何塑造接收方的后验信念，但通常忽略了显式的通信速率与计算约束。

本研究提出一种策略性高斯语义压缩模型，将这些视角统一起来。潜在高斯状态 X 支配任务语义，而编码器关注语义变量 $\Theta = BX + V$，其中 B 表示语义变换， V则体现语义噪声、偏好不确定性或任务变化。解码器旨在从率受限消息 M中推断出 X，而编码器则通过关于 Θ的二次语义损失来评估性能。双方的互动被建模为斯塔克尔伯格博弈：编码器在速率约束下承诺采用某种编码策略，解码器则通过根据自身目标形成贝叶斯ˆ最小均方误差估计 $X = \mathbb{E}[X \mid M]$ 作出最佳响应。该模型自然涵盖了三种具有实际意义的情形：(i) 直接编码，编码器可观测 X；(ii) 远程编码，编码器只能观测 Θ（或不完美的语义代理）；以及 (iii) 完全-

信息编码，编码器联合观测 (X, Θ)，并可利用语义噪声来塑造解码器的信念。第三种情形与高斯劝说和信息设计直接相关。

我们的主要技术贡献在于证明，这些战略语义压缩问题可以通过清晰的后验几何刻画来描述。由于解码器采用最小均方误差估计进行最优响应，编码器的语义失真主要取决于编码机制所产生的给定消息下 X 的后验协方差。对于高斯源，信息约束转化为后验不确定性的对数行列式界，揭示了一个根本性的熵预算：

$$\log\det K_X \geq \log\det \Sigma_X - 2R, \tag{1}$$

其中， K_X 是后验误差协方差， R 是可用率。

除了这些率失真和信息设计结果之外，该框架还为审视当代人工智能架构及其对能源效率的关注提供了有原则的视角。在基于Transformer的大语言模型（大语言模型）中，虽然没有显式的比特管道，但信息流受到有限注意力带宽、有限嵌入维度、受限上下文窗口以及有限推理计算的约束。这些架构瓶颈对内部信念形成和后验细化施加了隐式速率约束，与本文研究的显式速率约束相类似。由此产生的熵几何为扩展现象提供了分析性解释：增加深度、宽度或计算量会提高有效信息预算，并使后验信念更加精确；思维链推理则如同序贯率分配，在多个步骤中细化后验。同样，该理论量化了多模态编码器的优势：多种模态能够提高可恢复语义协方差，并消除远程编码固有的几何平均信息坍缩。这为多模态系统（如视觉语言模型）为何能在计算量没有相应增加的情况下实现更高的数据效率，提供了严谨的解释。

本文的贡献如下：(i) 提出一种战略高斯语义压缩框架，将目标不一致情形下的率失真理论与信息设计联系起来；(ii) 推导直接观测机制和远程观测机制下精确的战略率失真刻画，包括高斯最优性和语义注水解；(iii) 在全信息机制下构建并求解受率约束的高斯劝说问题，包括闭式对角解；(iv) 量化不同观测机制之间的性能差距，并

说明多模态如何弥合这些差距；(v) 将这些信息论极限与能量受限人工智能架构联系起来，把架构瓶颈解释为隐式速率约束。

本文结构如下。第二节回顾相关工作。第三节介绍策略语义压缩模型并定义策略性率失真函数。第四节给出直接编码器、远程编码器与全信息编码器的主要率失真结果以及受率约束的高斯劝说结果。第五节阐述多模态以及学习系统中计算受限推理的启示。第六节讨论局限性与未来方向，第七节作出结论。

II. 相关工作

本文处于率失真理论、策略通信（贝叶斯劝说）、语义压缩与现代机器学习的交叉领域。我们简要回顾最相关的已有工作，并相对于现有文献定位本文的贡献。

A. 率失真理论

率失真（RD）理论给出了在给定保真度准则下有损压缩的基本极限 [1]。对于具有二次失真的高斯源，率失真函数可通过反向注水得到闭式解 [2]。矢量高斯率失真以及间接（远程）信源编码问题也已得到广泛研究 [3], [4]。这些表述假定编码器与解码器共享同一失真度量，并未处理目标不一致问题。

我们的工作区别于经典率失真，允许编码器和解码器采用不同的二次目标函数，其中编码器失真定义在语义变量 $\Theta = BX + V$ 上。我们证明，这将导出带有对数行列式熵约束的后验协方差设计问题，并产生可推广经典率失真的语义注水定律。

B. 策略通信与贝叶斯劝说

策略通信与信号博弈在经济学和控制领域有着悠久的历史，始于空谈模型 [5], [6]。一种特殊情形是，发送方承诺采用某种信息结构，以影响接收方的后验信念与行动，即贝叶斯

劝说设定在 [7]中得到研究。后续工作将这些思想推广到一般状态空间以及具有二次收益的高斯环境，从而形成线性高斯劝说（线性二次高斯劝说） [8]–[11]。信号博弈受到工程学以及经济学研究者的广泛关注，例如见 [12]–[25]。这些模型（除少数考虑通信速率约束 [11], [19] 和量化 [20], [25]–[27]的工作外）通常假设信号传递不受限制，并不施加通信或速率约束。 [11], 研究了带有速率约束的高斯劝说，但仅限于标量域。本文所考虑的全信息设定可视为将 [11] 扩展到语义与多维设定。

C. 语义通信

近期关于语义通信的研究致力于超越符号级保真度，实现任务感知或意义感知传输 [28]–[32]。这些方法大多展示了经验上的收益，并为目标导向通信提供了依据；然而，它们并未研究基本极限的信息论刻画，也未探讨不同目标之间的策略失配。我们的研究在一个简化的多维高斯设定中提供了这样的刻画，指出后验熵是在率和观测约束下决定语义性能的核心资源。研究还将计算与缩放定律及多模态观测联系起来，架起信息论与现代机器学习缩放研究之间的桥梁，这是以往工作所缺乏的创新。

与我们的工作最为接近的研究，或许是 [33], 中对语义策略性率失真的探讨。该论文主要将 [11] 中的概念扩展到若干不同的语义场景，但未涉及本文所研究的具体情形。

D. 多模态学习与高效人工智能

多模态学习结合异构观测，以提升鲁棒性和数据效率 [34]。经验证据表明，多模态模型优于单模态模型，尤其是在接地和推理任务方面。此外，针对大型神经网络，已有研究记录了模型规模、计算量与性能之间的经验缩放定律 [35], [36]。

通过将架构瓶颈和计算限制解释为隐式信息率约束，我们的框架为这些现象提供了信息论解释。在

特别是，我们表明，多模态观测消除了远程编码固有的几何平均语义惩罚，而计算量的增加则对应于后验不确定性的指数级降低。

E. 总结

与以往研究不同，本文在单一的后验设计框架中统一了率失真理论、策略通信、贝叶斯劝说、语义压缩和神经缩放定律。核心对象是潜在语义变量的后验协方差，核心约束则是由率或计算量施加的熵预算。

与假设目标对齐、并关注某种共同失真度量下重构保真度的经典高斯率失真不同，本文研究语义失配和策略失配，并刻画速率约束下的最优后验几何。与通常允许不受限制信号传递的线性二次高斯信号传递和贝叶斯劝说模型不同，本文施加了明确的信息率约束，并推导由此产生的后验设计结构。最后，与近期主要依赖实证的语义通信和多模态学习方法不同，本文的结果从第一性原理出发，提供了多模态优势和计算驱动的缩放定律的信息论解释。据我们所知，这是首项在单一分析框架中统一率失真理论、高斯劝说、语义目标和计算受限学习架构的研究。

III. 问题表述

我们考虑一种策略通信场景，其中编码器和解码器具有不同的语义目标。潜在高斯状态 X 决定了所关注的语义内容，而编码器则通过一个派生的语义变量 $\Theta = BX + V$来评估重建结果。然而，解码器旨在估计状态 X 本身。编码器通过速率受限无噪信道与解码器通信，并预期解码器会采取贝叶斯最优响应。

符号说明：所有随机向量均定义在同一个概率空间上。所有对数均为以 e 为底的自然对数，率以奈特为单位。对于随机向量 Y，我们用 $\mathrm{Cov}(Y)$ 表示其协方差矩阵，并在

$$\Sigma_Y := \mathrm{Cov}(Y)$$

方便时使用。对于辅助随机变量 U， $\Sigma_{Y|U}$ 表示（平均）最小均方误差误差协方差:

$$\Sigma_{Y|U} = \mathbb{E}\big[(Y - \mathbb{E}[Y \mid U])(Y - \mathbb{E}[Y \mid U])^\top\big]. \tag{2}$$

$A \succeq 0$ （相应地， $A \succ 0$）表示 A 为对称半正定（相应地，正定）。对于对称A， B, $A \preceq B$ （Loewner 序）表示 $B - A \succeq 0$。我们用 tr(·) 和 det(·) 分别表示迹和行列式; 并用 $I(\cdot\,;\cdot)$ 和 $h(\cdot)$ 分别表示互信息和微分熵。diag(·) 和 blkdiag(·) 分别表示对角矩阵和块对角矩阵。 I 表示适当维数的单位矩阵。表达式 $Y \perp Z$ 表示随机变量 Y 与 Z 统计独立。

设 $X \in \mathbb{R}^k$ 满足

$$X \sim \mathcal{N}(0, \Sigma_X), \tag{3}$$

其中 $\Sigma_X \succ 0$。

编码 器通过线性高斯语义变换评估重构

$$\Theta = BX + V, \tag{4}$$

其中 $B \in \mathbb{R}^{m\times k}$ 已知，且

$$V \sim \mathcal{N}(0, \Sigma_V). \tag{5}$$

我们假设

$$V \perp X. \tag{6}$$

映射 $X \mapsto \Theta$ 反映了编码器的语义或对齐目标（例如任务相关性、监督信号）。

我们分析三种观测机制。编码器观测到

$$Z \in \{X,\ \Theta,\ (X, \Theta)\}. \tag{7}$$

- **直接编码器：** $Z = X$。编码器可以完美地观测到状态。
- **远程编码器：** $Z = \Theta$。编码器只能观测到 X 的语义代理。
- **全信息编码器：** $Z = (X,\ \Theta)$。编码器可以观测到状态和语义。

对于码长为 n 的情况，一个 $(n,\ R)$ 信源编码由一个编码器组成

$$f_n : \mathcal{Z}^n \to \{1, \ldots, \exp nR\}, \tag{8}$$

$$M = f_n(Z^n), \tag{9}$$

和一个解码器

$$g_n : \{1, \dots, \exp\{nr\}\} \to \mathbb{R}^{kn}, \tag{10}$$

$$\hat{X}^n = g_n(M). \tag{11}$$

速率约束为

$$R = \frac{1}{n} \log |\mathcal{M}_n|. \tag{12}$$

解码器最小化二次失真

$$D_d(f_n, g_n) = \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}\Big[(X_t - \hat{X}_t)^\top W_d (X_t - \hat{X}_t)\Big], \tag{13}$$

其中 $W_d \succ 0$。由于 X 是高斯分布，且目标函数严格凸，因此解码器的唯一最优响应是最小均方误差估计器

$$\hat{X}^n = \mathbb{E}[X^n \mid M]. \tag{14}$$

编码器失真为

$$D_e(f_n, g_n^\star) = \frac{1}{n} \sum_{t=1}^{n} \mathbb{E}\Big[(\Theta_t - \hat{X}_t)^\top W_e (\Theta_t - \hat{X}_t)\Big], \tag{15}$$

其中 $W_e \succ 0$。这里，我们有意将语义失配定义为相对于解码器的状态估计值 X 来衡量，因为在学习系统的下游决策管道中，解码器通常会暴露 X （或基于它采取行动），而不是单独进行语义重构 Θ。

若存在一系列编码器 $\{f_n\}$ ，使得式 (14) 成立且 ，则语义失真水平 D_e 在速率 R 下可达

$$\limsup_{n \to \infty} D_e(f_n, g_n^\star) \le D_e. \tag{16}$$

策略性率失真函数为

$$R(D_e) = \inf \{R : \ D_e \text{ is achievable at rate } R\}. \tag{17}$$

IV. 主要结果

现在，我们针对第三节介绍的三种编码器观测模型，刻画策略性高斯率失真函数。在直接机制和远程机制中，该问题可化简为在对数行列式速率约束下，对后验协方差进行凸优化，从而得到显式的语义注水解。在全信息机制中，编码器还可以通过语义噪声操纵后验互协方差，因此形成了一个受率约束的高斯说服问题，并且在可对角化情形下有闭式解。

A. 直接编码器： X的观测

我们首先考虑编码器仅观测到 X的情形，因此任何可容许消息都会产生满足 $0 \preceq K_X \preceq \Sigma_X$的后验误差协方差 K_X。

ˆ ˆ 利用正交分解 $X = X + e$ 与 $e \perp X$，语义失真可表示为一种形式，我们将在此设定中反复使用。我们先介绍一些给出此类辅助结果的引理。

引理 1. 编码器失真满足

$$D_e = \mathrm{tr}(W_e C_0) + \mathrm{tr}(\tilde{W} K_X). \tag{18}$$

其中， K_X 表示后验误差协方差

$$K_X = \Sigma_{X|M}, \tag{19}$$

语义偏移协方差表示为：

$$C_0 = (B - I)\Sigma_X (B - I)^\top + \Sigma_V, \tag{20}$$

有效权重矩阵为

$$\tilde{W} = W_e(B - I) + (B - I)^\top W_e + W_e. \tag{21}$$

证明。 回顾 $X \sim \mathcal{N}(0, \Sigma_X)$ 与 $\Theta = BX + V$，其中 $V \sim \mathcal{N}(0, \Sigma_V)$ 独立于 X。ˆ ˆ 解码器使用 $X = \mathbb{E}[X \mid M]$，我们将后验误差定义为 $e := X - X$。根据最小均方误差 ˆ 正交原理， $e \perp X$ 且

$$\mathrm{Cov}(e) = K_X,$$
$$\mathrm{Cov}(\hat{X}) = \Sigma_X - K_X.$$

根据定义，

$$D_e = \mathbb{E}\Big[(\Theta - \hat{X})^\top W_e (\Theta - \hat{X})\Big] = \mathrm{tr}\Big(W_e \,\mathrm{Cov}(\Theta - \hat{X})\Big).$$

ˆ 使用 $\Theta = BX + V = B(X + e) + V$，我们得到

$$\Theta - \hat{X} = (B - I)\hat{X} + Be + V.$$

ˆˆ 由于 $e \perp X$ 和 $V \perp (X,\ M)$，$(X,\ e,\ V)$ 之间的所有交叉协方差均为零，因此

$$\begin{aligned}\operatorname{Cov}(\Theta - \hat{X}) &= (B-I)\operatorname{Cov}(\hat{X})\,(B-I)^\top + B\operatorname{Cov}(e)\,B^\top + \Sigma_V \\ &= (B-I)(\Sigma_X - K_X)(B-I)^\top + BK_XB^\top + \Sigma_V.\end{aligned}$$

展开并化简后得到

$$\operatorname{Cov}(\Theta - \hat{X}) = C_0 + (BK_X + K_XB^\top - K_X).$$

乘以 W_e 并取迹，得到

$$D_e = \operatorname{tr}(W_eC_0) + \operatorname{tr}\big((W_eB + B^\top W_e - W_e)K_X\big)\,.$$

利用 $\tilde{W} = W_e(B-I) + (B-I)^\top W_e + W_e = W_eB + B^\top W_e - W_e$，得到 (18)。 □

引理 2. 设 $X \sim \mathcal{N}(0, \Sigma_X)$，并令 U 为任意随机变量。定义

$$K_X := \Sigma_{X|U}. \tag{22}$$

Then

$$I(X;U) \geq \frac{1}{2}\log\frac{\det\Sigma_X}{\det K_X}, \tag{23}$$

with 当且仅当 $X \mid U$ 几乎必然服从协方差为 K_X 的高斯分布时，等号成立。

证明。 我们有 $I(X;\ U) = h(X) - h(X \mid U)$。由于 $X \sim \mathcal{N}(0, \Sigma_X)$,

$$h(X) = \frac{1}{2}\log\big((2\pi e)^k \det\Sigma_X\big)\,.$$

对于每个 u，令 $\Sigma_{X|u}$ 为 $X \mid U = u$的协方差。于是

$$h(X \mid U = u) \leq \frac{1}{2}\log\big((2\pi e)^k \det\Sigma_{X|u}\big)\,,$$

当且仅当 $X \mid U = u$ 为高斯分布时，等号成立。取期望并应用对数行列式的凹性，得到

$$h(X \mid U) \leq \frac{1}{2}\log\big((2\pi e)^k \det K_X\big)\,,$$

这意味着 (23)。等号成立要求服从高斯分布，且条件协方差为常数。 □

引理 3. 设 $X \sim \mathcal{N}(0, \Sigma_X)$，并设 $0 \prec K_X \preceq \Sigma_X$。定义

$$\Sigma_Z^{-1} := K_X^{-1} - \Sigma_X^{-1}. \tag{24}$$

设 $Z \sim \mathcal{N}(0, \Sigma_Z)$ 与 X 独立，并定义 $U = X + Z$。则 $\Sigma_{X|U} = K_X$ 且

$$I(X;U) = \frac{1}{2} \log \frac{\det \Sigma_X}{\det K_X}. \tag{25}$$

证明。对于具有 $U = X + Z$ 的联合高斯 (X, U)，后验协方差为

$$\Sigma_{X|U} = (\Sigma_X^{-1} + \Sigma_Z^{-1})^{-1} = \left(\Sigma_X^{-1} + K_X^{-1} - \Sigma_X^{-1}\right)^{-1} = K_X.$$

互信息满足 $I(X; U) = h(X) - h(X \mid U)$，由此得到 (25) . □

定理 1。当编码器观测 X 时，策略性率失真函数为

$$R(D_e) = \min_{0 \preceq K_X \preceq \Sigma_X} \left\{ \frac{1}{2} \log \frac{\det \Sigma_X}{\det K_X} : \ \mathrm{tr}(\tilde{W} K_X) \le D_e' \right\}, \tag{26}$$

其中 $D_e' := D_e - \text{迹}(W_e C_0)$。对于每一个可行的 K_X，高斯测试信道 $U = X + Z$ 达到率 $\frac{1}{2} \log \frac{\text{行列式 } \Sigma_X}{\text{行列式 } K_X}$ 并诱导 $\Sigma_{X|U} = K_X$。

~证明。可实现性：任取满足 $0 \preceq K_X \preceq \Sigma_X$ 且迹$(W K_X) \le D_e'$ 的 K_X。由引理 3，存在一个高斯编码器 $U = X + Z$，它诱导 $\Sigma_{X|U} = K_X$ 并达到

$$R = I(X;U) = \frac{1}{2} \log \frac{\det \Sigma_X}{\det K_X}.$$

根据测试信道进行分块编码，可使失真任意接近迹$(W_e C_0)$ +~迹$(W K_X)$，并使率任意接近 R。对所有可行的 K_X 取下确界，即得 (26)。

逆定理：任何编码都会导出一个后验协方差 K_X，满足 $0 \preceq K_X \preceq \Sigma_X$ 和失真 (18)。此外，根据引理 2，

$$R \ge I(X;M) \ge \frac{1}{2} \log \frac{\det \Sigma_X}{\det K_X}.$$

因此，任何可达的 $(R, \ D_e)$ 都必须满足 (26) 中的约束，其中某个 K_X 满足上述条件。

□

注 1。在不受约束的信息设计中，发送方选择一种信息结构^，使其在贝叶斯可行性*[7]*的约束下导出后验均值 $X = \mathbb{E}[X \mid M]$ 的分布。对于这里考虑的高斯情形，这意味着二阶矩条件（例如参见 *[9], [10]*），

$$0 \preceq \Sigma_{\hat{X}\hat{X}} \preceq \Sigma_X, \tag{27}$$

连同

$$K_X = \Sigma_X - \Sigma_{\hat{X}\hat{X}}. \tag{28}$$

由于发送方的二次目标函数关于这些协方差块是仿射的，因此无约束问题是一个半定规划，如 *[9]* 所示。

施加显式通信速率约束 $I(X;\ M) \leq R$，通过对后验误差协方差施加熵预算来缩小贝叶斯可行集合:

$$\log\det K_X \geq \log\det \Sigma_X - 2R, \tag{29}$$

这与（1）中强调的约束相同。

注2。 程序（26）可用归一化变量 $D:=\Sigma_X^{-1/2} K_X \Sigma_X^{-1/2}$ 表示为

$$\min_{0\prec D\preceq I} \ \frac{1}{2}\log\frac{1}{\det D} \qquad \textit{subject to} \qquad \operatorname{tr}(A_X D) \leq D'_e, \tag{30}$$

其中，有效归一化权重为

$$A_X := \Sigma_X^{1/2} \tilde{W}\, \Sigma_X^{1/2}. \tag{31}$$

~因此，注水法所涉及的特征值几何结构由 A_X 决定，而非由 Σ_X 和 W 分别决定。若 A_X 存在非正特征值，则在速率最小化解中，这些方向绝不会被压缩，因为将相应的归一化后验特征值设为 *1*，既能降低率，也不会使语义约束收紧。

推论1。令 $A_X = U\Lambda U^\top$，其中

$$\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_k), \tag{32}$$

并定义指标集

$$\mathcal{I}_+ := \{i : \lambda_i > 0\}, \tag{33}$$

and

$$\mathcal{I}_0 := \{i : \lambda_i \leq 0\}. \tag{34}$$

那么，存在一个形如以下形式的(26)优化器

$$K_X^\star = \Sigma_X^{1/2} U \operatorname{diag}(d_1^\star, \ldots, d_k^\star)\, U^\top \Sigma_X^{1/2}, \tag{35}$$

与 $0 < d_i^\star \leq 1$ 和

$$d_i^\star = 1, \tag{36}$$

对于 $i \in \mathcal{I}_0$，而对于 $i \in \mathcal{I}_+$

$$d_i^\star = \min\left\{1,\ \frac{\nu^\star}{\lambda_i}\right\}. \tag{37}$$

若 $D_e' \geq$ 迹(A_X)，则 $K_X^\star = \Sigma_X$ 是可行的，且

$$R(D_e) = 0. \tag{38}$$

若 $D_e' <$ 迹(A_X)，则语义约束处于激活状态，且 $\nu^\star > 0$ 由以下条件唯一确定

$$\sum_{i\in\mathcal{I}_+} \min\{\lambda_i, \nu^\star\} = D_e' - \sum_{i\in\mathcal{I}_0} \lambda_i. \tag{39}$$

此外 ，当且仅当满足以下条件时，解处于内部区域（任意 $i \in \mathcal{I}_+$均未饱和）

$$\nu^\star < \min_{i\in\mathcal{I}_+} \lambda_i. \tag{40}$$

在该内部区域中，

$$\nu^\star = \frac{D_e' - \sum_{i\in\mathcal{I}_0} \lambda_i}{|\mathcal{I}_+|}, \tag{41}$$

且由此得到的率为

$$R(D_e) = \frac{1}{2}\sum_{i\in\mathcal{I}_+} \log\left(\frac{\lambda_i}{\nu^\star}\right). \tag{42}$$

˜**注3.** 尽管 $W_e \succ 0$，诱导矩阵 W 未必是半正定的。如果 A_X具有负特征值，那么沿这些方向降低后验不确定性可能会增加˜$\mathrm{tr}(WK_X)$，因此策略性率失真曲线未必会随可用率单调变化。排除这种异常的一项充分条件是

$$\tilde{W} \succeq 0, \tag{43}$$

等价地， $A_X \succeq 0$。

B. 远程编码器：仅观测 Θ

假设编码器观测到 $\Theta = BX + V$，其中 $X \sim \mathcal{N}(\mathbf{0}, \Sigma_X)$、 $V \sim \mathcal{N}(\mathbf{0}, \Sigma_V)$，且 ˆ $V \perp X$，而解码器以 $X = \mathbb{E}[X \mid M]$ 作出最佳响应。

Let

$$\Sigma_\Theta = \mathrm{Cov}(\Theta) = B\Sigma_X B^\top + \Sigma_V. \tag{44}$$

定义线性贝叶斯回归

$$L_X = \Sigma_X B^\top \Sigma_\Theta^{-1}. \tag{45}$$

$$L_V = \Sigma_V \Sigma_\Theta^{-1}. \tag{46}$$

令 K_Θ 表示 Θ的后验误差协方差:

$$K_\Theta = \Sigma_{\Theta|M}. \tag{47}$$

定义无限速率远程基准失真

$$D_\infty^{\mathrm{rem}} = \mathbb{E}\Big[\big(\Theta - \mathbb{E}[X \mid \Theta]\big)^\top W_e \big(\Theta - \mathbb{E}[X \mid \Theta]\big)\Big], \tag{48}$$

并定义

$$Q = L_X^\top \tilde{W} L_X + L_X^\top W_e L_V + L_V^\top W_e L_X. \tag{49}$$

定理 2. 策略性远程率失真函数为

$$R(D_e) = \min_{0 \preceq K_\Theta \preceq \Sigma_\Theta} \left\{ \frac{1}{2} \log \frac{\det \Sigma_\Theta}{\det K_\Theta} : \ \mathrm{tr}(Q K_\Theta) \le D_e - D_\infty^{\mathrm{rem}} \right\}. \tag{50}$$

对于每个可行的 K_Θ，高斯测试信道 $U = \Theta + Z$ （其中 $Z \perp \Theta$ 为高斯）均可达到最小值。

证明。 由于 $M = f_n(\Theta^n)$ 且 $|\mathcal{M}_n| \le \exp\{nr\}$,

$$nR \ge H(M) \ge I(\Theta^n; M).$$

根据标准的单字母化（时分复用），存在一个单字母对 (Θ, U)，其中 $U = (M, Q)$，使得 $\frac{1}{n} I(\Theta^n;\ M) = I(\Theta;\ U)$ 且 $K_\Theta = \Sigma_{\Theta|U}$。将引理 2 应用于 (Θ, U)，得到

$$R \ge I(\Theta; U) \ge \frac{1}{2} \log \frac{\det \Sigma_\Theta}{\det K_\Theta}.$$

^^令 $\Theta := \mathbb{E}[\Theta \mid M]$ 和 $e_\Theta := \Theta - \Theta$，因此 $\mathrm{Cov}(e_\Theta) = K_\Theta$。由于 $(X,\ V,\ \Theta)$ 联合服从高斯分布，条件期望是线性的:

$$\mathbb{E}[X \mid \Theta] = L_X \Theta,$$

$$\mathbb{E}[V \mid \Theta] = L_V \Theta.$$

由于 M 是 Θ 的函数，塔式性质给出

$$\hat{X} = \mathbb{E}[X \mid M] = L_X \hat{\Theta}.$$

定义新息 $N_X := X - L_X \Theta$ 和 $N_V := V - L_V \Theta$，它们与 Θ相互独立。然后，^最小均方误差为 $e := X - X = N_X + L_X e_\Theta$，并且可以证明

$$D_e = D_\infty^{\mathrm{rem}} + \mathrm{tr}(Q K_\Theta),$$

其中 D_∞^{rem}和 Q 分别由式 (48) 和 (49) 定义。

逆定理： 前述步骤表明，任意可达的 $(R,\ D_e)$ 都会导出 K_Θ，且 $0 \preceq K_\Theta \preceq \Sigma_\Theta$ 满足 (50) 中的约束。

可实现性： 固定任意满足 $0 \prec K_\Theta \preceq \Sigma_\Theta$ 的 K_Θ。定义 $\Sigma_Z^{-1} := K_\Theta^{-1} - \Sigma_\Theta^{-1}$，并选择与 Θ 独立的 $Z \sim \mathcal{N}(0,\ \Sigma_Z)$。令 $U := \Theta + Z$，由高斯恒等式可得 $\Sigma_{\Theta|U} = K_\Theta$，以及 $I(\Theta;\ U) = \frac{1}{2}\log \frac{\det \Sigma_\Theta}{\det K_\Theta}$。按照此测试信道进行编码，可达到失真 $D_\infty^{\mathrm{rem}} + \mathrm{tr}(QK_\Theta)$。对可行的 K_Θ 进行优化，即可证明该定理。

□

推论 2。 考虑高于无限速率远程基线的超额失真：

$$\Delta_{\mathrm{rem}}(R) = D_e^{(\mathrm{rem})}(R) - D_\infty^{\mathrm{rem}}. \tag{51}$$

假设 $Q \succ 0$，且 Q 和 Σ_Θ 同时对角化，即存在一个正交的 U，使得

$$U^\top \Sigma_\Theta U = \mathrm{diag}(\sigma_{\Theta,1}^2, \ldots, \sigma_{\Theta,k}^2), \tag{52}$$

and

$$U^\top Q\, U = \mathrm{diag}(q_1, \ldots, q_k), \tag{53}$$

其中 $q_i > 0$。那么，定理 *2* 中的一个最优后验协方差在此基下为对角矩阵，$K_\Theta^\star(R) = U\ \mathrm{diag}(k_1^\star(R) \ldots, k_k^\star(R))\ U^\top$，其中

$$k_i^\star(R) = \min\left\{\sigma_{\Theta,i}^2,\ \frac{\nu^\star(R)}{q_i}\right\}, \tag{54}$$

其中选择 $\nu^\star(R) > 0$，使速率约束处于激活状态：

$$\sum_{i=1}^{k} \log \frac{\sigma_{\Theta,i}^2}{k_i^\star(R)} = 2R. \tag{55}$$

在内部（未饱和）区域，对于所有 i，我们有 $k_i^\star(R) < \sigma_{\Theta,i}^2$

$$K_\Theta^\star(R) = \nu^\star(R)\, Q^{-1}, \tag{56}$$

with

$$\nu^\star(R) = \exp\left(\frac{1}{k}\left[\sum_{i=1}^{k} \log\bigl(q_i \sigma_{\Theta,i}^2\bigr) - 2R\right]\right), \tag{57}$$

且超额失真等于

$$\Delta_{\mathrm{rem}}(R) = \mathrm{tr}(QK_\Theta^\star(R)) = k\, \nu^\star(R). \tag{58}$$

证明。 对于固定的率 R，定理 2 等价于（相差常数 D_∞^{rem}）

$$\min_{0\prec K_\Theta\preceq\Sigma_\Theta} \ \mathrm{tr}(QK_\Theta) \quad \text{s.t.} \quad \frac{1}{2}\log\frac{\det\Sigma_\Theta}{\det K_\Theta}\le R.$$

等价地，该约束为 $\log\det K_\Theta \ge \log\det\Sigma_\Theta - 2R$。在同时对角化的条件下，写作 $K_\Theta = U\,\mathrm{diag}(k_i)\,U^\top$。于是，该问题化简为

$$\min_{0<k_i\le\sigma_{\Theta,i}^2} \ \sum_{i=1}^{k} q_i k_i \quad \text{s.t.} \quad \sum_{i=1}^{k}\log k_i \ge \sum_{i=1}^{k}\log\sigma_{\Theta,i}^2 - 2R.$$

KKT 条件给出 $k_i^\star = \min\{\sigma_{\Theta,i}^2, \nu/q_i\}$，且对数约束处于激活状态，从而得到 (54)–(58)。

□

注释 4。 推论 *2* 假设 $Q\succ 0$ 并且与 Σ_Θ 同时对角化，以便用标量参数 $\{q_i,\sigma_{\Theta,i}^2\}$表示解。在没有这一可交换性假设的情况下，相同的反向注水结构在对称矩阵的特征基中仍然成立

$$A_\Theta := \Sigma_\Theta^{1/2} Q\,\Sigma_\Theta^{1/2}. \tag{59}$$

如果 A_Θ 具有非正特征值，则相应的归一化后验特征值饱和为 *1*（这些方向上不进行压缩），率仅分配到正特征值子空间上。

C. 全信息编码器：受率约束的高斯劝说

当编码器观测到 $(X,\ \Theta)$ 时，它还可以操纵状态 X与语义噪声 V之间的后验相关性，从而引入策略自由度。回顾 $X\sim\mathcal{N}(0,\ \Sigma_X)$ 和 $V\sim\mathcal{N}(0,\ \Sigma_V)$ 相互独立，且 $\Theta = BX+V$。在全信息机制下，编码器观测到 $(X,\ \Theta)$ 并发送率受限消息 M。定义

$$W = \begin{bmatrix} X \\ V \end{bmatrix}, \tag{60}$$

具有先验协方差

$$\Sigma_W = \mathrm{blkdiag}(\Sigma_X,\Sigma_V), \tag{61}$$

以及后验误差协方差

$$K_W = \Sigma_{W|M}. \tag{62}$$

定理 3. 当编码器观测到 X、Θ 时，策略性率失真函数为

$$R(D_e) = \min_{0 \preceq K_W \preceq \Sigma_W} \left\{ \frac{1}{2} \log \frac{\det \Sigma_W}{\det K_W} : \ \mathrm{tr}(\bar{W} K_W) \leq D'_e \right\}, \tag{63}$$

其中 $D'_e := D_e - \mathrm{tr}(W_e C_0)$ 且

$$\bar{W} = \begin{bmatrix} \tilde{W} & W_e \\ W_e & 0 \end{bmatrix}. \tag{64}$$

可通过 高斯测试信道实现，其形式为 $U = W + Z$ 其中 Z 服从高斯分布 .

证明。 由于 $V = \Theta - BX$ 是 (X, Θ) 的确定性可逆线性变换，编码器实际上观测到 (X, V)，对于任意消息 M，

$$I(X, \Theta; M) = I(X, V; M) = I(W; M).$$

由于 $|\mathcal{M}_n| \leq \exp\{nr\}$，我们有 $H(M) \leq nR$，因此

$$I(W^n; M) \leq H(M) \leq nR.$$

我们采用标准率失真逆定理中的单字母化方法。利用链式法则和条件化降低熵，对于独立同分布的 W_t，

$$I(W^n; M) = \sum_{t=1}^{n} I(W_t; M \mid W^{t-1}) \geq \sum_{t=1}^{n} I(W_t; M).$$

令 $Q \sim$均匀分布$\{1, \ldots, n\}$与所有变量相互独立，并定义 $W := W_Q$ 和 $U := (M, Q)$。则

$$\begin{aligned} I(W; U) &= I(W_Q; M, Q) \\ &= \frac{1}{n} \sum_{t=1}^{n} I(W_t; M) \\ &\leq \frac{1}{n} I(W^n; M) \\ &\leq R. \end{aligned}$$

将引理 2 应用于高斯向量 W 和辅助变量 U：

$$I(W; U) \geq \frac{1}{2} \log \frac{\det \Sigma_W}{\det \Sigma_{W|U}}.$$

令 $K_W := \Sigma_{W|U}$。由于 $U = (M, Q)$，这恰好是经过时分复用后由编码诱导的逐字母后验误差协方差。综合可得

$$R \geq \frac{1}{2} \log \frac{\det \Sigma_W}{\det K_W},$$

等价地，

$$\log \det K_W \geq \log \det \Sigma_W - 2R.$$

最后，根据全协方差公式，$0 \preceq K_W \preceq \Sigma_W$。

^^^ 令 $X := \mathbb{E}[X \mid M]$，并定义最小均方误差 $e := X - X$。另定义 $V := \mathbb{E}[V \mid M]$ 和 $v := V - V$。于是

$$K_W = \Sigma_{W|M} = \mathrm{Cov}\left(\begin{bmatrix} e \\ v \end{bmatrix} \right) = \begin{bmatrix} K_X & K_{XV} \\ K_{VX} & K_V \end{bmatrix},$$

其中 $K_X = \mathrm{Cov}(e)$，且 $K_{XV} = \mathrm{Cov}(e, v)$。

^ 根据正交原理，e 与 M的任意可测函数正交，因此 $e \perp V$，进而

$$\mathrm{Cov}(e, V) = \mathrm{Cov}(e, \hat{V} + v) = \mathrm{Cov}(e, v) = K_{XV}.$$

由于独立性，$\mathrm{Cov}(X, V) = 0$，我们还有

$$0 = \mathrm{Cov}(X, V) = \mathrm{Cov}(\hat{X} + e, V) = \mathrm{Cov}(\hat{X}, V) + \mathrm{Cov}(e, V),$$

由此得到

$$\mathrm{Cov}(\hat{X}, V) = -K_{XV}.$$

接下来，

$$\Theta - \hat{X} = BX + V - \hat{X} = (B - I)\hat{X} + Be + V.$$

^使用 $e \perp X$，协方差为

$$\begin{aligned} \mathrm{Cov}(\Theta - \hat{X}) = (B - I)\mathrm{Cov}(\hat{X})(B - I)^\top + BK_X B^\top + \Sigma_V \\ + (B - I)\mathrm{Cov}(\hat{X}, V) + \mathrm{Cov}(V, \hat{X})(B - I)^\top \\ + B\mathrm{Cov}(e, V) + \mathrm{Cov}(V, e)B^\top. \end{aligned}$$

^^ 利用 $\mathrm{Cov}(X) = \Sigma_X - K_X$、$\mathrm{Cov}(X, V) = -K_{XV}$ 和 $\mathrm{Cov}(e, V) = K_{XV}$，交叉项可简化为

$$-(B - I)K_{XV} - K_{VX}(B - I)^\top + BK_{XV} + K_{VX}B^\top = K_{XV} + K_{VX}.$$

K_X项可按引理 1 化简：

$$(B-I)(\Sigma_X - K_X)(B-I)^\top + BK_XB^\top = (B-I)\Sigma_X(B-I)^\top + \big(BK_X + K_XB^\top - K_X\big).$$

因此

$$\mathrm{Cov}(\Theta - \hat{X}) = C_0 + \big(BK_X + K_XB^\top - K_X\big) + (K_{XV} + K_{VX}).$$

因此，

$$\begin{aligned} D_e &= \mathrm{tr}\Big(W_e \,\mathrm{Cov}(\Theta - \hat{X})\Big) \\ &= \mathrm{tr}(W_eC_0) + \mathrm{tr}\big((W_eB + B^\top W_e - W_e)K_X\big) + \mathrm{tr}(W_e(K_{XV} + K_{VX}))\,. \end{aligned}$$

定义 $\tilde{W} := W_eB + B^\top W_e - W_e$ 以及

$$\bar{W} := \begin{bmatrix} \tilde{W} & W_e \\ W_e & 0 \end{bmatrix},$$

我们有：

$$D_e = \mathrm{tr}(W_eC_0) + \mathrm{tr}(\bar{W}K_W).$$

逆定理： 上述分析表明，任何可达的 (R, D_e) 都会导出一个满足 $0 \preceq K_W \preceq \Sigma_W$ 且 $\log\det K_W \geq \log\det\Sigma_W - 2R$ 的矩阵 K_W，并达到 $D_e = \mathrm{tr}(W_eC_0) + \mathrm{tr}(WK_W)$。

可实现性： 令 $Q \sim \mathrm{Unif}\{1,\dots,n\}$ 与所有其他变量相互独立，定义 $W := W_Q$ 以及 $U := (M, Q)$。则

$$I(W;U) = I(W_Q; M, Q) \leq I(W^n; M) \leq H(M) \leq nR,$$

因此 $I(W;U) \leq R$。解码器重构 U^n并形成 $X^n = \mathbb{E}[X^n \mid U^n]$，这会产生所需的后验误差协方差 K_W，并实现失真 $\mathrm{tr}(W_eC_0) + \mathrm{tr}(WK_W)$。对可行的 K_W 进行优化即可证明该定理。 □

当 Σ_X， Σ_V， B 和 W_e 共享同一特征基时，定理 3 中的协方差设计问题会沿各个特征方向解耦。每个方向都可化为针对这对变量 $(x_i,\ v_i)$ 的二维后验设计问题，其行列式预算由标量注水定律分配。

假设1。 存在一个正交矩阵 U ，使得

$$U^\top \Sigma_X U = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_k^2), \tag{65}$$

$$U^\top \Sigma_V U = \mathrm{diag}(\tau_1^2, \ldots, \tau_k^2), \tag{66}$$

$$U^\top B U = \mathrm{diag}(b_1, \ldots, b_k), \tag{67}$$

$$U^\top W_e U = \mathrm{diag}(w_1, \ldots, w_k), \tag{68}$$

对于所有 i，均有 $\sigma_i > 0$、 $\tau_i > 0$ 和 $w_i > 0$ 。

备注5。 条件 $\tau_i > 0$ 确保对数行列式 Σ_V 有定义，并保证定理*4*中的每个 2×2 块都具有一个正特征值和一个负特征值。

定理4。 在假设*1*下，存在一个最优后验误差协方差 $K_W^\star(R)^-$，它在基 $U :=$ 分块对角(U, U)下跨 i 呈块对角形式:

$$K_W^\star(R) = \bar{U}\ \mathrm{blkdiag}\big(K_1^\star(R), \ldots, K_k^\star(R)\big)\ \bar{U}^\top, \tag{69}$$

包含 2×2 个块

$$K_i^\star(R) = D_i\, \tilde{K}_i^\star(R)\, D_i, \tag{70}$$

其中

$$D_i = \mathrm{diag}(\sigma_i, \tau_i). \tag{71}$$

每个归一化块可写为

$$\tilde{K}_i^\star(R) = R_i\, \mathrm{diag}(\delta_i^\star(R), 1)\, R_i^\top, \tag{72}$$

其中 $0 < \delta_i^\star(R) \le 1$ 且 R_i 对角化

$$A_i = \begin{bmatrix} \tilde{w}_i\, \sigma_i^2 & w_i\, \sigma_i \tau_i \\ w_i\, \sigma_i \tau_i & 0 \end{bmatrix}, \tag{73}$$

with

$$\tilde{w}_i = w_i(2b_i - 1). \tag{74}$$

记作 $R_i^\top A_i R_i = \mathrm{diag}(a_i^+, a_i^-)$ ，其中 $a_i^+ > 0 > a_i^-$ ，我们有

$$a_i^+ = \frac{1}{2}\left(\tilde{w}_i \sigma_i^2 + \sqrt{\tilde{w}_i^2 \sigma_i^4 + 4 w_i^2 \sigma_i^2 \tau_i^2}\right). \tag{75}$$

标量 $\delta_i^\star(R)$ 满足注水定律

$$\delta_i^\star(R) = \min\left\{1,\ \frac{\eta^\star(R)}{a_i^+}\right\}, \tag{76}$$

其中， $\eta^\star(R) > 0$ 的选取使归一化对数行列式约束取等号：

$$\sum_{i=1}^{k} \log \delta_i^\star(R) = -2R. \tag{77}$$

¯证明。 定义 $U :=$ 分块对角(U, U)，并进行变换

$$K' = \bar{U}^\top K_W \bar{U},$$

$$\Sigma' = \bar{U}^\top \Sigma_W \bar{U}.$$

根据假设1，我们有

$$\Sigma' = \mathrm{blkdiag}(\Sigma'_X, \Sigma'_V) = \mathrm{blkdiag}\big(\mathrm{diag}(\sigma_i^2), \mathrm{diag}(\tau_i^2)\big)\,.$$

由于迹和行列式在正交变换下保持不变，定理 3 中的协方差设计问题等价于

$$\min_{0 \preceq K' \preceq \Sigma'} \ \mathrm{tr}(\bar{W}' K') \quad \text{s.t.} \quad \log\det K' \geq \log\det \Sigma' - 2R,$$

其中， $\bar{W}' := \bar{U}^\top \bar{W}\, \bar{U}$在 i上呈块对角结构，包含 2×2个块

$$\bar{W}'_i = \begin{bmatrix} \tilde{w}_i & w_i \\ w_i & 0 \end{bmatrix}, \qquad \tilde{w}_i := w_i(2b_i - 1).$$

用 Σ'归一化：定义 $\tilde{K} := \Sigma'^{-1/2} K' \Sigma'^{-1/2}$。于是 $0 \preceq K' \preceq \Sigma'$ 变为 $0 \preceq \tilde{K} \preceq I$，且

$$\log\det K' = \log\det \Sigma' + \log\det \tilde{K},$$

$$\mathrm{tr}(\bar{W}' K') = \mathrm{tr}\Big(\Sigma'^{1/2} \bar{W}' \Sigma'^{1/2}\ \tilde{K}\Big)\,.$$

Let

$$A := \Sigma'^{1/2} \bar{W}' \Sigma'^{1/2},$$

其沿 i呈块对角结构，块为

$$A_i = \begin{bmatrix} \tilde{w}_i\, \sigma_i^2 & w_i\, \sigma_i \tau_i \\ w_i\, \sigma_i \tau_i & 0 \end{bmatrix}.$$

因此，归一化问题为

$$\min_{0 \preceq \tilde{K} \preceq I} \ \mathrm{tr}(A\tilde{K}) \quad \text{s.t.} \quad \log\det \tilde{K} \geq -2R.$$

˜接下来，我们将其化简为 k 个相互独立的 2×2 块。将 K 划分为 $k\times k$ 个大小为 2×2的块:

$$\tilde{K}=[\tilde{K}_{ij}]_{i,j=1}^{k}.$$

由于 A是块对角的,

$$\operatorname{tr}(A\tilde{K})=\sum_{i=1}^{k}\operatorname{tr}(A_i\tilde{K}_{ii}),$$

因此，非对角块不会影响目标函数。

˜对于可行性，注意到 $K\succeq 0$ 意味着行列式不等式

$$\det(\tilde{K})\le\prod_{i=1}^{k}\det(\tilde{K}_{ii}).$$

定义分块对角矩阵 $\tilde{K}^{\mathrm{bd}}:=\mathsf{blkdiag}(\tilde{K}_{11},\ldots,\tilde{K}_{kk})$。则 $\tilde{K}^{\mathrm{bd}}\preceq I$且

$$\det(\tilde{K}^{\mathrm{bd}})=\prod_{i=1}^{k}\det(\tilde{K}_{ii})\ge\det(\tilde{K}).$$

因此，只要 $\tilde{K}$ 可行， $\tilde{K}^{\mathrm{bd}}$ 就可行，并且能达到相同的目标函数值。因此，存在一个块对角的最优解:

$$\tilde{K}^{\star}=\operatorname{blkdiag}(\tilde{K}_1^{\star},\ldots,\tilde{K}_k^{\star}),\qquad 0\preceq\tilde{K}_i^{\star}\preceq I.$$

约束变为

$$\log\det\tilde{K}^{\star}=\sum_{i=1}^{k}\log\det(\tilde{K}_i^{\star})\ge -2R.$$

固定 i 并进行特征分解

$$A_i=R_i\operatorname{diag}(a_i^{+},a_i^{-})\,R_i^{\top},\qquad a_i^{+}>0>a_i^{-}.$$

特征值为

$$a_i^{\pm}=\frac{1}{2}\left(\tilde{w}_i\sigma_i^2\pm\sqrt{\tilde{w}_i^2\sigma_i^4+4w_i^2\sigma_i^2\tau_i^2}\right),$$

因此确实有 $a_i^{+}a_i^{-}=\det(A_i)=-w_i^2\sigma_i^2\tau_i^2<0$。

˜令 K_i 为任意可行的 2×2 半正定矩阵，其特征值为 $0<\kappa_{i,\,1}\le\kappa_{i,\,2}\le 1$。固定˜特征值（$\kappa_{i,\,1}$，$\kappa_{i,\,2}$）时，$\mathsf{tr}(A_iK_i)$ 的最小值可通过将较大的特征值$\kappa_{i,\,2}$ 与负特征值 a_i^{-} 对齐来实现。因此，我们可以将注意力限制在

$$\tilde{K}_i=R_i\operatorname{diag}(\kappa_{i,1},\kappa_{i,2})\,R_i^{\top},$$

由此得到代价 $a_i^{+}\kappa_{i,\,1}+a_i^{-}\kappa_{i,\,2}$。

由于 $a_i^- < 0$，增大 $\kappa_{i,\,2}$ 会严格降低目标函数，并增大行列式 $\det(K_i) = \kappa_{i,\,1}\kappa_{i,\,2}$，这在行列式下界的约束下仍可保持可行性。因此，在任何最优解处都必须有 $\kappa_{i,\,2} = 1$。记

$$\delta_i := \kappa_{i,1} = \det(\tilde{K}_i) \in (0, 1].$$

因此，最优块的形式为

$$\tilde{K}_i^\star = R_i \,\mathrm{diag}(\delta_i^\star, 1)\, R_i^\top, \qquad \delta_i^\star \in (0, 1],$$

其代价为 $\mathrm{tr}(A_i \tilde{K}_i^\star) = a_i^- + a_i^+ \delta_i^\star$。

汇总各块可得

$$\mathrm{tr}(A\tilde{K}^\star) = \sum_{i=1}^{k} \big(a_i^- + a_i^+ \delta_i\big),$$

$$\log\det(\tilde{K}^\star) = \sum_{i=1}^{k} \log \delta_i.$$

由于 $\sum_i a_i^-$ 为常数，归一化设计可化简为标量凸规划

$$\min_{0<\delta_i\le 1} \sum_{i=1}^{k} a_i^+ \delta_i \quad \text{s.t.} \quad \sum_{i=1}^{k} \log \delta_i \ge -2R.$$

该约束必须是紧约束：如果 $\sum_i \log \delta_i > -2R$，则可以稍微减小某些 δ_i（同时保持 $\delta_i \le 1$），从而严格降低目标函数。

KKT 条件给出注水解

$$\delta_i^\star(R) = \min\left\{1,\ \frac{\eta^\star(R)}{a_i^+}\right\},$$

其中，$\eta^\star(R) > 0$ 的选取使得（起作用的）约束成立：

$$\sum_{i=1}^{k} \log \delta_i^\star(R) = -2R.$$

撤销归一化后，即得到所述形式 (69)–(77)。 □

V. Implications in Learning Systems

A. 多模态编码器与远程惩罚的消除

我们研究一种遥感场景，其中编码器观测到潜在状态 X 的含噪模态（例如视觉/音频/文本特征），这些模态在给定 X 的条件下相互独立。

令 $X \sim \mathcal{N}(0, \Sigma_X)$，其中 $\Sigma_X \succ 0$，并假设编码器观测到 m 个条件独立的模态

$$Z_j = H_j X + W_j, \tag{78}$$

对于 $j = 1, \ldots, m$，其中 $W_j \sim \mathcal{N}(0, R_j)$ 相互独立，且与 (X, V) 独立。令$Z^{(m)} := (Z_1, \ldots, Z_m)$，并定义最小均方误差代理

$$S^{(m)} := \mathbb{E}[X \mid Z^{(m)}]. \tag{79}$$

定义

$$\Sigma_S^{(m)} := \mathrm{Cov}(S^{(m)}). \tag{80}$$

定义累积精度

$$J_m := \sum_{j=1}^{m} H_j^\top R_j^{-1} H_j \succeq 0. \tag{81}$$

定义归一化可恢复性矩阵

$$\Gamma_m := \Sigma_X^{-1/2} \, \Sigma_S^{(m)} \, \Sigma_X^{-1/2}. \tag{82}$$

等价地,

$$\Gamma_m = I - \Sigma_X^{-1/2} \Sigma_{X|Z^{(m)}} \Sigma_X^{-1/2}. \tag{83}$$

定义几何平均可恢复性因子

$$G_m := \left(\frac{\det \Sigma_S^{(m)}}{\det \Sigma_X} \right)^{1/k}. \tag{84}$$

等价地,

$$G_m = \left(\prod_{i=1}^{k} \gamma_i^{(m)} \right)^{1/k}, \tag{85}$$

其中 $\{\gamma_i^{(m)}\}$ 是 Γ_m的特征值。

定理 5. 后验协方差与可恢复协方差满足

$$\Sigma_{X|Z^{(m)}}^{-1} = \Sigma_X^{-1} + J_m, \tag{86}$$

$$\Sigma_S^{(m)} = \Sigma_X - \Sigma_{X|Z^{(m)}}. \tag{87}$$

此外，$\Sigma_{X|Z^{(m)}}$ 关于 m 在 ***Loewner*** 序下单调递减，而 $\Sigma_S^{(m)}$ 单调递增。

若 $J_m \succ 0$ （等价地，$\mathrm{Null}(J_m) = \{0\}$），则

$$0 \prec \Sigma_{X|Z^{(m)}} \prec \Sigma_X \quad \textit{and} \quad 0 \prec \Sigma_S^{(m)} \prec \Sigma_X, \tag{88}$$

因此 $G_m \in (0,1)$。

此外，若 $\lambda_{\min}(J_m) \to \infty$ 随 $m \to \infty$变化，则

$$\Sigma_{X|Z^{(m)}} \to 0, \qquad \Sigma_S^{(m)} \to \Sigma_X, \qquad G_m \to 1. \tag{89}$$

备注 6。累积精度矩阵 J_m 在线性高斯推断中发挥着核心作用。在矢量远程与间接高斯信源编码中，通常假设测量算子满秩（或诱导精度非奇异），否则信源的某些分量将永远无法观测到，间接编码惩罚也无法消失 *[37], [38]*。在线性滤波和传感器融合中，*[39]–[41]*, J_m 与映射到状态空间中的累积测量精度相一致，而诸如$J_m \succ 0$ 或 $\lambda_{\min}(J_m) \to \infty$ 之类的条件，则是可观测性和持续激励假设的静态对应形式，可确保误差协方差收缩。类似地，在贝叶斯线性反问题中，众所周知，后验收缩仅发生在似然信息子空间上，而与 J_m 的值域正交的方向则保留其先验方差 *[42]*。

在当前的语义压缩设定中，条件 $J_m \succ 0$ 确保潜在状态 X 的任何方向都不会永久对编码器隐藏，而更强的条件$\lambda_{\min}$（J_m）$\to \infty$则保证随着模态数量增加，后验协方差 $\Sigma_{X|Z^{(m)}}$ 会收敛至零。若不满足这些张成条件，则存在一个非平凡子空间，沿着该子空间，无论率或计算量如何，编码器都无法改进其估计；相对于直接编码基准的语义性能差距也无法消失。

证明。等式 (86)–(87) 是标准的线性高斯最小均方误差恒等式。

单调性成立，因为 $J_{m+1} = J_m + H_{m+1}^\top R_{m+1}^{-1} H_{m+1} \succeq J_m$，因此 $\Sigma_{X|Z^{(m+1)}}^{-1} \succeq \Sigma_{X|Z^{(m)}}^{-1}$，进而 $\Sigma_{X|Z^{(m+1)}} \preceq \Sigma_{X|Z^{(m)}}$。

若 $J_m \succ 0$，则 $\Sigma_X^{-1} + J_m \succ \Sigma_X^{-1}$，因此 $(\Sigma_X^{-1} + J_m)^{-1} \prec \Sigma_X$，得到 (88)。

若 $\lambda_{\min}(J_m) \to \infty$，则 $\lambda_{\min}(\Sigma_X^{-1} + J_m) \to \infty$，因此 $\|\Sigma_{X|Z^{(m)}}\| = \|(\Sigma_X^{-1} + J_m)^{-1}\| \to 0$，从而证明 (89)。关于 $\Sigma_S^{(m)}$ 和 G_m 的结论可由 (87) 以及 $\det(\cdot)$ 在 $\mathbb{S}_{++}^k$ 上的连续性得出。□

推论 3. 令 $D_e^{(m)}(R)$ 表示率为 R 时的最优语义失真，其中编码器观测到 $Z^{(m)} = (Z_1, \ldots, Z_m)$，$Z_j = H_j X + W_j$如定理 *5* 所述，且语义变量

为 $\Theta = BX + V$，其中 $V \perp (X, Z^{(m)})$。令 $D_e^{(\mathrm{dir})}(R)$ 表示编码器观测到 X 时对应的失真。

那么：

1) 对于每个 m,

$$D_e^{(m)}(R) = \mathrm{tr}(W_e C_0) + \mathrm{tr}\big(\tilde{W}\,\Sigma_{X|Z^{(m)}}\big) + \Delta_m(R), \tag{90}$$

其中

$$\Delta_m(R) := \min_{0 \prec K \preceq \Sigma_S^{(m)}} \left\{ \mathrm{tr}(\tilde{W}K) : \frac{1}{2}\log\frac{\det \Sigma_S^{(m)}}{\det K} \le R \right\}. \tag{91}$$

2) 若 $\lambda_{\min}(J_m) \to \infty$ 作为 $m \to \infty$ （等价地，$\Sigma_{X|Z^{(m)}} \to 0$），则对于每个固定的 R,

$$\lim_{m\to\infty} D_e^{(m)}(R) = D_e^{(\mathrm{dir})}(R). \tag{92}$$

3) 反之，若$\mathrm{Null}(J_\infty) \neq \{0\}$，其中 $J_\infty := \sum_{j\ge 1} H_j^\top R_j^{-1} H_j$ （在 *PSD* 意义下收敛），则 $\Sigma_{X|Z^{(m)}}$ 不会收敛到0，并且对于任意固定的率 R，与直接基准之间仍存在非零差距。

证明。由于 $V \perp (X,\ Z^{(m)})$ 且 $M = f(Z^{(m)})$，我们有 $V \perp M$，因此 $D_e = \mathrm{tr}(W_e C_0) + \tilde{}\mathrm{tr}(WK_X)$，其中 $K_X = \Sigma_{X|M}$。

对于高斯 $Z^{(m)}$，写出最小均方误差分解 $X = S^{(m)} + N^{(m)}$，其中 $S^{(m)} = \mathbb{E}[X \mid Z^{(m)}]$ 且 $N_{(m)} \perp Z^{(m)}$。对于任意 $M = f(Z^{(m)})$，$\hat{}X = \mathbb{E}[X \mid M] = \mathbb{E}[S^{(m)} \mid M]$，因此由全协方差公式，

$$\Sigma_{X|M} = \Sigma_{X|Z^{(m)}} + \Sigma_{S^{(m)}|M}.$$

令 $K := \Sigma_{S^{(m)}|M}$。速率约束意味着 $R \ge I(S^{(m)};\ M)$，而引理 2 给出 $I(S^{(m)};\ M) \ge_{1\ 2} \log$ 行列式 $\Sigma_S^{(m)}$ 行列式 K，从而得到 (91)。代入 D_e 得到 (90)。

若 $\lambda_{\min}(J_m) \to \infty$，则由定理 5 可得 $\Sigma_{X|Z^{(m)}} \to 0$ 和 $\Sigma_S^{(m)} \to \Sigma_X$。在此极限下，定义 $\Delta_m(R)$ 的优化是连续的（可行集收敛且目标函数为线性），因此 $\Delta_m(R) \to \Delta_{\mathrm{dir}}(R)$，由此得到 (92)。若 $\mathrm{Null}(J_\infty) \neq \{0\}$，则 $\Sigma_{X|Z^{(m)}}$在未观测子空间上仍保留一个非零分量，因此与直接基准之间的差距不可能趋于零。

□

注 7。定理 *2* 讨论了编码器观测到语义代理 $\Theta = BX + V$ 的情形，因此 M 可以依赖于 V，且编码器的目标函数包含额外的 X–V交叉项。相比之下，（78）中的多模态模型仅依赖于 X（以及传感器噪声），因此

$V \perp M$ 问题转化为对 X 的间接高斯编码。因此，多模态性缩小了与直接-X 基准之间的差距，但无法恢复从未被观测到的独立语义随机性V。

B. 将计算视为信息预算与语义精度的缩放

本小节将架构与推理时计算同信息预算联系起来，该预算决定模型能够将其对潜在状态的后验细化到多高的精度。核心思想是，每一层（或推理步骤）从输入（或记忆）中提取的新信息都有限，随后才会将细化后的信念传递下去。

令 $X \sim \mathcal{N}(0,\ \Sigma_X)$ 为潜在状态，并令 Z 表示与 X 相关的可用数据。表示学习器通过（可能具有随机性的）更新规则生成一系列内部状态 $H^{(0)},\ H^{(1)},\ \ldots,\ H^{(L)}$

$$H^{(\ell+1)} \sim p_\ell(\cdot \mid H^{(\ell)}, Z), \tag{93}$$

对于 $\ell = 0,\ \ldots,\ L-1$，使模型能够在每一步重新访问 Z。

我们将步骤 ℓ 的有限计算量解释为从 Z 中提取的增量信息的上限:

$$I\big(Z; H^{(\ell+1)} \mid H^{(\ell)}\big) \le R_\ell, \tag{94}$$

对于 $\ell = 0,\ \ldots,\ L-1$。

由于 $H^{(\ell+1)}$ 是由 $(H^{(\ell)},\ Z)$ 生成的，因此具有马尔可夫性质 $X - (H^{(\ell)},\ Z) - H^{(\ell+1)}$，由数据处理可得

$$I\big(X; H^{(\ell+1)} \mid H^{(\ell)}\big) \le R_\ell, \tag{95}$$

对于 $\ell = 0, \ldots, L-1$。

将步骤 ℓ 后的后验误差协方差定义为

$$K_X^{(\ell)} := \Sigma_{X|H^{(\ell)}}. \tag{96}$$

我们取

$$K_X^{(0)} = \Sigma_X. \tag{97}$$

定理 6. 令 $X \sim \mathcal{N}(0, \Sigma_X)$，并令 $H^{(0)}, \ldots, H^{(L)}$ 为由 $(H^{(\ell)}, Z)$生成的任意表征序列。假设对每个 ℓ, $X - (H^{(\ell)}, Z) - H^{(\ell+1)}$，(94) 均成立。那么

$$\log\det K_X^{(L)} \geq \log\det \Sigma_X - 2\sum_{\ell=0}^{L-1} R_\ell. \tag{98}$$

证明。 根据链式法则，

$$I(X; H^{(L)}) \leq I(X; H^{(0)}) + \sum_{\ell=0}^{L-1} I\big(X; H^{(\ell+1)} \mid H^{(\ell)}\big).$$

根据 $X - (H^{(\ell)}, Z) - H^{(\ell+1)}$ 下的数据处理性质及增量预算 (94)，我们有

$$\begin{aligned} I\big(X; H^{(\ell+1)} \mid H^{(\ell)}\big) &\leq I\big(Z; H^{(\ell+1)} \mid H^{(\ell)}\big) \\ &\leq R_\ell. \end{aligned}$$

因此，

$$I(X; H^{(L)}) \leq I(X; H^{(0)}) + \sum_{\ell=0}^{L-1} R_\ell.$$

由于 X 是高斯分布，引理 2 推出

$$I(X; H^{(L)}) \geq \frac{1}{2}\log\frac{\det \Sigma_X}{\det K_X^{(L)}}.$$

合并后得到 (98)。 □

注 8。 约束 (94) 应被视为一种建模抽象：计算不会创造关于 X 的新信息，这些信息并未包含在 Z中；但有限的计算能力和架构可能会阻止模型提取并表示这些信息。允许每一步依赖于 $(H^{(\ell)}, Z)$，则体现了一个在经验上很重要的事实：现代架构会反复重新关注输入和上下文。

定义总计算预算

$$R_{\text{tot}} := \sum_{\ell=0}^{L-1} R_\ell. \tag{99}$$

定理 6 意味着，在直接战略率失真问题中，R_{tot} 起到信息率预算的作用，因为任何最终后验 $K_X^{(L)}$ 都必须满足 $\frac{1}{2}\log\frac{\det \Sigma_X}{\det K_X^{(L)}} \leq R_{\text{tot}}$。

据此，定义

$$D_{\text{opt}}(R_{\text{tot}}) = \min_{0 \prec K \preceq \Sigma_X} \left\{ \operatorname{tr}(\tilde{W} K) : \frac{1}{2}\log\frac{\det \Sigma_X}{\det K} \leq R_{\text{tot}} \right\}. \tag{100}$$

定理 7. 令 R_{tot} 由(99)定义。那么，任何满足(95) 的机制都必须遵守

$$\operatorname{tr}(\tilde{W} K_X^{(L)}) \geq D_{\mathrm{opt}}(R_{\mathrm{tot}}), \tag{101}$$

并且，这一下界可由高斯测试信道达到。

此外，若优化器满足内部条件 $K^\star \prec \Sigma_X$，则唯一最优解为

$$K^\star(R_{\mathrm{tot}}) = \nu^\star(R_{\mathrm{tot}})\, \tilde{W}^{-1}, \tag{102}$$

with

$$\nu^\star(R_{\mathrm{tot}}) = \exp\!\left(\frac{1}{k} \big[\log\det \Sigma_X + \log\det \tilde{W} - 2R_{\mathrm{tot}} \big] \right), \tag{103}$$

且最优语义失真呈指数衰减：

$$D_{\mathrm{opt}}(R_{\mathrm{tot}}) = k\, (\det \Sigma_X\ \det \tilde{W})^{1/k}\ \exp\!\left(-2\frac{R_{\mathrm{tot}}}{k} \right). \tag{104}$$

证明。 可行性蕴含关系 $1\ 2\ \log$ 行列式 Σ_X 行列式 $K_X^{(L)} \leq R_{\mathrm{tot}}$ 可直接由定理 6 得出。由于 $D_{\mathrm{opt}}(R_{\mathrm{tot}})$ 是该可行集上 $\operatorname{tr}(WK)$ 的最小值，因此可得 (101)。

可实现性来自直接率失真证明中使用的同一高斯测试信道构造：对于任意可行的 K，均可构造一个高斯辅助变量 $U = X + Z$，使其诱导出 $\Sigma_{X|U} = K$ 和 $I(X;U) = \frac{1}{2}\log$ 行列式 Σ_X 行列式 $K \leq R_{\mathrm{tot}}$。

对于内部解，构造拉格朗日函数

$$\mathcal{L}(K, \nu) = \operatorname{tr}(\tilde{W} K) + \nu \left(\frac{1}{2} \log \frac{\det \Sigma_X}{\det K} - R_{\mathrm{tot}} \right),$$

其中有 $\nu \geq 0$。驻点条件给出 $\tilde{W} - \frac{\nu}{2} K^{-1} = 0$，因此 $K = \nu\, \tilde{W}^{-1}$ 在重新调整乘子之后。施加有效的信息约束可得到 (103)，代入 $\operatorname{tr}(WK^\star)$ 可得到 (104)。

□

注 9。 内部形式 (102)–(104) 要求 $W \succ 0$，且上界 $K \preceq \Sigma_X$ 处于非激活状态。一个方便的充要检验可以写成

$$A_X := \Sigma_X^{1/2} \tilde{W}\, \Sigma_X^{1/2}. \tag{105}$$

当 (102) 成立时，来自(103) 的乘子 $\nu^\star(R_{\mathrm{tot}})$ 必须满足

$$\nu^\star(R_{\mathrm{tot}}) < \lambda_{\min}(A_X), \tag{106}$$

这等价于 $K^\star(R_{\mathrm{tot}}) \prec \Sigma_X$。

如果 (106) 不成立，那么某些方向会在先验处饱和，优化器在 A_X 的特征基下呈现反向注水形式：存在 U，使得 $A_X = U\mathrm{diag}(\lambda_i)\,U^\top$

$$K^\star(R_{\mathrm{tot}}) = \Sigma_X^{1/2} U\,\mathrm{diag}(d_i^\star)\,U^\top \Sigma_X^{1/2}, \tag{107}$$

其中 $\lambda_i > 0$

$$d_i^\star = \min\left\{1,\ \frac{\nu^\star}{\lambda_i}\right\}, \tag{108}$$

且当 $\lambda_i \le 0$ 时， $d_i^\star = 1$ 。标量 $\nu^\star$ 的取值使速率约束处于激活状态:

$$\sum_{\lambda_i>0} \log\left(\frac{1}{d_i^\star}\right) = 2R_{\mathrm{tot}}. \tag{109}$$

在此区域内，指数标度（104）呈分段形式（每当某个模态达到饱和时，有效斜率就会改变）。

推论4。 若对所有 ℓ 都有 $R_\ell = R_0$ ，则在内部区域内，

$$D_{\mathrm{opt}}(L) = k(\det\Sigma_X\ \det\tilde{W})^{1/k}\exp\left(-2L\frac{R_0}{k}\right). \tag{110}$$

若推断过程产生中间状态 M_1、...、M_T，并满足增量预算 $I(X;\ M_t \mid M_{1:t-1}) \le R_{\mathrm{step}}$，则当 $R_{\mathrm{tot}} = TR_{\mathrm{step}}$ 时，同一指数律（104）仍然成立，只需将 L 替换为 T。

证明。 回顾总信息（计算）预算

$$R_{\mathrm{tot}} := \sum_{\ell=0}^{L-1} R_\ell. \tag{111}$$

~在定理 7 的假设下，且处于内部区域（即 $W \succ 0$ ，并且优化器满足 $K^\star(R_{\mathrm{tot}}) \prec \Sigma_X$）时，最优值遵循指数律

$$D_{\mathrm{opt}}(R_{\mathrm{tot}}) = k\,(\det\Sigma_X\ \det\tilde{W})^{1/k}\ \exp\left(-2\frac{R_{\mathrm{tot}}}{k}\right), \tag{112}$$

如 (104) 所述。若对所有 ℓ 均有 $R_\ell = R_0$ ，则由 (111) 得

$$R_{\mathrm{tot}} = \sum_{\ell=0}^{L-1} R_0 = LR_0. \tag{113}$$

将 (113) 代入 (112) 得到

$$D_{\mathrm{opt}}(L) = k(\det\Sigma_X\ \det\tilde{W})^{1/k}\exp\left(-2L\frac{R_0}{k}\right),$$

这就是 (110)。设一个推断过程生成中间状态 M_1，……， M_T，使得对于每一步

$$I(X; M_t \mid M_{1:t-1}) \le R_{\text{step}}. \tag{114}$$

根据互信息的链式法则及其非负性，

$$I(X; M_{1:T}) = \sum_{t=1}^{T} I(X; M_t \mid M_{1:t-1}) \le \sum_{t=1}^{T} R_{\text{step}} = TR_{\text{step}} =: R_{\text{tot}}. \tag{115}$$

因此，最终表示 $M_{1:T}$ 满足与定理 7 中相同的信息预算约束，总率为 $R_{\text{tot}} = TR_{\text{step}}$。将 $R_{\text{tot}} = TR_{\text{step}}$代入式（104），可得到相同的指数标度，只需将 L 替换为 T。

□

注 10。 上述抽象为现代人工智能系统提供了几种具体的解读。

- **深度作为累积率：** 如果每一层从上下文中提取的新增任务相关信息最多为 R_ℓ 比特，那么深度就会累加这些预算。
- **宽度和精度作为每层的率：** 一个粗略的代理指标是， R_ℓ 随可通过该层传播的自由度数量而变化：宽度、注意力带宽和数值精度。降低精度会减少 R_ℓ，从而将能源使用与后验熵联系起来。
- **思维链作为序贯率分配：** 如果每个推理步骤只能增加关于 X的有限信息，那么定理 6 和定理 7 预示着，不确定性会随着步骤数增加而成倍降低。
- **检索与工具使用作为率注入：**外部检索和额外模态通过提供超出原始 Z的新信息性输入，增加了有效的 R_{tot} 。

VI. 讨论与未来方向

本文针对语义失配、观测约束和信息速率限制下的策略性高斯率失真，构建了统一的信息论框架。除第三节和第四节中推导出的明确表征外，研究结果还揭示了对节能和数据高效人工智能更广泛的概念性启示，并为未来研究开辟了多个方向。

A. 后验几何作为统一的设计原则

本研究的一项核心概念性贡献，是指出后验协方差几何是资源约束下决定语义性能的基本要素。在所有编码器观测模型中，率或计算约束都会对后验误差协方差施加对数行列式界，而语义目标函数则要求加权精度。这便形成了熵减与语义对齐之间的几何权衡：在直接机制和远程机制中，该权衡体现为广义注水定律；在劝说机制中，则体现为后验设计。

B. 将计算视为信息率：对高效人工智能的启示

本研究的一个关键洞见是，现代学习系统中的计算和架构瓶颈构成了隐式的信息率约束。深度、宽度、注意力带宽、上下文长度和推理时计算都会限制跨层或跨时间步传播和精炼的互信息。从这一视角来看，神经缩放定律、思维链推理和迭代精炼可被视为分配额外信息率以降低后验熵的机制。

C. 多模态：应对语义维数灾难的良方

我们的差距分析表明，远程语义编码会遭受不可避免的几何平均惩罚，该惩罚随维数呈指数增长，除非语义代理在每个方向上都具有信息量。多模态观测由此成为一种有原则的解决方案。通过聚合互补模态，可恢复语义协方差得以增大，几何惩罚也随之消失，从而使性能接近直接基准，而无需按比例增加信息率或计算。

D. 策略通信与对齐考量

本文研究问题的策略性本质，凸显了对齐和目标不匹配在语义系统中的重要性。当编码器和解码器优化不同的目标函数时，编码器的最优策略不仅是忠实地压缩信息，还要以符合其语义优先级的方式塑造解码器的后验信念。在全信息机制下，这体现为速率约束下的高斯劝说。

E. 建模假设与局限性

本文的分析依赖若干简化假设，这些假设使得我们能够得到闭式刻画，同时也界定了结果的适用范围。

首先，我们将研究范围限定在高斯源以及具有二次目标函数的线性高斯语义模型上。这一选择使我们能够获得精确的后验表征（无需采用 Kamenica 和 Gentzkow [7] 的凹化方法），但现实世界中的语义变量可能是离散的，或具有重尾分布。将后验设计框架扩展到非高斯设定是一个重要方向。

其次，我们假设解码器通过最小均方误差估计做出最佳响应。虽然在二次损失和高斯假设下这是最优的，但实际系统可能会采用近似推断。研究战略性率失真解对次优解码器的鲁棒性，将进一步增强该方法的适用性。

第三，我们对计算即速率解释的采用抽象掉了硬件和算法效率的许多细节。纳入明确的能耗模型和硬件约束仍有待研究。

VII. 结论

本文提出了一个统一的战略语义压缩理论，涵盖率、观测和计算约束。通过将率失真理论与信息设计相结合，我们刻画了在目标不一致且资源有限的设定下语义推断的基本极限。由此得到的框架为直接编码、远程编码和全信息编码机制提供了显式解，量化了它们之间的性能差距，并解释了多模态观测和计算能力的提升如何缓解这些差距。

最后，本文的核心观点是，高效智能从根本上说是一个资源约束下的后验设计问题。通过明确后验几何，信息论可以在指导未来人工智能系统的开发中发挥核心作用，使其不仅能力更强，也能更高效地利用能源和数据，并更符合人类价值观。

参考文献

[1] C. E. Shannon，“具有保真度准则的离散信源编码定理”，载于*IRE* 全国会议记录，第4卷，1959年，第142–163页，第4部分。[2] T. Berger，《率失真理论：数据压缩的数学基础》。美国新泽西州恩格尔伍德克利夫斯：Prentice-Hall，1971年。

[3] J. K. Wolf 和 J. Ziv，“向有噪声接收方传输有噪声信息时的最小失真”，IEEE 信息论汇刊，第16卷，第4期，第406–411页，1970年。[4] H. S. Witsenhausen，“间接率失真问题”，IEEE 信息论汇刊，第26卷，第5期，第518–521页，1980年。[5] V. Crawford 和 J. Sobel，“策略性信息传输”，《计量经济学》：计量经济学会期刊，第1431–1451页，1982年。[6] J. Farrell 和 M. Rabin，“廉价言谈”，《经济展望杂志》，第103–118页，1996年。[7] E. Kamenica 和 M. Gentzkow，“贝叶斯劝说”，《美国经济评论》，第101卷，第6期，第2590–2615页，2011年。[8] M. Gentzkow 和 E. Kamenica，“贝叶斯劝说的罗斯柴尔德—斯蒂格利茨方法”，《美国经济评论》，第106卷，第5期，第597–601页，2016年5月，论文与会议录。[9] W. Tamura，“具有二次偏好的贝叶斯劝说”，2018年3月，可见 SSRN 1987877。 [在线]。网址：https://papers.ssrn.com/sol3/papers.cfm?abstract id=1987877[10] T. Ui，“线性二次高斯信息设计”，中央银行沟通研究项目，东京大学，中央银行沟通工作论文018，2020年3月。 [在线]。网址：https://repository.dl.itc.u-tokyo.ac.jp/record/ 54029/files/cb-wp018.pdf[11] E. Akyol、C. Langbort 和 T. Bas¸ar，“将策略性通信视为层级博弈的信息论方法”，IEEE 会议录，第105卷，第2期，第205–218页，2016年。[12] O. Candogan 和 K. Drakopoulos，“内容准确性的最优信号传递：参与度与错误信息”，《运筹学》，第68卷，第2期，第497–515页，2020年。[13] D. Bergemann 和 S. Morris，“信息设计：统一视角”，《经济文献杂志》，第57卷，第1期，第44–95页，2019年。[14] S. Dughmi 和 H. Xu，“算法贝叶斯劝说”，SIAM 计算杂志，第0期，第STOC16–68页，2019年。[15] R. Alonso 和 O. Cˆámara，“说服选民”，《美国经济评论》，第106卷，第11期，第3590–3605页，2016年。[16] B. T. Gould 和 P. N. Brown，“车对车通信的信息设计”，《交通研究C辑：新兴技术》，第150卷，文章编号104084，2023年。[17] A. S. Vora 和 A. A. Kulkarni，“通过噪声信道从策略性发送方提取信息”，载于2020年第59届IEEE决策与控制会议（CDC）。IEEE，2020年，第354–359页。[18] E. Kazıklı、S. Sarıtas¸、S. Gezici、T. Linder 和 S. Y¨uksel，“对数凹分布的信号博弈：区间数量与均衡性质”，IEEE 信息论汇刊，第68卷，第3期，第1731–1757页，2021年。[19] M. Le Treust 和 T. Tomala，“有限通信容量下的劝说”，《经济理论杂志》，第184卷，文章编号104940，2019年。[20] Y. C. Aybas¸ 和 E. T¨urkel，“粗粒度通信下的劝说”，arXiv预印本 arXiv:1910.13547，2019年。[21] M. Sayin、E. Akyol 和 T. Bas¸ar，“非合作通信与控制系统中的层级多阶段高斯信号博弈”，《自动化》，第107卷，第9–20页，2019年。[22] M. O. Sayin 和 T. Bas¸ar，“具有状态依赖二次成本度量的贝叶斯劝说”，IEEE 自动控制汇刊，第67卷，第3期，第1241–1252页，2021年。[23] F. Farhadi 和 D. Teneketzis，“动态信息设计：关于最优序贯信息披露的一个简单问题”，《动态博弈与应用》，第12卷，第2期，第443–484页，2022年。[24] T. Bas¸ar，“通过软政策诱导期望行为”，《国际博弈论评论》，文章编号2440002，2024年。[25] A. Anand 和 E. Akyol，“信道优化的策略量化”，IEEE 通信领域精选期刊，2025年。

[26] S. Dughmi、D. Kempe 和 R. Qiang，“有限通信下的说服”，载于《2016年ACM经济学与计算会议论文集》，2016年，第663–680页。[27] E. Akyol 和 A. Anand，“策略量化”，载于《2023年IEEE国际信息论研讨会（ISIT）》。IEEE，2023年，第543–548页。[28] J. Liu、W. Zhang 和 H. V. Poor，“用于表征语义信息的率失真框架”，载于《2021年IEEE国际信息论研讨会（ISIT）》。IEEE，2021年，第2894–2899页。[29] H. Xie、Z. Qin、G. Y. Li 和 B.-H. Juang，“深度学习赋能的语义通信系统”，《IEEE信号处理汇刊》，第69卷，第2663–2675页，2021年。[30] C. Dong、H. Liang、X. Xu、S. Han、B. Wang 和 P. Zhang，“基于语义切片模型传播的语义通信系统”，《IEEE 通信领域精选期刊》，第41卷，第1期，第202–213页，2023年1月。[31] Y. Shao、Q. Cao 和 D. G¨und¨uz，“语义通信理论”，《IEEE移动计算汇刊》，第23卷，第12期，第12 211–12228页，2024年。[32] D. G¨und¨uz、Z. Qin、I. E. Aguerri、H. S. Dhillon、Z. Yang、A. Yener、K. K. Wong 和 C.-B. Chae，“超越比特传输：上下文、语义与面向任务的通信”，《IEEE 通信领域精选期刊》，第41卷，第1期，第5–41页，2023年。[33] Y. Xiao、X. Zhang、Y. Li、G. Shi 和 T. Bas¸ar，“策略性语义通信的率失真理论”，载于《2022年IEEE信息论研讨会（ITW）》。IEEE，2022年，第279–284页。[34] T. Baltruˇsaitis、C. Ahuja 和 L.-P. Morency，“多模态机器学习：综述与分类”，《IEEE模式分析与机器智能汇刊》，第41卷，第2期，第423–443页，2019年。[35] J. Kaplan、S. McCandlish、T. Henighan、T. B. Brown、B. Chess、R. Child、S. Gray、A. Radford、J. Wu 和 D. Amodei，“神经语言模型的缩放定律”，arXiv预印本 arXiv:2001.08361，2020年。 [在线]。网址：https://arxiv.org/abs/2001.08361[36] J. Hoffmann、S. Borgeaud、A. Mensch、E. Buchatskaya、T. Cai、E. Rutherford、D. de Las Casas、L. A. Hendricks、J. Welbl、A. Clark、T. Hennigan、E. Noland、K. Millican、G. van den Driessche、B. Damoc、A. Guy、S. Osindero、K. Simonyan、E. Elsen、J. W. Rae、O. Vinyals 和 L. Sifre，“训练计算最优大型语言模型”，arXiv预印本 arXiv:2203.15556，2022年。 [在线]。网址：https://arxiv.org/abs/2203.15556[37] C. Tian 和 J. Chen，“具有互信息和失真约束、带解码器侧信息的远程向量高斯源编码”，《IEEE 信息论汇刊》，第55卷，第10期，第4676–4680页，2009年。[38] S. Rini、A. Kipnis、R. Song 和 A. J. Goldsmith，“高斯源的压缩估计编码方案”，《IEEE无线通信汇刊》，第18卷，第9期，第4344–4356页，2019年。[39] M. Bocquet、K. S. Gurumoorthy、A. Apte、A. Carrassi、C. Grudzien 和 C. K. Jones，“退化卡尔曼滤波误差协方差及其向不稳定子空间的收敛”，《SIAM不确定性量化期刊》，第5卷，第1期，第304–333页，2017年。[40] K. Plarre 和 F. Bullo，“关于具有间歇观测的可检测系统的卡尔曼滤波”，《IEEE 自动控制汇刊》，第54卷，第2期，第386–390页，2009年。[41] B. Sinopoli、L. Schenato、M. Franceschetti、K. Poolla、M. I. Jordan 和 S. S. Sastry，“间歇观测下的卡尔曼滤波”，《IEEE 自动控制汇刊》，第49卷，第9期，第1453–1464页，2004年。[42] G. Carere 和 H. C. Lie，“希尔伯特空间上线性高斯逆问题的最优低秩近似，第一部分：后验协方差近似”，arXiv预印本 arXiv:2411.01112，2024年。