

语用信息的数学理论

牛凯¹, 张平²

(1. 北京邮电大学泛网无线通信教育部重点实验室, 北京 100876; 2. 北京邮电大学网络与交换技术全国重点实验室, 北京 100876)

摘要: 本文提出语用信息的数学理论, 将通信、控制与决策纳入同一个框架。核心是“同效映射”(isoteleia mapping), 揭示了语用信息的本质——同效性: 不同语义路径若导向相同最优动作, 则在语用上等价。该映射诱导语法、语义、语用三层信息层级, 每层抽象都丢弃与任务无关的细节。在此基础上, 本文建立语用层面的信息论度量, 包括语用熵、上/下语用互信息、语用信道容量与语用率失真函数, 并证明 3 个编码定理——语用无损信源编码、语用信道编码与语用率失真编码。与经典和语义信息论相比, 该框架的关键在于以“信息对行动的有效性”为核心度量, 而非符号或语义保真度。

在信息度量之外, 本文引入语用信息价值 (VoI) 与语用信息成本 (CoI), 分别作为率失真函数与信道容量的决策论对偶。VoI 衡量信息带来的期望效用增益, CoI 衡量给定语用速率下所需的最小资源消耗。二者构成 Legendre-Fenchel 型对偶, 并通过拉格朗日框架实现跨层优化。由此得到语用效率上界, 即行为容量, 刻画资源受限智能系统在给定资源价格下的最大净效用。对于连续消息, 本文给出高斯信源与高斯信道的闭式表达, 包括语用容量、语用率失真及相应 VoI 与 CoI; 对于双边与单边抽象, 分别给出感知路径与表达路径的对偶关系。在动态场景中, 本文将框架扩展到部分可观测马尔可夫决策过程, 并推导信念状态下的贝尔曼方程。

语用信息论框架为任务导向通信、网络化控制、自主系统与具身智能提供统一数学语言, 将关注点从符号保真度转向信息对行动的有效性。具体而言, 语用熵刻画决策不确定性, 语用容量刻画可靠传递最优动作的最大速率, 语用率失真刻画给定任务失真下所需的最小速率, 而 VoI 与 CoI 则分别从收益与代价两侧给出资源分配的决策论依据。这些度量与经典和语义信息论保持层次兼容: 当同效映射退化为恒等映射时, 语用信息论退化为语义信息论; 当同义映射也退化为恒等映射时, 进一步退化为经典信息论。因此, 该框架在保持向后兼容的同时, 将信息论适用范围从符号传输扩展到目标导向的行为决策。

本文的基础贡献在于, 将语用信息从哲学概念推进为可定义、可计算、可证明的数学对象, 并在信息论中首次建立语法-语义-语用三层统一的严格层次。它通过同效映射与同义映射的复合, 把信息论极限、决策论目标与物理资源约束连接起来, 为任务导向通信、语义通信、6G、大模型与 AI Agent、具身智能、自动驾驶等通信与人工智能融合方向, 提供了以任务效用为中心的设计语言, 推动信息传输从“比特保真”走向“传递对行动真正有效的信息”, 具有重要的理论指导意义。

关键词: 语用信息论; 同效映射; 具象映射; 语用熵; 语用互信息; 语用信道容量; 语用率失真; 信息价值; 信息成本; 语用拉格朗日量; 行为容量

中图分类号: TP391

文献标志码: A

doi: 10.11959/j.issn.1000-436x.TXXB260559

收稿日期: 2026-09-01; **修回日期:** 2026-09-15

通信作者: 牛凯, niukai@bupt.edu.cn

基金项目: 国家重点研发计划项目 (No.2025YFF0514400, No.2025YFF0514404); 教育部基础学科和交叉学科突破计划项目 (No.JYB2025XDXM119); 国家自然科学基金重大项目 (No.62293481, No.62471054)

Foundation Items: The National Key Research and Development Program of China (No.2025YFF0514400, No.2025YFF0514404), The Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No.JYB2025XDXM119), The National Natural Science Foundation of China (No.62293481, No.62471054)

A mathematical theory of pragmatic information

Niu Kai¹, Zhang Ping²

1. Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications,
Beijing 100876, China

2. State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications,
Beijing 100876, China

Abstract: We propose a mathematical theory of pragmatic information that unifies communication, control, and decision-making within a coherent framework. At its core is the isoteleia mapping, which formalizes equifinality: distinct semantic paths that lead to the same optimal action are regarded as pragmatically equivalent. This mapping induces a three-tier hierarchy of syntactic, semantic, and pragmatic information, in which each successive abstraction discards distinctions that are irrelevant to the task. Building on this hierarchy, we develop a set of information-theoretic measures at the pragmatic level, including pragmatic entropy, up/down pragmatic mutual information, channel capacity, and rate-distortion. In addition, we prove three coding theorems—lossless source coding, channel coding, and rate-distortion—which generalize Shannon’s classical results. We then introduce pragmatic value of information (VoI) and pragmatic cost of information (CoI) as decision-theoretic duals to rate-distortion and capacity, respectively, and formulate a Lagrangian dual framework for cross-layer optimization. The resulting pragmatic efficiency bound $\mathcal{E}_p(\lambda) = \sup_R [\Phi_p(R) - \lambda \text{CoI}_p(R)]$ characterizes the maximum net utility attainable by a resource-constrained intelligent system under a given resource price, thereby providing a limit on purposeful behavior—a *behavioral capacity*. This limit can be viewed as a goal-directed analogue of Shannon’s symbol-level capacity under the stated utility and resource assumptions. Extensions to continuous messages yield closed-form expressions for Gaussian channels and sources, while dynamic settings are addressed through a Bellman equation for sequential decision-making. Finally, this framework provides a foundation for task-oriented communication, networked control, autonomous systems, and embodied AI, shifting the focus from symbol fidelity to the effectiveness of information in guiding actions, and offering a mathematical language for the design of next-generation intelligent systems.

Key words: pragmatic information theory, isoteleia mapping, reification mapping, pragmatic entropy, pragmatic mutual information, pragmatic channel capacity, pragmatic rate-distortion, value of information, cost of information, pragmatic lagrangian, behavioral capacity

I. Introduction

A. Background and Motivation

Classic information theory (CIT), established by C. E. Shannon in 1948^[1, 2], was a monumental achievement that laid the mathematical foundations for modern communication. By quantifying information through entropy, mutual information, channel capacity, and rate-distortion functions, CIT provided fundamental limits on data compression and reliable transmission over noisy channels. These theoretical bounds have guided the development of practical coding schemes over the past seven decades, with techniques such as Huffman coding, arithmetic coding, turbo codes, LDPC codes, and polar codes approaching the Shannon limits^[1, 2]. In parallel, the emergence of cybernetics and control the-

ory established complementary foundations for understanding and designing dynamical systems. Wiener’s cybernetics^[3] first unified the concepts of communication and control, while Tsien’s engineering cybernetics^[4] and Bellman’s dynamic programming^[5] provided powerful tools for optimal decision-making under uncertainty. Kalman’s filtering theory^[6] further bridged estimation and control, enabling real-time state estimation from noisy observations. Decision theory, particularly through the work of Stratonovich on the value of information (VoI)^[7, 8], addressed the fundamental question of how much utility can be gained from acquiring information before making a decision. Despite their shared intellectual roots, these three pillars—communication theory, control theory, and decision theory—have largely developed along independent tra-

jectories, with limited cross-fertilization among them.

Weaver ^[2, 9] first articulated that communication involves not only the accurate transmission of symbols (Level A) but also the conveyance of meaning (Level B) and the effectiveness of conduct (Level C). This layered view has motivated numerous efforts to formalize semantic information. Early attempts to formalize semantic information include Carnap and Bar-Hillel's logical approach based on propositional logic ^[10], Floridi's theory of strongly semantic information ^[11], and fuzzy entropy formulations by De Luca and Termini ^[12, 13] and Wu ^[14]. Zhong ^[15, 16] and Lu ^[17, 18] independently generalized Shannon's information theory to the semantic and pragmatic levels. Bao et al. ^[19] extended the logical framework to derive semantic source and channel coding theorems. More recently, Liu et al. ^[20] and others ^[21–23] investigated rate-distortion frameworks for semantic information. However, a systematic information-theoretic treatment of semantic communication remained elusive until the work of Niu and Zhang ^[24–26], who established a rigorous mathematical theory of semantic communication centered on the concept of synonymy, the principle that the same meaning can be expressed by many different syntactic realizations. Through the synonymous mapping, they introduced semantic entropy, up/down semantic mutual information, semantic channel capacity, and semantic rate-distortion functions, proving corresponding coding theorems that demonstrate performance gains over classical counterparts.

Despite these pioneering efforts, a systematic information-theoretic treatment of pragmatic information has remained largely absent. Early philosophical inquiries ^[27] and historical surveys ^[28] provided valuable conceptual foundations but stopped short of developing a rigorous mathematical framework. The quantitative formulation by Weinberger ^[29, 30], while introducing a formal definition of pragmatic information as the change in an agent's action distribution, was primarily developed within the context of biological evolution and did not establish fundamental coding theorems, rate-distortion limits, or channel capacity characterizations. More critically, none of these prior works

provided a unified framework that integrates pragmatic information with syntactic and semantic layers. Consequently, a comprehensive mathematical theory capable of quantifying, compressing, and reliably transmitting pragmatic information in engineered communication systems has been lacking until now.

The recent surge of interest in semantic communication, driven by advances in deep learning and the growing demand for intelligent communications, has further highlighted the limitations of the classical syntactic paradigm. Deep learning-based semantic communication systems have demonstrated significant performance improvements in tasks such as image transmission, speech communication, and text understanding ^[31–34]. These systems extract and transmit semantic features relevant to the task, moving beyond bit-level fidelity to meaning-level effectiveness. The convergence of communication and artificial intelligence has opened new possibilities for task-oriented communication systems that prioritize the end goal over exact symbol reconstruction. Zhang et al. ^[35] and Niu et al. ^[36] have articulated the vision of wisdom-evolutionary and primitive-concise communication networks for 6G, where semantic information plays a central role.

Equally important is the convergence of communication and control, extensively studied in networked control systems (NCS) and cyber-physical systems (CPS), where joint design of communication and control policies is essential for stability and performance under bandwidth constraints. Foundational results on this co-design have been established by Xiao et al. ^[37], with comprehensive surveys provided by Park et al. ^[38], Wang et al. ^[39], Zhao et al. ^[40], Negi and Chakraborty ^[41], and Lu and Guo ^[42]. Tishby and Polani ^[43] incorporated information-theoretic costs into the Bellman equation from a perception-action perspective, though their focus remained on action complexity rather than physical communication resources. Critically, these works treat information at the syntactic level, emphasizing signal fidelity over semantic or pragmatic content.

At a deeper level, modern information-driven systems, whether they are large language models interact-

ing with users, autonomous vehicles navigating complex environments, embodied AI agents performing physical tasks, or human learners acquiring knowledge, share a common operational structure that can be abstracted as a perception-communication-decision-execution (PCDE) closed loop. In such systems, information is not merely transmitted for its own sake; it is used to guide decisions and actions that ultimately determine system performance. The pragmatic dimension of information, its effect on conduct, is therefore paramount. The following examples illustrate the ubiquity of the PCDE loop and the central role of pragmatic information across diverse domains.

Large Language Models and AI Agents. In a typical interaction with a large language model (LLM) or AI agent, the user submits a query (perception), the model processes the query through its internal reasoning (communication), generates a response or takes an action (decision), and the user applies the response to their task (execution). The pragmatic value of the response lies not in its syntactic correctness but in its effectiveness in helping the user achieve their goal—whether solving a problem, generating code, or making a decision. The same response may be pragmatically equivalent or entirely different depending on the user’s context and intent. Figure 1 illustrates this PCDE loop, where pragmatic information bridges the gap between meaning and effective action.

Autonomous Driving. An autonomous vehicle operates within a complex PCDE loop: sensors (cameras, LiDAR, radar) perceive the surrounding environment (perception), the vehicle communicates with other vehicles and infrastructure via V2X (communication), the planning module decides on acceleration, braking, and steering (decision), and the actuators execute the planned actions (execution). The effectiveness of the system is measured by safety, efficiency, and comfort—all pragmatic objectives that depend on the correct interpretation of sensory data and the appropriate selection of actions. The same set of sensor readings may lead to different actions depending on the context (e.g., yielding to a pedestrian vs. proceeding through an intersection), highlighting the need for pragmatic reasoning.



Fig. 1: The perception-communication decision execution (PCDE) loop in large language model (LLM) and AI agent interactions. The user query (perception) is processed by the LLM (communication/understanding), which generates a response (decision) that drives the user’s subsequent action (execution). Pragmatic information resides in the effectiveness of the response in achieving the user’s intended goal.

Figure 2 depicts this loop, where pragmatic information—the optimal action class—is the key to safe and efficient navigation.

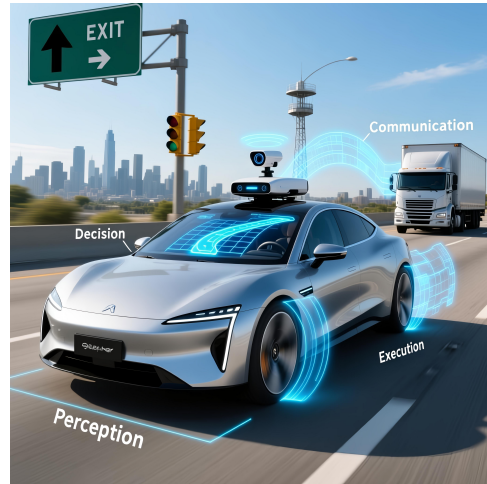


Fig. 2: The perception-communication-decision-execution loop in autonomous driving. Sensors perceive the environment (perception), the vehicle communicates with other vehicles and infrastructure (communication), the planning module decides on a trajectory (decision), and the actuators execute the maneuver (execution). Pragmatic information—the correct action to take—is the ultimate objective that transcends individual sensor measurements.

Embodied AI and Robotics. Embodied AI agents and robots operate in physical environments, where perception, communication, decision, and execution are tightly coupled. Sensors perceive the environment (perception), internal representations communicate state

and intent across modules (communication), the control policy decides on the next action (decision), and motors and actuators execute the motion (execution). The success of the system depends on the pragmatic effectiveness of the actions—whether the robot grasps the object, navigates to the target, or completes the task. The same sensory input may lead to different actions depending on the task goal, illustrating the context-dependent nature of pragmatic information. Figure 3 shows this loop, where pragmatic information is the bridge between perception and effective physical action.

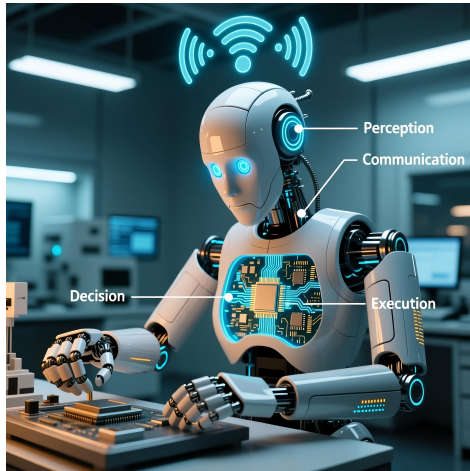


Fig. 3: The perception-communication-decision-execution loop in embodied AI and robotics. Sensors perceive the environment (perception), internal representations communicate state information (communication), the policy selects an action (decision), and the actuators interact with the physical world (execution). Pragmatic information is embodied in the action's effectiveness in achieving the task objective.

Human Learning and Education. Human learning follows a PCDE loop where instructional content is perceived (perception), communicated through explanation, demonstration, and feedback (communication), the learner decides on a learning strategy (decision), and applies it through practice and application (execution). The effectiveness of instruction depends on its pragmatic value—whether it enables the learner to perform better in real-world tasks. The same instructional content may have high pragmatic value for one learner but low value for another, depending on prior knowledge and learning context. Figure 4 illustrates this loop, where pragmatic information is the knowledge that translates into improved performance.

These four examples, spanning AI, autonomous



Fig. 4: The perception-communication-decision-execution loop in human learning. Instructional content is perceived (perception), communicated through explanation and feedback (communication), the learner decides on a learning strategy (decision), and applies it to practice (execution). Pragmatic information is the knowledge that effectively improves performance, beyond mere factual recall.

systems, robotics, and human cognition, share a common underlying structure: information flows through a perception-communication-decision-execution(PCDE) loop, and the ultimate measure of success is the effectiveness of the resulting actions. In each domain, pragmatic information, the information that makes a difference to decisions and actions, is the central concern. The value of information (VoI) quantifies the utility gain from information, while the cost of information (CoI) captures the resources needed to acquire and transmit it. Despite the ubiquity of this structure across diverse domains, no unified mathematical framework exists for describing, analyzing, and optimizing pragmatic information systems. Classical information theory emphasizes symbol fidelity; control theory emphasizes system stability; and decision theory emphasizes utility maximization. None provides a comprehensive account of the interplay among perception, communication, decision, and execution. A unified theory is therefore needed to bridge these disciplines and to provide a rigorous foundation for the design of modern information-driven systems.

B. Contributions and Organization

This paper presents a mathematical theory of pragmatic information that unifies communication, control, and decision-making within a single coherent frame-

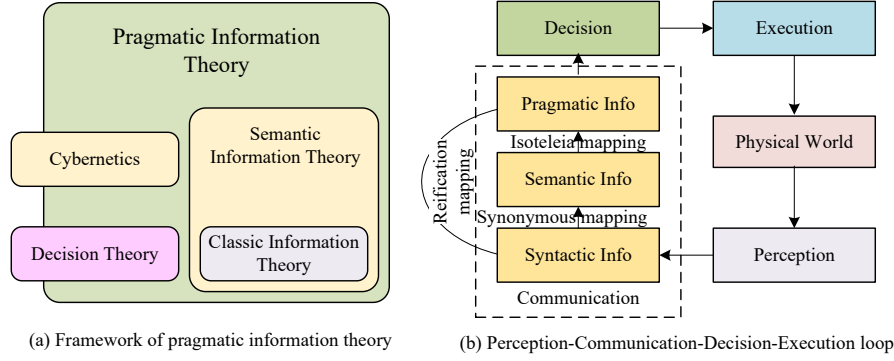


Fig. 5: Three-tier framework of pragmatic information theory and its integration with communication, control, and decision.

work. The proposed theory extends classical information theory by incorporating the pragmatic dimension of information, its effect on conduct, while subsuming semantic information theory and classical information theory as special cases. By introducing the isoteleia mapping, which formalizes the principle of equifinality (distinct semantic paths converging to the same optimal action), we establish a three-tier hierarchy of syntactic, semantic, and pragmatic information that provides a complete characterization of information in goal-directed systems.

Figure 5 illustrates the three-tier hierarchy of information and its integration with the PCDE (perception-communication-decision-execution) closed loop. As shown in Fig. 5(a), this unified framework positions pragmatic information theory as a natural generalization that subsumes both classical and semantic information theories, while simultaneously integrating decision theory and cybernetics into a single coherent structure. The three theories are nested and compatible: pragmatic information theory contains semantic information theory as a special case when the isoteleia mapping is trivial, and both contain classical information theory when the synonymous mapping is also trivial. This hierarchical compatibility ensures backward compatibility with existing theories while extending their reach into decision-making and control. Fig. 5(b) depicts the perception-communication-decision-execution (PCDE) loop, where syntactic information, at the bottom, is the domain of classical information theory and concerns the faithful transmission of symbols. Semantic information, the middle layer captured by the synonymous mapping, formalizes the principle that the

same meaning can be expressed by many different syntactic realizations. At the top lies pragmatic information, the focus of this paper, which concerns the effectiveness of information in guiding actions. The isoteleia mapping groups semantic classes that lead to the same optimal terminal action, embodying the principle of equifinality. Their composition, the reification mapping, directly bridges pragmatic purposes to syntactic signals, enabling the materialization of intent into physical communication. The PCDE loop connects the Physical World through perception, communication, decision, and execution, with pragmatic information serving as the essential link that closes the loop, answering not only “What signal should be sent?” (syntax) and “What does it mean?” (semantics), but fundamentally “What difference does this information make to the decisions and actions of the receiver?” (pragmatics).

Figure 6 illustrates the mathematical framework of pragmatic information theory.

The main contributions of this paper are summarized as follows:

1) Unified Pragmatic Information System Framework. We develop a comprehensive system model for pragmatic information, integrating Shannon’s communication paradigm with cybernetic control. The system is structured as a perception-communication-decision-execution (PCDE) closed loop, comprising ten core modules organized into two functional subsystems: the Cognitive-Regulation Subsystem (CRS), which handles reasoning, intent formulation, and value/cost evaluation, and the Symbolic-Conduction Subsystem (SCS), which handles encoding, transmission, and decoding. The

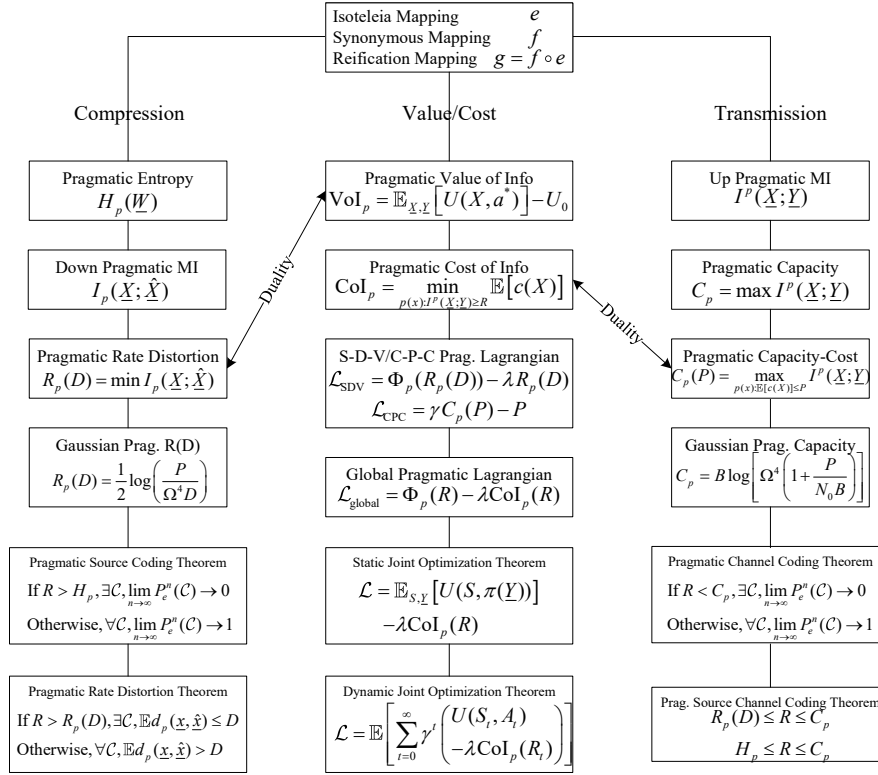


Fig. 6: The mathematical framework of pragmatic information theory.

reification mapping $g = f \circ e$ serves as the critical bridge between these subsystems, enabling the direct materialization of pragmatic intent into syntactic signals.

2) Information-Theoretic Measures for Pragmatic Information. We define a complete set of information-theoretic measures at the pragmatic level, including pragmatic entropy $H_p(\underline{W})$, up/down pragmatic mutual information $I^p(\underline{X}; \underline{Y})$ and $I_p(\underline{X}; \underline{Y})$, pragmatic channel capacity C_p , and pragmatic rate-distortion function $R_p(D)$. We establish the fundamental hierarchies that relate these measures to their syntactic and semantic counterparts: $H_p(\underline{W}) \leq H_s(\tilde{W}) \leq H(W)$, $C_p \geq C_s \geq C$, and $R_p(D) \leq R_s(D) \leq R(D)$. These hierarchies quantify the performance gains achievable by exploiting pragmatic abstraction.

3) Pragmatic Value of Information and Pragmatic Cost of Information. We introduce the pragmatic value of information (VoI) as the utility gain obtained from task-relevant information, and the pragmatic cost of information (CoI) as the minimum resource expenditure required to convey information at

a given pragmatic rate. We establish the dualities between VoI and the rate-distortion function, and between CoI and the channel capacity, and develop single-sided extensions for perceptual and expressive paths. These measures provide a decision-theoretic and economic interpretation of the information-theoretic limits.

4) Lagrangian Dual Framework and Pragmatic Shadow Price for Cross-Layer Optimization. We develop a unified Lagrangian dual framework that integrates the rate-distortion-value duality (down loop) and the capacity-cost duality (up loop) into a coherent cross-layer optimization principle. The global pragmatic Lagrangian $\mathcal{L}_{\text{global}}(R; \lambda) = \Phi_p(R) - \lambda \text{CoI}_p(R)$ expresses the net benefit as value minus cost, with the optimality condition $\lambda \text{CoI}'_p(R^*) = \Phi'_p(R^*)$ balancing marginal value against marginal cost. The resulting pragmatic efficiency functional $\mathcal{E}_p(\lambda) = \sup_R [\Phi_p(R) - \lambda \text{CoI}_p(R)]$ defines a fundamental behavioral limit for any resource-constrained intelligent system—a *behavioral capacity* that parallels Shannon's physical capacity, but replaces symbol fidelity with the effectiveness of information in guiding actions. Central to this frame-

work is the shadow price λ , which serves as a unified currency that converts physical resource consumption—such as transmit power, bandwidth, and latency—into utility-equivalent units. In static settings, λ determines the optimal operating point where marginal value equals marginal cost; in dynamic settings, it acts as a real-time scarcity signal that continuously adapts system behavior to varying channel conditions, power budgets, and task priorities. This framework provides a principled, practical, and dimensionally consistent approach for resource allocation and system design in task-oriented communication systems, bridging information-theoretic limits with decision-theoretic objectives without ad hoc heuristics.

5) Coding Theorems for Pragmatic Communication. We prove three fundamental coding theorems for pragmatic information systems: (i) the pragmatic lossless source coding theorem, establishing that the minimum rate for lossless pragmatic compression is the pragmatic entropy $H_p(\mathcal{W})$; (ii) the pragmatic channel coding theorem, establishing that the maximum reliable transmission rate is the pragmatic capacity C_p ; and (iii) the pragmatic rate-distortion coding theorem, establishing that the minimum rate for lossy pragmatic compression with distortion D is $R_p(D)$. These theorems generalize Shannon's classical coding theorems [1, 2] and the semantic coding theorems of Niu and Zhang [24, 25].

6) Continuous-Domain Pragmatic Information Measures. We extend the pragmatic information measures to continuous messages through the concept of isoteleic volumes Ω , which are the continuous analogues of pragmatic equivalence classes. We derive closed-form expressions for the pragmatic capacity of the Gaussian channel $C_p = \frac{1}{2} \log(\Omega^4(1 + P/\sigma^2))$ and the pragmatic rate-distortion function for Gaussian sources $R_p(D) = \frac{1}{2} \log(P/(\Omega^4 D))$, along with the corresponding CoI and VoI functions. We also derive the band-limited capacity $C_p = B \log(\Omega^4(1 + P/(N_0 B)))$.

7) Joint Optimization of Communication, Control, and Decision. We formulate the joint optimization of information, control, and decision-making in both static and dynamic settings. For the static case,

we derive the optimality conditions for the global pragmatic Lagrangian, showing that the optimal policy is the posterior-maximizing Bayes decision rule and the optimal rate satisfies marginal value equals marginal cost. For the dynamic case, we derive the Bellman equation for the dynamic pragmatic Lagrangian, providing a principled framework for co-designing communication and control in sequential decision-making systems.

8) Generalization and Unification. We demonstrate that the pragmatic information theory framework seamlessly subsumes classical communication systems (when semantic and pragmatic layers are trivialized), closed physical systems (with explicit utility functions), embodied and interactive systems (with multiple agents), and open social and cognitive systems (with implicit and evolving utility functions). This establishes the pragmatic information system as a primary reference model for analyzing and designing a system that involves goal-directed information processing under uncertainty.

The remainder of the paper is organized as follows. Section II presents the pragmatic information system model, including the system architecture, the three-tier hierarchy, the reification mapping, and the axiomatic foundations. Section III defines pragmatic entropy and its joint and conditional counterparts, establishing the fundamental hierarchies. Section IV introduces pragmatic relative entropy and up/down pragmatic mutual information. Section V defines pragmatic channel capacity and pragmatic rate-distortion functions, including single-sided extensions. Section VI establishes the pragmatic value of information and pragmatic cost of information, along with their Lagrangian dual framework. Section VII proves the pragmatic lossless source coding theorem. Section VIII proves the pragmatic channel coding theorem. Section IX proves the pragmatic rate-distortion coding theorem. Section X extends the pragmatic information measures to continuous messages, deriving the pragmatic capacity of Gaussian channels and the pragmatic rate-distortion function for Gaussian sources, along with the corresponding CoI and VoI functions. Section XI addresses the joint optimization of communication, control, and decision-

TABLE I: Principal symbols, meanings, and intuitive explanations

Symbol	Meaning and Intuitive Explanation
$\mathcal{X}, \mathcal{Y}, \mathcal{W}, \mathcal{V}$	Alphabets: collections of possible symbols
$X, Y, W, V / x, y, w, v$	Random variables / realizations at syntactic layer
$\tilde{X}, \tilde{Y}, \tilde{W}, \tilde{V} / \underline{X}, \underline{Y}, \underline{W}, \underline{V}$	Semantic layer (meanings) / Pragmatic layer (actions)
$P(\cdot)$ or $p(\cdot)$, $\mathbb{E}[\cdot]$, $\log(\cdot)$	Probability, expectation, log (bits/sebits/prabits)
$H(W), H_s(\tilde{W}), H_p(\underline{W})$	Entropy hierarchy: symbol $\rightarrow$ meaning $\rightarrow$ action uncertainty
$H(\cdot \cdot)$	Conditional entropy: residual uncertainty given context
$I(X; Y)$	Syntactic mutual information: symbol-level dependence
$I_s(\tilde{X}; \tilde{Y}), I^s(\tilde{X}; \tilde{Y})$	Down/Up semantic MI: meaning dependence (marginal/joint)
$I_p(\underline{X}; \underline{Y}), I^p(\underline{X}; \underline{Y})$	Down/Up pragmatic MI: action dependence (marginal/joint)
C, C_s, C_p	Capacity hierarchy: max reliable symbol/meaning/action rate
$R(D), R_s(D), R_p(D)$	Rate-distortion hierarchy: min symbol/meaning/action rate for distortion D
$f: \mathcal{W} \rightarrow 2^{\mathcal{W}}$	Synonymous mapping: one meaning $\rightarrow$ many signals
$e: \mathcal{W} \rightarrow 2^{\tilde{\mathcal{W}}}$	Isoteleia mapping: one action $\rightarrow$ many meanings
$g = f \circ e: \mathcal{W} \rightarrow 2^{\mathcal{W}}$	Reification mapping: one action $\rightarrow$ all realizing signals
$g^n, f^n, u^n, \mathcal{U}^n$	Sequential extensions: block-level mappings and sequences
Ω	Average isoteleic volume: avg size of action-equivalence set (continuous)
$\text{Vol}_p(R)$ or $\Phi_p(R)$	Pragmatic value of information: max utility gain from rate R
$\text{Col}_p(R)$	Pragmatic cost of information: min resource (power/bandwidth) for rate R
$\mathcal{L}_{\text{SDV}}, \mathcal{L}_{\text{CPC}}, \mathcal{L}_{\text{global}}$	Lagrangians: down-loop value–cost, up-loop rate–cost, global net benefit
λ, γ	Lagrange multipliers (shadow prices): convert resource cost to utility
$V(S)$	Value function in dynamic programming: optimal expected future reward

making in pragmatic information systems. Section XII concludes the paper.

II. Pragmatic Information System and Isoteleia Mapping

This section presents the theoretical foundations of the pragmatic information system. We establish a unified framework that integrates Shannon’s communication paradigm with cybernetic control, introducing the three-tier hierarchy of syntactic, semantic, and pragmatic information. Central to this framework is the Isoteleia Mapping, which formalizes the principle of equifinality, that is, distinct semantic paths converging to the same optimal action.

A. Notation Conventions

We adopt standard information-theoretic notation throughout. Calligraphic letters ($\mathcal{X}, \mathcal{Y}, \mathcal{W}, \mathcal{V}$) denote alphabets; uppercase letters (X, Y, W) are random variables, lowercase (x, y, w) their realizations. Semantic and pragmatic counterparts are marked with a tilde ($\tilde{W}$) and an underline ($\underline{W}$), respectively. Probabilities are written as $P(\cdot)$ or $p(\cdot)$, expectations as $\mathbb{E}[\cdot]$, and $\log(\cdot)$ denotes the logarithm to base 2. The main symbols are summarized in Table I; additional notation is introduced as needed in the text.

For sequences and conditional quantities, we adopt the usual conventions: $u^n = (u_1, \dots, u_n)$, $\mathcal{U}^n$ denotes the n -fold Cartesian product, and f^n is the sequential extension of a mapping f . Conditional entropy and mutual information are written as $H(\cdot|\cdot)$ and $I(\cdot; \cdot)$, with the convention $0 \log 0 = 0$. Syntactic quantities are measured in bits, semantic in sebits, and pragmatic in prabits.

B. Pragmatic Information System

The pragmatic information system is a perception-communication-decision-execution (PCDE) closed-loop architecture that integrates the classic Shannon communication paradigm with a cybernetic control framework. Its fundamental purpose is not merely to transmit symbols, but to ensure that the received information leads to terminal actions that maximize a predefined task utility. In this section, we present the system model in detail, describe its processing flow with state-space dynamics, formulate its axiomatic foundations, and discuss its broad applicability across diverse domains.

B.1 System Structure and Processing Flow

The system consists of ten core modules that operate in a sequential cascade, forming a complete perception – communication – decision – execution loop.

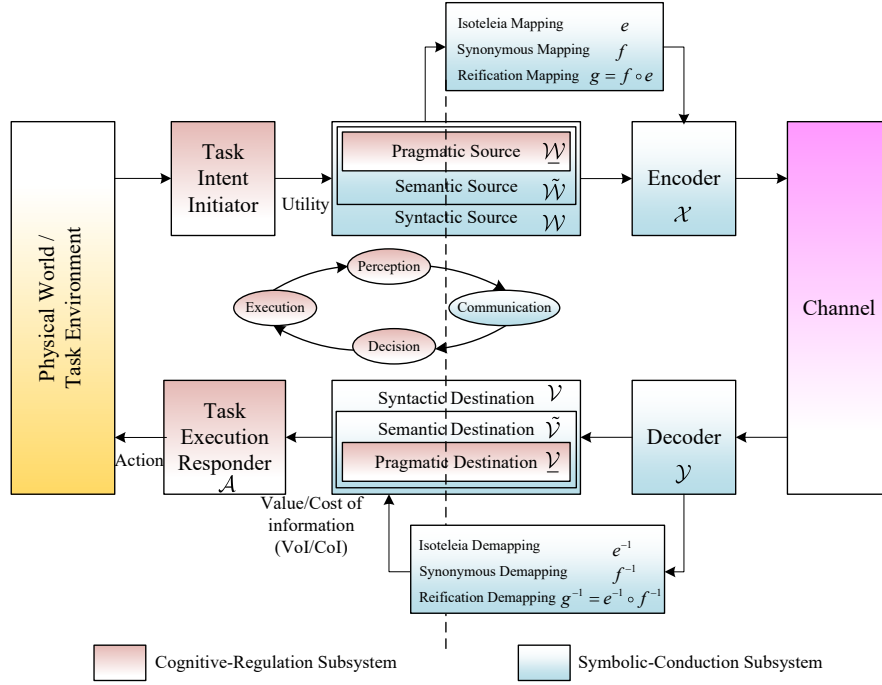


Fig. 7: The block diagram of Pragmatic Information System with PCDE loop, showing the interaction between the Physical World, the Cognitive-Regulation Subsystem (CRS), and the Symbolic-Conduction Subsystem (SCS).

These modules are: (1) Physical World / Task Environment, (2) Task Intent Initiator, (3) Source, (4) Reification Mapping, (5) Encoder, (6) Channel, (7) Decoder, (8) Demapping, (9) Destination, and (10) Task Execution Responder. Figure 7 provides a conceptual illustration of the overall architecture.

We now describe each module and the information flow that connects them.

Module 1: Physical World / Task Environment

The system operates within a Physical World characterized by a Plant with state $S_t \in \mathcal{S}$ at discrete time t , governed by:

$$S_{t+1} = F_m(S_t, A_t) + Z_t, \quad (1)$$

where $A_t \in \mathcal{A}$ is the action applied by the actuator, $F_m : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$ is the nominal transition dynamics, and Z_t represents exogenous disturbances. Observations are obtained through a sensor model:

$$Y_t = F_o(S_t) + V_t, \quad (2)$$

with $F_o : \mathcal{S} \rightarrow \mathcal{Y}$ the observation function and V_t measurement noise.

Module 2: Task Intent Initiator The Task Intent Initiator compares the observed state (or belief) $\hat{S}_t$

with a task goal G_t to generate a task intent:

$$T_t = \Phi(\hat{S}_t, G_t), \quad (3)$$

where Φ encodes the system's strategy. This module answers: “What should be done?”

Module 3: Source The Source is a three-tier hierarchical structure corresponding to the pragmatic, semantic, and syntactic levels [2, 9]. The Pragmatic Source $\underline{W}$ first maps the intent T_t to a pragmatic symbol $\underline{w} \in \underline{W}$ representing the optimal action class, answering: “What is the best terminal action?” This is then expanded by the Semantic Source $\tilde{W}$ into a semantic symbol $\tilde{w} \in \tilde{W}$ justifying the action, answering: “Why is this action justified?” Finally, the Syntactic Source $\mathcal{W}$ maps to a syntactic symbol $w \in \mathcal{W}$ for transmission, answering: “What signal should be sent?”

Module 4: Reification Mapping The Reification Mapping bridges the Pragmatic and Syntactic layers as the composition of the isoteleia and synonymous mappings:

$$g \triangleq f \circ e : \underline{W} \rightarrow \mathcal{P}(\mathcal{P}(\mathcal{W})), \quad (4)$$

where $e : \underline{W} \rightarrow \mathcal{P}(\tilde{W})$ maps a pragmatic symbol to all semantic justifications (equifinality), and $f : \tilde{W} \rightarrow$

$\mathcal{P}(\mathcal{W})$ maps a semantic symbol to all syntactic realizations (synonymy). Thus,

$$g(\underline{w}) = \bigcup_{\tilde{w} \in e(\underline{w})} f(\tilde{w}) \subset \mathcal{W}, \quad (5)$$

enumerating all syntactic realizations that fulfill a given pragmatic purpose.

Module 5: Encoder The Encoder maps a block of n syntactic symbols to channel input symbols:

$$\phi : \mathcal{W}^n \rightarrow \mathcal{X}^n, \quad (6)$$

performing a unified source-channel encoding operation designed to maximize end-to-end task utility.

Module 6: Channel The Channel is the physical medium characterized by the conditional distribution $P(Y_t|X_t)$, subject to resource constraints such as power, bandwidth, and delay.

Module 7: Decoder The Decoder recovers an estimate of the transmitted syntactic sequence:

$$\psi : \mathcal{Y}^n \rightarrow \mathcal{W}^n, \quad (7)$$

ensuring that the recovered symbols preserve semantic and pragmatic content for effective decision-making.

Module 8: Demapping The Demapping module reverses the Reification process: recovered syntactic symbols are first mapped to estimated semantic symbols via an inverse synonymous mapping, then to estimated pragmatic intentions via an inverse isoteleia mapping, recovering meaning and purpose from the received signals.

Module 9: Destination The Destination is the hierarchical counterpart to the Source. The Syntactic Destination $\mathcal{V}$ receives decoded syntactic symbols; the Semantic Destination $\tilde{\mathcal{V}}$ reconstructs semantic interpretations; and the Pragmatic Destination $\underline{\mathcal{V}}$ reconstructs the intended pragmatic purpose, answering: “*What was the intended purpose?*”

Module 10: Task Execution Responder The Task Execution Responder (Actuator) translates the recovered pragmatic intention $\hat{w}$ into a physical action:

$$A_t = \mu(\hat{w}_t), \quad (8)$$

where $\mu : \underline{\mathcal{V}} \rightarrow \mathcal{A}$ maps pragmatic intentions to actions, closing the perception-communication-decision-execution loop via Eq. (1).

The effectiveness of the entire process is evaluated by the Value of Information (VoI) and Cost of Information (CoI). The VoI quantifies the expected utility gain:

$$\text{VoI} = \mathbb{E}_{S,Y}[U(S_t, \mu(\hat{w}_t))] - \mathbb{E}_S[U(S_t, \mu(w_0^*))], \quad (9)$$

where w_0^* is the optimal action based on prior beliefs alone. The CoI quantifies the minimum resource expenditure required to convey pragmatic information at a given rate. Their joint consideration enables principled trade-off optimization between task performance and resource consumption, a central theme of the pragmatic Lagrangian framework developed in Section VI.

B.2 Functional Subsystems

The ten modules are naturally grouped into two functional subsystems: the **Cognitive-Regulation Subsystem (CRS)** and the **Symbolic-Conduction Subsystem (SCS)**.

The CRS handles high-level reasoning, intent formulation, and information value evaluation. It comprises the Task Intent Initiator (generating intent from state and goals), the Pragmatic Source (formalizing intent as pragmatic symbols), the Pragmatic Destination (reconstructing pragmatic intention), and the Task Execution Responder (translating intention into action). The isoteleia mapping e embodies equifinality by grouping semantic justifications that lead to the same optimal action. The CRS answers: “*What should be done, and is it worth doing?*”

The SCS handles encoding, transmission, and decoding. It comprises the Semantic Source (generating semantic symbols from pragmatic intent), the Syntactic Source (generating syntactic symbols from semantic meanings), the Encoder (converting symbols to physical signals), the Decoder (recovering symbols from received signals), and the Syntactic and Semantic Destinations (reconstructing meanings along the reverse path). The synonymous mapping f captures synonymy by mapping each meaning to multiple syntactic realizations. The SCS answers: “*How should it be expressed and transmitted?*”

The reification mapping $g = f \circ e$ serves as the critical bridge between these subsystems, allowing CRS intent to be directly materialized into syntactic signals

by the SCS.

B.3 Axiomatic Foundations

The operation of the pragmatic information system is governed by three fundamental axioms that ensure its mathematical tractability, general applicability, and consistency with both information theory and control theory.

Axiom 1: The Probability Axiom The system operates in a world where all observations, states, and intentions are subject to irreducible uncertainty. This uncertainty is quantified by probability measures. Formally, for any time t , the joint distribution over the system variables is:

$$P(S_t, W_t, Y_t) \in \Delta(\mathcal{S} \times \mathcal{W} \times \mathcal{Y}), \quad (10)$$

where Δ denotes the space of all probability distributions. The channel is characterized by the conditional distribution $P(Y_t|X_t)$. This axiom ensures that all information-theoretic quantities—entropy, mutual information, channel capacity, rate-distortion functions, and VoI/CoI—are well-defined and computable.

Axiom 2: The Decision Axiom All deliberate actions of the system are governed by the principle of maximizing expected utility. Formally, the **utility function** is defined as:

$$U : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}, \quad (11)$$

which assigns a real-valued scalar to each state-action pair (S_t, A_t) , representing the immediate reward or cost associated with taking action A_t in state S_t . For any decision node where the system has access to information I , it chooses an action a^* such that:

$$a^* = \arg \max_{a \in \mathcal{A}} \mathbb{E}_{P(S_t|I)}[U(S_t, a)]. \quad (12)$$

This axiom underpins the isoteleia mapping, as it defines the equivalence classes of semantic states that lead to the same optimal action (the same *telos*). It also underpins the Value of Information (VoI), which measures the utility gain obtained from additional information, and the Pragmatic Lagrangian Functional, which formalizes the trade-off between information value and communication cost.

Axiom 3: The Markov Axiom The system's dynamics satisfy the Markov property:

$$P(S_{t+1}|S_t, S_{t-1}, \dots, A_t, A_{t-1}, \dots) = P(S_{t+1}|S_t, A_t). \quad (13)$$

This property allows the system to be modeled as a Partially Observable Markov Decision Process (POMDP). Furthermore, the Markov property implies that the system can be decomposed across time: by removing the time indices, the dynamic control problem can be reduced to a static communication and encoding problem. In this static formulation, the channel coding and source coding theorems can be applied directly to characterize the fundamental limits of information transmission, treating each time step independently while preserving the overall structure of the pragmatic system. This temporal decomposition forms the basis for the coding theorems developed in subsequent sections.

B.4 Generality and Scope: A Unified Framework for Information-Driven Systems

The pragmatic information system provides a mathematical language for describing a class of systems that involve goal-directed information processing under uncertainty. The framework spans a continuous spectrum of systems, ranging from pure symbol transmission to tightly coupled physical control loops, and further to open-ended social and cognitive interactions.

We organize the discussion into four conceptual classes, ordered according to the *degree of closure* and the *nature of the utility function*. At one extreme lies the classical communication paradigm, where utility is defined solely by symbol fidelity. Moving along the spectrum, we encounter closed physical systems with well-defined utility functions; embodied and interactive systems with multiple agents; and finally, open social and cognitive systems where utility is implicit, dynamic, and subject to interpretation. This progression reflects increasing complexity in the relationship between information, meaning, and value.

Case I: Classical Communication Systems

When the Semantic and Pragmatic layers are trivialized—i.e., when utility depends solely on symbol fidelity (mean squared error, Hamming dis-

tance, or bit error rate)—the isoteleia and synonymous mappings reduce to identity mappings. The three-tier source collapses to a single syntactic source, and the system reverts to the familiar source-channel separation architecture. All classical results, including Shannon’s source coding, channel coding, and rate-distortion theorems, emerge as special cases. The framework thus serves not as a replacement but as a rigorous generalization of classical information theory, extending its reach from “how accurately can symbols be transmitted?” to “how effectively does the received meaning affect conduct in the desired way?”

Case II: Closed Physical Systems Here the Physical World is governed by well-defined physical laws, the utility function is explicitly specified and measurable (e.g., tracking error, energy consumption, or task success rate), and system behavior is constrained by strict physical and temporal limits. The framework naturally embeds the control-theoretic feedback loop via the Plant, Task Intent Initiator, and Actuator. Representative domains include networked control systems, robotics and industrial automation, autonomous vehicles and intelligent transportation, UAV swarm control, smart grids, and healthcare cyber-physical systems. In all these cases, VoI provides a rigorous metric for quantifying the trade-off between communication resource consumption and closed-loop performance, while the isoteleia mapping captures how diverse sensory inputs converge to the same control or navigation command, enabling robust decision-making under noise and delays.

Case III: Embodied and Interactive Systems In this class, the physical world remains central, but the system now involves multiple agents (human or artificial) with possibly distinct sensors, actuators, and world-views. Utility may be shared or conflicting, and communication becomes strategic or collaborative rather than a mere data pipe. Representative domains include embodied AI and autonomous agents, multi-agent systems and collaborative AI, human-machine interaction (HMI), and intelligent decision support systems. The framework provides a reference model for signaling games, negotiation protocols, and shared intention-

ality. The isoteleia mapping captures how different perceptual or contextual states lead to the same joint action or collaborative outcome, while VoI quantifies the benefit of sharing information among agents with potentially asymmetric information and conflicting objectives.

Case IV: Open Social and Cognitive Systems

Here the physical world recedes into the background, and the focus shifts to internal cognitive processes or emergent social dynamics. Utility is no longer a single measurable objective but may be implicit, dynamic, subjective, or even contradictory across agents. Representative domains include human learning and education, human communication and language, social and economic systems (where agents exchange signals to coordinate actions under uncertainty), biological and cognitive systems (aligning with predictive processing and active inference), and recommendation and content delivery systems. The framework may provide a useful starting point for modeling information exchange in open social and cognitive systems, but the present theory relies on explicit probabilistic and utility assumptions. Relaxing these assumptions is left for future work. The isoteleia mapping may offer a conceptual analogy for the human cognitive ability to generalize from diverse experiences to coherent principles, while VoI suggests a way to quantify the effectiveness of communication in reducing uncertainty and improving decisions under subjective and evolving objectives.

The Unity of the Framework Despite the apparent diversity across these four cases, they can be discussed within a common conceptual core: a state space $\mathcal{S}$, a utility function $U : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$, a three-tier information hierarchy (Syntactic, Semantic, Pragmatic), communication constraints, and a VoI metric guiding resource allocation and decision-making. The spectrum from Case I to Case IV is not a rigid partition but a continuum: classical communication systems can be enhanced with semantic and pragmatic layers; closed physical systems can involve human operators; and social systems can incorporate physical sensors and actuators. The pragmatic information system can serve as a reference model for analyzing a class of goal-directed systems, ranging from classical communi-

cation and closed-loop control to embodied and interactive systems. Extensions to social, economic, and cognitive domains are conceptually appealing but are not developed formally in this paper and remain open research directions.

C. Design Principles of Pragmatic Information Systems

Building upon the system architecture and the axiomatic foundations established in the preceding subsections, we now articulate three fundamental design principles that guide the analysis, synthesis, and operation of pragmatic information systems. These principles are not arbitrary engineering heuristics; they are direct consequences of the mathematical structure of the pragmatic framework and the nature of information as it flows from the physical world through semantic interpretation to pragmatic action.

Principle I: Pragmatic Imperceptibility and Isoteleia Pragmatic information, the ultimate purpose or terminal action that a communication system is intended to achieve, is not directly observable. It cannot be read from the received signal as one would read a bit string. Instead, pragmatic information must be inferred or evaluated indirectly through the lens of a utility function, which measures the consequences of actions taken in response to received information. Consequently, the fundamental characteristic of pragmatic information is *Isoteleia*: different semantic interpretations, and even different syntactic realizations, are pragmatically equivalent if they lead to the same optimal terminal action under the given utility. This is precisely the equivalence relation captured by the isoteleia mapping $e : \mathcal{W} \rightarrow \mathcal{P}(\tilde{\mathcal{W}})$, which partitions the semantic space into classes of meanings that share the same optimal action. The Value of Information (VoI) serves as the primary metric for quantifying pragmatic information, measuring the utility gain obtained from information. This principle implies that the system must be designed around the utility function, not around the fidelity of symbol reconstruction; performance metrics should be defined in terms of the quality of the resulting actions; and communication resources should be allocated according to the marginal Value of Information they pro-

vide.

Principle II: Semantic Imperceptibility and Synonymy Semantic information, the meaning or interpretation conveyed by a signal, is also not directly observable. It cannot be extracted from the signal by simple demodulation or decoding. Instead, semantic information must be inferred from the observed syntactic symbols through a process of pattern recognition, contextual reasoning, and knowledge-based interpretation. Consequently, the fundamental characteristic of semantic information is *synonymy*: the same meaning can be expressed by many different syntactic realizations. This equivalence relation is captured by the synonymous mapping $f : \tilde{\mathcal{W}} \rightarrow \mathcal{P}(\mathcal{W})$, which partitions the syntactic space into equivalence classes of symbols that share the same semantic interpretation. This principle implies that the system should exploit synonymy by allowing lossy compression at the syntactic level, as long as the semantic content is preserved; that the decoder must incorporate contextual information to resolve ambiguities; and that the encoder should extract and transmit semantic features relevant to the task, rather than attempting to reconstruct the original signal.

Principle III: Backward Compatibility and Unification The pragmatic information system must be backward compatible with both classical communication systems and classical control systems. This means that the pragmatic framework is not a replacement for existing theories but a strict generalization that contains them as special cases. When the pragmatic and semantic layers are trivialized, that is, when the utility function depends solely on symbol fidelity and the isoteleia and synonymous mappings become identity mappings, the pragmatic framework must collapse exactly into the classical Shannon communication paradigm. Similarly, when communication is ignored, the framework must collapse into classical control theory. This principle imposes that the three-tier hierarchy must be implemented in a modular fashion, allowing each layer to be independently analyzed, optimized, or bypassed; that the system must include explicit mechanisms to detect when the utility function reduces to syntactic fidelity and automatically switch to classical operation; and that the

encoder and decoder must support both pragmatic encoding and syntactic encoding to interface with legacy infrastructure.

Together, these three principles establish the philosophical and mathematical foundation for the design of pragmatic information systems. They ensure that the system is simultaneously *task-effective* (Principle I), *meaning-aware* (Principle II), and *infrastructure-compatible* (Principle III).

D. Pragmatic Information and Isoteleia Mapping

We now turn to the purely structural account of the mappings that bridge the three layers of information: Syntactic, Semantic, and Pragmatic.

D.1 The Nature of Pragmatic Information

In classical information theory, information is defined syntactically: it concerns the statistical properties of symbols and their transmission, independent of meaning or purpose. Semantic information, as formalized in the framework of Niu and Zhang [24, 25], extends this by considering the meaning conveyed by symbols: it concerns the relationship between symbols and their interpretations.

Pragmatic information, in contrast, concerns the *use* or *purpose* of information. It addresses the fundamental question: “*What difference does this information make to the decisions and actions of the receiver?*” Pragmatic information is inherently value-laden: it is defined not by what symbols mean, but by what they *achieve*. Two messages that convey different syntactic symbols and even different semantic meanings may carry the same pragmatic information if they lead the receiver to the same optimal terminal action. Conversely, two messages that convey identical semantic content may carry different pragmatic information if they are interpreted in different contexts or lead to different decisions.

The defining characteristic of pragmatic information is therefore **Isoteleia**, the principle of *equifinality*: distinct semantic paths converging to the same ultimate end, or *telos*. This principle captures the essence of goal-directed communication: the receiver’s objective is not to reconstruct the sender’s exact message, but to extract from it the guidance necessary to achieve a de-

sired outcome.

D.2 Foundations: Power Sets, Equivalence Relations, and Fibers

To rigorously formalize the relationships among syntactic, semantic, and pragmatic information, by using the concepts in [44], we introduce the mathematical structures of power sets, equivalence relations, and fibers. Let $\mathcal{W}$ be the **syntactic alphabet** (the set of all possible physical symbols or signals). A **semantic partition** of $\mathcal{W}$ is a collection of disjoint subsets whose union is $\mathcal{W}$, where each subset groups symbols that share the same meaning. Equivalently, a semantic partition is induced by an equivalence relation $\sim_s$ on $\mathcal{W}$: $w_1 \sim_s w_2$ iff they convey the same semantic meaning. The quotient set $\mathcal{W} / \sim_s$ is the **semantic alphabet**, denoted by $\tilde{\mathcal{W}}$.

The collection of all subsets of $\mathcal{W}$ is its **power set**, denoted $\mathcal{P}(\mathcal{W})$. The semantic partition Π_{sem} is a subset of $\mathcal{P}(\mathcal{W})$:

$$\Pi_{sem} \triangleq \mathcal{W} / \sim_s = \{\tilde{w}_1, \tilde{w}_2, \dots, \tilde{w}_K\} \subset \mathcal{P}(\mathcal{W}). \quad (14)$$

Each element $\tilde{w}_k \in \Pi_{sem}$ is a **semantic fiber** (or synonymous class): it is a subset of $\mathcal{W}$ containing all syntactic symbols that carry the same meaning.

Similarly, a **pragmatic partition** is a collection of disjoint subsets of the semantic alphabet, induced by an equivalence relation $\sim_p$ on $\tilde{\mathcal{W}}$: $\tilde{w}_i \sim_p \tilde{w}_j$ iff they lead to the same optimal terminal action. The quotient set $\tilde{\mathcal{W}} / \sim_p$ is the **pragmatic alphabet**, denoted by $\underline{\mathcal{W}}$. The pragmatic partition Π_{prag} is a subset of the power set of $\tilde{\mathcal{W}}$:

$$\begin{aligned} \Pi_{prag} &\triangleq \tilde{\mathcal{W}} / \sim_p \\ &= \{\underline{w}_1, \underline{w}_2, \dots, \underline{w}_L\} \subset \mathcal{P}(\tilde{\mathcal{W}}) \subset \mathcal{P}(\mathcal{P}(\mathcal{W})). \end{aligned} \quad (15)$$

Thus, each element $\underline{w}_l \in \Pi_{prag}$ is a set of semantic fibers, i.e., a subset of $\mathcal{P}(\mathcal{W})$, forming a **second-order power set**. The fibers at each level form a hierarchy: syntactic fibers are individual elements $w \in \mathcal{W}$; semantic fibers are subsets $\tilde{w} \subset \mathcal{W}$; and pragmatic fibers are subsets $\underline{w} \subset \tilde{\mathcal{W}} \subset \mathcal{P}(\mathcal{W})$, i.e., collections of semantic fibers. This hierarchy is the mathematical backbone of pragmatic information theory.

D.3 Formal Definitions of the Mappings

We now formally define the two fundamental mappings—the synonymous mapping and the isoteleia mapping—along with their composite, the reification mapping.

Definition 1 (Synonymous Mapping): Let $\mathcal{W}$ be the syntactic alphabet and $\tilde{\mathcal{W}}$ be the semantic alphabet. Let $\sim_s$ be an equivalence relation on $\mathcal{W}$ defined by meaning identity. The **Synonymous Mapping**

$$f : \tilde{\mathcal{W}} \rightarrow 2^{\mathcal{W}} \quad (16)$$

is defined for each semantic symbol $\tilde{w} \in \tilde{\mathcal{W}}$ as:

$$f(\tilde{w}) \triangleq \{w \in \mathcal{W} \mid w \sim_s \tilde{w}\} \subset \mathcal{W}. \quad (17)$$

Here, $2^{\mathcal{W}}$ denotes the power set of $\mathcal{W}$. The mapping is *one-to-many* in the forward direction: a single meaning corresponds to a set of syntactic realizations.

Definition 2 (Isoteleia Mapping): Let $\tilde{\mathcal{W}}$ be the semantic alphabet and let $U : \mathcal{W} \times \mathcal{A} \rightarrow \mathbb{R}$ be a utility function. For each semantic symbol $\tilde{w} \in \tilde{\mathcal{W}}$, define its induced optimal action as:

$$a^*(\tilde{w}) \triangleq \arg \max_{a \in \mathcal{A}} \mathbb{E}_{W \sim P(W|\tilde{w})} [U(W, a)]. \quad (18)$$

Define an equivalence relation $\sim_p$ on $\tilde{\mathcal{W}}$ by $\tilde{w}_i \sim_p \tilde{w}_j \iff a^*(\tilde{w}_i) = a^*(\tilde{w}_j)$. The **Isoteleia Mapping**

$$e : \underline{\mathcal{W}} \rightarrow 2^{\tilde{\mathcal{W}}} \quad (19)$$

is defined for each pragmatic class $\underline{w} \in \underline{\mathcal{W}} \triangleq \tilde{\mathcal{W}} / \sim_p$ as:

$$e(\underline{w}) \triangleq \{\tilde{w} \in \tilde{\mathcal{W}} \mid a^*(\tilde{w}) = \underline{w}\} \subset \tilde{\mathcal{W}}. \quad (20)$$

The mapping is *one-to-many* in the forward direction: a single pragmatic purpose corresponds to a set of semantic justifications.

Definition 3 (Dynamic Compatible Isoteleia Mapping): For sequential decision-making problems (Section XI), the optimal action must account for future consequences. Let $V : \Delta(\mathcal{S}) \rightarrow \mathbb{R}$ be the value function over belief states. We define the *dynamic optimal action* for a semantic symbol $\tilde{w}$ as:

$$a^*(\tilde{w}, V) \triangleq \arg \max_{a \in \mathcal{A}} \mathbb{E}_{S|\tilde{w}} [U(S, a) +$$

$$\gamma \sum_{s'} P(s'|S, a) V(b'_{s'})], \quad (21)$$

where $b'_{s'}$ is the updated belief after transitioning to s' . Consequently, the dynamic isoteleia mapping is:

$$e_V(\underline{w}) \triangleq \{\tilde{w} \in \tilde{\mathcal{W}} \mid a^*(\tilde{w}, V) = \underline{w}\}. \quad (22)$$

When $\gamma = 0$ (static decision), we have $e_V(\underline{w}) \equiv e(\underline{w})$, recovering the static Definition 2. All dynamic Bellman equations in Section XI-B implicitly adopt this conditional extension of the isoteleia mapping.

Definition 4 (Reification Mapping): The **Reification Mapping**

$$g : \underline{\mathcal{W}} \rightarrow 2^{\mathcal{W}} \quad (23)$$

is defined as the composition of the isoteleia and synonymous mappings:

$$g(\underline{w}) \triangleq \bigcup_{\tilde{w} \in e(\underline{w})} f(\tilde{w}) \subset \mathcal{W}. \quad (24)$$

This mapping directly associates a pragmatic purpose with the complete set of syntactic symbols that can realize it.

Figure 8 illustrates the three-tier quotient mapping structure. The isoteleia mapping $e : \underline{\mathcal{W}} \rightarrow 2^{\tilde{\mathcal{W}}}$ groups semantic classes that lead to the same optimal action, forming pragmatic equivalence classes. The synonymous mapping $f : \tilde{\mathcal{W}} \rightarrow 2^{\mathcal{W}}$ associates each semantic meaning with a set of syntactic realizations. Their composition, the reification mapping $g = f \circ e : \underline{\mathcal{W}} \rightarrow 2^{\mathcal{W}}$, directly links a pragmatic purpose to all possible syntactic signals. The resulting pragmatic classes are evaluated by the utility function U and mapped to physical actions $A_t \in \mathcal{A}$, forming the closed loop of pragmatic information processing.

D.4 The Five Relations Between Layers

The three mappings give rise to five fundamental relations. The isoteleia mapping e is a pragmatic-to-semantic relation: a single pragmatic class corresponds to multiple semantic classes. Conversely, each semantic class maps to exactly one pragmatic class via the inverse isoteleia, ensuring the partition is well-defined. The synonymous mapping f relates semantic to syntactic: a single meaning corresponds to multiple syntactic

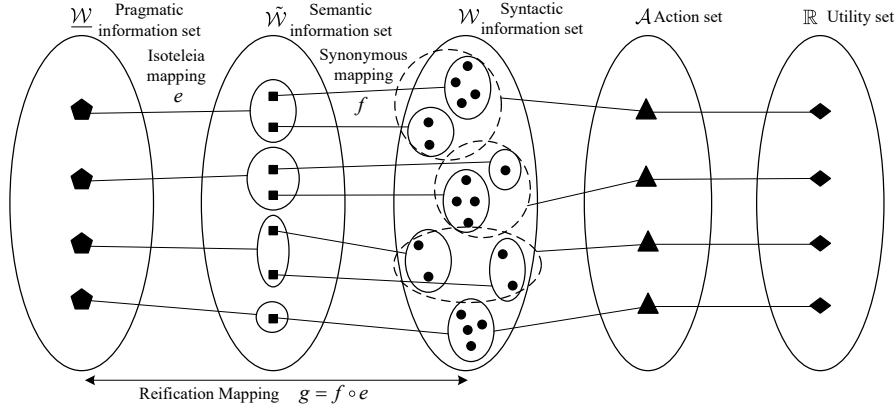


Fig. 8: The hierarchical mapping from pragmatic classes to semantic classes and syntactic symbols, with correspondence to the action space $\mathcal{A}$ and the utility evaluation U .

symbols, while each syntactic symbol belongs to exactly one semantic class via the inverse synonymous mapping. Finally, the reification mapping g directly relates pragmatic purposes to syntactic signals, associating a single purpose with the union of all syntactic symbols in its associated semantic classes.

D.5 Consistency Condition: Semantic Refinement

A critical condition must be satisfied for the isoteleia mapping to be well-defined: every semantic class $\tilde{w}$ must be *pragmatically homogeneous*, meaning that for any two syntactic symbols $w_1, w_2 \in \tilde{w}$, we must have $a^*(w_1) = a^*(w_2)$. If this fails, the semantic equivalence is too coarse, and we must *refine* the semantic partition by splitting $\tilde{w}$ into smaller classes, each homogeneous with respect to the optimal action. This operation, termed **semantic refinement**, is defined as:

$$\tilde{w} \rightarrow \{\tilde{w}^{(1)}, \tilde{w}^{(2)}, \dots\}, \quad \tilde{w}^{(j)} \triangleq \{w \in \tilde{w} \mid a^*(w) = a_j\}. \quad (25)$$

Only after such refinement is the isoteleia mapping uniquely defined.

D.6 Quotient-Space Structure

The syntactic, semantic, and pragmatic layers are successive quotients of the same original set $\mathcal{W}$: $\mathcal{W} \xrightarrow{\sim s} \tilde{\mathcal{W}} \xrightarrow{\sim p} \underline{\mathcal{W}}$. Syntactic information is the finest description, retaining all symbol distinctions. Semantic information is the quotient by meaning equivalence, discarding differences that do not affect meaning. Pragmatic information is the quotient by action equivalence, discarding semantic differences that do not affect optimal actions. Thus, the three layers are different lev-

els of abstraction of the same underlying signal, with the pragmatic layer representing the coarsest, most task-relevant description. This quotient-space perspective provides the mathematical foundation for all subsequent information-theoretic analyses.

III. Pragmatic Entropy

In this section, we begin with pragmatic entropy, the fundamental measure of decision uncertainty, and then extend to joint and conditional pragmatic entropies, establishing their compatibility with syntactic and semantic counterparts.

A. Formal Definition of Pragmatic Entropy

Pragmatic entropy is the fundamental measure of decision uncertainty in a pragmatic information system. Unlike syntactic entropy, which measures uncertainty about which symbol will be transmitted, or semantic entropy, which measures uncertainty about which meaning is conveyed, pragmatic entropy measures the uncertainty about which terminal action will be optimal, that is, the uncertainty that directly affects the system's ability to achieve its task objectives. This measure is derived from the utility function U through the Isoteleia Mapping $e : \underline{\mathcal{W}} \rightarrow 2^{\tilde{\mathcal{W}}}$, which groups semantic symbols into pragmatic equivalence classes based on their induced optimal actions.

Let $\mathcal{W}$ be the syntactic alphabet with probability mass function $P(W)$, and let $\tilde{\mathcal{W}}$ be the semantic alphabet induced by the Synonymous Mapping $f : \tilde{\mathcal{W}} \rightarrow 2^{\mathcal{W}}$. Let the Isoteleia Mapping $e : \underline{\mathcal{W}} \rightarrow 2^{\tilde{\mathcal{W}}}$ partition the semantic space according to the utility function

$U : \mathcal{W} \times \mathcal{A} \rightarrow \mathbb{R}$, as defined in Eq. (11).

For each pragmatic symbol $\underline{w} \in \underline{\mathcal{W}}$, its probability is given by:

$$P(\underline{w}) = \sum_{\tilde{w} \in e(\underline{w})} P(\tilde{w}) = \sum_{\tilde{w} \in e(\underline{w})} \sum_{w \in f(\tilde{w})} P(w). \quad (26)$$

Definition 5 (Pragmatic Entropy): The **pragmatic entropy** of a pragmatic source $\underline{\mathcal{W}}$ is defined as:

$$H_p(\underline{W}) \triangleq - \sum_{\underline{w} \in \underline{\mathcal{W}}} P(\underline{w}) \log P(\underline{w}), \quad (27)$$

where $P(\underline{w})$ is given by Eq. (26). The unit of pragmatic entropy is the **pragmatic bit (Prabit)**, reflecting that it measures uncertainty about terminal actions rather than about symbols or meanings.

Equivalently, pragmatic entropy can be expressed directly in terms of the syntactic distribution and the composite Reification Mapping $g = f \circ e : \underline{\mathcal{W}} \rightarrow 2^{\mathcal{W}}$:

$$H_p(\underline{W}) = - \sum_{\underline{w} \in \underline{\mathcal{W}}} \left(\sum_{w \in g(\underline{w})} P(w) \right) \log \left(\sum_{w \in g(\underline{w})} P(w) \right). \quad (28)$$

The pragmatic entropy quantifies the minimum average number of pragmatic bits required to specify the optimal terminal action. It represents the irreducible decision uncertainty that must be resolved through communication and control.

We now establish the fundamental properties of pragmatic entropy. These properties are direct consequences of the quotient-space structure and the deterministic nature of the Isoteleia and Synonymous mappings.

Lemma 1 (Non-negativity of Pragmatic Entropy): The pragmatic entropy is non-negative:

$$H_p(\underline{W}) \geq 0. \quad (29)$$

Equality holds if and only if the pragmatic alphabet has a single element, i.e., there is only one possible optimal action regardless of the state.

Proof. By definition, $P(\underline{w}) \in [0, 1]$ for all $\underline{w} \in \underline{\mathcal{W}}$, and $\sum_{\underline{w}} P(\underline{w}) = 1$. Thus $-\log P(\underline{w}) \geq 0$, and the weighted sum is non-negative. Equality holds if and

only if $P(\underline{w}) = 1$ for some $\underline{w}$ and $P(\underline{w}') = 0$ for all $\underline{w}' \neq \underline{w}$, which means the pragmatic alphabet has cardinality 1. $\square$

Lemma 2 (Pragmatic Entropy Hierarchy): Let $H(W)$ denote the syntactic entropy (classical Shannon entropy), $H_s(\tilde{W})$ denote the semantic entropy as defined in semantic information theory, and $H_p(\underline{W})$ denote the pragmatic entropy. Then:

$$H_p(\underline{W}) \leq H_s(\tilde{W}) \leq H(W). \quad (30)$$

Proof. We prove the two inequalities separately.

First inequality: $H_p(\underline{W}) \leq H_s(\tilde{W})$.

Since the Isoteleia Mapping $e : \underline{\mathcal{W}} \rightarrow 2^{\tilde{\mathcal{W}}}$ partitions the semantic alphabet $\tilde{\mathcal{W}}$ into disjoint equivalence classes, the pragmatic symbol $\underline{W}$ is a deterministic function of the semantic symbol $\tilde{W}$. That is, there exists a deterministic function $q : \tilde{\mathcal{W}} \rightarrow \underline{\mathcal{W}}$ such that $\underline{W} = q(\tilde{W})$. By the data processing inequality for entropy, the entropy of a function of a random variable cannot exceed the entropy of the original variable:

$$H_p(\underline{W}) = H(q(\tilde{W})) \leq H(\tilde{W}) = H_s(\tilde{W}). \quad (31)$$

Equality holds if and only if the isoteleia mapping is injective (no two semantic classes are merged into the same pragmatic class).

Second inequality: $H_s(\tilde{W}) \leq H(W)$. This is the fundamental result of semantic information theory, established in [24, 25].

Combining both inequalities yields the desired hierarchy. $\square$

This hierarchy captures the essence of pragmatic information theory: each successive layer of abstraction reduces uncertainty by discarding distinctions that are irrelevant to the current level of analysis. The syntactic layer retains all symbol distinctions; the semantic layer discards distinctions that do not affect meaning; and the pragmatic layer discards distinctions that do not affect the optimal terminal action.

We now establish the maximum entropy principle for pragmatic entropy, which characterizes the upper bound of decision uncertainty given the pragmatic alphabet.

Lemma 3 (Maximum Pragmatic Entropy): Let $\underline{\mathcal{W}}$ be the pragmatic alphabet with cardinality $|\underline{\mathcal{W}}| = L$. The

pragmatic entropy satisfies:

$$H_p(\underline{W}) \leq \log L. \quad (32)$$

Equality holds if and only if the pragmatic symbols are uniformly distributed, i.e., $P(\underline{w}) = 1/L$ for all $\underline{w} \in \underline{\mathcal{W}}$.

Proof. Applying the standard entropy maximization argument, we know that the entropy is maximized when the distribution is uniform. Therefore, $H_p(\underline{W}) \leq \log L$, with equality if and only if $P(\underline{w}) = 1/L$ for all $\underline{w} \in \underline{\mathcal{W}}$. $\square$

Corollary 1 (Hierarchy of Maximum Entropies): Let $L_p = |\underline{\mathcal{W}}|$, $L_s = |\tilde{\mathcal{W}}|$, and $L = |\mathcal{W}|$ denote the cardinalities of the pragmatic, semantic, and syntactic alphabets, respectively. Since the Isoteleia and Synonymous mappings are quotient mappings, we have $L_p \leq L_s \leq L$. Consequently:

$$\log L_p \leq \log L_s \leq \log L. \quad (33)$$

Combining with Lemma 2, we obtain the full hierarchy of maximum entropies:

$$H_p(\underline{W}) \leq \log L_p \leq \log L_s \leq \log L. \quad (34)$$

Remark 1: The maximum pragmatic entropy $\log L_p$ is strictly less than the maximum semantic entropy $\log L_s$ whenever the Isoteleia Mapping is non-injective (i.e., when multiple semantic classes merge into the same pragmatic class). This reflects the fundamental principle that the goal-directed compression of semantic distinctions reduces the theoretical upper bound of decision uncertainty. In the extreme case where all semantic classes map to a single pragmatic class (i.e., the same action is optimal for all states), we have $L_p = 1$ and $H_p(\underline{W}) = 0$, indicating that there is no decision uncertainty whatsoever.

The maximum pragmatic entropy lemma provides a theoretical limit on the amount of decision uncertainty that can exist in a pragmatic information system. It also serves as a design guideline: to maximize the information content of pragmatic communication, the system should be designed such that the pragmatic symbols are as uniformly distributed as possible, within the constraints imposed by the utility function and the environment.

Example 1 (Autonomous Vehicle Decision-Making):

We now present a concrete numerical example to illustrate the calculation and interpretation of pragmatic entropy, comparing it with syntactic and semantic entropies. Consider an autonomous vehicle approaching an intersection. The vehicle's perception system observes three possible traffic light states, which serve as the syntactic alphabet $\mathcal{W} = \{w_1, w_2, w_3\}$ with probabilities $P(W) = \{0.6, 0.3, 0.1\}$, representing green, yellow, and red lights, respectively.

The vehicle's semantic interpretation groups traffic lights into two meaning classes: $\tilde{w}_1 = \text{"Proceed with caution"}$ (green and yellow) with $P(\tilde{w}_1) = 0.9$, and $\tilde{w}_2 = \text{"Stop"}$ (red) with $P(\tilde{w}_2) = 0.1$. The decision utility $U(w, a)$ for actions $a_1 = \text{Go}$ and $a_2 = \text{Stop}$ is given in Table II.

TABLE II: Utility function $U(w, a)$ for autonomous vehicle decision example.

State w	Go (a_1)	Stop (a_2)
w_1 (green)	+10	-5
w_2 (yellow)	-2	+1
w_3 (red)	-100	+5

For each semantic class, the expected utility is $E[U(\text{Go})|\tilde{w}_1] = (0.6/0.9) * 10 + (0.3/0.9) * (-2) = 6$, $E[U(\text{Stop})|\tilde{w}_1] = (0.6/0.9) * (-5) + (0.3/0.9) * (+1) = -3$, $E[U(\text{Go})|\tilde{w}_2] = (0.1/0.1) * (-100) = -100$, and $E[U(\text{Stop})|\tilde{w}_2] = (0.1/0.1) * (+5) = +5$. Thus, the optimal action for $\tilde{w}_1$ is Go ($6 > -3$), while for $\tilde{w}_2$ it is Stop ($-100 < +5$). The Isoteleia Mapping induces two pragmatic classes: $\underline{w}_1 = \text{"Go"}$ containing $\tilde{w}_1$ with $P(\underline{w}_1) = 0.9$, and $\underline{w}_2 = \text{"Stop"}$ containing $\tilde{w}_2$ with $P(\underline{w}_2) = 0.1$.

The syntactic entropy is $H(W) = -[0.6 \log 0.6 + 0.3 \log 0.3 + 0.1 \log 0.1] \approx 1.295$ bits. The semantic entropy is $H_s(\tilde{W}) = -[0.9 \log 0.9 + 0.1 \log 0.1] \approx 0.469$ sebits. The pragmatic entropy is $H_p(\underline{W}) = -[0.9 \log 0.9 + 0.1 \log 0.1] \approx 0.469$ prabits. In this example, $H_p = H_s$ because each semantic class maps to a distinct pragmatic class (the Isoteleia Mapping is injective), yet the hierarchy $H_p = 0.469 < 1.295 = H(W)$ still holds.

The syntactic entropy of 1.295 bits represents the uncertainty about the specific traffic light color ob-

served. The semantic entropy of 0.469 sebits captures the uncertainty about whether the vehicle should “Proceed with caution” or “Stop” based on the meaning of the observation. The pragmatic entropy of 0.469 prabits reflects the uncertainty about the optimal terminal action, which ultimately determines the vehicle’s behavior. The reduction from syntactic to pragmatic entropy (from 1.295 to 0.469 bits) demonstrates that the vehicle does not need to distinguish between green and yellow lights to make the correct decision; it only needs to distinguish between “Go” and “Stop” situations.

This example demonstrates how the pragmatic entropy measure provides a principled way to quantify the minimum decision uncertainty that must be resolved through communication and control, thereby establishing the theoretical foundation for resource allocation and system design in pragmatic information systems.

B. Pragmatic Joint Entropy and Conditional Entropy

Building upon the definition of pragmatic entropy for a single variable, we now extend the framework to multiple semantic/pragmatic variables. This extension is essential for analyzing the dependencies and information flows between different components of a pragmatic information system, such as the relationship between the transmitted pragmatic intention and the received pragmatic interpretation.

To formalize the joint behavior of two pragmatic sources, we first extend the Isoteleia Mapping to the joint and conditional cases. Let $\underline{\mathcal{W}}$ and $\underline{\mathcal{V}}$ be two pragmatic alphabets, with associated semantic alphabets $\tilde{\mathcal{W}}$ and $\tilde{\mathcal{V}}$, respectively. Let the corresponding syntactic alphabets be $\mathcal{W}$ and $\mathcal{V}$, with a joint distribution $P(W, V)$.

Definition 6 (Joint Isoteleia Mapping): The **joint Isoteleia Mapping**

$$e_{wv} : \underline{\mathcal{W}} \times \underline{\mathcal{V}} \rightarrow 2^{\tilde{\mathcal{W}} \times \tilde{\mathcal{V}}} \quad (35)$$

is defined by its preimage: for each pair of pragmatic symbols $(\underline{w}, \underline{v}) \in \underline{\mathcal{W}} \times \underline{\mathcal{V}}$, the image is the set of all semantic pairs $(\tilde{w}, \tilde{v})$ that jointly induce the same optimal terminal actions:

$$e_{wv}(\underline{w}, \underline{v}) \triangleq \{(\tilde{w}, \tilde{v}) \in \tilde{\mathcal{W}} \times \tilde{\mathcal{V}} \mid \arg \max_{a \in \mathcal{A}} \mathbb{E}[U(W, a) | \tilde{w}]\}$$

$$= \underline{w}, \arg \max_{b \in \mathcal{B}} \mathbb{E}[U(V, b) | \tilde{v}] = \underline{v}\}. \quad (36)$$

This mapping partitions the joint semantic space $\tilde{\mathcal{W}} \times \tilde{\mathcal{V}}$ into equivalence classes according to the pair of optimal actions.

Definition 7 (Conditional Isoteleia Mapping): Given a specific pragmatic symbol $\underline{w} \in \underline{\mathcal{W}}$, the **conditional Isoteleia Mapping**

$$e_{v|w}(\underline{v} | \underline{w}) : \underline{\mathcal{V}} \rightarrow 2^{\tilde{\mathcal{V}}} \quad (37)$$

is defined as the set of semantic symbols $\tilde{v}$ that, together with the given $\underline{w}$, lead to the optimal action $\underline{v}$:

$$e_{v|w}(\underline{v} | \underline{w}) \triangleq \left\{ \tilde{v} \in \tilde{\mathcal{V}} \mid \arg \max_{b \in \mathcal{B}} \mathbb{E}[U(V, b) | \tilde{v}, \underline{w}] = \underline{v} \right\}. \quad (38)$$

This mapping captures the pragmatic equivalence of the second variable conditioned on the first pragmatic class.

The joint and conditional synonymous mappings have been established in the semantic information theory literature [24, 25]. The joint synonymous mapping partitions the syntactic product space according to meaning identity, while the conditional synonymous mapping partitions the syntactic space of one variable given the semantic class of the other. We will leverage these existing definitions in the subsequent entropy formulations.

With the joint and conditional isoteleia mappings defined, we can now introduce the corresponding entropy measures.

Definition 8 (Pragmatic Joint Entropy): Let $(\underline{W}, \underline{V})$ be a pair of pragmatic random variables derived from the syntactic pair (W, V) via the joint isoteleia mapping e_{wv} . The **pragmatic joint entropy** is defined as the Shannon entropy of the joint pragmatic distribution:

$$H_p(\underline{W}, \underline{V}) \triangleq - \sum_{\underline{w} \in \underline{\mathcal{W}}} \sum_{\underline{v} \in \underline{\mathcal{V}}} P(\underline{w}, \underline{v}) \log P(\underline{w}, \underline{v}), \quad (39)$$

where the joint pragmatic probabilities are obtained by aggregating over all semantic and syntactic fibers:

$$\begin{aligned} P(\underline{w}, \underline{v}) &= \sum_{(\tilde{w}, \tilde{v}) \in e_{wv}(\underline{w}, \underline{v})} P(\tilde{w}, \tilde{v}) \\ &= \sum_{(\tilde{w}, \tilde{v}) \in e_{wv}(\underline{w}, \underline{v})} \sum_{(w, v) \in f_{wv}(\tilde{w}, \tilde{v})} P(w, v), \end{aligned} \quad (40)$$

with f_{uv} denoting the joint Synonymous Mapping.

Definition 9 (Pragmatic Conditional Entropy): There are two natural variants of pragmatic conditional entropy, depending on whether the conditioning variable is syntactic or pragmatic.

(a) Pragmatic conditional entropy of $\underline{V}$ given the syntactic variable $\underline{W}$:

$$H_p(\underline{V}|\underline{W}) \triangleq - \sum_{w \in \mathcal{W}} \sum_{v \in \mathcal{V}} P(w)P(v|w) \log P(v|w), \quad (41)$$

where $P(v|w) = \sum_{\tilde{v} \in e(v)} P(\tilde{v}|w)$ is the pragmatic distribution of $\underline{V}$ conditioned on a specific syntactic observation w .

(b) Pragmatic conditional entropy of $\underline{V}$ given the pragmatic variable $\underline{W}$:

$$H_p(\underline{V}|\underline{W}) \triangleq - \sum_{w \in \mathcal{W}} \sum_{v \in \mathcal{V}} P(w, v) \log P(v|w), \quad (42)$$

where $P(v|w) = P(w, v)/P(w)$.

Similarly, one can define $H_p(\underline{W}|\underline{V})$ and $H_p(\underline{W}|\underline{V})$ by symmetry. The choice between conditioning on syntactic or pragmatic variables depends on the application: $H_p(\underline{V}|\underline{W})$ measures the residual decision uncertainty after observing the exact syntactic signal, while $H_p(\underline{V}|\underline{W})$ measures the uncertainty of one decision given the other decision.

Analogous to semantic entropy, pragmatic entropy does not satisfy the exact additive chain rule of classical entropy; instead, it satisfies a series of inequalities that reflect the coarsening of information. We first establish the chain rule for a pair of pragmatic variables.

Theorem 1 (Chain Rule for Binary Pragmatic Entropy): For any two pragmatic variables $\underline{W}$ and $\underline{V}$ derived from syntactic variables W and V via the isoteleia and synonymous mappings, the following chain of inequalities holds:

$$H_p(\underline{W}) + H_p(\underline{V}|\underline{W}) \leq H_p(\underline{W}, \underline{V}) \leq H(W) + H_p(\underline{V}|\underline{W}), \quad (43)$$

and symmetrically,

$$H_p(\underline{V}) + H_p(\underline{W}|\underline{V}) \leq H_p(\underline{W}, \underline{V}) \leq H(V) + H_p(\underline{W}|\underline{V}). \quad (44)$$

Proof. Let $g_W : \mathcal{W} \rightarrow 2^{\mathcal{W}}$ and $g_V : \mathcal{V} \rightarrow 2^{\mathcal{V}}$ be the reification mappings, so that $\underline{W} = g_W^{-1}(W)$ and $\underline{V} = g_V^{-1}(V)$ are deterministic functions of W and V , respectively.

For any $\underline{w} \in \mathcal{W}$, define its preimage set

$$\mathcal{W}_{\underline{w}} \triangleq \{w \in \mathcal{W} : g_W(w) = \underline{w}\},$$

and similarly $\mathcal{V}_{\underline{v}} \triangleq \{v \in \mathcal{V} : g_V(v) = \underline{v}\}$. The aggregated probabilities are

$$p(\underline{w}) = \sum_{w \in \mathcal{W}_{\underline{w}}} p(w), \quad p(\underline{w}, \underline{v}) = \sum_{w \in \mathcal{W}_{\underline{w}}} \sum_{v \in \mathcal{V}_{\underline{v}}} p(w, v).$$

For a fixed $\underline{w}$, the conditional distribution of $\underline{V}$ given $\underline{W} = \underline{w}$ is

$$p(\underline{v} | \underline{w}) = \frac{p(\underline{w}, \underline{v})}{p(\underline{w})} = \sum_{w \in \mathcal{W}_{\underline{w}}} \frac{p(w)}{p(\underline{w})} \cdot p(v | w), \quad (45)$$

where $p(\underline{v} | \underline{w}) \triangleq \sum_{v \in \mathcal{V}_{\underline{v}}} p(v | w)$. Thus, for each $\underline{w}$,

the distribution $p(\cdot | \underline{w})$ is a mixture of the distributions $\{p(\cdot | w) : w \in \mathcal{W}_{\underline{w}}\}$ with weights $\frac{p(w)}{p(\underline{w})}$.

Since the Shannon entropy is concave,

$$H(p(\cdot | \underline{w})) \geq \sum_{w \in \mathcal{W}_{\underline{w}}} \frac{p(w)}{p(\underline{w})} H(p(\cdot | w)). \quad (46)$$

Multiplying by $p(\underline{w})$ and summing over $\underline{w}$:

$$\sum_{\underline{w}} p(\underline{w}) H(p(\cdot | \underline{w})) \geq \sum_{\underline{w}} \sum_{w \in \mathcal{W}_{\underline{w}}} p(w) H(p(\cdot | w)). \quad (47)$$

The left-hand side is $H_p(\underline{V} | \underline{W})$, and the right-hand side is $H_p(\underline{V} | \underline{W})$. Hence,

$$H_p(\underline{V} | \underline{W}) \geq H_p(\underline{V} | \underline{W}). \quad (48)$$

Since $\underline{W}$ and $\underline{V}$ are ordinary random variables, the classical chain rule applies:

$$H_p(\underline{W}, \underline{V}) = H_p(\underline{W}) + H_p(\underline{V} | \underline{W}). \quad (49)$$

Substituting (48) into (49) yields the left inequality of (43):

$$H_p(\underline{W}, \underline{V}) \geq H_p(\underline{W}) + H_p(\underline{V} | \underline{W}). \quad (50)$$

For the right inequality of (43), since $\underline{W}$ is a function of W ,

$$\begin{aligned} H_p(\underline{W}, \underline{V}) &= H(\underline{W}, \underline{V}) \leq H(W, \underline{V}) = H(W) + H(\underline{V} | W) \\ &= H(W) + H_p(\underline{V} | W), \end{aligned} \quad (51)$$

which proves the right inequality.

The symmetric chain (44) follows by interchanging W and V , noting that $\underline{V}$ is a function of V , and applying the same mixture-entropy argument to $p(\cdot | \underline{v})$. This completes the proof. $\square$

The chain rule (Eq. 43) reveals that pragmatic joint entropy lies between two bounds: the lower bound is the sum of the marginal pragmatic entropy of $\underline{W}$ and the pragmatic uncertainty of $\underline{V}$ given the syntactic observation W ; the upper bound is the sum of the syntactic entropy of W and the same pragmatic conditional term. This intermediate position reflects the fact that pragmatic abstraction reduces uncertainty, but not as much as semantic abstraction.

Theorem 2 (Hierarchy of Binary Joint Entropies): For any two syntactic variables W and V with their associated semantic variables $\tilde{W}$, $\tilde{V}$ and pragmatic variables $\underline{W}$, $\underline{V}$, the following inequality holds:

$$H_p(\underline{W}, \underline{V}) \leq H_s(\tilde{W}, \tilde{V}) \leq H(W, V). \quad (52)$$

Moreover, equality in the first inequality holds if and only if the isoteleia mapping is injective on the joint semantic space (i.e., no two distinct semantic pairs are merged into the same pragmatic pair). Equality in the second holds if and only if the joint synonymous mapping is injective (no two distinct syntactic pairs share the same semantic meaning).

Proof. We prove the two inequalities separately.

First inequality: $H_p(\underline{W}, \underline{V}) \leq H_s(\tilde{W}, \tilde{V})$.

By definition, the pragmatic variables are deterministic functions of the semantic variables via the joint Isoteleia mapping:

$$(\underline{W}, \underline{V}) = e_{uv}^{-1}(\tilde{W}, \tilde{V}), \quad (53)$$

where e_{uv}^{-1} is the inverse of the isoteleia mapping that collapses each semantic equivalence class to a single pragmatic symbol. Since $(\underline{W}, \underline{V})$ is a deterministic func-

tion of $(\tilde{W}, \tilde{V})$, the joint entropy of a function of a random variable cannot exceed the joint entropy of the original variable. Applying the data processing inequality for entropy:

$$H_p(\underline{W}, \underline{V}) = H(e_{uv}^{-1}(\tilde{W}, \tilde{V})) \leq H(\tilde{W}, \tilde{V}) = H_s(\tilde{W}, \tilde{V}). \quad (54)$$

Equality holds if and only if the mapping is injective, i.e., no two distinct semantic pairs are merged.

Second inequality: $H_s(\tilde{W}, \tilde{V}) \leq H(W, V)$.

Similarly, the semantic variables are deterministic functions of the syntactic variables via the joint synonymous mapping:

$$(\tilde{W}, \tilde{V}) = f_{uv}^{-1}(W, V), \quad (55)$$

where f_{uv}^{-1} collapses each syntactic equivalence class to a single semantic symbol. Hence,

$$H_s(\tilde{W}, \tilde{V}) = H(f_{uv}^{-1}(W, V)) \leq H(W, V). \quad (56)$$

Equality holds if and only if the joint synonymous mapping is injective.

Combining both inequalities yields the desired hierarchy. $\square$

We now extend this result to sequences of pragmatic variables. To do so, we first define the sequential extension of the joint reification mapping.

Definition 10 (Sequential Joint Reification Mapping): Let $g_{WV} : \mathcal{W} \times \mathcal{V} \rightarrow 2^{\mathcal{W} \times \mathcal{V}}$ be the joint Reification Mapping that partitions the joint syntactic space $\mathcal{W} \times \mathcal{V}$ into pragmatic equivalence classes. Its n -th extension

$$g_{WV}^n : \mathcal{W}^n \times \mathcal{V}^n \rightarrow 2^{\mathcal{W}^n \times \mathcal{V}^n} \quad (57)$$

is defined by

$$g_{WV}^n(\underline{w}^n, \underline{v}^n) \triangleq \prod_{i=1}^n g_{WV}(\underline{w}_i, \underline{v}_i), \quad (58)$$

where $\underline{w}^n = (\underline{w}_1, \dots, \underline{w}_n)$ and $\underline{v}^n = (\underline{v}_1, \dots, \underline{v}_n)$ are pragmatic sequences, and the product denotes the Cartesian product of sets. In other words, a pragmatic sequence pair $(\underline{w}^n, \underline{v}^n)$ corresponds to the set of all syntactic sequence pairs (w^n, v^n) such that for every position i , $(w_i, v_i) \in g_{WV}(\underline{w}_i, \underline{v}_i)$.

The marginal sequential pragmatic variables $\underline{W}^n$ and $\underline{V}^n$ are then obtained by projecting from the joint pragmatic sequence. The probability of a pragmatic sequence pair is given by

$$P(\underline{w}^n, \underline{v}^n) = \sum_{(w^n, v^n) \in g_{WV}^n(\underline{w}^n, \underline{v}^n)} P(w^n, v^n). \quad (59)$$

Theorem 3 (Chain Rule for Sequential Pragmatic Entropy): Let $(\underline{W}_1, \underline{W}_2, \dots, \underline{W}_n)$ be a sequence of pragmatic variables derived from the syntactic sequence $(W_1, W_2, \dots, W_n)$ via the sequential joint Reification Mapping. Define the sequential partial entropy

$$\tilde{H}_p(\underline{W}_1^m, W_{m+1}^n) \triangleq \sum_{k=1}^m H_p(\underline{W}_k | \underline{W}_1^{k-1}) + \sum_{k=m+1}^n H_p(W_k | \underline{W}_1^m, W_{m+1}^{k-1}), \quad (60)$$

where the notation $\underline{W}_1^m$ denotes the block $(\underline{W}_1, \dots, \underline{W}_m)$, and W_{m+1}^n denotes the block $(W_{m+1}, \dots, W_n)$. Then the following chain of inequalities holds:

$$\sum_{k=1}^n H_p(\underline{W}_k | W_1^{k-1}) \leq H_p(\underline{W}_1^n) \leq \tilde{H}_p(\underline{W}_1^{n-1}, W_n) \leq \dots \leq \tilde{H}_p(\underline{W}_1, W_2^n) \leq H_p(W_1^n), \quad (61)$$

where the last term $H_p(W_1^n)$ denotes the syntactic entropy of the full sequence.

Proof. The proof follows by iteratively applying the binary chain rule to each adjacent pair in the sequence. Starting from the full pragmatic joint entropy $H_p(\underline{W}_1^n)$, we use the binary chain rule to bound it in terms of the entropy of the first $n - 1$ variables and the last variable, either conditioned on the syntactic or pragmatic prefix. By gradually replacing pragmatic conditioning with syntactic conditioning, we obtain the chain of inequalities. The details are analogous to the semantic sequential chain rule established in [24, 25], with the substitution of pragmatic entropies. $\square$

Theorem 4 (Hierarchy of Sequential Joint Entropies): Let $W^n = (W_1, \dots, W_n)$ be a syntactic sequence, with associated semantic sequence $\tilde{W}^n$ and pragmatic sequence $\underline{W}^n$ obtained by applying the Synonymous and

Isoteleia mappings element-wise (or, equivalently, by the sequential joint Reification Mapping). Then:

$$H_p(\underline{W}^n) \leq H_s(\tilde{W}^n) \leq H(W^n), \quad (62)$$

where $H_p(\underline{W}^n)$, $H_s(\tilde{W}^n)$, and $H(W^n)$ denote the joint entropies of the entire sequences.

Proof. The proof follows directly from the element-wise deterministic mappings. Since the sequential joint reification mapping is defined as the product of the individual mappings, the entire pragmatic sequence $\underline{W}^n$ is a deterministic function of the semantic sequence $\tilde{W}^n$, and $\tilde{W}^n$ is a deterministic function of the syntactic sequence W^n . Therefore:

$$H_p(\underline{W}^n) = H(e^{-n}(\tilde{W}^n)) \leq H(\tilde{W}^n) = H_s(\tilde{W}^n), \quad (63)$$

and

$$H_s(\tilde{W}^n) = H(f^{-n}(W^n)) \leq H(W^n). \quad (64)$$

Combining these yields the desired hierarchy. $\square$

Corollary 2 (Hierarchy of Sequential Conditional Entropies): For any m with $1 \leq m \leq n$, the sequential partial entropies also satisfy the same ordering:

$$\tilde{H}_p(\underline{W}_1^m, W_{m+1}^n) \leq \tilde{H}_s(\tilde{W}_1^m, W_{m+1}^n) \leq H(W^n), \quad (65)$$

where the partial entropies are defined as in the chain rule (Eq. (60)). This follows because the partial entropies are linear combinations of conditional entropies, each of which obeys the single-variable hierarchy.

Remark 2: The hierarchy theorems reveal a layered structure of uncertainty: syntactic entropy captures all symbol-level distinctions; semantic entropy removes meaningless syntactic variations; and pragmatic entropy further removes decision-irrelevant semantic distinctions. Each successive coarsening reduces the joint entropy, providing a principled foundation for multi-level source coding and compression in pragmatic communication systems.

Example 2 (Multi-Vehicle Autonomous Driving Scenario): We now present a concrete numerical example to illustrate the computation and interpretation of pragmatic joint and conditional entropies, with particular

emphasis on distinguishing pragmatic entropy from semantic entropy. Consider two autonomous vehicles, Vehicle A and Vehicle B, approaching an intersection. Each vehicle must decide whether to “Go” or “Stop” based on its observed traffic light state and, potentially, the state of the other vehicle.

Let the syntactic state of Vehicle A be $W \in \{w_1, w_2, w_3, w_4\}$ representing green, yellow, red-short, and red-long, respectively; Vehicle B’s state V has the same interpretation. The joint probability distribution $P(W, V)$ is given in Table III, with marginals $P(w_1) = 0.23$, $P(w_2) = 0.22$, $P(w_3) = 0.25$, $P(w_4) = 0.30$, and similarly for V .

TABLE III: Joint probability distribution $P(W, V)$ for two vehicles.

$P(W, V)$	v_1 (green)	v_2 (yellow)	v_3 (red-short)	v_4 (red-long)
w_1 (green)	0.12	0.06	0.03	0.02
w_2 (yellow)	0.06	0.10	0.04	0.02
w_3 (red-short)	0.03	0.04	0.12	0.06
w_4 (red-long)	0.02	0.02	0.06	0.20

The synonymous mapping groups the syntactic states into three semantic classes: $\tilde{w}_1 = \tilde{v}_1 =$ “Proceed with caution” (green and yellow) with probability 0.45, $\tilde{w}_2 = \tilde{v}_2 =$ “Stop-short” (red-short) with probability 0.25, and $\tilde{w}_3 = \tilde{v}_3 =$ “Stop-long” (red-long) with probability 0.30. Under a safety-first utility, the optimal action for $\tilde{w}_1$ is “Go”, while for both $\tilde{w}_2$ and $\tilde{w}_3$ it is “Stop”. The isoteleia mapping therefore merges the two stop classes into a single pragmatic class: $\underline{w}_1 =$ “Go” containing $\tilde{w}_1$ with probability 0.45, and $\underline{w}_2 =$ “Stop” containing $\tilde{w}_2$ and $\tilde{w}_3$ with probability 0.55. The same mapping applies to Vehicle B.

Aggregating the syntactic probabilities yields the semantic joint distribution $P(\tilde{W}, \tilde{V})$ (Table IV) and the pragmatic joint distribution $P(\underline{W}, \underline{V})$ (Table V). The syntactic joint entropy is $H(W, V) \approx 3.6314$ bits; the semantic joint entropy is $H_s(\tilde{W}, \tilde{V}) \approx 2.7563$ sebits; and the pragmatic joint entropy is $H_p(\underline{W}, \underline{V}) \approx 1.7509$ prabits, confirming the hierarchy $H_p \leq H_s \leq H$. Marginal entropies are $H_s(\tilde{W}) \approx 1.5395$ bits and $H_p(\underline{W}) \approx 0.9928$ prabits.

The pragmatic conditional entropy is $H_p(\underline{V}|\underline{W}) = H_p(\underline{W}, \underline{V}) - H_p(\underline{W}) = 0.7581$ prabits, while

TABLE IV: Joint semantic distribution $P(\tilde{W}, \tilde{V})$.

$P(\tilde{W}, \tilde{V})$	$\tilde{v}_1$ (Proceed)	$\tilde{v}_2$ (Stop-short)	$\tilde{v}_3$ (Stop-long)
$\tilde{w}_1$ (Proceed)	0.34	0.07	0.04
$\tilde{w}_2$ (Stop-short)	0.07	0.12	0.06
$\tilde{w}_3$ (Stop-long)	0.04	0.06	0.20

TABLE V: Joint pragmatic distribution $P(\underline{W}, \underline{V})$.

$P(\underline{W}, \underline{V})$	$\underline{v}_1$ (Go)	$\underline{v}_2$ (Stop)
$\underline{w}_1$ (Go)	0.34	0.11
$\underline{w}_2$ (Stop)	0.11	0.44

$H_p(\underline{V}|\underline{W}) \approx H_p(\underline{W}, \underline{V}) - H(W) \approx 2.7328 - 1.9893 = 0.7435$ prabits. The binary chain rule is verified: $H_p(\underline{W}) + H_p(\underline{V}|\underline{W}) = 0.9928 + 0.7435 = 1.7363 \leq H_p(\underline{W}, \underline{V}) = 1.7509$, and $H_p(\underline{W}, \underline{V}) = 1.7509 \leq H(W) + H_p(\underline{V}|\underline{W}) = 1.9893 + 0.7435 = 2.7328$. These results (summarized in Table VI) demonstrate that pragmatic abstraction, by merging semantically distinct but action-equivalent states, reduces joint uncertainty beyond semantic abstraction.

TABLE VI: Comparison of joint and conditional entropies.

Entropy Measure	Value (bits/sebits/prabits)
Syntactic Joint Entropy $H(W, V)$	3.6314
Semantic Joint Entropy $H_s(\tilde{W}, \tilde{V})$	2.7563
Pragmatic Joint Entropy $H_p(\underline{W}, \underline{V})$	1.7509
Pragmatic Marginal Entropy $H_p(\underline{W})$	0.9928
Pragmatic Conditional Entropy $H_p(\underline{V} \underline{W})$	0.7581
Pragmatic Conditional Entropy $H_p(\underline{V} \underline{W})$	0.7435

This example illustrates the key insight: **pragmatic entropy is strictly less than semantic entropy when the isoteleia mapping merges distinct semantic classes that share the same optimal action.** The two types of “stop” (short red and long red) are semantically distinct but pragmatically equivalent, resulting in a joint pragmatic entropy (1.7509 prabits) that is significantly lower than the joint semantic entropy (2.7563 sebits). This reduction quantifies the decision-irrelevant information that can be discarded in a pragmatic communication system.

IV. Pragmatic Relative Entropy and Pragmatic Mutual Information

We now introduce the information-theoretic analogues of relative entropy and mutual information at the pragmatic level. These measures quantify the divergence between probability distributions and the statisti-

cal dependence between variables, respectively, when the comparison and dependency are evaluated at the level of terminal actions rather than symbols or meanings.

A. Pragmatic Relative Entropy

In classical information theory, the relative entropy (Kullback-Leibler divergence) [45] between two probability mass functions $p(w)$ and $q(w)$ over a syntactic alphabet $\mathcal{W}$ is defined as $D(p\|q) = \sum_w p(w) \log \frac{p(w)}{q(w)}$. Semantic information theory [24, 25] extends this by introducing semantic relative entropies that aggregate probabilities over synonymous sets, thereby measuring divergence at the meaning level. In pragmatic information theory, we further aggregate over pragmatic equivalence classes induced by the isoteleia mapping, so that the divergence is evaluated with respect to the terminal actions that the variables induce. This yields three forms of pragmatic relative entropy: full, partial (Type I), and partial (Type II), which correspond to different ways of combining the pragmatic aggregation with the syntactic or semantic distributions.

Let $\mathcal{W}$ be a syntactic alphabet with two probability mass functions $p(w)$ and $q(w)$, $w \in \mathcal{W}$. Let $\tilde{\mathcal{W}}$ be the associated semantic alphabet induced by the synonymous mapping $f : \tilde{\mathcal{W}} \rightarrow 2^{\mathcal{W}}$, and let $\underline{\mathcal{W}}$ be the pragmatic alphabet induced by the isoteleia mapping $e : \underline{\mathcal{W}} \rightarrow 2^{\tilde{\mathcal{W}}}$ according to a given utility function $U : \mathcal{W} \times \mathcal{A} \rightarrow \mathbb{R}$. The pragmatic partition of $\mathcal{W}$ is given by the equivalence relation $\sim_p$ defined in Eq. (15), which groups syntactic symbols that lead to the same optimal action. Equivalently, the pragmatic classes $\underline{w} \in \underline{\mathcal{W}}$ are the atoms of the partition $\Pi_{\text{prag}} = \mathcal{W} / \sim_p$. For each pragmatic class $\underline{w}$, its probability under a distribution p is defined as

$$p_p(\underline{w}) \triangleq \sum_{w \in g(\underline{w})} p(w), \quad (66)$$

where $g = f \circ e : \underline{\mathcal{W}} \rightarrow 2^{\mathcal{W}}$ is the reification mapping. Similarly, we define $q_p(\underline{w})$ for the distribution q .

Definition 11 (Pragmatic Relative Entropy): Given two probability mass functions $p(w)$ and $q(w)$ on the syntactic alphabet $\mathcal{W}$, and the pragmatic partition in-

duced by the reification mapping g , we define the following three forms of pragmatic relative entropy:

(a) Full Pragmatic Relative Entropy:

$$D_p(p_p\|q_p) \triangleq \sum_{\underline{w} \in \underline{\mathcal{W}}} p_p(\underline{w}) \log \frac{p_p(\underline{w})}{q_p(\underline{w})}. \quad (67)$$

(b) Partial Pragmatic Relative Entropy (Type I):

$$D_p(p_p\|q) \triangleq \sum_{\underline{w} \in \underline{\mathcal{W}}} \sum_{w \in g(\underline{w})} p(w) \log \frac{p_p(\underline{w})}{q(w)}. \quad (68)$$

(c) Partial Pragmatic Relative Entropy (Type II):

$$D_p(p\|q_p) \triangleq \sum_{\underline{w} \in \underline{\mathcal{W}}} \sum_{w \in g(\underline{w})} p(w) \log \frac{p(w)}{q_p(\underline{w})}. \quad (69)$$

These definitions mirror the semantic relative entropies [24, 25] but with the aggregation performed over pragmatic classes instead of semantic classes. The full pragmatic relative entropy compares the pragmatic marginal distributions directly. The partial Type I compares the aggregated pragmatic distribution of p against the fine-grained syntactic distribution of q , while the partial Type II compares the fine-grained syntactic distribution of p against the aggregated pragmatic distribution of q . As in the semantic case, $D_p(p\|q_p)$ may be negative; in practice, one can take its positive part $(D_p(p\|q_p))^+$.

We now establish the fundamental properties of pragmatic relative entropy. These properties are analogous to those of semantic relative entropy, with the pragmatic partition replacing the semantic partition.

Theorem 5 (Pragmatic Information Inequality): Let p and q be two probability mass functions on $\mathcal{W}$. Given the Reification Mapping g , the following inequalities hold:

$$\begin{cases} D_p(p_p\|q_p) \geq 0, \\ D_p(p_p\|q) \geq D_p(p_p\|p), \\ D_p(p\|q_p) \geq D_p(p\|p_p), \end{cases} \quad (70)$$

with equality in the first inequality if and only if $p_p(\underline{w}) = q_p(\underline{w})$ for all $\underline{w} \in \underline{\mathcal{W}}$; equality in the second if

and only if $p(w) = q(w)$ for all w within each pragmatic class; equality in the third if and only if $p_p(\underline{w}) = q_p(\underline{w})$ for all $\underline{w}$.

Proof. For the first inequality, by Jensen's inequality applied to the negative of $D_p(p_p \| q_p)$, we have

$$\begin{aligned} -D_p(p_p \| q_p) &= \sum_{\underline{w}} p_p(\underline{w}) \log \frac{q_p(\underline{w})}{p_p(\underline{w})} \\ &\leq \log \sum_{\underline{w}} p_p(\underline{w}) \frac{q_p(\underline{w})}{p_p(\underline{w})} \end{aligned} \quad (71)$$

$$= \log \sum_{\underline{w}} q_p(\underline{w}) = 0. \quad (72)$$

Thus $D_p(p_p \| q_p) \geq 0$, with equality iff $p_p = q_p$.

For the second inequality, we compute

$$\begin{aligned} D_p(p_p \| q) - D_p(p_p \| p) &= \sum_{\underline{w}} \sum_{w \in g(\underline{w})} p(w) \log \frac{p_p(w)}{q(w)} - \\ &\quad \sum_{\underline{w}} \sum_{w \in g(\underline{w})} p(w) \log \frac{p_p(w)}{p(w)} \end{aligned} \quad (73)$$

$$= \sum_{\underline{w}} \sum_{w \in g(\underline{w})} p(w) \log \frac{p(w)}{q(w)} \quad (74)$$

$$= D(p \| q) \geq 0. \quad (75)$$

The last inequality is the classical Gibbs inequality, with equality iff $p = q$ pointwise. Hence $D_p(p_p \| q) \geq D_p(p_p \| p)$.

For the third inequality,

$$\begin{aligned} D_p(p \| q_p) - D_p(p \| p_p) &= \sum_{\underline{w}} \sum_{w \in g(\underline{w})} p(w) \log \frac{p(w)}{q_p(w)} - \\ &\quad \sum_{\underline{w}} \sum_{w \in g(\underline{w})} p(w) \log \frac{p(w)}{p_p(w)} \end{aligned} \quad (76)$$

$$= \sum_{\underline{w}} \sum_{w \in g(\underline{w})} p(w) \log \frac{p_p(w)}{q_p(w)} \quad (77)$$

$$= D_p(p_p \| q_p) \geq 0. \quad (78)$$

Thus $D_p(p \| q_p) \geq D_p(p \| p_p)$. This completes the proof. $\square$

Corollary 3 (Ordering of Pragmatic Relative Entropies): The three forms of pragmatic relative entropy satisfy the following chain of inequalities:

$$D_p(p \| q_p) \leq D_p(p_p \| q_p) \leq D_p(p_p \| q). \quad (79)$$

Proof. Detailed derivations are analogous to the semantic case shown in [24, 25]. $\square$

Corollary 4 (Relation to Classical and Semantic Relative Entropies): Let $D(p \| q)$ denote the classical relative entropy, and let $D_s(p_s \| q_s)$, $D_s(p_s \| q)$, $D_s(p \| q_s)$ denote the full and partial semantic relative entropies as defined in [24, 25]. Then the following relationships hold:

$$D_p(p \| q_p) \leq D_s(p \| q_s) \leq D(p \| q) \leq D_s(p_s \| q) \leq D_p(p_p \| q), \quad (80)$$

where p_s and q_s denote the semantic aggregations, and p_p, q_p denote the pragmatic aggregations. More precisely, we have the two separate chains:

$$D_p(p \| q_p) \leq D_s(p \| q_s) \leq D(p \| q), \quad (81)$$

and

$$D(p \| q) \leq D_s(p_s \| q) \leq D_p(p_p \| q). \quad (82)$$

Proof. The semantic relative entropies are defined by aggregating over the semantic partition, which is finer than the pragmatic partition (since multiple semantic classes may merge into one pragmatic class). By the data processing inequality for relative entropy [45] (or by applying the chain of inequalities analogous to the semantic case), the coarser aggregation (pragmatic) yields a smaller or equal divergence when the first argument is aggregated and the second is fine-grained, and a larger or equal divergence when the first is fine-grained and the second is aggregated. The classical relative entropy lies between the semantic partial entropies as established in [24, 25]. Combining these facts yields the desired hierarchy. $\square$

Theorem 6 (Convexity): The pragmatic relative entropies $D_p(p_p \| q_p)$, $D_p(p_p \| q)$, and $D_p(p \| q_p)$ are convex in the pair of distributions (p, q) . That is, for any two pairs (p_1, q_1) and (p_2, q_2) , and for $0 \leq \theta \leq 1$, we have

$$\begin{cases}
D_p(\theta p_{p,1} + (1-\theta)p_{p,2} \| \theta q_{p,1} + (1-\theta)q_{p,2}) \\
\leq \theta D_p(p_{p,1} \| q_{p,1}) + (1-\theta)D_p(p_{p,2} \| q_{p,2}), \\
D_p(\theta p_{p,1} + (1-\theta)p_{p,2} \| \theta q_1 + (1-\theta)q_2) \\
\leq \theta D_p(p_{p,1} \| q_1) + (1-\theta)D_p(p_{p,2} \| q_2), \\
D_p(\theta p_1 + (1-\theta)p_2 \| \theta q_{p,1} + (1-\theta)q_{p,2}) \\
\leq \theta D_p(p_1 \| q_{p,1}) + (1-\theta)D_p(p_2 \| q_{p,2}).
\end{cases} \quad (83)$$

Proof. The proof follows from the joint convexity of the classical KL divergence and the fact that the pragmatic aggregation is a linear operation on probabilities. Each of the three pragmatic relative entropies can be expressed as a linear combination of classical KL divergences between appropriate fine-grained or aggregated distributions, and the convexity is inherited. Detailed derivations are analogous to the semantic case presented in [24, 25]. $\square$

Remark 3: The pragmatic relative entropy measures the difference between two probability distributions at the level of terminal actions. The full pragmatic relative entropy $D_p(p_p \| q_p)$ quantifies how much the decision-oriented distributions differ. The partial measures allow comparison between a decision-level distribution and a syntactic-level distribution, which is useful in scenarios where one belief is expressed in actions (e.g., a control policy) and the other in raw observations. The hierarchy in Eq. (80) shows that pragmatic abstraction—by discarding distinctions that do not affect the optimal action—reduces the divergence when comparing fine-grained against coarse-grained distributions, and increases the divergence when comparing coarse-grained against fine-grained, perfectly mirroring the information-theoretic processing inequality for relative entropy under coarse-graining.

B. Pragmatic Mutual Information

In semantic information theory, two distinct forms of mutual information—the up and down semantic mutual information—were introduced to account for the non-additive nature of semantic entropy [24, 25]. Analogously, we define up and down pragmatic mutual information, which capture the reduction in decision uncer-

tainty from different perspectives. Furthermore, we extend the framework to four single-sided pragmatic mutual information measures that characterize the information flow when only one side of the communication link is coarsened to the pragmatic level. These measures are essential for analyzing perception, expression, and reconstruction in task-oriented systems.

B.1 Definitions of Up and Down Pragmatic Mutual Information

Let $\mathcal{W}$ and $\mathcal{V}$ be two syntactic alphabets with a joint probability mass function $P(W, V)$, and let the associated pragmatic alphabets be $\underline{\mathcal{W}}$ and $\underline{\mathcal{V}}$ induced by the joint reification mapping

$$g_{WV} : \mathcal{W} \times \mathcal{V} \rightarrow 2^{\mathcal{W} \times \mathcal{V}}, \quad (84)$$

which partitions the joint syntactic space $\mathcal{W} \times \mathcal{V}$ into pragmatic equivalence classes. The pragmatic variables are then defined as $\underline{W} = g_{WV}^{-1}(W, V)$ and $\underline{V} = g_{WV}^{-1}(W, V)$. More precisely, the pragmatic pair $(\underline{W}, \underline{V})$ is the unique pragmatic class to which the syntactic pair (W, V) belongs. Equivalently, the joint reification mapping directly induces the joint pragmatic distribution by aggregating probabilities over the joint syntactic cells:

$$P(\underline{w}, \underline{v}) = \sum_{(w,v) \in g_{WV}(\underline{w}, \underline{v})} P(w, v). \quad (85)$$

The marginal pragmatic distributions are obtained by further summing over the other variable:

$$P(\underline{w}) = \sum_{\underline{v}'} P(\underline{w}, \underline{v}'), \quad P(\underline{v}) = \sum_{\underline{w}'} P(\underline{w}', \underline{v}). \quad (86)$$

Definition 12 (Pragmatic Mutual Information): The **up pragmatic mutual information** is defined as

$$\begin{aligned}
I^p(\underline{W}; \underline{V}) &\triangleq \sum_{w \in \mathcal{W}} \sum_{v \in \mathcal{V}} P(w, v) \log \frac{P(\underline{w}, \underline{v})}{P(\underline{w})P(\underline{v})} \\
&= H(W) + H(V) - H_p(\underline{W}, \underline{V}), \quad (87)
\end{aligned}$$

where $(\underline{w}, \underline{v}) = g_{WV}(w, v)$ is the pragmatic class assigned to the syntactic pair (w, v) .

The **down pragmatic mutual information** is defined as

$$I_p(\underline{W}; \underline{V}) \triangleq \sum_{w \in \mathcal{W}} \sum_{v \in \mathcal{V}} P(w, v) \log \frac{P(w)P(v)}{P(\underline{w}, \underline{v})}$$

$$= H_p(\underline{W}) + H_p(\underline{V}) - H(W, V), \quad (88)$$

We adopt the canonical definition for all subsequent rate-distortion and Lagrangian formulations: $I_p(\underline{W}; \underline{V}) = \max \{0, H_p(\underline{W}) + H_p(\underline{V}) - H(W, V)\} = (I_p)^+$. The positive-part operation ensures that the rate-distortion function $R_p(D)$ is strictly convex and bounded below by 0, resolving the ill-posed minimization that would otherwise arise.

The up pragmatic mutual information measures the reduction in the syntactic uncertainty of W and V when their joint pragmatic structure is taken into account. It can be interpreted as the maximum amount of information about the terminal actions that can be extracted from the syntactic variables. The down pragmatic mutual information measures the reduction in pragmatic uncertainty when the joint syntactic structure is known; it can be seen as the minimum amount of pragmatic dependence that survives after coarse-graining.

We now establish the fundamental relationships among the pragmatic, semantic, and syntactic mutual information measures. For reference, the semantic mutual informations are defined as [24, 25]:

$$I^s(\tilde{W}; \tilde{V}) = H(W) + H(V) - H_s(\tilde{W}, \tilde{V}), \quad (89)$$

$$I_s(\tilde{W}; \tilde{V}) = H_s(\tilde{W}) + H_s(\tilde{V}) - H(W, V). \quad (90)$$

Theorem 7 (Hierarchy of Mutual Information): For any two syntactic variables W, V with associated semantic variables $\tilde{W}, \tilde{V}$ and pragmatic variables $\underline{W}, \underline{V}$, the following chain of inequalities holds:

$$\begin{aligned} I_p(\underline{W}; \underline{V}) &\leq I_s(\tilde{W}; \tilde{V}) \leq I(W; V) \\ &\leq I^s(\tilde{W}; \tilde{V}) \leq I^p(\underline{W}; \underline{V}). \end{aligned} \quad (91)$$

Proof. The inequalities follow directly from the entropy hierarchy established in Section III. Specifically, $I_p \leq I_s$ because pragmatic variables are deterministic functions of semantic variables, implying $H_p(\underline{W}) \leq H_s(\tilde{W})$ and $H_p(\underline{V}) \leq H_s(\tilde{V})$; $I_s \leq I$ because $H_s(\tilde{W}) \leq H(W)$ and $H_s(\tilde{V}) \leq H(V)$; $I \leq I^s$ because $H_s(\tilde{W}, \tilde{V}) \leq H(W, V)$; and $I^s \leq I^p$ because $H_p(\underline{W}, \underline{V}) \leq H_s(\tilde{W}, \tilde{V})$. Combining these yields the desired chain. $\square$

Beyond the inequality chain, we can derive explicit additive decompositions that reveal the incre-

mental contributions of semantic and pragmatic coarse-graining.

Theorem 8 (Additive Decomposition of Up Pragmatic Mutual Information): The up pragmatic mutual information can be decomposed into the up semantic mutual information plus a pragmatic increment:

$$I^p(\underline{W}; \underline{V}) = I^s(\tilde{W}; \tilde{V}) + \Delta_p^{up}(\underline{W}, \underline{V}), \quad (92)$$

where the pragmatic increment is

$$\Delta_p^{up}(\underline{W}, \underline{V}) \triangleq H_s(\tilde{W}, \tilde{V}) - H_p(\underline{W}, \underline{V}) \geq 0. \quad (93)$$

Furthermore, the up pragmatic mutual information can also be decomposed relative to the classical mutual information:

$$I^p(\underline{W}; \underline{V}) = I(W; V) + \Delta_s^{up}(\tilde{W}, \tilde{V}) + \Delta_p^{up}(\underline{W}, \underline{V}), \quad (94)$$

where

$$\Delta_s^{up}(\tilde{W}, \tilde{V}) \triangleq H(W, V) - H_s(\tilde{W}, \tilde{V}) \geq 0, \quad (95)$$

is the semantic increment (the reduction in joint entropy due to semantic coarse-graining).

Proof. From the definitions, $I^p = H(W) + H(V) - H_p(\underline{W}, \underline{V})$. Adding and subtracting $H_s(\tilde{W}, \tilde{V})$:

$$\begin{aligned} I^p &= [H(W) + H(V) - H_s(\tilde{W}, \tilde{V})] + \\ &\quad [H_s(\tilde{W}, \tilde{V}) - H_p(\underline{W}, \underline{V})] \\ &= I^s + \Delta_p^{up}. \end{aligned} \quad (96)$$

For the second decomposition, add and subtract $H(W, V)$:

$$\begin{aligned} I^p &= [H(W) + H(V) - H(W, V)] \\ &\quad + [H(W, V) - H_s(\tilde{W}, \tilde{V})] \\ &\quad + [H_s(\tilde{W}, \tilde{V}) - H_p(\underline{W}, \underline{V})] \\ &= I + \Delta_s^{up} + \Delta_p^{up}. \end{aligned} \quad (97)$$

The non-negativity follows from the entropy hierarchy. $\square$

Theorem 9 (Additive Decomposition of Down Pragmatic Mutual Information): The down pragmatic mutual information can be decomposed into the down semantic mutual information minus a pragmatic loss:

$$I_p(\underline{W}; \underline{V}) = I_s(\tilde{W}; \tilde{V}) - \Delta_p^{down}(\underline{W}, \underline{V}), \quad (98)$$

where the pragmatic loss is

$$\Delta_p^{down}(\underline{W}, \underline{V}) \triangleq [H_s(\tilde{W}) - H_p(\underline{W})] + [H_s(\tilde{V}) - H_p(\underline{V})] \geq 0. \quad (99)$$

Furthermore, the down pragmatic mutual information can be decomposed relative to the classical mutual information:

$$I_p(\underline{W}; \underline{V}) = I(W; V) - \Delta_s^{down}(\tilde{W}, \tilde{V}) - \Delta_p^{down}(\underline{W}, \underline{V}), \quad (100)$$

where

$$\Delta_s^{down}(\tilde{W}, \tilde{V}) \triangleq [H(W) - H_s(\tilde{W})] + [H(V) - H_s(\tilde{V})] \geq 0, \quad (101)$$

is the semantic loss.

Proof. From $I_p = H_p(\underline{W}) + H_p(\underline{V}) - H(W, V)$. Adding and subtracting $H_s(\tilde{W})$ and $H_s(\tilde{V})$:

$$\begin{aligned} I_p &= [H_s(\tilde{W}) + H_s(\tilde{V}) - H(W, V)] \\ &\quad - [H_s(\tilde{W}) - H_p(\underline{W})] \\ &\quad - [H_s(\tilde{V}) - H_p(\underline{V})] \\ &= I_s - \Delta_p^{down}. \end{aligned} \quad (102)$$

For the classical decomposition, add and subtract $H(W)$ and $H(V)$:

$$\begin{aligned} I_p &= [H(W) + H(V) - H(W, V)] \\ &\quad - [H(W) - H_s(\tilde{W})] \\ &\quad - [H(V) - H_s(\tilde{V})] \\ &\quad - [H_s(\tilde{W}) - H_p(\underline{W})] \\ &\quad - [H_s(\tilde{V}) - H_p(\underline{V})] \\ &= I - \Delta_s^{down} - \Delta_p^{down}. \end{aligned} \quad (103)$$

The non-negativity follows from the entropy hierarchy. $\square$

Remark 4: The additive decompositions reveal a clear structural interpretation: the up pragmatic mutual information *adds* the information gains from each layer of

abstraction, while the down pragmatic mutual information *subtracts* the information losses. The classical mutual information serves as the central reference point. The inequalities $I_p \leq I_s \leq I \leq I^s \leq I^p$ follow immediately from the non-negativity of the increments/losses.

B.2 Single-Sided Pragmatic Mutual Information

In many engineering scenarios, the coarse-graining to the pragmatic level occurs on only one side of the communication link. For instance, in perception tasks, the source W is coarsened to $\underline{W}$ (the action), while the observation V remains syntactic. Conversely, in expressive or reconstructive tasks, the source remains syntactic while the reconstruction or observation is coarsened to $\underline{V}$ (the action label). To capture these asymmetric situations, we define four single-sided pragmatic mutual information measures.

Let W and V be syntactic variables with joint distribution $P(W, V)$, and let $\underline{W}$ and $\underline{V}$ be the corresponding pragmatic variables induced by the respective reification mappings. The following four measures are defined.

Definition 13 (Single-Sided Pragmatic Mutual Information):

1) **Perceptual Pragmatic Mutual Information (PPMI):**

$$\begin{aligned} I_p(\underline{W}; V) &\triangleq H_p(\underline{W}) + H(V) - H(\underline{W}, V) \\ &= H_p(\underline{W}) - H_p(\underline{W}|V) = I(\underline{W}; V), \end{aligned} \quad (104)$$

This measure quantifies how much information the syntactic observation V carries about the pragmatic action $\underline{W}$.

2) **Prospective Perceptual Pragmatic Mutual Information (PPPMI):**

$$I^p(\underline{W}; V) \triangleq H(W) + H(V) - H(\underline{W}, V), \quad (105)$$

This measure provides an upper bound on the perceptual information when the syntactic entropy of the source is retained in the marginal term.

3) **Expressive Pragmatic Mutual Information (EPMI):**

$$\begin{aligned} I_p(W; \underline{V}) &\triangleq H(W) + H_p(\underline{V}) - H(W, \underline{V}) \\ &= H(W) - H(W|\underline{V}) = I(W; \underline{V}), \end{aligned} \quad (106)$$

This measure quantifies how much information about the syntactic source W is contained in the pragmatic label $\underline{V}$.

4) Prospective Expressive Pragmatic Mutual Information (PEPMI):

$$I^P(W; \underline{V}) \triangleq H(W) + H(V) - H(W, \underline{V}), \quad (107)$$

This measure provides an upper bound on the expressive information when the syntactic entropy of the observation is retained.

The naming reflects their roles: “Perceptual” refers to the direction from observations to actions; “Expressive” refers to the direction from actions to source reconstruction. The prefix “Prospective” indicates that the measure uses the syntactic (larger) marginal entropy, thus providing an upper bound.

These four measures, together with the classical mutual information $I(W; V)$ and the two bilateral pragmatic mutual informations $I_p(\underline{W}; \underline{V})$ and $I^P(\underline{W}; \underline{V})$, form a complete lattice of information quantities that cover all combinations of coarse-graining (none, one side, both sides) and marginal entropy type (syntactic or pragmatic).

Theorem 10 (Hierarchy of Single-Sided and Bilateral Measures): For any syntactic variables W, V with associated pragmatic variables $\underline{W}, \underline{V}$, the following inequalities hold:

$$\begin{aligned} I_p(\underline{W}; \underline{V}) &\leq I_p(\underline{W}; V) \leq I(W; V) \\ &\leq I^P(\underline{W}; V) \leq I^P(\underline{W}; \underline{V}), \end{aligned} \quad (108)$$

$$\begin{aligned} I_p(\underline{W}; \underline{V}) &\leq I_p(W; \underline{V}) \leq I(W; V) \\ &\leq I^P(W; \underline{V}) \leq I^P(W; V). \end{aligned} \quad (109)$$

Moreover, the cross comparisons between the left chain (perceptual) and the right chain (expressive) are generally incomparable, except that any measure from the lower (pragmatic marginal) side is bounded above by any measure from the upper (syntactic marginal) side:

$$\begin{aligned} I_p(\underline{W}; V) &\leq I(W; V) \leq I^P(W; \underline{V}), \quad \text{and} \\ I_p(W; \underline{V}) &\leq I(W; V) \leq I^P(\underline{W}; V). \end{aligned} \quad (110)$$

Proof. The left chain in Eq. (108) follows from four successive inequalities. First, $I_p(\underline{W}; \underline{V}) \leq I_p(\underline{W}; V)$ be-

cause $\underline{V}$ is a function of V (data processing inequality). Second, $I_p(\underline{W}; V) \leq I(W; V)$ because $\underline{W}$ is a function of W . Third, $I(W; V) \leq I^P(\underline{W}; V)$ since $I^P(\underline{W}; V) = I(W; V) + [H(W) - H(\underline{W})] - H(W|\underline{W}, V) \geq I(W; V)$. Fourth, $I^P(\underline{W}; V) \leq I^P(\underline{W}; \underline{V})$ because $I^P(\underline{W}; \underline{V}) = I^P(\underline{W}; V) + [H(V) - H(\underline{V})] - H(V|\underline{W}, \underline{V}) \geq I^P(\underline{W}; V)$. The right chain in Eq. (109) follows by symmetry.

The cross bounds between perceptual and expressive measures are direct consequences of the fact that any single-sided measure with a pragmatic marginal entropy (PPMI or EPMI) is bounded above by the classical mutual information, while any single-sided measure with a syntactic marginal entropy (PPPMI or PEPMI) is bounded below by the classical mutual information. The incomparability between PPMI and EPMI (and similarly between PPPMI and PEPMI) can be demonstrated by constructing examples where one exceeds the other, depending on the relative losses of coarse-graining on each side. $\square$

B.3 Sequential Chain Rules for Pragmatic Mutual Information

We now extend the pragmatic mutual information to sequences, including the single-sided measures. Let (W^n, V^n) be a pair of syntactic sequences, and let $(\underline{W}^n, \underline{V}^n)$ be the corresponding pragmatic sequences obtained by applying the joint Reification Mapping element-wise. Define the sequential partial entropy as in Section III.B:

$$\begin{aligned} \tilde{H}_p(\underline{W}_1^m, W_{m+1}^n) &\triangleq \sum_{k=1}^m H_p(\underline{W}_k | \underline{W}_1^{k-1}) \\ &+ \sum_{k=m+1}^n H_p(W_k | \underline{W}_1^m, W_{m+1}^{k-1}), \end{aligned} \quad (111)$$

with analogous notation for mixed partial entropies involving $\underline{V}$ and V .

Theorem 11 (Sequential Chain Rule for Down Pragmatic Mutual Information): For a syntactic sequence pair (W^n, V^n) and its pragmatic counterpart, the down pragmatic mutual information satisfies the following chain of inequalities:

$$I_p(\underline{W}^n; \underline{V}^n) \leq I_p(\underline{W}_1^{n-1}, W_n; \underline{V}^n) \leq \dots$$

$$\leq I_p(\underline{W}_1, W_2^n; \underline{V}^n) \leq I(W^n; V^n), \quad (112)$$

where

$$\begin{aligned} I_p(\underline{W}_1^m, W_{m+1}^n; \underline{V}^n) &\triangleq \tilde{H}_p(\underline{W}_1^m, W_{m+1}^n) \\ &+ H_p(\underline{V}^n) - H(W^n, V^n). \end{aligned} \quad (113)$$

Proof. The proof follows the same steps as the semantic sequential chain rule (Theorem 8 in [24, 25]), with the semantic entropies replaced by pragmatic entropies. By the hierarchy of sequential entropies (Theorem 3 in Section III-B), we have

$$\tilde{H}_p(\underline{W}_1^m, W_{m+1}^n) \leq \tilde{H}_p(\underline{W}_1^{m-1}, W_m^n) \leq \dots \leq H(W^n). \quad (114)$$

Substituting these into the definition of I_p for the mixed blocks yields the desired chain. $\square$

Similarly, for the up pragmatic mutual information we have:

Theorem 12 (Sequential Chain Rule for Up Pragmatic Mutual Information):

$$\begin{aligned} I(W^n; V^n) &\leq I^p(\underline{W}_1, W_2^n; \underline{V}^n) \leq \dots \\ &\leq I^p(\underline{W}_1^{n-1}, W_n; \underline{V}^n) \leq I^p(\underline{W}^n; \underline{V}^n), \end{aligned} \quad (115)$$

where

$$\begin{aligned} I^p(\underline{W}_1^m, W_{m+1}^n; \underline{V}^n) &\triangleq H(W^n) + H(V^n) \\ &- \tilde{H}_p(\underline{W}_1^m, W_{m+1}^n, \underline{V}^n), \end{aligned} \quad (116)$$

with $\tilde{H}_p$ denoting the corresponding mixed sequential entropy.

Proof. Analogous to the semantic case, using the fact that replacing a syntactic variable by its pragmatic counterpart increases the up mutual information due to the decrease in the joint entropy term. $\square$

For the single-sided measures, we can also derive sequential chain rules. For example, the perceptual pragmatic mutual information $I_p(\underline{W}^n; V^n)$ can be bounded by sequentially replacing W 's with $\underline{W}$'s:

Theorem 13 (Sequential Chain Rule for Perceptual Pragmatic Mutual Information):

$$\begin{aligned} I_p(\underline{W}^n; V^n) &\leq I_p(\underline{W}_1^{n-1}, W_n; V^n) \leq \dots \\ &\leq I_p(\underline{W}_1, W_2^n; V^n) \leq I(W^n; V^n), \end{aligned} \quad (117)$$

where the mixed quantities are defined analogously using the corresponding conditional entropies.

Similarly, for the expressive pragmatic mutual information $I_p(W^n; \underline{V}^n)$, we have

$$\begin{aligned} I_p(W^n; \underline{V}^n) &\leq I_p(W^n; \underline{V}_1^{n-1}, V_n) \leq \dots \\ &\leq I_p(W^n; \underline{V}_1, V_2^n) \leq I(W^n; V^n). \end{aligned} \quad (118)$$

The up versions (PPPMI and PEPMI) satisfy the reverse chain inequalities, starting from the classical mutual information and increasing towards the upper bounds.

Theorem 14 (Sequential Chain Rule for Prospective Perceptual Pragmatic Mutual Information):

$$\begin{aligned} I(W^n; V^n) &\leq I^p(\underline{W}_1, W_2^n; V^n) \leq \dots \\ &\leq I^p(\underline{W}_1^{n-1}, W_n; V^n) \leq I^p(\underline{W}^n; V^n). \end{aligned} \quad (119)$$

Analogous chains hold for PEPMI by replacing V 's with $\underline{V}$'s in the mixed blocks.

Finally, the hierarchy of sequential mutual informations extends to all seven measures, with the same partial order as in the single-letter case, for any fixed n .

Example 3 (Multi-Vehicle Scenario): We now revisit the multi-vehicle autonomous driving scenario from Section III-B to compute the classical, semantic, and pragmatic mutual informations, and also the four single-sided measures. Recall the joint syntactic distribution $P(W, V)$ given in Table III. The syntactic states are traffic light colors, with four possible values: green, yellow, red-short, red-long. The semantic mapping groups green and yellow into “Proceed”, red-short into “Stop-short”, and red-long into “Stop-long”. The pragmatic mapping further merges the two stop classes into a single “Stop” action, as the utility function treats them identically. The joint Reification Mapping g_{WV} thus

maps the 4×4 syntactic pairs to a 2×2 pragmatic partition: “Go” (green/yellow combined) and “Stop” (both red types combined).

The syntactic joint distribution is given in Table III and the corresponding semantic and pragmatic joint distributions are given in Tables IV and V, respectively.

From earlier computations in Example 2, we can calculate the mixed entropies $H(W, \underline{V}) = H(\underline{W}, V) \approx 2.7328$ bits. Further, we need $H(\underline{W})$ and $H(\underline{V})$ which are both 0.9928 (since marginal pragmatic entropies equal). Now compute the single-sided measures:

1. PPMI: $I_p(\underline{W}; V) = H_p(\underline{W}) + H(V) - H(\underline{W}, V) = 0.9928 + 1.9893 - 2.7328 = 0.2493$ bits. 2. PPPMI: $I^p(\underline{W}; V) = H(W) + H(V) - H(\underline{W}, V) = 1.9893 + 1.9893 - 2.7328 = 1.2458$ bits. 3. EPMI: $I_p(W; \underline{V}) = H(W) + H_p(\underline{V}) - H(W, \underline{V}) = 1.9893 + 0.9928 - 2.7328 = 0.2493$ bits (same as PPMI due to symmetry in this example, but in general they differ). 4. PEPMI: $I^p(W; \underline{V}) = H(W) + H(V) - H(W, \underline{V}) = 1.9893 + 1.9893 - 2.7328 = 1.2458$ bits (same as PPPMI).

The classical mutual information $I(W; V) = 0.3472$ bits. The down bilateral $I_p(\underline{W}; \underline{V}) = 0$ (positive part). The up bilateral $I^p(\underline{W}; \underline{V}) = 2.2277$ prabits. The results are summarized in Table VII.

TABLE VII: Comparison of all mutual information measures for the two-vehicle scenario.

Mutual Information	Value (bits/sebits/prabits)
Down Pragmatic $I_p(\underline{W}; \underline{V})$	0 (positive part)
Down Semantic $I_s(\tilde{W}; \tilde{V})$	0 (positive part)
Perceptual Pragmatic $I_p(\underline{W}; V)$	0.2493 prabits
Expressive Pragmatic $I_p(W; \underline{V})$	0.2493 prabits
Classical $I(W; V)$	0.3472 bits
Prospective Perceptual $I^p(\underline{W}; V)$	1.2458 bits
Prospective Expressive $I^p(W; \underline{V})$	1.2458 bits
Up Semantic $I^s(\tilde{W}; \tilde{V})$	1.2223 sebits
Up Pragmatic $I^p(\underline{W}; \underline{V})$	2.2277 prabits

We observe that the single-sided measures lie between the down bilateral and the classical mutual information (for the lower ones) or between the classical and the up bilateral (for the upper ones), confirming the lattice structure. The equality of PPMI and EPMI in this symmetric example is coincidental; in general they differ.

This example demonstrates the utility of the single-sided measures in quantifying the information loss or

gain when only one side of the communication is coarsened to the pragmatic level. They provide a finer-grained analysis of task-oriented information flow compared to the bilateral measures alone.

V. Pragmatic Channel Capacity and Pragmatic Rate Distortion

We introduce the pragmatic channel capacity and rate-distortion function as the fundamental limits of pragmatic communication, paralleling their classical and semantic counterparts. These quantities characterize the maximum reliable transmission rate and the minimum rate for a given task distortion, respectively, and reveal performance gains from pragmatic abstraction. We further extend to single-sided perceptual and expressive measures for asymmetric sensing-actuation systems.

A. Pragmatic Channel Capacity

The pragmatic channel capacity quantifies the maximum rate at which pragmatic information—the optimal terminal actions—can be reliably transmitted over a given communication channel. It is defined as the supremum of the up pragmatic mutual information between the input pragmatic variable and the output pragmatic variable, maximized over the input distribution and the joint reification mapping.

Consider a discrete memoryless channel with input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$, and transition probability $P(Y|X)$. Let the associated semantic alphabets be $\tilde{\mathcal{X}}$ and $\tilde{\mathcal{Y}}$, and the pragmatic alphabets be $\underline{\mathcal{X}}$ and $\underline{\mathcal{Y}}$, induced by the joint reification mapping $g_{XY} : \underline{\mathcal{X}} \times \underline{\mathcal{Y}} \rightarrow 2^{\mathcal{X} \times \mathcal{Y}}$, which partitions the joint syntactic space into pragmatic equivalence classes.

Definition 14 (Pragmatic Channel Capacity): The **pragmatic channel capacity** is defined as

$$C_p \triangleq \max_{p(x)} \max_{g_{XY}} I^p(\underline{X}; \underline{Y}), \quad (120)$$

where $I^p(\underline{X}; \underline{Y}) = H(X) + H(Y) - H_p(\underline{X}, \underline{Y})$ is the up pragmatic mutual information, and the maximization is over all input distributions $p(x)$ and all joint reification mappings g_{XY} .

For reference, the semantic channel capacity [24, 25] and the classical (syntactic) channel capacity

ity [1, 45] are defined as:

$$\begin{cases} C_s \triangleq \max_{p(x)} \max_{f_{XY}} I^s(\tilde{X}; \tilde{Y}), \\ C \triangleq \max_{p(x)} I(X; Y), \end{cases} \quad (121)$$

where f_{XY} denotes the joint synonymous mapping.

The following theorem establishes the hierarchy among the three capacities.

Theorem 15 (Hierarchy of Channel Capacities): The pragmatic, semantic, and syntactic channel capacities satisfy:

$$C \leq C_s \leq C_p. \quad (122)$$

Proof. The inequality $C \leq C_s$ is a known result of semantic information theory [24, 25], following from the fact that $I(X; Y) \leq I^s(\tilde{X}; \tilde{Y})$ for any joint synonymous mapping.

To prove $C_s \leq C_p$, note that for any input distribution and any joint Synonymous Mapping f_{XY} , we can define a corresponding joint Reification Mapping g_{XY} that is coarser than f_{XY} (i.e., it merges semantic classes that lead to the same optimal action). For such g , we have $H_p(\underline{X}, \underline{Y}) \leq H_s(\tilde{X}, \tilde{Y})$, hence

$$\begin{aligned} I^p(\underline{X}; \underline{Y}) &= H(X) + H(Y) - H_p(\underline{X}, \underline{Y}) \\ &\geq H(X) + H(Y) - H_s(\tilde{X}, \tilde{Y}) = I^s(\tilde{X}; \tilde{Y}). \end{aligned} \quad (123)$$

Therefore, the maximum over g_{XY} of I^p is at least the maximum over f_{XY} of I^s , yielding $C_p \geq C_s$. $\square$

Beyond the inequality, we can quantify the capacity gains explicitly.

Theorem 16 (Additive Decomposition of Pragmatic Capacity): The pragmatic channel capacity can be decomposed as

$$C_p = C + \Delta_{ps}^C, \quad (124)$$

where

$$\Delta_{ps}^C \triangleq \Delta_s^C + \Delta_p^C \geq 0, \quad (125)$$

with

$$\begin{cases} \Delta_s^C \triangleq \max_{p(x)} \max_{f_{XY}} [H(X, Y) - H_s(\tilde{X}, \tilde{Y})] \geq 0, \\ \Delta_p^C \triangleq \max_{p(x)} \max_{g_{XY}} [H_s(\tilde{X}, \tilde{Y}) - H_p(\underline{X}, \underline{Y})] \geq 0. \end{cases} \quad (126)$$

Equivalently, the pragmatic capacity can also be expressed relative to the semantic capacity as

$$C_p = C_s + \Delta_p^C. \quad (127)$$

Proof. From the additive relation of the up mutual information:

$$I^p = I^s + \Delta_p^{up}, \quad (128)$$

with $\Delta_p^{up} = H_s(\tilde{X}, \tilde{Y}) - H_p(\underline{X}, \underline{Y})$. Similarly, $I^s = I + \Delta_s^{up}$, where $\Delta_s^{up} = H(X, Y) - H_s(\tilde{X}, \tilde{Y})$. Combining these and taking the maximum over $p(x)$, f_{XY} , and g_{XY} yields the decomposition. $\square$

Remark 5: The classical capacity C is the fundamental physical limit determined by the channel transition probability $P(Y|X)$, representing the maximum rate at which syntactic symbols can be reliably transmitted. The additional semantic gain Δ_s^C and pragmatic gain Δ_p^C are not violations of this physical limit; rather, they reflect the fact that semantic and pragmatic communication systems relax the requirement of exact symbol reconstruction. By allowing errors that do not affect meaning (semantic) or that do not affect the optimal terminal action (pragmatic), the system can effectively transmit more “useful” information per physical bit. The total pragmatic capacity $C_p = C + \Delta_{ps}^C$ thus represents the capacity measured in pragmatic bits—the maximum rate of reliably conveying terminal actions—which can exceed the syntactic bit rate C because each pragmatic bit may correspond to multiple syntactic bits that are semantically or pragmatically equivalent.

B. Pragmatic Rate Distortion

The pragmatic rate-distortion function characterizes the minimum rate required to describe a source such that the resulting distortion, measured in terms of task utility loss, does not exceed a given threshold. Unlike classical rate distortion, which uses a syntactic distortion measure (e.g., mean squared error), the pragmatic distortion is defined directly on the terminal actions.

Let $X \sim p(x)$ be a syntactic source, and let $\tilde{X}$ and $\underline{X}$ be the associated semantic and pragmatic variables induced by the Synonymous and Isoteleia mappings. The decoder produces a reconstruction $\hat{X}$ with associated semantic $\hat{\tilde{X}}$ and pragmatic $\hat{\underline{X}}$, where the reconstructions

are obtained via a test channel $p(\hat{x}|x)$. The pragmatic distortion measure $d_p(\underline{x}, \hat{x})$ quantifies the cost of representing the pragmatic symbol $\underline{x}$ by $\hat{x}$; in practice, it can be defined as the utility loss:

$$d_p(\underline{x}, \hat{x}) = \max_a \mathbb{E}[U(X, a)|\underline{x}] - \max_a \mathbb{E}[U(X, a)|\hat{x}], \quad (129)$$

or any other non-negative function that reflects the decision degradation.

Definition 15 (Pragmatic Rate Distortion): The **pragmatic rate-distortion function** is defined as

$$R_p(D) \triangleq \min_{g_X, g_{\hat{X}}} \min_{p(\hat{x}|x): \mathbb{E}[d_p(\underline{X}, \hat{X})] \leq D} I_p(\underline{X}; \hat{X}), \quad (130)$$

where $I_p(\underline{X}; \hat{X}) = H_p(\underline{X}) + H_p(\hat{X}) - H(X, \hat{X})$ is the down pragmatic mutual information ($I_p(\underline{X}; \hat{X}) = (I_p)^+$ is non-negative), and the minimization is over all joint reification mappings g_X and $g_{\hat{X}}$ for the source and reconstruction, and over all test channels satisfying the distortion constraint.

The semantic rate-distortion function $R_s(D)$ [24, 25] and the classical rate-distortion function $R(D)$ [1, 45] are defined analogously:

$$\begin{cases} R_s(D) \triangleq \min_{f_X, f_{\hat{X}}} \min_{p(\hat{x}|x): \mathbb{E}[d_s(\tilde{X}, \hat{X})] \leq D} I_s(\tilde{X}; \hat{X}), \\ R(D) \triangleq \min_{p(\hat{x}|x): \mathbb{E}[d(X, \hat{X})] \leq D} I(X; \hat{X}), \end{cases} \quad (131)$$

where d_s is a semantic distortion measure (e.g., based on meaning), and d is a syntactic distortion measure (e.g., squared error). The following hierarchy holds.

Theorem 17 (Hierarchy of Rate-Distortion Functions): For any $D \geq 0$,

$$R_p(D) \leq R_s(D) \leq R(D). \quad (132)$$

Proof. The inequality $R_s(D) \leq R(D)$ is a known result of semantic information theory [24, 25], following from the fact that $I_s(\tilde{X}; \hat{X}) \leq I(X; \hat{X})$.

To prove $R_p(D) \leq R_s(D)$, note that for any Synonymous Mappings $f_X, f_{\hat{X}}$ and any test channel, we can define coarser Reification Mappings $g_X, g_{\hat{X}}$ that merge semantic classes with identical optimal actions. For such coarser mappings, we have

$$I_p(\underline{X}; \hat{X}) = I_s(\tilde{X}; \hat{X}) - \Delta_p^{down}, \quad (133)$$

where $\Delta_p^{down} = [H_s(\tilde{X}) - H_p(\underline{X})] + [H_s(\hat{X}) - H_p(\hat{X})] \geq 0$. Therefore, the minimum over $g_X, g_{\hat{X}}$ of I_p is no larger than the minimum over $f_X, f_{\hat{X}}$ of I_s , provided the distortion constraints are compatible (which holds by defining d_p appropriately from d_s). Hence $R_p(D) \leq R_s(D)$. $\square$

The gains can also be expressed additively.

Theorem 18 (Additive Decomposition of Pragmatic Rate Distortion): The pragmatic rate distortion function can be written as

$$R_p(D) = R(D) - \Delta_{ps}^R(D), \quad (134)$$

where

$$\Delta_{ps}^R(D) \triangleq \Delta_s^R(D) + \Delta_p^R(D) \geq 0, \quad (135)$$

with

$$\begin{cases} \Delta_s^R(D) \triangleq \max_{f_X, f_{\hat{X}}} \{ [H(X) - H_s(\tilde{X})] \\ \quad + [H(\hat{X}) - H_s(\hat{X})] \} \geq 0, \\ \Delta_p^R(D) \triangleq \max_{g_X, g_{\hat{X}}} \{ [H_s(\tilde{X}) - H_p(\underline{X})] \\ \quad + [H_s(\hat{X}) - H_p(\hat{X})] \} \geq 0. \end{cases} \quad (136)$$

Equivalently, the pragmatic rate distortion can be expressed relative to the semantic rate distortion as

$$R_p(D) = R_s(D) - \Delta_p^R(D). \quad (137)$$

Proof. From the additive relation $I_p = I_s - \Delta_p^{down}$, where Δ_p^{down} is the pragmatic loss defined above, and $I_s = I - \Delta_s^{down}$, combining these and taking the minimum over the test channel and the mappings yields the decomposition. $\square$

Remark 6: The classical rate-distortion function $R(D)$ represents the minimum syntactic bit rate required to achieve a given syntactic distortion. The semantic loss $\Delta_s^R(D)$ and pragmatic loss $\Delta_p^R(D)$ reflect the reductions in rate achievable by allowing distortions at the semantic and pragmatic levels, respectively. These reductions reflect the fact that the distortion measure itself has been redefined to capture task-relevant fidelity rather

than symbol-level fidelity. By tolerating syntactic or semantic errors that do not affect the terminal action, the pragmatic rate-distortion function $R_p(D)$ can be significantly lower than $R(D)$, providing a principled theoretical foundation for task-oriented compression.

C. Single-Sided Pragmatic Achievable Rates and Rate-Distortion Functions

The bilateral pragmatic capacity and rate-distortion functions defined above assume that both the source and the reconstruction (or input and output) are coarsened to the pragmatic level. However, in many practical systems—such as perception, sensing, or actuation—only one side of the communication link is subject to pragmatic abstraction. To address such asymmetric scenarios, we introduce four single-sided measures: two rate-distortion functions (perceptual and expressive) and two achievable rates (prospective perceptual and prospective expressive). These measures provide a finer-grained analysis of task-oriented information flows when either the observation or the action label is coarsened.

C.1 Perceptual Pragmatic Rate-Distortion Function

The perceptual pragmatic rate-distortion function captures the fundamental limit of compressing an observation V to extract information about the pragmatic action $\underline{W}$, where the source W itself remains syntactic. It is defined using the down perceptual pragmatic mutual information $I_p(\underline{W}; V)$.

Definition 16 (Perceptual Pragmatic Rate-Distortion Function): Let $\underline{W}$ be the pragmatic variable associated with the source W , and let V be an observation. The **perceptual pragmatic rate-distortion function** is defined as

$$R_p^{\text{perc}}(D) \triangleq \min_{P(\hat{W}|V): \mathbb{E}[d_p(\underline{W}, \hat{W})] \leq D} I_p(\underline{W}; V), \quad (138)$$

where $\hat{W}$ is an estimate of $\underline{W}$ based on V , and d_p is a distortion measure on pragmatic symbols.

This function quantifies the minimum rate (in bits per observation) required to achieve a given perceptual distortion D when the receiver is only interested in the optimal action class. It is particularly relevant in perception systems where the goal is to infer the correct action from sensory data, and the raw sensory stream

can be compressed by exploiting the action equivalence classes.

C.2 Expressive Pragmatic Rate-Distortion Function

The expressive pragmatic rate-distortion function captures the fundamental limit of representing a syntactic source W using a pragmatic label $\underline{V}$ (e.g., a control command or action label) that is then used to reconstruct the original source. It is defined using the down expressive pragmatic mutual information $I_p(W; \underline{V})$.

Definition 17 (Expressive Pragmatic Rate-Distortion Function): Let W be a syntactic source, and let $\underline{V}$ be a pragmatic variable (action label) that is a function of W . The **expressive pragmatic rate-distortion function** is defined as

$$R_p^{\text{expr}}(D) \triangleq \min_{P(\hat{W}|\underline{V}): \mathbb{E}[d_s(W, \hat{W})] \leq D} I_p(W; \underline{V}), \quad (139)$$

where $\hat{W}$ is a reconstruction of W based on $\underline{V}$, and d_s is a distortion measure on the syntactic symbols (or on the corresponding semantics).

This function quantifies the minimum rate (in bits per action label) required to achieve a given reconstruction distortion when the only information available about the source is the pragmatic action label. It is relevant in actuation or control systems where a low-bandwidth command channel must convey enough information to reconstruct the intended state or message.

C.3 Prospective Perceptual Pragmatic Achievable Rate

The prospective perceptual pragmatic achievable rate is the dual of the perceptual rate-distortion function. It quantifies the maximum rate at which information about the pragmatic action $\underline{W}$ can be reliably conveyed through an observation V , when the receiver has access to the full syntactic observation but is restricted to using only the pragmatic information. It is defined using the up perceptual pragmatic mutual information $I^p(\underline{W}; V)$.

Definition 18 (Prospective Perceptual Pragmatic Achievable Rate): Let $\underline{W}$ be the pragmatic variable and V the observation. The **prospective perceptual pragmatic achievable rate** is defined as

$$C_p^{\text{p-perc}}(P) \triangleq \max_{P(v|\underline{w}): \mathbb{E}[c(\underline{W})] \leq P} I^p(\underline{W}; V), \quad (140)$$

where $c(W)$ is a cost function on the pragmatic source (e.g., sensing power or bandwidth), and the maximization is over all conditional distributions $p(v|w)$ satisfying the resource constraint.

This achievable rate represents the maximum amount of pragmatic information that can be extracted from the observation when the observation itself is not coarsened. It provides an upper bound on the rate at which a sensor can convey actionable information, given a resource budget. The superscript “p-perc” denotes “prospective perceptual”.

C.4 Prospective Expressive Pragmatic Achievable Rate

The prospective expressive pragmatic achievable rate is the dual of the expressive rate-distortion function. It quantifies the maximum rate at which a pragmatic label $\underline{V}$ can convey information about the syntactic source W , when the receiver has access to the full syntactic source but is restricted to using only the pragmatic label. It is defined using the up expressive pragmatic mutual information $I^P(W; \underline{V})$.

Definition 19 (Prospective Expressive Pragmatic Achievable Rate): Let W be the syntactic source and $\underline{V}$ the pragmatic label. The **prospective expressive pragmatic achievable rate** is defined as

$$C_p^{\text{p-expr}}(P) \triangleq \max_{p(\underline{v}|w): \mathbb{E}[c(W)] \leq P} I^P(W; \underline{V}), \quad (141)$$

where $c(W)$ is a cost function on the source (e.g., transmit power), and the maximization is over all conditional distributions $p(\underline{v}|w)$ satisfying the resource constraint.

This achievable rate represents the maximum rate at which a command or action label can convey information about the source, given a resource budget. It is relevant in scenarios where a low-dimensional action space is used to encode high-dimensional source information, and the system designer wants to know the best possible expressive performance under resource limits. The superscript “p-expr” denotes “prospective expressive”.

The four single-sided measures are related through the data processing inequality and the mutual information hierarchy, yielding the following bounds:

$$\begin{cases} R_p^{\text{perc}}(D) \geq R_p(D), \\ R_p^{\text{expr}}(D) \geq R_p(D), \\ C_p^{\text{p-perc}}(P) \leq C_p(P), \\ C_p^{\text{p-expr}}(P) \leq C_p(P). \end{cases} \quad (142)$$

These inequalities indicate that coarsening only one side preserves more syntactic distinctions, thus requiring higher rates for the same distortion and yielding lower achievable rates for the same resource.

Such single-sided measures are particularly useful in asymmetric systems, including:

- *Perceptual* (e.g., cameras, LIDAR): inferring actions from observations without coarsening the observation.
- *Expressive* (e.g., remote control, teleoperation): conveying intent via low-dimensional action labels.
- *Edge AI*: feature extraction at sensors (perceptual) and reconstruction at the central processor (expressive).

Together with their bilateral counterparts, these measures provide a complete pragmatic information theoretic framework for designing task-oriented communication systems with asymmetric constraints.

We conclude this section with a conceptual clarification that is essential for interpreting the hierarchies above.

Remark 7: Throughout this paper, the pragmatic rate R_p and pragmatic capacity C_p are *syntax-layer* rate measures (bits per channel use), while their reliability criterion is defined at the pragmatic layer (action correctness). Classical capacity C corresponds to the criterion of *message correctness*; pragmatic capacity C_p corresponds to the criterion of *action correctness*. Since the latter is strictly weaker, $C_p \geq C$ and $I^p > H(X)$ are admissible and do not contradict the data processing inequality. The same logic applies to the pragmatic rate-distortion function $R_p(D)$: it measures the minimum syntax-layer rate required to satisfy a pragmatic distortion criterion, hence $R_p(D) \leq R_s(D) \leq R(D)$ is consistent with the data processing inequality and does not imply that pragmatic distortion is less informative than

syntactic distortion. The hierarchy $H_p \leq H_s \leq H$ is a statement about source uncertainty, whereas $C_p \geq C_s \geq C$ and $R_p(D) \leq R_s(D) \leq R(D)$ are statements about syntax-layer rates under progressively relaxed criteria.

VI. Pragmatic Cost of Information and Pragmatic Value of Information

In this section, we introduce the pragmatic value of information (VoI) and the pragmatic cost of information (CoI) as the benefit and resource-cost duals to the rate-distortion and capacity functions, respectively. We then combine these dualities into a Lagrangian framework for cross-layer optimization of pragmatic communication systems.

A. Pragmatic Value of Information

The concept of the value of information was originally introduced by Stratonovich [7, 8] in the context of statistical decision theory, where it measures the maximum expected utility gain that a decision-maker can obtain by acquiring additional information before making a decision. Classical VoI is defined as the difference between the expected utility with the observation and the expected utility without it, providing a quantitative basis for evaluating whether information acquisition is worthwhile.

The pragmatic value of information extends this classical notion to the pragmatic level by considering both the state and the observation through the lens of the isoteleia mapping. Unlike the classical VoI, which treats the raw state X and raw observation Y as the basis for decision-making, the pragmatic VoI recognizes that what ultimately matters is not the exact state or the exact observation, but their pragmatic implications—the optimal actions they induce. Both the state and the observation are therefore coarsened to their pragmatic equivalence classes via the isoteleia mapping, reflecting the principle of equifinality: distinct states or observations that lead to the same optimal action are pragmatically equivalent.

Let $X \in \mathcal{X}$ be the state variable, $A \in \mathcal{A}$ the action, and $U(X, A)$ the utility function. Let $\underline{X}$ and $\underline{Y}$ be the pragmatic variables induced by the isoteleia mappings $e_X : \mathcal{X} \rightarrow 2^{\tilde{\mathcal{X}}}$ and $e_Y : \mathcal{Y} \rightarrow 2^{\tilde{\mathcal{Y}}}$, respectively (com-

posed with the synonymous mappings). The pragmatic variable $\underline{X}$ groups states that lead to the same optimal action, while $\underline{Y}$ groups observations that are pragmatically equivalent. The receiver, upon observing the pragmatic observation $\underline{Y}$, chooses the action that maximizes the expected utility conditioned on $\underline{Y}$, where the expectation is taken over the pragmatic state $\underline{X}$:

$$a^*(\underline{Y}) = \arg \max_{a \in \mathcal{A}} \mathbb{E}_{\underline{X}|\underline{Y}}[U(X, a)], \quad (143)$$

where X is any representative state in the pragmatic class $\underline{X}$, and the expectation is well-defined because all states in the same pragmatic class yield the same optimal action.

The baseline utility, without any information, is obtained by choosing the prior-optimal action based on the pragmatic state distribution:

$$a_0^* = \arg \max_{a \in \mathcal{A}} \mathbb{E}_{\underline{X}}[U(X, a)], \quad U_0 = \mathbb{E}_{\underline{X}}[U(X, a_0^*)]. \quad (144)$$

Definition 20 (Pragmatic Value of Information): The pragmatic value of information is defined as the expected utility gain obtained when the receiver makes decisions based on the pragmatic observation $\underline{Y}$, where both the state and the observation have been coarsened to their pragmatic equivalence classes:

$$\text{VoI}_p \triangleq \mathbb{E}_{\underline{X}, \underline{Y}}[U(X, a^*(\underline{Y}))] - U_0, \quad (145)$$

where the expectation is taken over the joint distribution of the pragmatic state and pragmatic observation, and X is any representative state in the pragmatic class $\underline{X}$. This measures the utility gain when both the state and the observation are evaluated at the pragmatic level, that is, when decisions are based on the optimal action classes rather than on raw states or observations.

This bilateral coarse-graining is essential for a proper pragmatic characterization of information value. By coarsening both the state and the observation, the pragmatic VoI captures the decision-relevant information that survives after discarding all distinctions that do not affect the optimal action. The classical VoI, which operates on raw states and observations, is a special case of the pragmatic VoI when the isoteleia mappings are trivial (identity).

The pragmatic VoI satisfies the following basic properties.

Theorem 19 (Non-negativity): $\text{VoI}_p \geq 0$, with equality if and only if the pragmatic observation $\underline{Y}$ provides no useful information for improving the decision, i.e., $a^*(\underline{Y}) = a_0^*$ almost surely.

Proof. For any realization of $\underline{Y}$, the optimal posterior action yields an expected utility at least as high as the prior-optimal action:

$$\max_a \mathbb{E}_{\underline{X}|\underline{Y}}[U(X, a)] \geq \mathbb{E}_{\underline{X}|\underline{Y}}[U(X, a_0^*)]. \quad (146)$$

Taking expectation over $\underline{Y}$ and subtracting $U_0 = \mathbb{E}_{\underline{X}}[U(X, a_0^*)]$ yields the result. $\square$

Theorem 20 (Monotonicity under Refinement): If the pragmatic partition of the observation is refined (i.e., $\underline{Y}_2$ is a finer pragmatic partition than $\underline{Y}_1$), then

$$\text{VoI}_p(\underline{Y}_1) \leq \text{VoI}_p(\underline{Y}_2). \quad (147)$$

This follows directly from the data processing inequality for utility-based decisions, as a finer pragmatic observation provides no less information about the pragmatic state than a coarser one.

Theorem 21 (Concavity in Rate): When the pragmatic observation is obtained through a channel with rate R , the maximum achievable pragmatic value, defined as

$$\Phi_p(R) \triangleq \max_{P(\underline{Y}|\underline{X}): I_p(\underline{X}; \underline{Y}) \leq R} \text{VoI}_p, \quad (148)$$

where $I_p(\underline{X}; \underline{Y})$ is the down pragmatic mutual information between the pragmatic state and the pragmatic observation, is a non-decreasing concave function of R .

Lemma 4 (Constructive Concavity of the Upper Concave Envelope): Let

$$\mathcal{R} = \{(R, V) \mid R \geq 0, V \leq \Phi_p(R)\}$$

be the achievable rate-value region of the pragmatic information system, where $\Phi_p(R)$ is the original (possibly non-concave) value function. Define its upper concave envelope (i.e., the upper boundary of the convex hull) as

$$\overline{\Phi}_p(R) \triangleq \sup\{V \mid (R, V) \in \text{Conv}(\mathcal{R})\}, \quad (149)$$

where $\text{Conv}(\mathcal{R})$ denotes the convex hull of $\mathcal{R}$. Then $\overline{\Phi}_p(R)$ is a non-decreasing concave function on $R \geq 0$,

and it satisfies $\overline{\Phi}_p(R) \geq \Phi_p(R)$ for all R . If $\mathcal{R}$ is already convex, then equality holds.

Proof. Let (R_1, V_1) and (R_2, V_2) be any two points in $\text{Conv}(\mathcal{R})$ lying on the upper boundary, i.e., $V_1 = \overline{\Phi}_p(R_1)$ and $V_2 = \overline{\Phi}_p(R_2)$. For any $\lambda \in [0, 1]$, consider the convex combination

$$\begin{aligned} (R_\lambda, V_\lambda) &= \lambda(R_1, V_1) + (1 - \lambda)(R_2, V_2) \\ &= (\lambda R_1 + (1 - \lambda)R_2, \lambda V_1 + (1 - \lambda)V_2). \end{aligned}$$

By definition of the convex hull, $(R_\lambda, V_\lambda) \in \text{Conv}(\mathcal{R})$. Hence, by the supremum property of $\overline{\Phi}_p(R_\lambda)$, we have

$$\overline{\Phi}_p(\lambda R_1 + (1 - \lambda)R_2) \geq \lambda \overline{\Phi}_p(R_1) + (1 - \lambda) \overline{\Phi}_p(R_2),$$

which is exactly the definition of concavity. Monotonicity is obvious because a larger rate allows a larger set of achievable values; and since the convex hull contains the original set, $\overline{\Phi}_p(R) \geq \Phi_p(R)$. If $\mathcal{R}$ is itself convex, its convex hull coincides with $\mathcal{R}$, so equality holds. $\square$

Remark 8: The key distinction between the pragmatic VoI and the classical VoI lies in the level of abstraction at which both the state and the observation are evaluated. Classical VoI treats every distinction in the state and observation as potentially valuable; pragmatic VoI recognizes that only distinctions that affect the optimal action matter. By coarsening both the state and the observation to their pragmatic equivalence classes, the pragmatic VoI provides a more concise and task-relevant measure of information value.

The pragmatic value of information is the natural decision-theoretic dual to the pragmatic rate-distortion function. Recall from Section V-B that the pragmatic rate-distortion function is defined as $R_p(D) = \min_{P(\hat{X}|\underline{X}): \mathbb{E}[d_p(\underline{X}, \hat{X})] \leq D} I_p(\underline{X}; \hat{X})$, where d_p is a distortion measure defined on pragmatic symbols.

The duality manifests through the following variational relationship.

Theorem 22 (Variational Duality): For a given utility function U , there exists a concave function Ψ_p such that

$$\Phi_p(R) = \max_{D \geq 0} \{\Psi_p(D) - \lambda R_p(D)\}, \quad (150)$$

where $\lambda > 0$ is a Lagrange multiplier. Conversely,

$$R_p(D) = \max_{\lambda \geq 0} \{ \lambda D - \Phi_p^*(\lambda) \}, \quad (151)$$

where Φ_p^* is the concave conjugate of Φ_p . This establishes a Legendre – Fenchel duality between the value-rate function $\Phi_p(R)$ and the rate-distortion function $R_p(D)$.

Example 4 (Autonomous Vehicle (Continued)): We revisit the two-vehicle autonomous driving scenario from Sections III.B and IV.B. The state X is the traffic light state (four possible values), and the observation Y is the corresponding signal. The pragmatic mapping merges “Stop-short” and “Stop-long” into a single “Stop” action. The utility is 10 for correct action, 0 otherwise. The baseline utility (prior-optimal action “Stop”) is $U_0 = 11$ (combined for two vehicles). The computed pragmatic VoI is $\text{VoI}_p \approx 0.5$ utils. This low value reflects the fact that the binary pragmatic observation (Go/Stop) resolves little decision uncertainty. In contrast, the classical VoI (using full syntactic signal) is about 4.2 utils.

The duality with rate-distortion can be illustrated by noting that to achieve a pragmatic distortion D (e.g., probability of wrong action), the required rate $R_p(D)$ must be at least the rate that enables the corresponding VoI. For instance, if we require $\text{VoI}_p \geq 4.0$, the rate must exceed the value $R_p(D^*)$ where D^* is the distortion corresponding to that utility level.

B. Pragmatic Cost of Information

While the value of information quantifies the benefit, the pragmatic cost of information quantifies the resource expenditure required to convey information at a given pragmatic rate. Unlike VoI, which depends on the utility function and the task, CoI is determined solely by the physical communication channel and the pragmatic abstraction (the isoteleic volume).

Consider a communication channel with input X , output Y , and a joint reification mapping $g_{XY} : \underline{X} \times \underline{Y} \rightarrow 2^{\underline{X} \times \underline{Y}}$ that defines the pragmatic equivalence classes. Let the resource consumption be measured by a cost function $c(x)$ on the channel input (e.g., transmit power $\mathbb{E}[|X|^2]$, or for discrete channels, the number of channel uses or energy per symbol).

Lemma 5 (Concavity of the Pragmatic Capacity – Re-

source Function): Let $P \geq 0$ denote the available communication resource (e.g., power) and define the pragmatic capacity under resource constraint P as

$$C_p(P) \triangleq \max_{p(x) : \mathbb{E}[c(X)] \leq P} I^p(\underline{X}; \underline{Y}), \quad (152)$$

where $c(X)$ is a non-negative cost function and $I^p(\underline{X}; \underline{Y})$ is the up-pragmatic mutual information. Then $C_p(P)$ is a non-decreasing concave function of P .

Proof. For fixed channel and pragmatic mapping, the mutual information $I^p(\underline{X}; \underline{Y})$ is a concave function of the input distribution $p(x)$ (this follows from the joint convexity of the KL divergence and the fact that the pragmatic aggregation is a linear operation on probabilities). The resource constraint $\mathbb{E}[c(X)] \leq P$ defines a convex set of input distributions.

Let $P_1, P_2 \geq 0$ and $\lambda \in [0, 1]$. Choose distributions p_1 and p_2 that attain the maxima for P_1 and P_2 , respectively, i.e.,

$$C_p(P_i) = I^p(p_i) \quad (i = 1, 2), \quad \mathbb{E}_{p_i}[c(X)] \leq P_i.$$

Consider the mixture distribution $p_\lambda = \lambda p_1 + (1 - \lambda)p_2$. Its average cost is

$$\begin{aligned} \mathbb{E}_{p_\lambda}[c(X)] &= \lambda \mathbb{E}_{p_1}[c(X)] + (1 - \lambda) \mathbb{E}_{p_2}[c(X)] \\ &\leq \lambda P_1 + (1 - \lambda) P_2, \end{aligned}$$

so p_λ is feasible for the resource level $\lambda P_1 + (1 - \lambda) P_2$. By concavity of I^p in the distribution,

$$\begin{aligned} I^p(p_\lambda) &\geq \lambda I^p(p_1) + (1 - \lambda) I^p(p_2) \\ &= \lambda C_p(P_1) + (1 - \lambda) C_p(P_2). \end{aligned}$$

Since $C_p(\lambda P_1 + (1 - \lambda) P_2)$ is the maximum over all feasible distributions, we have

$$\begin{aligned} C_p(\lambda P_1 + (1 - \lambda) P_2) &\geq I^p(p_\lambda) \\ &\geq \lambda C_p(P_1) + (1 - \lambda) C_p(P_2). \end{aligned}$$

Thus C_p is concave. Monotonicity follows immediately from the fact that a larger resource set contains smaller ones. $\square$

The pragmatic cost of information is defined as the inverse of this capacity-resource relationship.

Definition 21 (Pragmatic Cost of Information): For a given pragmatic rate R , the pragmatic cost of information is the minimum resource consumption required to achieve that rate:

$$\text{CoI}_p(R) \triangleq \min_{p(x): I^p(\underline{X}; \underline{Y}) \geq R} \mathbb{E}[c(X)]. \quad (153)$$

This definition is completely general: the resource $c(X)$ can be power, bandwidth, time, number of channel uses, or any other cost measure. It does not depend on the utility function or the task; it is purely a property of the physical channel and the pragmatic abstraction.

The pragmatic cost of information satisfies the following fundamental properties.

Theorem 23 (Non-negativity and Monotonicity): $\text{CoI}_p(R) \geq 0$, with $\text{CoI}_p(0) = 0$. It is strictly increasing in R for $R > 0$.

Lemma 6 (Convexity of the Inverse of an Increasing Concave Function): Let $f : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ be a strictly increasing concave function. Then its inverse function f^{-1} is convex on the range of f .

Proof. Let $R_1, R_2 \geq 0$ be in the range of f , and let $P_1 = f^{-1}(R_1)$, $P_2 = f^{-1}(R_2)$. For any $\lambda \in [0, 1]$, define

$$R_\lambda = \lambda R_1 + (1 - \lambda) R_2, \quad P_\lambda = \lambda P_1 + (1 - \lambda) P_2.$$

By concavity of f ,

$$f(P_\lambda) \geq \lambda f(P_1) + (1 - \lambda) f(P_2) = \lambda R_1 + (1 - \lambda) R_2 = R_\lambda.$$

Since f is strictly increasing, applying f^{-1} to both sides preserves the inequality:

$$P_\lambda \geq f^{-1}(R_\lambda).$$

Substituting the definition of P_λ yields

$$\lambda f^{-1}(R_1) + (1 - \lambda) f^{-1}(R_2) \geq f^{-1}(\lambda R_1 + (1 - \lambda) R_2),$$

which is exactly the definition of convexity. $\square$

Theorem 24 (Convexity of the Pragmatic Cost of Information): Let $C_p(P)$ be the pragmatic capacity as a function of the resource P . Assume that $C_p(P)$ is concave and strictly increasing in P . Then the pragmatic cost of information $\text{CoI}_p(R) = C_p^{-1}(R)$ is a convex function of R on the achievable rate region.

Proof. By Lemma 6, the inverse of any strictly increasing concave function is convex. Applying this result with $f = C_p$ immediately gives that $\text{CoI}_p(R) = C_p^{-1}(R)$ is convex.

For completeness, we restate the argument in detail. Let R_1, R_2 be two rates and let $P_i = \text{CoI}_p(R_i)$ for $i = 1, 2$. For any $\lambda \in [0, 1]$, set

$$R_\lambda = \lambda R_1 + (1 - \lambda) R_2, \quad P_\lambda = \lambda P_1 + (1 - \lambda) P_2.$$

Because C_p is concave,

$$\begin{aligned} C_p(P_\lambda) &\geq \lambda C_p(P_1) + (1 - \lambda) C_p(P_2) \\ &= \lambda R_1 + (1 - \lambda) R_2 = R_\lambda. \end{aligned}$$

Since C_p is strictly increasing, its inverse is also strictly increasing, so

$$P_\lambda \geq C_p^{-1}(R_\lambda) = \text{CoI}_p(R_\lambda).$$

Therefore,

$$\lambda \text{CoI}_p(R_1) + (1 - \lambda) \text{CoI}_p(R_2) \geq \text{CoI}_p(\lambda R_1 + (1 - \lambda) R_2),$$

which proves convexity. $\square$

Remark 9: The convexity established above is weak convexity. Strict convexity would require $C_p(P)$ to be strictly concave (i.e., no linear segments) and the strict increasing assumption to hold globally. In practice, when zero-cost capacity plateaus exist, $\text{CoI}_p(R)$ may have linear segments and hence is convex but not strictly convex.

Theorem 25 (Duality with Pragmatic Channel Capacity): The pragmatic cost of information and the pragmatic channel capacity are strict inverses:

$$\text{CoI}_p(R) = C_p^{-1}(R), \quad C_p(P) = \text{CoI}_p^{-1}(P). \quad (154)$$

Thus, they form a perfect mathematical dual pair: one gives the maximum rate for a given resource, the other gives the minimum resource for a given rate.

C. Lagrangian Duality

Having established the two dual pairs, (rate-distortion, value) and (capacity, cost), we now combine them into a unified Lagrangian framework. This framework allows cross-layer optimization where the rate R serves as the common variable linking the source (decision) side and the channel (transmission) side.

C.1 Source-Distortion-Value Lagrangian (Down Loop)

The down loop concerns the trade-off between the distortion (or decision quality) and the rate required to achieve it. The Lagrangian for this loop is formulated by combining the rate-distortion function $R_p(D)$ and the value-rate function $\Phi_p(R)$.

Definition 22 (Source-Distortion-Value Pragmatic Lagrangian): For a given distortion level D , the down-loop Lagrangian is

$$\mathcal{L}_{SDV}(D; \lambda_d) \triangleq \Phi_p(R_p(D)) - \lambda_d R_p(D), \quad (155)$$

where $\lambda_d > 0$ is the Lagrange multiplier representing the marginal cost of rate (in utility units per bit).

Maximizing $\mathcal{L}_{SDV}$ over D yields the optimal operating point that balances the value obtained from reducing distortion against the rate cost. The first-order condition is:

$$\frac{d\Phi_p}{dR} \cdot R'_p(D) - \lambda_d R'_p(D) = 0 \quad \Rightarrow \quad \frac{d\Phi_p}{dR} = \lambda_d, \quad (156)$$

since $R'_p(D) \neq 0$. Thus, at optimality, the marginal value of rate equals the marginal rate cost.

C.2 Channel-Power-Cost Lagrangian (Up Loop)

The up loop concerns the trade-off between the rate and the physical resource cost. Using the pragmatic capacity $C_p(P)$ and the cost function $P(R) = \text{CoI}_p(R)$, we define:

Definition 23 (Channel-Power-Cost Pragmatic Lagrangian): For a given resource budget P , the up-loop Lagrangian is

$$\mathcal{L}_{CPC}(P; \gamma) \triangleq \gamma C_p(P) - P, \quad (157)$$

where $\gamma > 0$ is the Lagrange multiplier representing the shadow price of the resource (in utility units per unit resource).

Equivalently, in terms of rate R , using $P = \text{CoI}_p(R)$, the up-loop Lagrangian becomes

$$\mathcal{L}_{CPC}(R; \gamma) = \gamma R - \text{CoI}_p(R). \quad (158)$$

Maximizing over P (or R) gives the first-order condition:

$$\gamma C'_p(P) = 1 \quad \Rightarrow \quad \gamma = \frac{1}{C'_p(P)} = \text{CoI}'_p(R). \quad (159)$$

Thus, at optimality, the shadow price of the resource equals the marginal cost of increasing the rate.

C.3 Global Cross-Layer Lagrangian and Behavioral Capacity

The two loops are coupled through the common rate R : the down loop requires a certain rate to achieve a given utility, while the up loop supplies that rate at a certain resource cost. The global optimization problem is to maximize the net benefit (value minus cost) over the rate. To unify the two layers, we introduce a global Lagrangian that incorporates a shadow price λ (in utility units per unit resource) to convert physical resource cost into utility-equivalent cost:

$$\mathcal{L}_{\text{global}}(R; \lambda) = \Phi_p(R) - \lambda \text{CoI}_p(R), \quad (160)$$

where $\lambda > 0$ is the Lagrange multiplier representing the shadow price of the physical resource (e.g., power). It serves as the conversion factor between resource consumption and utility.

Theorem 26 (Concavity of the Global Pragmatic Lagrangian): Assume that the pragmatic value function $\Phi_p(R)$ is concave and the pragmatic cost function $\text{CoI}_p(R)$ is convex on $R \geq 0$. Then, for every fixed Lagrange multiplier $\lambda \geq 0$, the global pragmatic Lagrangian

$$\mathcal{L}_{\text{global}}(R; \lambda) \triangleq \Phi_p(R) - \lambda \text{CoI}_p(R)$$

is a concave function of R .

Proof. Let $R_1, R_2 \geq 0$ and let $\theta \in [0, 1]$. Define $R_\theta = \theta R_1 + (1 - \theta)R_2$.

By concavity of Φ_p ,

$$\Phi_p(R_\theta) \geq \theta \Phi_p(R_1) + (1 - \theta) \Phi_p(R_2). \quad (161)$$

By convexity of CoI_p ,

$$\text{CoI}_p(R_\theta) \leq \theta \text{CoI}_p(R_1) + (1 - \theta) \text{CoI}_p(R_2). \quad (162)$$

Multiplying (162) by $-\lambda$ (which is non-positive) reverses the inequality:

$$-\lambda \text{CoI}_p(R_\theta) \geq -\lambda \theta \text{CoI}_p(R_1) - \lambda (1 - \theta) \text{CoI}_p(R_2). \quad (163)$$

Adding (161) and (163) yields

$$\begin{aligned}
 \mathcal{L}_{\text{global}}(R_\theta; \lambda) &= \Phi_p(R_\theta) - \lambda \text{CoI}_p(R_\theta) \\
 &\geq \theta [\Phi_p(R_1) - \lambda \text{CoI}_p(R_1)] \\
 &\quad + (1 - \theta) [\Phi_p(R_2) - \lambda \text{CoI}_p(R_2)] \\
 &= \theta \mathcal{L}_{\text{global}}(R_1; \lambda) + (1 - \theta) \mathcal{L}_{\text{global}}(R_2; \lambda).
 \end{aligned} \tag{164}$$

Inequality (164) is precisely the definition of concavity. $\square$

Corollary 5: Under the assumptions of Theorem 26, if $\mathcal{L}_{\text{global}}$ is differentiable, then the first-order condition

$$\Phi'_p(R^*) = \lambda \text{CoI}'_p(R^*) \tag{165}$$

is sufficient for R^* to be a global maximizer of $\mathcal{L}_{\text{global}}(\cdot; \lambda)$.

Proof. For a differentiable concave function, any point at which the derivative vanishes is a global maximum. Hence (165) yields the global optimum. $\square$

This condition states that the marginal value of increasing the rate (in utility per bit) must equal the marginal cost of increasing the rate (in utility per bit), where the marginal cost is the product of the resource shadow price and the marginal resource consumption per bit. When the physical cost $\text{CoI}_p(R)$ is already expressed in utility units (i.e., the shadow price is normalized to unity), we recover the simplified form $\mathcal{L}_{\text{global}}(R) = \Phi_p(R) - \text{CoI}_p(R)$ and $\Phi'_p(R^*) = \text{CoI}'_p(R^*)$. In general, however, the explicit λ is required to maintain dimensional consistency.

Theorem 27 (Inner Duality of the Two Lagrangians): Let λ_d be the multiplier from the down-loop (marginal cost of rate) and γ be the multiplier from the up-loop (shadow price of resource). At the global optimum defined by (165), the following relationships hold:

$$\gamma = \lambda, \quad \lambda_d = \lambda \text{CoI}'_p(R^*) = \Phi'_p(R^*). \tag{166}$$

Thus, the global multiplier λ coincides with the up-loop shadow price γ , while the down-loop marginal rate cost λ_d equals the marginal value of rate, which is also $\lambda \text{CoI}'_p(R^*)$. This establishes a consistent duality between the decision-theoretic and physical-layer perspectives.

Proof. From the down-loop optimality condition, $\Phi'_p(R^*) = \lambda_d$. From the up-loop condition, $\gamma = \text{CoI}'_p(R^*)$. The global condition (165) gives $\Phi'_p(R^*) = \lambda \text{CoI}'_p(R^*)$. Combining, we get $\lambda_d = \lambda \gamma$, and since $\gamma = \text{CoI}'_p(R^*)$, it follows that $\lambda_d = \lambda \gamma$. To satisfy the global condition, we must have $\lambda = \gamma$ (otherwise the marginal value equation would not hold for all R). Hence, $\lambda = \gamma$ and $\lambda_d = \lambda \text{CoI}'_p(R^*) = \Phi'_p(R^*)$. $\square$

Definition 24 (Pragmatic Efficiency Bound (Behavioral Capacity)): For a given resource shadow price $\lambda \geq 0$, the *pragmatic efficiency functional* $\mathcal{E}_p(\lambda)$ is defined as the supreme net benefit achievable by the system:

$$\mathcal{E}_p(\lambda) \triangleq \sup_{R \geq 0} [\Phi_p(R) - \lambda \cdot \text{CoI}_p(R)]. \tag{167}$$

This quantity represents the maximum net utility that an intelligent agent can extract from its environment per unit of time, after accounting for the physical cost of sensing, communication, and computation. It constitutes the *behavioral capacity* of the system under resource pricing λ , and serves as the master performance metric for any resource-constrained goal-directed system.

Remark 10: When the pragmatic rate R is continuous and the function $\Phi_p(R) - \lambda \text{CoI}_p(R)$ is strictly concave (which holds under the assumptions of Theorem 26), the supremum in (167) is attained at a unique point R^* satisfying the marginal condition $\Phi'_p(R^*) = \lambda \text{CoI}'_p(R^*)$. Consequently, in all operational contexts, we may write

$$\mathcal{E}_p(\lambda) = \max_{R \geq 0} \mathcal{L}_{\text{global}}(R; \lambda),$$

where the maximum is understood in the sense of the unique global maximizer. In the absence of resource constraints ($\lambda = 0$), $\mathcal{E}_p(0)$ reduces to $\sup_R \Phi_p(R)$, i.e., the system's performance under unlimited resources.

The Lagrangian dual framework provides a principled method for resource allocation in pragmatic systems:

1) **Decision-driven design:** Given a target utility level, the down-loop Lagrangian determines the minimum rate required. The corresponding distortion D and

rate R are obtained by solving $\max_D \mathcal{L}_{\text{SDV}}$.

2) **Resource-limited design:** Given a resource budget P , the up-loop Lagrangian determines the maximum achievable rate. The optimal rate is found by maximizing $\mathcal{L}_{\text{CPC}}$.

3) **Global optimum:** When both utility and cost are considered, for a given shadow price λ (or equivalently, a resource budget constraint), the rate R^* from (165) yields the maximal net benefit. This rate can be implemented by choosing the appropriate distortion level (via source coding) and resource allocation (via channel coding) such that $R_p(D^*) = C_p(P^*) = R^*$.

The framework also enables adaptive operation: as the channel quality or task requirements change, the system can re-optimize the Lagrangian to adjust the operating point.

In summary, the Lagrangian duality between the source-distortion-value and channel-power-cost Lagrangians provides a complete mathematical characterization of the fundamental trade-offs in pragmatic communication systems, unifying Shannon's physical-layer limits with the decision-theoretic value of information, while maintaining dimensional consistency through the explicit shadow price λ .

D. Single-Sided Pragmatic Cost of Information and Pragmatic Value of Information

The single-sided pragmatic rate-distortion functions and achievable rates defined in Section V-C have natural dual counterparts in terms of information value and cost. These single-sided VoI and CoI measures capture the utility gains and resource expenditures when only one side of the communication link is subject to pragmatic abstraction. They provide a finer-grained analysis for asymmetric systems, such as sensing-to-action (perceptual) and command-to-execution (expressive) scenarios.

D.1 Perceptual Pragmatic Value of Information

The perceptual pragmatic value of information quantifies the maximum expected utility gain obtainable from an observation V when the decision is based on the pragmatic action $\underline{W}$. It is the dual of the perceptual pragmatic rate-distortion function $R_p^{\text{perc}}(D)$, and is defined using the down perceptual pragmatic mutual in-

formation $I_p(\underline{W}; V)$.

Definition 25 (Perceptual Pragmatic Value of Information): For a given rate constraint R , the perceptual pragmatic value of information is defined as

$$\text{VoI}_p^{\text{perc}}(R) \triangleq \max_{P(V|\underline{W}): I_p(\underline{W}; V) \leq R} \text{VoI}_p, \quad (168)$$

where VoI_p is the pragmatic value defined in Eq. (145), and the maximization is over all conditional distributions $P(V|\underline{W})$ satisfying the rate constraint.

This quantity represents the maximum utility gain achievable when the observation is constrained to carry at most R bits of information about the pragmatic action. It is non-decreasing and concave in R , and satisfies the variational duality:

$$\text{VoI}_p^{\text{perc}}(R) = \max_{D \geq 0} \{ \Psi_p^{\text{perc}}(D) - \lambda R_p^{\text{perc}}(D) \}, \quad (169)$$

where Ψ_p^{perc} is a concave function determined by the utility, and $\lambda > 0$ is the Lagrange multiplier.

D.2 Expressive Pragmatic Value of Information

The expressive pragmatic value of information quantifies the maximum expected utility gain obtainable from a pragmatic label $\underline{V}$ when reconstructing the syntactic source W . It is the dual of the expressive pragmatic rate-distortion function $R_p^{\text{expr}}(D)$, and is defined using the down expressive pragmatic mutual information $I_p(W; \underline{V})$.

Definition 26 (Expressive Pragmatic Value of Information): For a given rate constraint R , the expressive pragmatic value of information is defined as

$$\text{VoI}_p^{\text{expr}}(R) \triangleq \max_{P(\underline{V}|W): I_p(W; \underline{V}) \leq R} \text{VoI}_p, \quad (170)$$

where the maximization is over all conditional distributions $P(\underline{V}|W)$ satisfying the rate constraint, and VoI_p is evaluated with respect to the reconstructed source.

This quantity measures the utility gain achieved by using the pragmatic label as a compressed representation of the source, subject to a rate limit. It is also non-decreasing and concave in R , and satisfies the dual variational relationship with $R_p^{\text{expr}}(D)$.

D.3 Prospective Perceptual Pragmatic Cost of Information

The prospective perceptual pragmatic cost of information quantifies the minimum resource expenditure re-

quired to achieve a given level of perceptual pragmatic information rate. It is the dual of the prospective perceptual pragmatic achievable rate $C_p^{\text{p-perc}}(P)$, and is defined using the up perceptual pragmatic mutual information $I^p(W; V)$.

Definition 27 (Prospective Perceptual Pragmatic Cost of Information): For a given pragmatic rate R , the prospective perceptual pragmatic cost of information is defined as

$$\text{CoI}_p^{\text{p-perc}}(R) \triangleq \min_{P(V|W): I^p(W; V) \geq R} \mathbb{E}[c(W)], \quad (171)$$

where $c(W)$ is the cost function on the pragmatic source (e.g., sensing power), and the minimization is over all conditional distributions $P(V|W)$ satisfying the rate constraint.

This quantity represents the minimum resource cost required to extract at least R bits of pragmatic information from the observation. It is non-decreasing and convex in R , and is the strict inverse of the prospective perceptual achievable rate:

$$\text{CoI}_p^{\text{p-perc}}(R) = (C_p^{\text{p-perc}})^{-1}(R). \quad (172)$$

D.4 Prospective Expressive Pragmatic Cost of Information

The prospective expressive pragmatic cost of information quantifies the minimum resource expenditure required to achieve a given level of expressive pragmatic information rate. It is the dual of the prospective expressive pragmatic achievable rate $C_p^{\text{p-expr}}(P)$, and is defined using the up expressive pragmatic mutual information $I^p(W; \underline{V})$.

Definition 28 (Prospective Expressive Pragmatic Cost of Information): For a given pragmatic rate R , the prospective expressive pragmatic cost of information is defined as

$$\text{CoI}_p^{\text{p-expr}}(R) \triangleq \min_{P(\underline{V}|W): I^p(W; \underline{V}) \geq R} \mathbb{E}[c(W)], \quad (173)$$

where $c(W)$ is the cost function on the source (e.g., transmit power), and the minimization is over all conditional distributions $P(\underline{V}|W)$ satisfying the rate constraint.

This quantity represents the minimum resource cost required to convey at least R bits of source information through the pragmatic label. It is non-decreasing and convex in R , and is the strict inverse of the prospective expressive achievable rate:

$$\text{CoI}_p^{\text{p-expr}}(R) = (C_p^{\text{p-expr}})^{-1}(R). \quad (174)$$

D.5 Relationships and Duality Structure

The four single-sided VoI and CoI measures form a complete duality framework that parallels the bilateral case, with the following key relationships:

1) **Rate-Distortion $\leftrightarrow$ Value Duality**: For both perceptual and expressive paths, the value function is the Legendre–Fenchel dual of the corresponding rate-distortion function:

$$\begin{cases} \text{VoI}_p^{\text{perc}}(R) = \max_{D \geq 0} \{ \Psi_p^{\text{perc}}(D) - \lambda R_p^{\text{perc}}(D) \}, \\ \text{VoI}_p^{\text{expr}}(R) = \max_{D \geq 0} \{ \Psi_p^{\text{expr}}(D) - \lambda R_p^{\text{expr}}(D) \}. \end{cases} \quad (175)$$

2) **Capacity $\leftrightarrow$ Cost Duality**: For both prospective perceptual and expressive paths, the cost function is the inverse of the corresponding achievable rate:

$$\begin{cases} \text{CoI}_p^{\text{p-perc}}(R) = (C_p^{\text{p-perc}})^{-1}(R), \\ \text{CoI}_p^{\text{p-expr}}(R) = (C_p^{\text{p-expr}})^{-1}(R). \end{cases} \quad (176)$$

3) **Comparison with Bilateral Counterparts**: Single-sided measures are bounded by their bilateral counterparts:

$$\begin{cases} \text{VoI}_p^{\text{perc}}(R) \leq \text{VoI}_p(R), & \text{VoI}_p^{\text{expr}}(R) \leq \text{VoI}_p(R), \\ \text{CoI}_p^{\text{p-perc}}(R) \geq \text{CoI}_p(R), & \text{CoI}_p^{\text{p-expr}}(R) \geq \text{CoI}_p(R), \end{cases} \quad (177)$$

where $\text{VoI}_p(R)$ and $\text{CoI}_p(R)$ are the bilateral functions defined in Sections VI.A and VI.B.

Table VIII consolidates all bilateral and single-sided pragmatic information measures, their defining mutual informations, and their applicability in symmetric versus asymmetric systems. The perceptual path (sensing) coarsens only the source side, whereas the expressive path (actuation) coarsens only the reconstruction side. The prospective variants use the up mutual

TABLE VIII: Summary of bilateral and single-sided pragmatic information measures.

Measure	Abbrev.	Definition / Expression	Scenario
Bilateral VoI	$\text{VoI}_p(R)$	$\max_{P(\underline{Y} \underline{X}): I_p \leq R} \text{VoI}_p$	Symmetric, both ends coarsened
Bilateral CoI	$\text{CoI}_p(R)$	$\min_{P(x): I_p^p \geq R} \mathbb{E}[c(X)]$	Symmetric, both ends coarsened
Perceptual VoI	$\text{VoI}_p^{\text{perc}}(R)$	$\max_{P(V \underline{W}): I_p(W; V) \leq R} \text{VoI}_p$	Sensing: observation to action
Expressive VoI	$\text{VoI}_p^{\text{expr}}(R)$	$\max_{P(\underline{V} \underline{W}): I_p(W; \underline{V}) \leq R} \text{VoI}_p$	Actuation: command to reconstruction
Prospective Perceptual CoI	$\text{CoI}_p^{\text{p-perc}}(R)$	$\min_{P(V \underline{W}): I_p^p(W; V) \geq R} \mathbb{E}[c(W)]$	Sensing with full syntactic observation
Prospective Expressive CoI	$\text{CoI}_p^{\text{p-expr}}(R)$	$\min_{P(\underline{V} \underline{W}): I_p^p(W; \underline{V}) \geq R} \mathbb{E}[c(W)]$	Actuation with full syntactic source

information and thus provide upper bounds on achievable rates.

Figure 9 illustrates the duality structure of pragmatic information measures, encompassing both bilateral and single-sided quantities. At the core lie the two fundamental dual pairs: $R_p(D) \leftrightarrow \text{VoI}_p(R)$ and $C_p(P) \leftrightarrow \text{CoI}_p(R)$. Extending this duality to asymmetric scenarios, $R_p^{\text{perc}}(D) \leftrightarrow \text{VoI}_p^{\text{perc}}(R)$ and $R_p^{\text{expr}}(D) \leftrightarrow \text{VoI}_p^{\text{expr}}(R)$ capture the rate–value trade-off when only one side is coarsened. On the capacity–cost side, $C_p^{\text{p-perc}}(P) \leftrightarrow \text{CoI}_p^{\text{p-perc}}(R)$ and $C_p^{\text{p-expr}}(P) \leftrightarrow \text{CoI}_p^{\text{p-expr}}(R)$ complete the duality lattice, which collectively characterizes trade-offs between rate, distortion, capacity, cost, and value in both symmetric and asymmetric pragmatic communication systems.

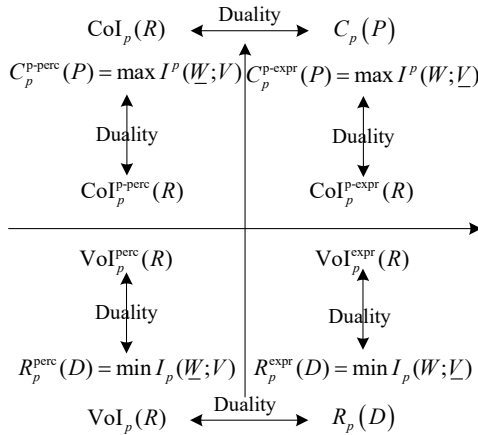


Fig. 9: Duality of pragmatic information measures.

The single-sided VoI and CoI measures provide a precise toolkit for analyzing asymmetric pragmatic systems:

- **Perceptual systems (sensing):** $\text{VoI}_p^{\text{perc}}$ measures the utility gain of a sensor under a rate constraint, while $\text{CoI}_p^{\text{p-perc}}$ quantifies the sensing power required to achieve a certain perceptual information

rate.

- **Expressive systems (actuation):** $\text{VoI}_p^{\text{expr}}$ measures the utility gain of a command channel under a rate constraint, while $\text{CoI}_p^{\text{p-expr}}$ quantifies the transmit power required to convey a certain expressive information rate.
- **Joint optimization:** The single-sided measures can be combined with the bilateral Lagrangian framework to design systems where sensing and actuation have different resource constraints and utility contributions.

Together with the bilateral value and cost functions, these single-sided counterparts complete the pragmatic information-theoretic spectrum, enabling a comprehensive analysis of task-oriented communication systems with asymmetric constraints.

VII. Pragmatic Lossless Source Coding

In this section, we investigate the problem of pragmatic lossless source coding. We extend the asymptotic equipartition property (AEP) to the pragmatic domain and introduce the pragmatic typical set, the isoteleia typical set, and the reified typical set. These concepts serve as the mathematical foundation for proving the pragmatic lossless source coding theorem.

A. Asymptotic Equipartition Property and Pragmatic Typical Set

Specifically, we establish the syntactic, semantic, and pragmatic AEP theorems. Based on these results, we define the pragmatic typical set, the isoteleia typical set, and the reified typical set, and analyze their cardinality and probability properties. The hierarchical structure of these typical sets reveals that pragmatic abstraction, by merging semantic distinctions that do not affect the optimal terminal action, further reduces the number

of typical sequences compared to semantic abstraction, thereby enabling more efficient lossless compression at the pragmatic level.

To establish the asymptotic equipartition property at the pragmatic level, we first introduce the sequential extensions of the fundamental mappings that bridge the syntactic, semantic, and pragmatic layers.

Let $\mathcal{W}$ be the syntactic alphabet with probability mass function $P(W)$, and let $\tilde{\mathcal{W}}$ be the semantic alphabet induced by the synonymous mapping $f : \tilde{\mathcal{W}} \rightarrow 2^{\mathcal{W}}$. Let $\underline{\mathcal{W}}$ be the pragmatic alphabet induced by the isoteleia mapping $e : \underline{\mathcal{W}} \rightarrow 2^{\tilde{\mathcal{W}}}$ according to a given utility function $U : \mathcal{W} \times \mathcal{A} \rightarrow \mathbb{R}$. The reification mapping $g = f \circ e : \underline{\mathcal{W}} \rightarrow 2^{\mathcal{W}}$ directly bridges the pragmatic and syntactic layers.

Definition 29 (Sequential Synonymous Mapping): Let $f : \tilde{\mathcal{W}} \rightarrow 2^{\mathcal{W}}$ be the synonymous mapping. Its n -th sequential extension $f^n : \tilde{\mathcal{W}}^n \rightarrow 2^{\mathcal{W}^n}$ is defined by

$$\begin{aligned} f^n(\tilde{w}^n) &\triangleq \prod_{k=1}^n f(\tilde{w}_k) \\ &= \{w^n \in \mathcal{W}^n : w_k \in f(\tilde{w}_k), k = 1, \dots, n\}, \end{aligned} \quad (178)$$

where $\tilde{w}^n = (\tilde{w}_1, \dots, \tilde{w}_n)$.

Definition 30 (Sequential Isoteleia Mapping): Let $e : \underline{\mathcal{W}} \rightarrow 2^{\tilde{\mathcal{W}}}$ be the isoteleia mapping. Its n -th sequential extension $e^n : \underline{\mathcal{W}}^n \rightarrow 2^{\tilde{\mathcal{W}}^n}$ is defined by

$$\begin{aligned} e^n(\underline{w}^n) &\triangleq \prod_{k=1}^n e(\underline{w}_k) \\ &= \{\tilde{w}^n \in \tilde{\mathcal{W}}^n : \tilde{w}_k \in e(\underline{w}_k), k = 1, \dots, n\}, \end{aligned} \quad (179)$$

where $\underline{w}^n = (\underline{w}_1, \dots, \underline{w}_n)$.

Definition 31 (Sequential Reification Mapping): Let $g = f \circ e : \underline{\mathcal{W}} \rightarrow 2^{\mathcal{W}}$ be the reification mapping. Its n -th sequential extension $g^n : \underline{\mathcal{W}}^n \rightarrow 2^{\mathcal{W}^n}$ is defined as the composition of the sequential isoteleia and synonymous mappings:

$$g^n(\underline{w}^n) \triangleq (f^n \circ e^n)(\underline{w}^n) = \bigcup_{\tilde{w}^n \in e^n(\underline{w}^n)} f^n(\tilde{w}^n) = \prod_{k=1}^n g(\underline{w}_k). \quad (180)$$

Equivalently,

$$g^n(\underline{w}^n) = \{w^n \in \mathcal{W}^n : w_k \in g(\underline{w}_k), k = 1, \dots, n\}. \quad (181)$$

A.1 Three-Tier AEP Theorems

We now state the AEP at each of the three layers. The syntactic AEP is the classical result [1, 45]. The semantic AEP was established in [24, 25]. The pragmatic AEP follows analogously.

Theorem 28 (Three-Tier AEP): Let $(W_1, W_2, \dots)$ be an i.i.d. syntactic sequence drawn according to $P(W)$, and let $(\tilde{W}_1, \tilde{W}_2, \dots)$ and $(\underline{W}_1, \underline{W}_2, \dots)$ be the associated semantic and pragmatic sequences under the sequential Synonymous Mapping f^n and the sequential Reification Mapping $g^n = f^n \circ e^n$, respectively. Then the following convergences hold in probability:

$$\begin{cases} -\frac{1}{n} \log P(W_1, \dots, W_n) \rightarrow H(W), \\ -\frac{1}{n} \log P(\tilde{W}_1, \dots, \tilde{W}_n) \rightarrow H_s(\tilde{W}), \\ -\frac{1}{n} \log P(\underline{W}_1, \dots, \underline{W}_n) \rightarrow H_p(\underline{W}), \end{cases} \quad (182)$$

where

$$\begin{aligned} P(W^n) &= \prod_{k=1}^n P(W_k), \quad P(\tilde{w}^n) = \prod_{k=1}^n P(\tilde{w}_k), \\ P(\tilde{w}_k) &= \sum_{w_k \in f(\tilde{w}_k)} P(w_k), \end{aligned}$$

and

$$P(\underline{w}^n) = \prod_{k=1}^n P(\underline{w}_k), \quad P(\underline{w}_k) = \sum_{w_k \in g(\underline{w}_k)} P(w_k).$$

The syntactic convergence is the classical result [1, 45]; the semantic and pragmatic convergences follow by applying the weak law of large numbers to the aggregated probabilities over synonymous and reified equivalence classes, respectively.

The following two AEP theorems characterize the asymptotic relationship between adjacent layers; they follow directly from the three AEP theorems above.

Theorem 29 (Isoteleia AEP):

$$-\frac{1}{n} [\log P(\tilde{W}^n) - \log P(\underline{W}^n)] \rightarrow H_s(\tilde{W}) - H_p(\underline{W}) \quad (183)$$

in probability.

Theorem 30 (Reified AEP):

$$-\frac{1}{n} [\log P(W^n) - \log P(\underline{W}^n)] \rightarrow H(W) - H_p(\underline{W}) \quad (184)$$

in probability.

A.2 Definitions of the Hierarchical Typical Sets

We now define the typical sets at each layer and the equivalence classes that partition them.

Definition 32 (Syntactic, Semantic, and Pragmatic Typical Sets): For $\epsilon > 0$, the syntactically typical set $A_\epsilon^{(n)}$, the semantically typical set $\tilde{A}_\epsilon^{(n)}$, and the pragmatic typical set $\underline{A}_\epsilon^{(n)}$ are respectively defined as

$$\begin{cases} A_\epsilon^{(n)} \triangleq \left\{ w^n \in \mathcal{W}^n : \left| -\frac{1}{n} \log P(w^n) - H(W) \right| < \epsilon \right\}, \\ \tilde{A}_\epsilon^{(n)} \triangleq \left\{ \tilde{w}^n \in \tilde{\mathcal{W}}^n : \left| -\frac{1}{n} \log P(\tilde{w}^n) - H_s(\tilde{W}) \right| < \epsilon \right\}, \\ \underline{A}_\epsilon^{(n)} \triangleq \left\{ \underline{w}^n \in \underline{\mathcal{W}}^n : \left| -\frac{1}{n} \log P(\underline{w}^n) - H_p(\underline{W}) \right| < \epsilon \right\}. \end{cases} \quad (185)$$

Definition 33 (Synonymous Typical Set): For a given semantic typical sequence $\tilde{w}^n \in \tilde{A}_\epsilon^{(n)}$, the synonymous typical set $S_\epsilon^{(n)}(\tilde{w}^n)$ is defined as the set of syntactic sequences that map to $\tilde{w}^n$ under f^n and are syntactically typical:

$$\begin{aligned} S_\epsilon^{(n)}(\tilde{w}^n) \triangleq & \left\{ w^n \in \mathcal{W}^n : w^n \in f^n(\tilde{w}^n), \right. \\ & \left| -\frac{1}{n} \log P(w^n) - H(W) \right| < \epsilon, \\ & \left| -\frac{1}{n} \log \frac{P(w^n)}{P(\tilde{w}^n)} - (H(W) - H_s(\tilde{W})) \right| < \epsilon \left. \right\}. \end{aligned} \quad (186)$$

The collection $\{S_\epsilon^{(n)}(\tilde{w}^n)\}_{\tilde{w}^n \in \tilde{A}_\epsilon^{(n)}}$ forms a partition of $A_\epsilon^{(n)}$:

$$A_\epsilon^{(n)} = \bigcup_{\tilde{w}^n \in \tilde{A}_\epsilon^{(n)}} S_\epsilon^{(n)}(\tilde{w}^n), \quad (187)$$

with disjointness for distinct $\tilde{w}^n$.

Definition 34 (Isoteleia Typical Set): For a given pragmatic typical sequence $\underline{w}^n \in \underline{A}_\epsilon^{(n)}$, the Isoteleia typical set $E_\epsilon^{(n)}(\underline{w}^n)$ is defined as the set of semantic sequences that map to $\underline{w}^n$ under e^n and are semantically typical:

$$\begin{aligned} E_\epsilon^{(n)}(\underline{w}^n) \triangleq & \left\{ \tilde{w}^n \in \tilde{\mathcal{W}}^n : \tilde{w}^n \in e^n(\underline{w}^n), \right. \\ & \left| -\frac{1}{n} \log P(\tilde{w}^n) - H_s(\tilde{W}) \right| < \epsilon, \\ & \left| -\frac{1}{n} \log \frac{P(\tilde{w}^n)}{P(\underline{w}^n)} - (H_s(\tilde{W}) - H_p(\underline{W})) \right| < \epsilon \left. \right\}. \end{aligned} \quad (188)$$

The collection $\{E_\epsilon^{(n)}(\underline{w}^n)\}_{\underline{w}^n \in \underline{A}_\epsilon^{(n)}}$ forms a partition of $\tilde{A}_\epsilon^{(n)}$:

$$\tilde{A}_\epsilon^{(n)} = \bigcup_{\underline{w}^n \in \underline{A}_\epsilon^{(n)}} E_\epsilon^{(n)}(\underline{w}^n), \quad (189)$$

with disjointness for distinct $\underline{w}^n$.

Definition 35 (Reified Typical Set): For a given pragmatic typical sequence $\underline{w}^n \in \underline{A}_\epsilon^{(n)}$, the Reified typical set $B_\epsilon^{(n)}(\underline{w}^n)$ is defined as the set of syntactic sequences that map to $\underline{w}^n$ under the composite mapping $g^n = f^n \circ e^n$ and are syntactically typical:

$$\begin{aligned} B_\epsilon^{(n)}(\underline{w}^n) \triangleq & \left\{ w^n \in \mathcal{W}^n : w^n \in g^n(\underline{w}^n), \right. \\ & \left| -\frac{1}{n} \log P(w^n) - H(W) \right| < \epsilon, \\ & \left| -\frac{1}{n} \log \frac{P(w^n)}{P(\underline{w}^n)} - (H(W) - H_p(\underline{W})) \right| < \epsilon \left. \right\}. \end{aligned} \quad (190)$$

The collection $\{B_\epsilon^{(n)}(\underline{w}^n)\}_{\underline{w}^n \in \underline{A}_\epsilon^{(n)}}$ forms a partition of $A_\epsilon^{(n)}$:

$$A_\epsilon^{(n)} = \bigcup_{\underline{w}^n \in \underline{A}_\epsilon^{(n)}} B_\epsilon^{(n)}(\underline{w}^n), \quad (191)$$

with disjointness for distinct $\underline{w}^n$.

The hierarchical relationships among the pragmatic, semantic, and syntactic typical sets, as well as their corresponding equivalence classes, are illustrated schematically in Fig. 10. In this two-level hierarchical partition: the synonymous mapping first refines the

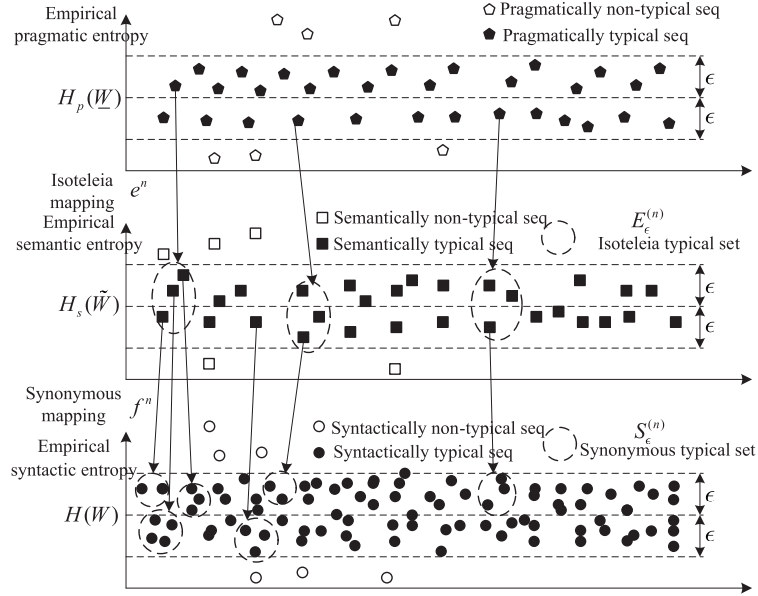


Fig. 10: Hierarchical structure of syntactic, semantic, and pragmatic typical sets.

syntactic typical set into synonym classes; the isoteleia mapping then groups these classes into coarser Reified classes, each corresponding to a single pragmatic action. This nested structure directly reflects the composition $g^n = f^n \circ e^n$ and underpins the entropy reduction $H_p(\underline{W}) \leq H_s(\tilde{W}) \leq H(W)$. Moreover, each reified typical set is the union of synonymous typical sets over all semantic sequences in the corresponding isoteleia typical set:

$$B_\epsilon^{(n)}(\underline{w}^n) = \bigcup_{\tilde{w}^n \in E_\epsilon^{(n)}(\underline{w}^n)} S_\epsilon^{(n)}(\tilde{w}^n).$$

A.3 Properties of the Typical Sets

The syntactic and semantic typical sets satisfy the standard properties established in [1, 45] and [24, 25], respectively. For completeness, we list them below without proof.

Theorem 31 (Properties of the Syntactic Typical Set):

For sufficiently large n :

- 1) For any $w^n \in A_\epsilon^{(n)}$,
$$2^{-n(H(W)+\epsilon)} \leq P(w^n) \leq 2^{-n(H(W)-\epsilon)}. \quad (192)$$
- 2) $\Pr\{A_\epsilon^{(n)}\} > 1 - \epsilon$.
- 3) $(1 - \epsilon)2^{n(H(W)-\epsilon)} \leq |A_\epsilon^{(n)}| \leq 2^{n(H(W)+\epsilon)}$.

Theorem 32 (Properties of the Semantic Typical Set):

For sufficiently large n :

- 1) For any $\tilde{w}^n \in \tilde{A}_\epsilon^{(n)}$,

$$2^{-n(H_s(\tilde{W})+\epsilon)} \leq P(\tilde{w}^n) \leq 2^{-n(H_s(\tilde{W})-\epsilon)}. \quad (193)$$

- 2) $\Pr\{\tilde{A}_\epsilon^{(n)}\} > 1 - \epsilon$.

$$3) (1 - \epsilon)2^{n(H_s(\tilde{W})-\epsilon)} \leq |\tilde{A}_\epsilon^{(n)}| \leq 2^{n(H_s(\tilde{W})+\epsilon)}.$$

We now establish the properties of the pragmatic typical set.

Theorem 33 (Properties of the Pragmatic Typical Set):

For sufficiently large n :

- 1) For any $\underline{w}^n \in \underline{A}_\epsilon^{(n)}$,

$$2^{-n(H_p(\underline{W})+\epsilon)} \leq P(\underline{w}^n) \leq 2^{-n(H_p(\underline{W})-\epsilon)}. \quad (194)$$

- 2) $\Pr\{\underline{A}_\epsilon^{(n)}\} > 1 - \epsilon$.

$$3) (1 - \epsilon)2^{n(H_p(\underline{W})-\epsilon)} \leq |\underline{A}_\epsilon^{(n)}| \leq 2^{n(H_p(\underline{W})+\epsilon)}.$$

The proof of Theorem 33 follows the standard AEP counting argument. For the sake of readability, the detailed summation steps are provided in Appendix A.

Next, we state and prove the properties of the isoteleia typical set.

Theorem 34 (Properties of the Isoteleia Typical Set):

For any $\underline{w}^n \in \underline{A}_\epsilon^{(n)}$, for sufficiently large n :

- 1) For any $\tilde{w}^n \in E_\epsilon^{(n)}(\underline{w}^n)$,

$$2^{-n(H_s(\tilde{W})-H_p(\underline{W})+\epsilon)} \leq \frac{P(\tilde{w}^n)}{P(\underline{w}^n)} \leq 2^{-n(H_s(\tilde{W})-H_p(\underline{W})-\epsilon)}. \quad (195)$$

$$2) \\ 2^{n(H_s(\tilde{W})-H_p(\underline{W})-\epsilon)} \leq |E_\epsilon^{(n)}(\underline{w}^n)| \leq 2^{n(H_s(\tilde{W})-H_p(\underline{W})+\epsilon)}. \quad (196)$$

The details of the proof are provided in Appendix B.

The synonymous typical set properties are known from [24, 25], we restate them for reference.

Theorem 35 (Properties of the Synonymous Typical Set): For any $\tilde{w}^n \in \tilde{A}_\epsilon^{(n)}$, for sufficiently large n :

1) For any $w^n \in S_\epsilon^{(n)}(\tilde{w}^n)$,

$$2^{-n(H(W)-H_s(\tilde{W})+\epsilon)} \leq \frac{P(w^n)}{P(\tilde{w}^n)} \leq 2^{-n(H(W)-H_s(\tilde{W})-\epsilon)}. \quad (197)$$

2)

$$2^{n(H(W)-H_s(\tilde{W})-\epsilon)} \leq |S_\epsilon^{(n)}(\tilde{w}^n)| \leq 2^{n(H(W)-H_s(\tilde{W})+\epsilon)}. \quad (198)$$

Finally, we prove the properties of the Reified typical set.

Theorem 36 (Properties of the Reified Typical Set): For any $\underline{w}^n \in \underline{A}_\epsilon^{(n)}$, for sufficiently large n :

1) For any $w^n \in B_\epsilon^{(n)}(\underline{w}^n)$,

$$2^{-n(H(W)-H_p(\underline{W})+\epsilon)} \leq \frac{P(w^n)}{P(\underline{w}^n)} \leq 2^{-n(H(W)-H_p(\underline{W})-\epsilon)}. \quad (199)$$

2)

$$2^{n(H(W)-H_p(\underline{W})-\epsilon)} \leq |B_\epsilon^{(n)}(\underline{w}^n)| \leq 2^{n(H(W)-H_p(\underline{W})+\epsilon)}. \quad (200)$$

The details of the proof are provided in Appendix C.

Corollary 6 (Hierarchy of Typical Set Sizes): The cardinalities of the typical sets satisfy

$$|\underline{A}_\epsilon^{(n)}| \leq |\tilde{A}_\epsilon^{(n)}| \leq |A_\epsilon^{(n)}|, \quad (201)$$

and the size of a Reified typical set is the product of the sizes of the corresponding Isoteleia and Synonymous typical sets:

$$|B_\epsilon^{(n)}(\underline{w}^n)| = |E_\epsilon^{(n)}(\underline{w}^n)| \cdot |S_\epsilon^{(n)}(\tilde{w}^n)|, \quad (202)$$

for any $\tilde{w}^n \in E_\epsilon^{(n)}(\underline{w}^n)$, reflecting the composition $g^n = f^n \circ e^n$.

Remark 11: The hierarchy $|\underline{A}_\epsilon^{(n)}| \leq |\tilde{A}_\epsilon^{(n)}|$ reflects the fact that the isoteleia mapping merges semantic distinctions that do not affect the optimal action, thereby reducing the number of distinct pragmatic typical sequences. Similarly, the synonymous mapping reduces the number of semantic typical sequences relative to syntactic ones. The reified typical set $B_\epsilon^{(n)}(\underline{w}^n)$ directly collects all syntactic sequences that realize a given pragmatic purpose, and its size is approximately $2^{n(H(W)-H_p(\underline{W}))}$, which is larger than a synonymous typical set by a factor of $2^{n(H_s(\tilde{W})-H_p(\underline{W}))}$. This additional freedom is the source of compression gains in pragmatic lossless source coding.

B. Pragmatic Lossless Source Coding Theorem

We establish the fundamental limit of pragmatic lossless source coding using a three-tier architecture that refines symbols through meanings to actions, yet only the pragmatic index is transmitted. The code maps each pragmatic typical sequence to a unique codeword, allowing arbitrary syntactic and semantic realizations. Using random coding and typical set properties, we prove that pragmatic entropy is the limit: rates above it are achievable with vanishing error, while rates below fail. We also relate this to the down pragmatic value of information, showing that the full decision utility can only be obtained when the rate reaches the pragmatic entropy.

B.1 Three-Tier Pragmatic Source Code Model

We consider a discrete memoryless syntactic source W with alphabet $\mathcal{W}$ and probability mass function $P(W)$. Let $\tilde{W}$ and $\underline{W}$ be the associated semantic and pragmatic variables induced by the synonymous mapping f and the isoteleia mapping e , respectively, with the reification mapping $g = f \circ e$. The source generates an i.i.d. sequence $W^n = (W_1, \dots, W_n)$ according to $P(W^n) = \prod_{k=1}^n P(W_k)$. The semantic sequence $\tilde{W}^n$ and pragmatic sequence $\underline{W}^n$ are obtained by applying the sequential mappings f^n and g^n element-wise.

In the pragmatic lossless source coding scenario, the encoder may operate in three stages corresponding to the three layers of information, but in the order from pragmatic to semantic to syntactic: it first identifies

the pragmatic purpose (optimal action class) from the source, then selects a semantic interpretation that justifies that action, and finally chooses a syntactic realization that conveys that meaning. However, since the reification mapping $g^n = f^n \circ e^n$ directly gives the pragmatic class of any syntactic sequence, the encoder can equivalently apply g^n directly and then encode the resulting pragmatic index. The decoder, upon receiving the index, outputs an arbitrary syntactic sequence from the corresponding reified typical set $B_\epsilon^{(n)}(\underline{w}^n)$, thereby recovering the pragmatic purpose without needing to reconstruct the exact syntactic or semantic details.

Definition 36 (Pragmatic Lossless Source Code): An (M, n) pragmatic lossless source code consists of:

1) A pragmatic index set $\mathcal{J}_p = \{1, 2, \dots, M_p\}$, where each index $i_p \in \mathcal{J}_p$ corresponds to a distinct pragmatic purpose (optimal action class). Additionally, a reification index set $\mathcal{J}_r = \{1, 2, \dots, M_r\}$, where each index $i_r \in \mathcal{J}_r$ enumerates the distinct syntactic realizations within each reified typical set. The total number of syntactic sequences that can be represented is $M = M_p \cdot M_r$.

2) An encoding function that operates in three stages:

(a) **Pragmatic-to-semantic mapping (isoteleia):**

For a given pragmatic index i_p , the encoder first selects a semantic interpretation from the corresponding isoteleia typical set $E_\epsilon^{(n)}(\underline{w}_{i_p})$, where the index j enumerates the semantic sequences within that set.

(b) **Semantic-to-syntactic mapping (synonymous):**

For the chosen semantic sequence $\tilde{w}^j$, the encoder then chooses a syntactic realization from the synonymous typical set $S_\epsilon^{(n)}(\tilde{w}^j)$, where the index k enumerates the syntactic sequences within that set.

(c) **Reification index selection:** The encoder finally selects a reification index $i_r \in \mathcal{J}_r$, which corresponds to a specific syntactic sequence w^n from the reified typical set $B_\epsilon^{(n)}(\underline{w}_{i_p})$. The reified typical set is the union of all synonymous typical sets corresponding to semantic sequences in the isoteleia typical set for that pragmatic index.

Equivalently, the encoding function $\phi : \mathcal{W}^n \rightarrow \mathcal{J}_p \times \mathcal{J}_r$

maps each syntactic sequence w^n to a pair of indices (i_p, i_r) such that $w^n \in B_\epsilon^{(n)}(\underline{w}_{i_p})$ and i_r identifies the specific syntactic sequence within that reified typical set. Pragmatically, only the pragmatic index i_p needs to be transmitted to preserve the action; the reification index i_r is optional and determines syntactic fidelity.

3) A decoding function $\psi : \mathcal{J}_p \rightarrow \mathcal{W}^n$ that, upon receiving an index i_p , outputs a representative syntactic sequence $\hat{w}^n$ belonging to the reified typical set $B_\epsilon^{(n)}(\underline{w}_{i_p})$. If the reification index i_r is also transmitted, the decoder can output the exact syntactic sequence identified by i_r ; otherwise, it may choose any syntactic sequence from the corresponding synonymous typical sets, as all share the same pragmatic meaning.

The pragmatic code rate is defined as $R = \frac{1}{n} \log_2 M_p$ (prabits per source symbol), and the reification rate is defined as $R_r = \frac{1}{n} \log_2 M_r$ (bits per source symbol). The total syntactic code rate is therefore $R_{\text{syn}} = R + R_r = \frac{1}{n} \log_2 M$. The error probability for a given code is

$$P_e^{(n)} = \Pr\{\psi(\phi(W^n)) \notin B_\epsilon^{(n)}(\underline{W}^n)\}, \quad (203)$$

i.e., the probability that the decoded syntactic sequence does not belong to the correct reified typical set, and hence the pragmatic purpose is not preserved.

This coding architecture is illustrated in Fig. 11. The encoder first applies the Reification Mapping g^n to determine the pragmatic class of the observed syntactic sequence, then encodes that class index. The decoder recovers the index and outputs an arbitrary syntactic sequence from the corresponding reified typical set, which indicates that the decoder's output may differ syntactically from the input but preserves the pragmatic meaning. The intermediate semantic layer is implicit in the composition $g^n = f^n \circ e^n$, and the system can be viewed as a cascade of synonymous and isoteleia mappings.

Theorem 37 (Pragmatic Lossless Source Coding Theorem): Let W be a discrete memoryless syntactic source with associated pragmatic variable $\underline{W}$ having pragmatic entropy $H_p(\underline{W})$. For any $\epsilon > 0$ and pragmatic code rate R :

1) **Achievability:** If $R > H_p(\underline{W})$, then for suffi-

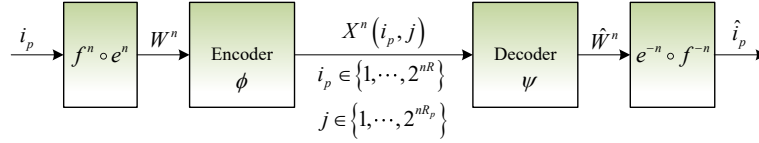


Fig. 11: Block diagram of pragmatic lossless source coding. The encoder applies the reification mapping g^n to map the pragmatic class to its syntactic sequence and encodes the index. The decoder outputs a representative syntactic sequence from the corresponding reified typical set.

ciently large n , there exists an $(2^{n(R+R_p)}, n)$ pragmatic lossless source code with error probability $P_e^{(n)} < \epsilon$, where R_p is the rate of the reified typical set (i.e., the number of syntactic realizations per pragmatic class).

2) **Converse:** If $R < H_p(\underline{W})$, then for any $(2^{n(R+R_p)}, n)$ pragmatic lossless source code, the error probability $P_e^{(n)}$ tends to 1 as $n \rightarrow \infty$.

Moreover, the pragmatic Value of Information (VoI) achieved by the compressed index is maximized (i.e., equals the VoI obtained from the full pragmatic variable $\underline{W}$) if and only if $R \geq H_p(\underline{W})$.

Proof. Achievability: Fix $\epsilon > 0$ and choose $R > H_p(\underline{W})$. By the properties of the pragmatic typical set (Theorem 33), the pragmatic typical set $\underline{A}_\epsilon^{(n)}$ satisfies

$$(1 - \epsilon)2^{n(H_p(\underline{W}) - \epsilon)} \leq |\underline{A}_\epsilon^{(n)}| \leq 2^{n(H_p(\underline{W}) + \epsilon)}.$$

Since $R > H_p(\underline{W})$, we can select n large enough such that $2^{nR} \geq |\underline{A}_\epsilon^{(n)}|$. We construct a one-to-one mapping from each pragmatic typical sequence $\underline{w}^n \in \underline{A}_\epsilon^{(n)}$ to a distinct pragmatic index $i_p \in \{1, \dots, 2^{nR}\}$. For any syntactic sequence w^n , the encoder computes its pragmatic class $\underline{w}^n$. If $\underline{w}^n \in \underline{A}_\epsilon^{(n)}$, it sends the assigned index i_p ; otherwise, it sends a default index (e.g., 1).

Each pragmatic index i_p corresponds to a reified typical set $B_\epsilon^{(n)}(\underline{w}_{i_p})$ of size approximately 2^{nR_p} , where $R_p = H(W) - H_p(\underline{W})$ is the maximum rate of the reified typical set. The decoder receives the index i_p and outputs an arbitrary syntactic sequence from the corresponding reified typical set $B_\epsilon^{(n)}(\underline{w}_{i_p})$ (which is non-empty by definition). This ensures that the pragmatic purpose is correctly recovered whenever $\underline{W}^n \in \underline{A}_\epsilon^{(n)}$. By Theorem 33, $\Pr\{\underline{W}^n \in \underline{A}_\epsilon^{(n)}\} > 1 - \epsilon$, so the error probability is bounded by ϵ . Hence, for any $R > H_p(\underline{W})$, there exists a code with arbitrarily small error probability.

The total syntactic code rate is $R_{syn} = R + R_p$. When $R_p = 0$ (i.e., each pragmatic class is repre-

sented by a single syntactic sequence), the total rate approaches $H_p(\underline{W})$. When R_p increases to its maximum value $H(W) - H_p(\underline{W})$, the total rate approaches the syntactic entropy $H(W)$, recovering the classical lossless source coding limit. The intermediate values of R_p correspond to different levels of syntactic redundancy preservation within each pragmatic equivalence class.

Converse: Assume a code with pragmatic rate $R < H_p(\underline{W})$ and error probability $P_e^{(n)}$. Let $\mathcal{J}_p$ be the set of pragmatic indices, with $|\mathcal{J}_p| = 2^{nR}$. Define the decoding sets $\mathcal{D}_{i_p} = \{w^n : \phi(w^n) = i_p\}$, and their corresponding pragmatic classes $\underline{\mathcal{D}}_{i_p} = \{\underline{w}^n : g^n(w^n) = \underline{w}^n \text{ for some } w^n \in \mathcal{D}_{i_p}\}$. Since each decoding set is assigned to a single pragmatic index, the decoder can correctly recover the pragmatic purpose only if the true $\underline{W}^n$ belongs to $\underline{\mathcal{D}}_{i_p}$. The total number of distinct pragmatic classes that can be correctly decoded is at most $|\mathcal{J}_p| = 2^{nR}$. To achieve small error probability, these classes must cover most of the probability mass of $\underline{W}^n$. By the properties of the pragmatic typical set, the number of pragmatic typical sequences is approximately $2^{nH_p(\underline{W})}$. If $R < H_p(\underline{W})$, then $2^{nR} < (1 - \epsilon)2^{n(H_p(\underline{W}) - \epsilon)}$ for large n , so at most a fraction of the pragmatic typical sequences can be assigned distinct indices. Consequently, for any code, the probability that the true pragmatic sequence falls into an unassigned class or is mapped to a wrong index is at least $1 - 2^{-n(H_p(\underline{W}) - R - \epsilon)}$, which tends to 1 as $n \rightarrow \infty$. Thus, $P_e^{(n)} \rightarrow 1$. This proves the converse.

Connection to Pragmatic VoI: The pragmatic Value of Information achieved by the compressed index I_p is defined as

$$\text{VoI}_p(I_p) = \mathbb{E}_{W, I_p} [U(W, a^*(\underline{W}(I_p)))] - U_0,$$

where $a^*(\underline{W}(I_p))$ is the optimal action based on the recovered pragmatic class, and U_0 is the baseline utility without any information. Since I_p is a function of the

encoded index, it carries at most nR bits of information about the source. By the data processing inequality, the mutual information between the pragmatic variable $\underline{W}$ and I_p is bounded by

$$I(\underline{W}; I_p) \leq H(I_p) \leq R.$$

To achieve the maximum pragmatic value $\text{VoI}_p^{\max}$ —i.e., the same utility as if the true pragmatic variable $\underline{W}$ were directly observed—the index I_p must preserve all pragmatically relevant distinctions. This requires that $\underline{W}$ be recoverable from I_p , which in turn necessitates $I(\underline{W}; I_p) = H_p(\underline{W})$. However, this equality can hold only if $R \geq H_p(\underline{W})$. If $R < H_p(\underline{W})$, then $I(\underline{W}; I_p) \leq R < H_p(\underline{W})$, implying that some uncertainty about the optimal action class remains unresolved. Consequently, the expected utility conditioned on I_p is strictly less than the utility obtained under perfect pragmatic knowledge, and the achievable VoI is strictly below $\text{VoI}_p^{\max}$. Conversely, when $R \geq H_p(\underline{W})$, the encoder can simply transmit a lossless description of the pragmatic index, and the decoder recovers $\underline{W}$ perfectly, thereby attaining the full pragmatic VoI. Thus, the minimum rate for full pragmatic value extraction is exactly $H_p(\underline{W})$. $\square$

Remark 12: The pragmatic lossless source coding theorem generalizes Shannon’s classical theorem. The hierarchy $H_p(\underline{W}) \leq H_s(\tilde{W}) \leq H(W)$ quantifies the compression gains from pragmatic abstraction. The reified typical set $B_\epsilon^{(n)}(\underline{w}^n)$ admits a two-level decomposition via the composition $g^n = f^n \circ e^n$: it partitions into isoteleia typical sets $E_\epsilon^{(n)}(\underline{w}^n)$ (grouping semantic justifications that lead to the same action) and, within each, synonymous typical sets $S_\epsilon^{(n)}(\tilde{w}^n)$ (grouping syntactic realizations of the same meaning). This structure underpins the additional rate R_p , which offers a flexible trade-off: setting $R_p = 0$ achieves the pure pragmatic limit, while increasing R_p preserves syntactic details within each pragmatic class, recovering the classical limit as $R_p \rightarrow H(W) - H_p(\underline{W})$.

The optimal encoding strategy compresses directly to the pragmatic index, discarding all syntactic and semantic distinctions that do not affect the optimal action—consistent with the design principles of pragmatic

information systems (Section II). The pragmatic VoI provides a direct performance metric: for a given rate R , the achievable utility is maximized when the code preserves pragmatic equivalence classes, establishing a principled trade-off between communication rate and task performance in applications such as autonomous driving, remote control, and industrial automation.

VIII. Pragmatic Channel Coding

In this section, we study pragmatic information transmission over noisy channels. Extending the joint AEP via the joint reification mapping, we define the jointly reified typical set that groups syntactic input-output pairs by their common pragmatic purpose. Using random coding and joint typical decoding, we prove the pragmatic channel coding theorem: the pragmatic capacity equals the maximum achievable rate and extends the classical capacity by tolerating errors that do not affect optimal actions.

A. Jointly Asymptotic Equipartition Property and Jointly Reified Typical Set

To establish the fundamental limits of pragmatic channel coding, we first extend the asymptotic equipartition property to the joint distribution of channel input and output sequences at the pragmatic level. Unlike the semantic approach, which relies on the joint synonymous mapping, we directly employ the joint reification mapping, which is the composition of the joint synonymous and joint isoteleia mappings, to aggregate syntactic sequence pairs into pragmatic equivalence classes. This approach provides a direct characterization of the equivalence of input-output pairs with respect to the optimal terminal actions.

Consider a discrete memoryless channel with input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$, and transition probability $P(Y|X)$. Let $\tilde{\mathcal{X}}$ and $\tilde{\mathcal{Y}}$ be the associated semantic alphabets, and let $\underline{\mathcal{X}}$ and $\underline{\mathcal{Y}}$ be the pragmatic alphabets induced by the isoteleia mappings $e_X : \underline{\mathcal{X}} \rightarrow 2^{\tilde{\mathcal{X}}}$ and $e_Y : \underline{\mathcal{Y}} \rightarrow 2^{\tilde{\mathcal{Y}}}$, respectively. The synonymous mappings $f_X : \tilde{\mathcal{X}} \rightarrow 2^{\mathcal{X}}$ and $f_Y : \tilde{\mathcal{Y}} \rightarrow 2^{\mathcal{Y}}$ map semantic symbols to sets of syntactic symbols. The reification mappings $g_X = f_X \circ e_X : \underline{\mathcal{X}} \rightarrow 2^{\mathcal{X}}$ and $g_Y = f_Y \circ e_Y : \underline{\mathcal{Y}} \rightarrow 2^{\mathcal{Y}}$ directly bridge the pragmatic and syntactic layers.

For the joint behavior of channel input and output, we define the joint reification mapping, the joint synonymous mapping, and the joint isoteleia mapping. Their sequential extensions are defined element-wise.

Definition 37 (Joint Mappings for Sequences): Let $f_{XY} : \tilde{\mathcal{X}} \times \tilde{\mathcal{Y}} \rightarrow 2^{\mathcal{X} \times \mathcal{Y}}$ be the joint synonymous mapping, and let $e_{XY} : \underline{\mathcal{X}} \times \underline{\mathcal{Y}} \rightarrow 2^{\tilde{\mathcal{X}} \times \tilde{\mathcal{Y}}}$ be the joint isoteleia mapping. The joint reification mapping $g_{XY} = f_{XY} \circ e_{XY} : \underline{\mathcal{X}} \times \underline{\mathcal{Y}} \rightarrow 2^{\mathcal{X} \times \mathcal{Y}}$ is their composition. Their n -th sequential extensions are defined by:

$$\begin{cases} f_{XY}^n(\tilde{x}^n, \tilde{y}^n) \triangleq \prod_{k=1}^n f_{XY}(\tilde{x}_k, \tilde{y}_k), \\ e_{XY}^n(\underline{x}^n, \underline{y}^n) \triangleq \prod_{k=1}^n e_{XY}(\underline{x}_k, \underline{y}_k), \\ g_{XY}^n(\underline{x}^n, \underline{y}^n) \triangleq (f_{XY}^n \circ e_{XY}^n)(\underline{x}^n, \underline{y}^n) = \prod_{k=1}^n g_{XY}(\underline{x}_k, \underline{y}_k). \end{cases} \quad (204)$$

Thus, g_{XY}^n partitions the joint syntactic space $\mathcal{X}^n \times \mathcal{Y}^n$ into equivalence classes according to the pragmatic pair $(\underline{x}^n, \underline{y}^n)$.

We now define the relevant typical sets. Let $P(X, Y) = P(X)P(Y|X)$ denote the joint distribution induced by the channel. The syntactically jointly typical set $A_\epsilon^{(n)}$ is defined as in classical information theory [1, 45]:

$$\begin{aligned} A_\epsilon^{(n)} \triangleq \{ (x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \\ \left| -\frac{1}{n} \log P(x^n) - H(X) \right| < \epsilon, \\ \left| -\frac{1}{n} \log P(y^n) - H(Y) \right| < \epsilon, \\ \left| -\frac{1}{n} \log P(x^n, y^n) - H(X, Y) \right| < \epsilon \}. \end{aligned} \quad (205)$$

For the pragmatic layer, we define the pragmatically jointly typical set, which aggregates probabilities over reified equivalence classes.

Definition 38 (Pragmatically Jointly Typical Set): The pragmatically jointly typical set $\underline{A}_\epsilon^{(n)}$ is the set of pragmatic sequence pairs $(\underline{x}^n, \underline{y}^n) \in \underline{\mathcal{X}}^n \times \underline{\mathcal{Y}}^n$ such that

$$\underline{A}_\epsilon^{(n)} \triangleq \{ (\underline{x}^n, \underline{y}^n) : \left| -\frac{1}{n} \log P(\underline{x}^n) - H_p(\underline{X}) \right| < \epsilon, \quad$$

$$\begin{aligned} \left| -\frac{1}{n} \log P(\underline{y}^n) - H_p(\underline{Y}) \right| < \epsilon, \\ \left| -\frac{1}{n} \log P(\underline{x}^n, \underline{y}^n) - H_p(\underline{X}, \underline{Y}) \right| < \epsilon \}, \end{aligned} \quad (206)$$

where

$$\begin{aligned} P(\underline{x}^n, \underline{y}^n) &= \prod_{k=1}^n P(\underline{x}_k, \underline{y}_k), \quad P(\underline{x}_k, \underline{y}_k) \\ &= \sum_{(x_k, y_k) \in g_{XY}(\underline{x}_k, \underline{y}_k)} P(x_k, y_k). \end{aligned} \quad (207)$$

The marginal pragmatic probabilities $P(\underline{x}^n)$ and $P(\underline{y}^n)$ are obtained by summing over the other variable.

For a given pragmatic pair, the jointly reified typical set collects all syntactic pairs that share that pragmatic interpretation and are syntactically jointly typical. This set is naturally decomposed into synonymous typical sets corresponding to each semantic interpretation in the isoteleia typical set.

Definition 39 (Jointly Reified Typical Set): For a given pragmatic typical pair $(\underline{x}^n, \underline{y}^n) \in \underline{A}_\epsilon^{(n)}$, the jointly reified typical set $B_\epsilon^{(n)}(\underline{x}^n, \underline{y}^n)$ with the syntactically jointly typical sequences (x^n, y^n) is defined as the set of n -sequence pairs $(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n$ such that the following conditions hold:

$$\begin{aligned} B_\epsilon^{(n)}(\underline{x}^n, \underline{y}^n) &= \left\{ (x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \right. \\ &\left| -\frac{1}{n} \log P(x^n) - H(X) \right| < \epsilon, \\ &\left| -\frac{1}{n} \log P(y^n) - H(Y) \right| < \epsilon, \\ &\left| -\frac{1}{n} \log P(x^n, y^n) - H(X, Y) \right| < \epsilon, \\ &\left| -\frac{1}{n} \log P(\underline{x}^n) - H_p(\underline{X}) \right| < \epsilon, \\ &\left| -\frac{1}{n} \log P(\underline{y}^n) - H_p(\underline{Y}) \right| < \epsilon, \\ &\left| -\frac{1}{n} \log P(\underline{x}^n, \underline{y}^n) - H_p(\underline{X}, \underline{Y}) \right| < \epsilon, \\ &\left| -\frac{1}{n} \log P((\underline{x}^n, \underline{y}^n) \rightarrow (x^n, y^n)) \right. \\ &\quad \left. - (H(X, Y) - H_p(\underline{X}, \underline{Y})) \right| < \epsilon, \left. \right\} \end{aligned} \quad (208)$$

where

$$P((\underline{x}^n, \underline{y}^n) \rightarrow (x^n, y^n)) \triangleq \begin{cases} \frac{P(x^n, y^n)}{P(\underline{x}^n, \underline{y}^n)}, & \text{if } (x^n, y^n) \in g_{XY}^n(\underline{x}^n, \underline{y}^n), \\ 0, & \text{otherwise.} \end{cases} \quad (209)$$

The first three conditions ensure syntactic joint typicality of the sequence pair (x^n, y^n) . The fourth through sixth conditions ensure pragmatic joint typicality of the sequence pair $(\underline{x}^n, \underline{y}^n)$. The seventh condition enforces the reification typicality, i.e., the difference between the syntactic joint entropy and the pragmatic joint entropy is ϵ -close to the entropy reduction $H(X, Y) - H_p(\underline{X}, \underline{Y})$.

Equivalently, the jointly reified typical set can be expressed as the union of jointly synonymous typical sets over all semantic pairs in the corresponding isoteleia typical set:

$$B_\epsilon^{(n)}(\underline{x}^n, \underline{y}^n) = \bigcup_{(\tilde{x}^n, \tilde{y}^n) \in E_\epsilon^{(n)}(\underline{x}^n, \underline{y}^n)} S_\epsilon^{(n)}(\tilde{x}^n, \tilde{y}^n), \quad (210)$$

where $E_\epsilon^{(n)}(\underline{x}^n, \underline{y}^n)$ denotes the isoteleia typical set, and $S_\epsilon^{(n)}(\tilde{x}^n, \tilde{y}^n)$ denotes the jointly synonymous typical set. Thus, the jointly reified typical set partitions the syntactically jointly typical set according to the equivalence classes induced by the composition of the joint isoteleia and joint synonymous mappings, i.e., the joint reification mapping.

Under the sequential joint reification mapping g_{XY}^n , the syntactically jointly typical set $A_\epsilon^{(n)}$ can be partitioned into jointly reified typical sets:

$$A_\epsilon^{(n)} = \bigcup_{(\underline{x}^n, \underline{y}^n) \in \underline{A}_\epsilon^{(n)}} B_\epsilon^{(n)}(\underline{x}^n, \underline{y}^n), \quad (211)$$

with disjointness for distinct pragmatic typical pairs.

Fig. 12 illustrates the hierarchical relationship among the syntactically jointly typical set $A_\epsilon^{(n)}$, the pragmatically jointly typical set $\underline{A}_\epsilon^{(n)}$, and the jointly reified typical sets $B_\epsilon^{(n)}(\underline{x}^n, \underline{y}^n)$. The joint reification mapping $g_{XY}^n = f_{XY}^n \circ e_{XY}^n$ maps each syntactic sequence pair (x^n, y^n) to its corresponding pragmatic pair $(\underline{x}^n, \underline{y}^n)$. Under the joint reification mapping, $A_\epsilon^{(n)}$ is partitioned

into disjoint jointly reified typical sets, each corresponding to a distinct pragmatic typical pair $(\underline{x}^n, \underline{y}^n)$ in $\underline{A}_\epsilon^{(n)}$. Each jointly reified typical set $B_\epsilon^{(n)}(\underline{x}^n, \underline{y}^n)$ is itself the union of jointly synonymous typical sets $S_\epsilon^{(n)}(\tilde{x}^n, \tilde{y}^n)$ over all semantic pairs in the isoteleia typical set $E_\epsilon^{(n)}(\underline{x}^n, \underline{y}^n)$, reflecting the composition $g_{XY}^n = f_{XY}^n \circ e_{XY}^n$. The dashed circles within each reified set represent the individual synonymous typical sets, which group syntactic pairs sharing the same semantic interpretation. This nested structure visually demonstrates that the pragmatic abstraction merges semantically distinct but pragmatically equivalent pairs, thereby reducing the number of equivalence classes from $|A_\epsilon^{(n)}|$ to $|\underline{A}_\epsilon^{(n)}|$, consistent with the entropy reduction $H_p(\underline{X}, \underline{Y}) \leq H_s(\tilde{X}, \tilde{Y}) \leq H(X, Y)$ established in Section III.

We now establish the AEP properties of these typical sets. The syntactically joint AEP is the classical result [1, 45]; we restate it for reference.

Theorem 38 (Syntactically Joint AEP): Let (X^n, Y^n) be a sequence pair of length n drawn i.i.d. according to $P(x^n, y^n)$. Then for sufficiently large n :

- 1) $\Pr\{(X^n, Y^n) \in A_\epsilon^{(n)}\} > 1 - \epsilon$.
- 2) $(1 - \epsilon)2^{n(H(X, Y) - \epsilon)} \leq |A_\epsilon^{(n)}| \leq 2^{n(H(X, Y) + \epsilon)}$.
- 3) If $(\dot{X}^n, \dot{Y}^n) \sim P(x^n)P(y^n)$ are independent sequences with the same marginals, then

$$(1 - \epsilon)2^{-n(I(X; Y) + 3\epsilon)} \leq \Pr\{(\dot{X}^n, \dot{Y}^n) \in A_\epsilon^{(n)}\} \leq 2^{-n(I(X; Y) - 3\epsilon)}.$$

The pragmatically joint AEP follows by applying the weak law of large numbers to the aggregated pragmatic probabilities.

Theorem 39 (Pragmatically Joint AEP for Transmission): Let $(\underline{X}^n, \underline{Y}^n)$ be a pragmatic sequence pair of length n drawn i.i.d. according to $P(\underline{x}^n, \underline{y}^n)$, with the associated syntactic sequence pair (X^n, Y^n) under the joint reification mapping g_{XY}^n . Then for sufficiently large n :

- 1) $\Pr\{(\underline{X}^n, \underline{Y}^n) \in \underline{A}_\epsilon^{(n)}\} > 1 - \epsilon$.
- 2) For any $(\underline{x}^n, \underline{y}^n) \in \underline{A}_\epsilon^{(n)}$,
 $2^{-n(H_p(\underline{X}, \underline{Y}) + \epsilon)} \leq P(\underline{x}^n, \underline{y}^n) \leq 2^{-n(H_p(\underline{X}, \underline{Y}) - \epsilon)}$.
- 3) $(1 - \epsilon)2^{n(H_p(\underline{X}, \underline{Y}) - \epsilon)} \leq |\underline{A}_\epsilon^{(n)}| \leq 2^{n(H_p(\underline{X}, \underline{Y}) + \epsilon)}$.
- 4) If $(\dot{X}^n, \dot{Y}^n) \sim P(x^n)P(y^n)$ are independent sequences with the same marginals as X^n and Y^n , and

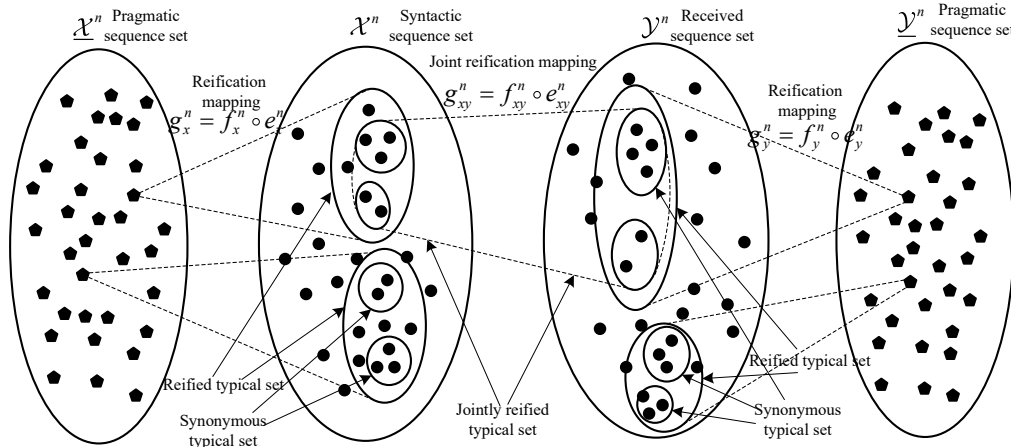


Fig. 12: The joint reification mapping g_{XY}^n maps pragmatic classes ($\underline{x}^n, \underline{y}^n$) to syntactic pairs (x^n, y^n). Each pragmatic typical pair corresponds to a jointly reified typical set $B_\epsilon^{(n)}(\underline{x}^n, \underline{y}^n)$, which is a union of synonymous typical sets over the isoteleia typical set.

($\underline{\dot{X}}^n, \underline{\dot{Y}}^n$) are the corresponding pragmatic sequences, then

$$(1 - \epsilon) 2^{-n(IP(\underline{X}; \underline{Y}) + 3\epsilon)} \leq \Pr\{(\underline{\dot{X}}^n, \underline{\dot{Y}}^n) \in \underline{A}_\epsilon^{(n)}\} \leq 2^{-n(IP(\underline{X}; \underline{Y}) - 3\epsilon)}. \quad (212)$$

The details of the proof are provided in Appendix D.

Finally, we establish the properties of the jointly reified typical set and its decomposition.

Theorem 40 (Properties of the Jointly Reified Typical Set): For any pragmatic typical pair $(\underline{x}^n, \underline{y}^n) \in \underline{A}_\epsilon^{(n)}$, for sufficiently large n :

1) For any $(x^n, y^n) \in B_\epsilon^{(n)}(\underline{x}^n, \underline{y}^n)$,

$$2^{-n(H(X, Y) - H_p(\underline{X}, \underline{Y}) + \epsilon)} \leq \frac{P(x^n, y^n)}{P(\underline{x}^n, \underline{y}^n)} \leq 2^{-n(H(X, Y) - H_p(\underline{X}, \underline{Y}) - \epsilon)}$$

2)

$$2^{n(H(X, Y) - H_p(\underline{X}, \underline{Y}) - \epsilon)} \leq |B_\epsilon^{(n)}(\underline{x}^n, \underline{y}^n)| \leq 2^{n(H(X, Y) - H_p(\underline{X}, \underline{Y}) + \epsilon)}$$

The details of the proof are provided in Appendix E.

Corollary 7 (Decomposition of the Jointly Reified Typical Set): The jointly reified typical set can be decomposed as the union of synonymous typical sets over the isoteleia typical set:

$$B_\epsilon^{(n)}(\underline{x}^n, \underline{y}^n) = \bigcup_{(\tilde{x}^n, \tilde{y}^n) \in E_\epsilon^{(n)}(\underline{x}^n, \underline{y}^n)} S_\epsilon^{(n)}(\tilde{x}^n, \tilde{y}^n).$$

Consequently, its cardinality satisfies

$$|B_\epsilon^{(n)}(\underline{x}^n, \underline{y}^n)| \approx |E_\epsilon^{(n)}(\underline{x}^n, \underline{y}^n)| \cdot |S_\epsilon^{(n)}(\tilde{x}^n, \tilde{y}^n)|, \quad (213)$$

where the approximation holds for any $\tilde{x}^n, \tilde{y}^n$ in the isoteleia typical set, and the product reflects the composition $g_{XY}^n = f_{XY}^n \circ e_{XY}^n$.

Remark 13: The jointly reified typical set $B_\epsilon^{(n)}(\underline{x}^n, \underline{y}^n)$ groups together all syntactically jointly typical pairs that share the same pragmatic interpretation. Its decomposition into synonymous and isoteleia typical sets reveals the two-level abstraction: the isoteleia mapping merges semantic differences that do not affect the optimal actions, while the synonymous mapping merges syntactic differences that do not affect the meaning. The cardinality of a jointly reified typical set is approximately $2^{n(H(X, Y) - H_p(\underline{X}, \underline{Y}))}$, which is larger than a single synonymous typical set by a factor of $2^{n(H_s(\tilde{X}, \tilde{Y}) - H_p(\underline{X}, \underline{Y}))}$, reflecting the additional syntactic freedom gained when only the pragmatic purpose needs to be preserved. When randomly selecting an independent pair of syntactic sequences, the probability that they fall into the same jointly reified typical set (i.e., are pragmatically jointly typical) is about $2^{-nIP(\underline{X}; \underline{Y})}$, which is not larger than the classical joint typicality probability $2^{-nI(X; Y)}$, reflecting the pragmatic capacity gain.

B. Pragmatic Channel Coding Theorem

We establish the fundamental limit for pragmatic transmission over noisy channels using a two-level code: the encoder selects a pragmatic index (optimal action) and then a syntactic codeword from its reified typical set; the decoder recovers only the index, tolerat-

ing syntactic errors. Via random coding and joint typical decoding based on the jointly reified typical set, we prove that the pragmatic channel capacity is the maximum achievable rate. We also connect this result to the pragmatic cost of information.

Definition 40 (Pragmatic Channel Code): An (M_p, M_r, n) pragmatic channel code consists of:

1) A pragmatic index set $\mathcal{J}_p = \{1, 2, \dots, M_p\}$, where each index $i_p \in \mathcal{J}_p$ corresponds to a distinct pragmatic purpose (optimal action class).

2) A reification index set $\mathcal{J}_r = \{1, 2, \dots, M_r\}$, where each index $i_r \in \mathcal{J}_r$ enumerates the syntactic codewords within each reified typical set.

3) An encoding function $\phi : \mathcal{J}_p \times \mathcal{J}_r \rightarrow \mathcal{X}^n$ that maps each pair (i_p, i_r) to a codeword $x^n(i_p, i_r)$ such that for each pragmatic index i_p , the set of codewords $\{x^n(i_p, i_r) : i_r \in \mathcal{J}_r\}$ lies entirely within the corresponding reified typical set $B_\epsilon^{(n)}(\underline{x}_{i_p}^n, \underline{y}_{i_p}^n)$ for some typical output pragmatic sequence $\underline{y}_{i_p}^n$. (The decoder will recover i_p without needing the exact i_r .)

4) A decoding function $\psi : \mathcal{Y}^n \rightarrow \mathcal{J}_p$ that maps the received sequence to a pragmatic index.

The pragmatic code rate is $R_p = \frac{1}{n} \log_2 M_p$ (prabits per channel use), and the reification rate is $R_r = \frac{1}{n} \log_2 M_r$ (bits per channel use). The total syntactic rate is

$$R_{\text{tot}} = R_p + R_r = \frac{1}{n} \log_2 (M_p M_r). \quad (214)$$

For a given code, the error probability is defined as

$$P_e^{(n)} = \Pr\{\psi(Y^n) \neq i_p\}, \quad (215)$$

where i_p is the intended pragmatic index, and Y^n is the channel output when the corresponding codeword is transmitted.

Figure 13 depicts the process of pragmatic channel coding. The encoder first selects a pragmatic index $i_p \in \mathcal{J}_p$ corresponding to the intended optimal action

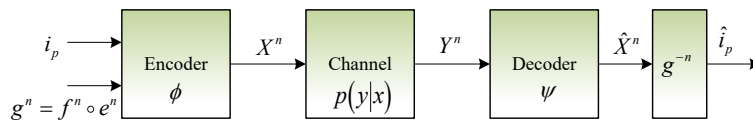


Fig. 13: Block diagram of pragmatic channel coding. The encoder selects a pragmatic index i_p and a reification index i_r , then transmits the corresponding codeword. The decoder recovers only the pragmatic index $\hat{i}_p$, ignoring syntactic variations within the reified typical set.

class. Given i_p , the encoder then chooses a reification index $i_r \in \mathcal{J}_r$, which selects a specific syntactic codeword from the reified typical set $B_\epsilon^{(n)}(\underline{x}_{i_p}^n, \underline{y}_{i_p}^n)$. The transmitted codeword $x^n(i_p, i_r)$ is sent over the channel. The decoder recovers only the pragmatic index $\hat{i}_p$ via jointly typical decoding based on the jointly reified typical set, ignoring syntactic variations within the reified typical set. Thus, the two-level structure separates the pragmatic purpose from its syntactic realizations, enabling reliable transmission at rates up to the pragmatic channel capacity.

Theorem 41 (Pragmatic Channel Coding Theorem): Let C_p be the pragmatic channel capacity defined in Eq. (120) of Section V-A, i.e.,

$$C_p = \max_{p(x)} \max_{g_{XY}} I^p(\underline{X}; \underline{Y}), \quad (216)$$

where $I^p(\underline{X}; \underline{Y}) = H(X) + H(Y) - H_p(\underline{X}, \underline{Y})$ is the up pragmatic mutual information. For any $\epsilon > 0$ and total rate R_{tot} :

1) **Achievability:** If $R_{\text{tot}} < C_p$, then for sufficiently large n , there exists an $(2^{n(R_p + R_r)}, n)$ pragmatic channel code with error probability $P_e^{(n)} < \epsilon$.

2) **Converse:** If $R_{\text{tot}} > C_p$, then for any $(2^{n(R_p + R_r)}, n)$ pragmatic channel code, the error probability $P_e^{(n)}$ is bounded away from zero for sufficiently large n .

Consequently, the pragmatic cost of information $\text{CoI}_p(R)$ is finite for $R \leq C_p$ and becomes unbounded for $R > C_p$, establishing C_p as the fundamental threshold for resource-limited pragmatic communication.

Proof. Achievability: Fix $\epsilon > 0$ and choose rates R_p, R_r such that $R_{\text{tot}} = R_p + R_r < C_p$. By the definition of C_p , there exists an input distribution $p(x)$ and a joint reification mapping g_{XY} such that $I^p(\underline{X}; \underline{Y}) > R_{\text{tot}}$. Let $R = I^p(\underline{X}; \underline{Y}) - \delta$ for some small $\delta > 0$, so that $R_{\text{tot}} < R < C_p$.

We construct a random codebook $\mathcal{C}$ of size

$2^{n(R_p+R_r)}$ by drawing each codeword independently according to $P(X^n) = \prod_{k=1}^n P(X_k)$. These codewords are then partitioned uniformly into 2^{nR_p} groups, each corresponding to a pragmatic index i_p , with 2^{nR_r} codewords per group. This partition is based on the reification mapping: all codewords in group i_p are chosen from the reified typical set $B_\epsilon^{(n)}(x_{i_p}^n, y_{i_p}^n)$ for some typical output pragmatic sequence.

The encoder, upon receiving a pragmatic message i_p , selects one of the 2^{nR_r} codewords in group i_p (any choice is sufficient, as all share the same pragmatic purpose) and transmits it. The decoder uses the jointly typical decoding rule based on the jointly reified typical set: upon receiving Y^n , it searches for the unique pragmatic index i_p such that there exists a codeword x^n in group i_p with $(x^n, Y^n) \in B_\epsilon^{(n)}(x_{i_p}^n, y_{i_p}^n)$, i.e., the pair is jointly reified typical. If no such index exists, or if multiple exist, an error is declared.

We now analyze the error probability. Assume codeword $x^n(1, 1)$ (from group $i_p = 1$) is transmitted. Define the following events:

- E_0 : The transmitted codeword and the received sequence are not jointly reified typical. By the jointly reified typical set properties (Theorem 39), the probability of this event tends to 0 as $n \rightarrow \infty$, i.e., $P(E_0) \leq \epsilon/2$ for large n .
- E_i for $i_p \neq 1$: There exists some codeword in a wrong pragmatic group that is jointly reified typical with the received sequence. For a fixed wrong group i_p , the number of codewords is 2^{nR_r} , and each is independent of the received sequence. By the probability bound in Theorem 39, the probability that a given codeword from a wrong group is jointly typical with Y^n is at most $2^{-n(IP(\underline{X}; \underline{Y})-3\epsilon)}$. Hence, by the union bound,

$$P(E_i) \leq 2^{nR_r} \cdot 2^{-n(IP(\underline{X}; \underline{Y})-3\epsilon)} = 2^{-n(IP(\underline{X}; \underline{Y})-R_r-3\epsilon)}.$$

Summing over all $2^{nR_p} - 1$ wrong groups gives

$$\begin{aligned} P\left(\bigcup_{i_p \neq 1} E_i\right) &\leq 2^{nR_p} \cdot 2^{-n(IP(\underline{X}; \underline{Y})-R_r-3\epsilon)} \\ &= 2^{-n(IP(\underline{X}; \underline{Y})-R_p-R_r-3\epsilon)} \end{aligned}$$

$$= 2^{-n(IP(\underline{X}; \underline{Y})-R_{\text{tot}}-3\epsilon)}. \quad (217)$$

Since $R_{\text{tot}} < IP(\underline{X}; \underline{Y})$, this probability tends to 0 as $n \rightarrow \infty$. Therefore, the total error probability $P_e^{(n)} \leq P(E_0) + \sum_{i_p \neq 1} P(E_i) < \epsilon$ for sufficiently large n . This proves the achievability.

Converse: Assume a code with total rate $R_{\text{tot}} > C_p$ and error probability $P_e^{(n)}$. Let W_p be the uniformly distributed pragmatic message, and let $\hat{W}_p$ be the decoder's estimate. By Fano's inequality,

$$H(W_p | \hat{W}_p) \leq 1 + P_e^{(n)} \log M_p = 1 + P_e^{(n)} nR_p.$$

Using the data processing inequality and the fact that $\underline{X}^n$ is a function of W_p and the codebook, we have

$$\begin{aligned} nR_p &= H(W_p) \leq I(W_p; Y^n) + H(W_p | Y^n) \\ &\leq I(\underline{X}^n; Y^n) + 1 + P_e^{(n)} nR_p. \end{aligned}$$

Now, by the definition of pragmatic capacity and the chain rule for up pragmatic mutual information (Theorem 12), we have $I(\underline{X}^n; Y^n) \leq IP(\underline{X}^n; \underline{Y}^n) \leq nC_p$. Thus,

$$nR_p \leq nC_p + 1 + P_e^{(n)} nR_p \implies R_p(1 - P_e^{(n)}) \leq C_p + \frac{1}{n}.$$

Since the total rate $R_{\text{tot}} = R_p + R_r$, and $R_r \geq 0$, we get

$$R_{\text{tot}} - C_p \leq P_e^{(n)} R_p + \frac{1}{n} + R_r. \quad (218)$$

Letting $n \rightarrow \infty$ and assuming $P_e^{(n)} \rightarrow 0$, we obtain $R_{\text{tot}} \leq C_p$, a contradiction. Hence, if $R_{\text{tot}} > C_p$, the error probability cannot tend to zero. This proves the converse.

Connection to Pragmatic Cost of Information:

The pragmatic cost of information $\text{CoI}_p(R)$ is the minimum resource required to achieve rate R . Since reliable transmission is possible iff $R \leq C_p$, $\text{CoI}_p(R)$ is finite for $R \leq C_p$ and increases to infinity as $R \rightarrow C_p$. Thus, C_p is the threshold beyond which no finite resource can guarantee reliable pragmatic communication. $\square$

Remark 14: The pragmatic channel coding theorem generalizes Shannon's classical theorem (when isoteleia and synonymous mappings are trivial, $C_p =$

C) and the semantic theorem (when only the synonymous mapping is nontrivial, $C_p = C_s$). By tolerating errors that do not affect the optimal action, pragmatic coding achieves a larger capacity. The two-level code structure separates the pragmatic purpose from its syntactic realizations, allowing the encoder to exploit the redundancy in syntactic representations while ensuring that the decoder recovers the essential pragmatic information. The jointly reified typical set serves as the foundation for both encoding and decoding, enabling reliable communication at rates up to C_p . The theorem also establishes a direct link between pragmatic capacity and the pragmatic cost of information, highlighting the resource trade-offs inherent in task-oriented communication.

IX. Pragmatic Lossy Source Coding

In this section, we investigate the fundamental limits of pragmatic lossy source coding, where the goal is to compress a source while preserving its pragmatic utility rather than its exact syntactic or semantic content. We extend the asymptotic equipartition property to the joint distribution of the source and its reconstruction at the pragmatic level. By introducing the reification mapping for both the source and the reconstruction sequences, we define the jointly reified typical set, which characterizes the equivalence classes of source-reconstruction pairs that share the same pragmatic interpretation. Using random coding and jointly typical encoding based on this typical set, we prove the pragmatic rate-distortion coding theorem, which establishes that the pragmatic rate-distortion function $R_p(D)$ is the fundamental limit for task-oriented compression.

A. Pragmatic Distortion and Jointly Typical Set

To establish the fundamental limits of pragmatic lossy source coding, we first extend the asymptotic equipartition property to the joint distribution of the source sequence and its reconstruction at the pragmatic level. Unlike the semantic approach, which relies on the joint synonymous mapping, we directly employ the reification mapping, which is the composition of the synonymous and isoteleia mappings, to aggregate syntactic sequences into pragmatic equivalence classes.

This approach provides a direct characterization of the equivalence of source and reconstruction pairs with respect to the optimal terminal actions.

Consider a discrete memoryless source X with alphabet $\mathcal{X}$ and probability mass function $P(X)$. Let $\tilde{X}$ be the associated semantic variable and $\underline{X}$ be the pragmatic variable induced by the isoteleia mapping $e_X : \mathcal{X} \rightarrow 2^{\tilde{X}}$ and the synonymous mapping $f_X : \tilde{X} \rightarrow 2^{\mathcal{X}}$, with the reification mapping $g_X = f_X \circ e_X : \mathcal{X} \rightarrow 2^{\mathcal{X}}$. Similarly, let $\hat{X}$ be the reconstruction variable with associated semantic $\tilde{\hat{X}}$ and pragmatic $\underline{\hat{X}}$, induced by the reification mapping $g_{\hat{X}} : \hat{\mathcal{X}} \rightarrow 2^{\hat{\mathcal{X}}}$.

For lossy source coding, the encoder produces a reconstruction sequence $\hat{X}^n$ based on a test channel $P(\hat{X}^n|X^n)$. The pragmatic distortion measure $d_p(\underline{x}, \underline{\hat{x}})$ quantifies the loss of utility when a pragmatic symbol $\underline{x}$ is represented by $\underline{\hat{x}}$. In practice, it can be defined as the utility loss:

$$d_p(\underline{x}, \underline{\hat{x}}) = \max_a \mathbb{E}[U(X, a)|\underline{x}] - \max_a \mathbb{E}[U(X, a)|\underline{\hat{x}}], \quad (219)$$

or any other non-negative function that reflects the degradation in decision quality. The average pragmatic distortion is then

$$\bar{d}_p = \mathbb{E}[d_p(\underline{X}, \underline{\hat{X}})] = \sum_{\underline{x}, \underline{\hat{x}}} P(\underline{x}, \underline{\hat{x}}) d_p(\underline{x}, \underline{\hat{x}}). \quad (220)$$

We now define the relevant typical sets. Let $P(X, \hat{X}) = P(X)P(\hat{X}|X)$ denote the joint distribution induced by the test channel. The syntactically jointly typical set $A_\epsilon^{(n)}$ is defined as in classical information theory [1, 45]:

$$A_\epsilon^{(n)} \triangleq \left\{ (x^n, \hat{x}^n) \in \mathcal{X}^n \times \hat{\mathcal{X}}^n : \begin{aligned} & \left| -\frac{1}{n} \log P(x^n) - H(X) \right| < \epsilon, \\ & \left| -\frac{1}{n} \log P(\hat{x}^n) - H(\hat{X}) \right| < \epsilon, \\ & \left| -\frac{1}{n} \log P(x^n, \hat{x}^n) - H(X, \hat{X}) \right| < \epsilon \end{aligned} \right\}. \quad (221)$$

For the pragmatic layer, we define the pragmatically jointly typical set, which aggregates probabilities over reified equivalence classes.

Definition 41 (Pragmatically Jointly Typical Set for Source and Reconstruction): The pragmatically jointly typical set $\underline{A}_\epsilon^{(n)}$ is the set of pragmatic sequence pairs $(\underline{x}^n, \underline{\hat{x}}^n) \in \underline{\mathcal{X}}^n \times \underline{\hat{\mathcal{X}}}^n$ such that

$$\begin{aligned} \underline{A}_\epsilon^{(n)} \triangleq \{(\underline{x}^n, \underline{\hat{x}}^n) : & \left| -\frac{1}{n} \log P(\underline{x}^n) - H_p(\underline{X}) \right| < \epsilon, \\ & \left| -\frac{1}{n} \log P(\underline{\hat{x}}^n) - H_p(\underline{\hat{X}}) \right| < \epsilon, \\ & \left| -\frac{1}{n} \log P(\underline{x}^n, \underline{\hat{x}}^n) - H_p(\underline{X}, \underline{\hat{X}}) \right| < \epsilon \}, \end{aligned} \quad (222)$$

where

$$\begin{aligned} P(\underline{x}^n, \underline{\hat{x}}^n) &= \prod_{k=1}^n P(\underline{x}_k, \underline{\hat{x}}_k), \\ P(\underline{x}_k, \underline{\hat{x}}_k) &= \sum_{(x_k, \hat{x}_k) \in g_{X\hat{X}}(\underline{x}_k, \underline{\hat{x}}_k)} P(x_k, \hat{x}_k), \end{aligned} \quad (223)$$

with $g_{X\hat{X}} = g_X \times g_{\hat{X}}$ being the joint reification mapping for the source and reconstruction. The marginal pragmatic probabilities are obtained by summing over the other variable.

For a given pragmatic pair, the jointly reified typical set collects all syntactic source reconstruction pairs that share that pragmatic interpretation and are syntactically jointly typical.

Definition 42 (Jointly Reified Typical Set for Source and Reconstruction): For a given pragmatic typical pair $(\underline{x}^n, \underline{\hat{x}}^n) \in \underline{A}_\epsilon^{(n)}$, the jointly reified typical set $B_\epsilon^{(n)}(\underline{x}^n, \underline{\hat{x}}^n)$ is defined as the set of syntactic pairs $(x^n, \hat{x}^n) \in \mathcal{X}^n \times \hat{\mathcal{X}}^n$ such that the following conditions hold:

$$\begin{aligned} B_\epsilon^{(n)}(\underline{x}^n, \underline{\hat{x}}^n) &= \{(\underline{x}^n, \underline{\hat{x}}^n) : \left| -\frac{1}{n} \log P(\underline{x}^n) - H(\underline{X}) \right| < \epsilon, \\ & \left| -\frac{1}{n} \log P(\underline{\hat{x}}^n) - H(\underline{\hat{X}}) \right| < \epsilon, \\ & \left| -\frac{1}{n} \log P(\underline{x}^n, \underline{\hat{x}}^n) - H(\underline{X}, \underline{\hat{X}}) \right| < \epsilon, \\ & \left| -\frac{1}{n} \log \frac{P(\underline{x}^n)}{P(\underline{\hat{x}}^n)} - (H(\underline{X}) - H_p(\underline{X})) \right| < \epsilon, \\ & \left| -\frac{1}{n} \log \frac{P(\underline{\hat{x}}^n)}{P(\underline{x}^n)} - (H(\underline{\hat{X}}) - H_p(\underline{\hat{X}})) \right| < \epsilon, \\ & \left| -\frac{1}{n} \log \frac{P(\underline{x}^n, \underline{\hat{x}}^n)}{P(\underline{x}^n)} - (H(\underline{X}, \underline{\hat{X}}) - H(\underline{X})) \right| < \epsilon, \\ & \left| -\frac{1}{n} \log \frac{P(\underline{x}^n, \underline{\hat{x}}^n)}{P(\underline{\hat{x}}^n)} - (H(\underline{X}, \underline{\hat{X}}) - H(\underline{\hat{X}})) \right| < \epsilon \}. \end{aligned}$$

$$H_p(\underline{X}, \underline{\hat{X}}) \Big| < \epsilon,$$

$$(x^n, \hat{x}^n) \in g_{X\hat{X}}^n(\underline{x}^n, \underline{\hat{x}}^n) \}. \quad (224)$$

The first three conditions ensure syntactic joint typicality. The fourth and fifth conditions enforce individual reification typicality for the source and reconstruction sequences. The sixth condition enforces joint reification typicality. The last condition guarantees that the syntactic pair maps to the given pragmatic pair under the joint reification mapping.

Under the sequential joint reification mapping $g_{X\hat{X}}^n$, the syntactically jointly typical set $A_\epsilon^{(n)}$ can be partitioned into jointly reified typical sets:

$$A_\epsilon^{(n)} = \bigcup_{(\underline{x}^n, \underline{\hat{x}}^n) \in \underline{A}_\epsilon^{(n)}} B_\epsilon^{(n)}(\underline{x}^n, \underline{\hat{x}}^n), \quad (225)$$

with disjointness for distinct pragmatic typical pairs.

As shown in Fig. 14, the joint reification mapping $g_{X\hat{X}}^n = f_{X\hat{X}}^n \circ e_{X\hat{X}}^n$ first collapses syntactic sequence pairs into pragmatic equivalence classes according to the optimal actions, yielding the pragmatically jointly typical set $\underline{A}_\epsilon^{(n)}$. For each pragmatic pair, the preimage is the jointly reified typical set $B_\epsilon^{(n)}$, which collects all syntactic pairs sharing the same pragmatic purpose. Within $B_\epsilon^{(n)}$, the isoteleia mapping further groups semantic interpretations that lead to the same action, while the synonymous mapping collects syntactic realizations of each meaning. Thus, $B_\epsilon^{(n)}$ is the union of synonymous typical sets over the isoteleia typical set. This hierarchical coarsening reduces the number of equivalence classes from $|A_\epsilon^{(n)}|$ to $|\underline{A}_\epsilon^{(n)}|$, consistent with the entropy hierarchy $H_p(\underline{X}, \underline{\hat{X}}) \leq H_s(\tilde{X}, \tilde{\hat{X}}) \leq H(X, \hat{X})$, and provides the basis for the pragmatic rate-distortion theorem.

We now establish the AEP properties of these typical sets. The syntactically joint AEP is the classical result [1, 45], we restate it for reference.

Theorem 42 (Syntactically Joint AEP): Let $(X^n, \hat{X}^n)$ be a sequence pair of length n drawn i.i.d. according to $P(x^n, \hat{x}^n)$. Then for sufficiently large n :

- 1) $\Pr\{(X^n, \hat{X}^n) \in A_\epsilon^{(n)}\} > 1 - \epsilon$.
- 2) $(1 - \epsilon)2^{n(H(X, \hat{X}) - \epsilon)} \leq |A_\epsilon^{(n)}| \leq 2^{n(H(X, \hat{X}) + \epsilon)}$.
- 3) If $(\hat{X}^n, \hat{\hat{X}}^n) \sim P(\hat{x}^n)P(\hat{\hat{x}}^n)$ are independent se-

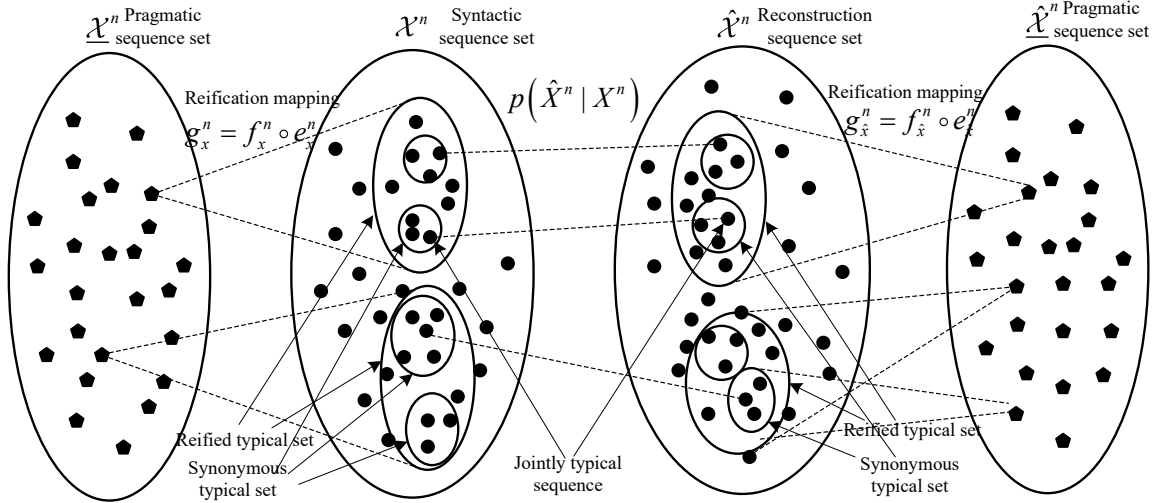


Fig. 14: Hierarchical typical sets for pragmatic lossy source coding. The joint reification mapping $g_{X\hat{X}}^n = f_{X\hat{X}}^n \circ e_{X\hat{X}}^n$ partitions the syntactic joint typical set $A_\epsilon^{(n)}$ into disjoint jointly reified typical sets $B_\epsilon^{(n)}(\underline{x}^n, \underline{\hat{x}}^n)$, each corresponding to a pragmatic pair in $\underline{A}_\epsilon^{(n)}$. Each $B_\epsilon^{(n)}$ is further decomposed into synonymous typical sets $S_\epsilon^{(n)}$ over the isoteleia typical sets $E_\epsilon^{(n)}$.

quences with the same marginals, then

$$(1 - \epsilon)2^{-n(I(X;\hat{X})+3\epsilon)} \leq \Pr\{(\hat{X}^n, \hat{X}^n) \in A_\epsilon^{(n)}\} \leq 2^{-n(I(X;\hat{X})-3\epsilon)}.$$

The pragmatically joint AEP follows by applying the weak law of large numbers to the aggregated pragmatic probabilities.

Theorem 43 (Pragmatically Joint AEP for Compression): Let $(\underline{X}^n, \underline{\hat{X}}^n)$ be a pragmatic sequence pair of length n drawn i.i.d. according to $P(\underline{x}^n, \underline{\hat{x}}^n)$, with the associated syntactic sequence pair $(X^n, \hat{X}^n)$ under the joint reification mapping $g_{X\hat{X}}^n$. Then for sufficiently large n :

- 1) $\Pr\{(\underline{X}^n, \underline{\hat{X}}^n) \in \underline{A}_\epsilon^{(n)}\} > 1 - \epsilon$.
- 2) For any $(\underline{x}^n, \underline{\hat{x}}^n) \in \underline{A}_\epsilon^{(n)}$,
 $2^{-n(H_p(\underline{X}, \underline{\hat{X}})+\epsilon)} \leq P(\underline{x}^n, \underline{\hat{x}}^n) \leq 2^{-n(H_p(\underline{X}, \underline{\hat{X}})-\epsilon)}$.
- 3) $(1 - \epsilon)2^{n(H_p(\underline{X}, \underline{\hat{X}})-\epsilon)} \leq |\underline{A}_\epsilon^{(n)}| \leq 2^{n(H_p(\underline{X}, \underline{\hat{X}})+\epsilon)}$.
- 4) If $(\hat{X}^n, \hat{X}^n) \sim P(x^n)P(\hat{x}^n)$ are independent sequences with the same marginals as X^n and $\hat{X}^n$, and $(\underline{\hat{X}}^n, \underline{\hat{X}}^n)$ are the corresponding pragmatic sequences, then

$$(1 - \epsilon)2^{-n(I_p(\underline{X}; \underline{\hat{X}})+3\epsilon)} \leq \Pr\{(\underline{\hat{X}}^n, \underline{\hat{X}}^n) \in \underline{A}_\epsilon^{(n)}\} \leq 2^{-n(I_p(\underline{X}; \underline{\hat{X}})-3\epsilon)}.$$

The details of the proof are provided in Appendix F.

Finally, we establish the properties of the jointly reified typical set.

Theorem 44 (Properties of the Jointly Reified Typical Set): For any pragmatic typical pair $(\underline{x}^n, \underline{\hat{x}}^n) \in \underline{A}_\epsilon^{(n)}$, for sufficiently large n :

- 1) For any $(x^n, \hat{x}^n) \in B_\epsilon^{(n)}(\underline{x}^n, \underline{\hat{x}}^n)$,

$$2^{-n(H(X, \hat{X})-H_p(\underline{X}, \underline{\hat{X}})+\epsilon)} \leq \frac{P(x^n, \hat{x}^n)}{P(\underline{x}^n, \underline{\hat{x}}^n)} \leq 2^{-n(H(X, \hat{X})-H_p(\underline{X}, \underline{\hat{X}})-\epsilon)}.$$

- 2)

$$2^{n(H(X, \hat{X})-H_p(\underline{X}, \underline{\hat{X}})-\epsilon)} \leq |B_\epsilon^{(n)}(\underline{x}^n, \underline{\hat{x}}^n)| \leq 2^{n(H(X, \hat{X})-H_p(\underline{X}, \underline{\hat{X}})+\epsilon)}.$$

Proof. Property (1) follows directly from the definition of $B_\epsilon^{(n)}(\underline{x}^n, \underline{\hat{x}}^n)$. The cardinality bounds are obtained by the same summation argument as in Theorem 36 of Section VIII-A. $\square$

Remark 15: The jointly reified typical set $B_\epsilon^{(n)}(\underline{x}^n, \underline{\hat{x}}^n)$ groups together all syntactically jointly typical source-reconstruction pairs that share the same pragmatic interpretation. Its definition directly employs the reification mapping $g_{X\hat{X}}^n = f_{X\hat{X}}^n \circ e_{X\hat{X}}^n$, which is the composition of the joint synonymous and joint isoteleia mappings. Consequently, the jointly reified typical set can be decomposed as the union of jointly synonymous typical sets over the isoteleia typical set:

$$B_\epsilon^{(n)}(\underline{x}^n, \underline{\hat{x}}^n) = \bigcup_{(\tilde{x}^n, \tilde{\hat{x}}^n) \in E_\epsilon^{(n)}(\underline{x}^n, \underline{\hat{x}}^n)} S_\epsilon^{(n)}(\tilde{x}^n, \tilde{\hat{x}}^n),$$

where $E_\epsilon^{(n)}(\underline{x}^n, \underline{\hat{x}}^n)$ is the isoteleia typical set and $S_\epsilon^{(n)}(\tilde{x}^n, \tilde{\hat{x}}^n)$ is the jointly synonymous typical set. The cardinality of a jointly reified typical set is approximately $2^{n(H(X, \hat{X}) - H_p(\underline{X}, \underline{\hat{X}}))}$, which is larger than a single synonymous typical set by a factor of $2^{n(H_s(\hat{X}, \underline{\hat{X}}) - H_p(\underline{X}, \underline{\hat{X}}))}$, reflecting the additional syntactic freedom gained when only the pragmatic purpose needs to be preserved.

B. Pragmatic Rate-Distortion Coding Theorem

We now investigate the fundamental limit of pragmatic lossy source coding. As depicted in Fig. 15, with the help of the reification mapping $g^n = f^n \circ e^n$, a pragmatic index i_p is mapped into the syntactic source sequence $X^n(i)$. Considering the distortion requirement, the encoder selects a suitable reconstruction codeword $\hat{X}^n(j)$ from a pragmatic class to represent the source, then sends the indices to the receiver. The decoder outputs the reconstruction, and after demapping recovers the estimated pragmatic index.

For a lossy source coding system, let $p(\hat{X}|X)$ be the test channel and $\mathcal{X}, \hat{\mathcal{X}}$ denote the source and reconstruction syntactic alphabets. The conditional distribution for the n -th extension is

$$p(\hat{x}^n|x^n) = \prod_{k=1}^n p(\hat{x}_k|x_k). \quad (226)$$

Definition 43 (Pragmatic Lossy Source Code): An (M, n) code for pragmatic lossy source coding consists of the following parts:

(1) A pragmatic index set $\mathcal{J}_p = \{1, \dots, M_p\}$ and a source reification index set $\mathcal{J}_r = \{1, \dots, M_r\}$; a reconstruction reification index set $\mathcal{J}_c = \{1, \dots, M_c\}$. The syntactic source and reconstruction index sets are $\mathcal{J} = \{1, \dots, M\}$ and $\mathcal{J}' = \{1, \dots, M'\}$.

(2) A reification mapping $g_X^n : \underline{\mathcal{X}}^n \rightarrow \mathcal{X}^n$ generates the set of source sequences, namely the pragmatic

sourcebook $\mathcal{S} = \{X^n(1), \dots, X^n(M')\}$. This book is partitioned into reified source subsets $\mathcal{S}_r$ according to the pragmatic index.

(3) An encoding function $\phi : \mathcal{X}^n \rightarrow \hat{\mathcal{X}}^n$ generates the reconstruction codebook

$\mathcal{C} = \{\hat{X}^n(1), \dots, \hat{X}^n(M)\}$, partitioned into pragmatic codeword subsets $\mathcal{C}_p$ via the reification mapping g_X^n .

(4) A decoding function $\psi : \hat{\mathcal{X}}^n \rightarrow \hat{\mathcal{X}}^n$ outputs the reconstruction $\hat{X}^n$.

(5) After demapping, $g_X^n(\hat{X}^n) = \hat{i}_p$, the estimated pragmatic index is obtained. Both ψ and g_X^n are deterministic.

The quotient sets $\mathcal{S}/g_X^n = \{\mathcal{S}_r\}$ and $\mathcal{C}/g_X^n = \{\mathcal{C}_p\}$ have sizes $|\mathcal{S}/g_X^n| = M_r$ and $|\mathcal{C}/g_X^n| = M_p$. Define the pragmatic code rate $R = \frac{1}{n} \log_2 M_p$, the source reification rate $R_r = \frac{1}{n} \log_2 M_r$, and the reconstruction reification rate $R_c = \frac{1}{n} \log_2 M_c$. We assume each reified set has equal size: $|\mathcal{S}_r| = M'/M_r = 2^{nR_r}$ and $|\mathcal{C}_p| = M/M_p = 2^{nR_c}$. The total syntactic rate is $R_{\text{tot}} = R + R_c$.

Theorem 45 (Pragmatic Rate-Distortion Coding Theorem): Given an i.i.d. syntactic source $X \sim p(x)$ with associated pragmatic variable $\underline{X}$ under the reification mapping g_X , and a bounded pragmatic distortion measure $d_p(\underline{x}, \underline{\hat{x}})$, define the pragmatic rate-distortion function as

$$R_p(D) = \min_{g_X, \hat{g}_X} \min_{\substack{p(\hat{x}|x): \\ \mathbb{E}[d_p(\underline{X}, \underline{\hat{X}})] \leq D}} I_p(\underline{X}; \underline{\hat{X}}), \quad (227)$$

where $I_p(\underline{X}; \underline{\hat{X}}) = H_p(\underline{X}) + H_p(\underline{\hat{X}}) - H(X, \hat{X})$ is the down pragmatic mutual information.

For any $\epsilon > 0$:

1) **Achievability**: If $R > R_p(D)$, then for sufficiently large n , there exists an $(2^{n(R+R_c)}, n)$ code such that the average pragmatic distortion satisfies $\bar{d}_p \leq D + \epsilon$.

2) **Converse**: If $R < R_p(D)$, then for any

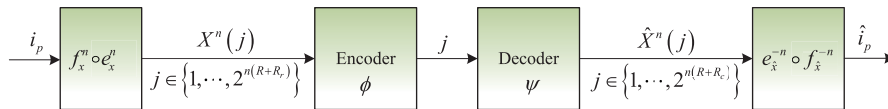


Fig. 15: Block diagram of pragmatic lossy source coding.

$(2^{n(R+R_c)}, n)$ code, the average distortion $\bar{d}_p > D - \epsilon$ for sufficiently large n .

Moreover, the pragmatic Value of Information (VoI) achieved by the reconstruction is maximized (i.e., equals the VoI obtained from the full pragmatic variable $\underline{X}$) if and only if $R \geq R_p(D)$. If $R < R_p(D)$, the achievable pragmatic VoI is strictly less than the full value, and the gap is determined by the rate shortfall.

Proof. Achievability: Fix $\epsilon > 0$ and choose a test channel $p(\hat{x}|x)$ such that $\mathbb{E}[d_p(\underline{X}, \hat{\underline{X}})] \leq D/(1 + \epsilon)$. Let $I_p = I_p(\underline{X}; \hat{\underline{X}})$ and select a pragmatic rate R satisfying $R_p(D) < R < I_p - 3\epsilon$.

Generate a pragmatic sourcebook $\mathcal{S}$ of size $2^{n(R+R_r)}$ by drawing sequences independently from $p(x^n) = \prod_{k=1}^n p(x_k)$. Partition it uniformly into 2^{nR} groups $\mathcal{S}_r(i_r)$ according to the reification mapping $g_{\hat{X}}^n$, each of size 2^{nR_r} , such that $\mathcal{S}_r(i_r) \subset B_\epsilon^{(n)}(\underline{x}_{i_r})$ for some representative pragmatic sequence.

Generate a reconstruction codebook $\mathcal{C}$ of size $2^{n(R+R_c)}$ independently from $p(\hat{x}^n) = \prod_{k=1}^n p(\hat{x}_k)$, where $p(\hat{x}) = \sum_x p(x)p(\hat{x}|x)$. Partition it uniformly into 2^{nR} groups $\mathcal{C}_p(i_p)$ according to $g_{\hat{X}}^n$, each of size 2^{nR_c} , with $\mathcal{C}_p(i_p) \subset B_\epsilon^{(n)}(\hat{\underline{x}}_{i_p})$.

Given a source sequence X^n , the encoder first finds the unique source reification group $\mathcal{S}_r(w_r)$ containing it. Then, based on the distortion criterion, it searches for a pragmatic index w_p and a codeword $\hat{X}^n(w_p, l) \in \mathcal{C}_p(w_p)$ such that $(X^n, \hat{X}^n(w_p, l)) \in B_\epsilon^{(n)}$ (jointly reified typical). If more than one such pair exists, choose the smallest index; if none, set $w_p = 1$ and $l = 1$. The encoder sends the indices (w_p, l) . Similar to the classical random coding argument [46], we analyze the probability of encoding error and the expected distortion.

Define the error event $\mathcal{E}$ as the event that the chosen pair is not jointly reified typical. It can be decomposed into

$$\mathcal{E} = \mathcal{E}_1 \cup \mathcal{E}_2,$$

where $\mathcal{E}_1 = \{X^n \notin A_\epsilon^{(n)}(X^n)\}$, and

$$\mathcal{E}_2 = \{X^n \in \mathcal{S}_r(w_r), \hat{X}^n(w_p, l) \in \mathcal{C}_p(w_p),$$

$$(X^n, \hat{X}^n(w_p, l)) \notin B_\epsilon^{(n)}, \forall w_p, l\}.$$

By the AEP, $\Pr(\mathcal{E}_1) \rightarrow 0$ as $n \rightarrow \infty$. For $\mathcal{E}_2$, using the fact that $\mathcal{S}_r(w_r) \subset B_\epsilon^{(n)}(\underline{x}_{w_r})$ and $\mathcal{C}_p(w_p) \subset B_\epsilon^{(n)}(\hat{\underline{x}}_{w_p})$, we have

$$\begin{aligned} P(\mathcal{E}_2) &= \sum_{x^n \in A_\epsilon^{(n)}} p(x^n) \prod_{w_p=1}^{2^{nR}} \Pr((X^n, \hat{X}^n(w_p, l)) \\ &\quad \notin B_\epsilon^{(n)} \mid X^n = x^n) \\ &\leq \sum_{x^n \in A_\epsilon^{(n)}} p(x^n) \left[1 - \Pr((X^n, \hat{X}^n(1, l)) \in B_\epsilon^{(n)})\right]^{2^{nR}}. \end{aligned} \quad (228)$$

Since $x^n \in A_\epsilon^{(n)}$ and $\hat{X}^n(1, l)$ is drawn independently from $p(\hat{x}^n)$, by the pragmatically joint AEP (Theorem 43), the probability that the pair is jointly reified typical is bounded below by

$$\Pr((X^n, \hat{X}^n(1, l)) \in B_\epsilon^{(n)}) \geq (1 - \epsilon)2^{-n(I_p + 3\epsilon)}, \quad (229)$$

for sufficiently large n . This bound follows from the fact that the probability that an independent pair falls into the pragmatically jointly typical set $\underline{A}_\epsilon^{(n)}$ is at least $(1 - \epsilon)2^{-n(I_p + 3\epsilon)}$, and $B_\epsilon^{(n)} \subseteq \underline{A}_\epsilon^{(n)}$.

Using $(1 - x)^m \leq e^{-mx}$ for $x \in [0, 1]$ and $m \geq 0$, we obtain

$$\begin{aligned} P(\mathcal{E}_2) &\leq \left(1 - (1 - \epsilon)2^{-n(I_p + 3\epsilon)}\right)^{2^{nR}} \\ &\leq \exp\left(-(1 - \epsilon)2^{n(R - I_p - 3\epsilon)}\right). \end{aligned} \quad (230)$$

Since $R < I_p - 3\epsilon$, the exponent tends to $-\infty$, hence $P(\mathcal{E}_2) \rightarrow 0$. Therefore the total encoding error probability $P_e^{(n)} \rightarrow 0$.

Conditioned on the success event $\mathcal{E}^c$ (i.e., the pair is jointly reified typical), the average distortion is bounded by

$$\mathbb{E}[d_p(\underline{X}^n, \hat{\underline{X}}^n) \mid \mathcal{E}^c] \leq (1 + \epsilon)\mathbb{E}[d_p(\underline{X}, \hat{\underline{X}})] \leq D.$$

Let $d_{\max}$ be the maximum distortion. The overall expected distortion satisfies

$$\mathbb{E}[d_p(\underline{X}^n, \hat{\underline{X}}^n)] \leq P_e^{(n)}d_{\max} + (1 - P_e^{(n)})D \leq D + \epsilon, \quad (231)$$

for sufficiently large n . This proves achievability.

Converse: Assume a sequence of $(2^{n(R+R_c)}, n)$ codes with average distortion $\bar{d}_p \leq D$. Let W_p be the uniform pragmatic message (index i_p), and let $\hat{W}_p$ be its estimate at the decoder. We will show that $R \geq R_p(D)$ must hold.

First, by the data processing inequality and the fact that $W_p \rightarrow X^n \rightarrow \hat{X}^n$ forms a Markov chain (where $\hat{X}^n$ is the reconstruction), we have

$$I(W_p; X^n) \geq I(\hat{X}^n(W_p); X^n), \quad (232)$$

where $\hat{X}^n(W_p)$ denotes the pragmatic reconstruction associated with the chosen codeword. Since W_p is a function of $\hat{X}^n(W_p)$, we have $H(W_p) \leq I(W_p; X^n) + H(W_p | X^n)$. However, a more direct route is used in [24].

We proceed as follows:

$$nR = H(W_p) \geq I(W_p; X^n) \quad (233)$$

$$\geq I(\hat{X}^n(W_p); X^n) \quad (234)$$

$$\geq I_p(X^n; \hat{X}^n). \quad (235)$$

Inequality (234) follows from the data processing inequality because $\hat{X}^n$ is a function of W_p (through the codebook). Inequality (235) is a pragmatic analogue of Lemma 10 in [24, 25], which states that the mutual information between the syntactic source and the pragmatic reconstruction is at least the down pragmatic mutual information between the pragmatic source and the pragmatic reconstruction. Indeed, by the definition of down pragmatic mutual information and the fact that the reconstruction variable is a function of the channel output, one can verify that $I(X^n; \hat{X}^n) \geq I_p(X^n; \hat{X}^n)$ (this follows from the entropy hierarchy and the data processing inequality).

Since the source is memoryless, the down pragmatic mutual information satisfies

$$I_p(X^n; \hat{X}^n) = \sum_{k=1}^n I_p(\underline{X}_k; \hat{\underline{X}}_k), \quad (236)$$

because the pragmatic variables are also memoryless due to the deterministic reification mapping.

By the definition of the pragmatic rate-distortion function $R_p(D)$, for each k , we have

$$I_p(\underline{X}_k; \hat{\underline{X}}_k) \geq R_p(\mathbb{E}[d_p(\underline{X}_k, \hat{\underline{X}}_k)]). \quad (237)$$

Summing over k and using the convexity of $R_p(D)$, we obtain

$$\begin{aligned} \sum_{k=1}^n I_p(\underline{X}_k; \hat{\underline{X}}_k) &\geq \sum_{k=1}^n R_p(\mathbb{E}[d_p(\underline{X}_k, \hat{\underline{X}}_k)]) \\ &\geq nR_p\left(\frac{1}{n} \sum_{k=1}^n \mathbb{E}[d_p(\underline{X}_k, \hat{\underline{X}}_k)]\right). \end{aligned} \quad (238)$$

The average distortion over the entire sequence is $\bar{d}_p =$

$$\frac{1}{n} \sum_{k=1}^n \mathbb{E}[d_p(\underline{X}_k, \hat{\underline{X}}_k)] \leq D \text{ by assumption. Since } R_p(D)$$

is non-increasing, $R_p(\bar{d}_p) \geq R_p(D)$. Therefore,

$$I_p(X^n; \hat{X}^n) \geq nR_p(\bar{d}_p) \geq nR_p(D). \quad (239)$$

Combining (233)–(235) with (239), we obtain

$$nR \geq nR_p(D) \implies R \geq R_p(D). \quad (240)$$

Thus, if $R < R_p(D)$, the assumption $\bar{d}_p \leq D$ leads to a contradiction. Hence, for any code with $R < R_p(D)$, the average distortion must satisfy $\bar{d}_p > D - \epsilon$ for sufficiently large n . This completes the converse.

Connection to Pragmatic Value of Information

The pragmatic Value of Information achieved by the reconstruction $\hat{X}$ is

$$\text{VoI}_p(\hat{X}) = \mathbb{E}_{X, \hat{X}} [U(X, a^*(\hat{X}))] - U_0, \quad (241)$$

where $a^*(\hat{X})$ is the optimal action based on the reconstructed pragmatic class. The full VoI (as if $\underline{X}$ were directly observed) is attained if and only if the reconstruction preserves all pragmatic distinctions, which requires $I_p(\underline{X}; \hat{\underline{X}}) = H_p(\underline{X})$. This occurs only when $R \geq H_p(\underline{X})$. More generally, for a given distortion level D , the full VoI is attainable iff $R \geq R_p(D)$. If $R < R_p(D)$, the loss in pragmatic information reduces the achievable utility. $\square$

Remark 16: The pragmatic rate-distortion theorem generalizes both the classical rate-distortion theorem [1, 45] and the semantic rate-distortion theorem [24, 25]. When the reification mapping g is trivial (identity), $R_p(D) = R(D)$; when only the synonymous mapping

f is non-trivial, we recover $R_p(D) = R_s(D)$. In general, the hierarchy $R_p(D) \leq R_s(D) \leq R(D)$ holds. Although our code uses three levels (pragmatic + two reification rates), the reification mapping $g = f \circ e$ implicitly includes the semantic layer. The extra reification rates R_r and R_c allow flexible trade-offs: setting them to zero achieves the pure pragmatic compression limit $R_p(D)$, while increasing them preserves additional syntactic or semantic details, recovering the classical or semantic limits. This modularity is consistent with the design principles of pragmatic information systems.

X. Pragmatic Information Measure of Continuous Message

In this section, we extend the pragmatic information measures to continuous messages, completing the theoretical framework with both bilateral and single-sided characterizations. Building upon the isoteleic volume formalism, we derive the pragmatic capacity and rate-distortion functions for Gaussian channels and sources, respectively. More importantly, we introduce the corresponding pragmatic cost of information (CoI) and pragmatic value of information (VoI) for both bilateral and single-sided cases, establishing the full duality spectrum: CoI is the inverse of the achievable rate functions, while VoI is the Legendre-Fenchel dual of the rate-distortion functions. These results provide a complete information-theoretic and decision-theoretic characterization of task-oriented communication over continuous alphabets, encompassing perceptual (observation-to-action) and expressive (action-to-reconstruction) asymmetric scenarios.

A. Pragmatic Entropy and Pragmatic Mutual Information for Continuous Message

In this subsection, we extend the pragmatic information measures, such as pragmatic entropy and pragmatic mutual information, to the domain of continuous messages. Unlike the syntactic layer, which deals with exact signal values, or the semantic layer, which deals with meanings, the pragmatic layer is concerned with the consequences of actions. For continuous signals, the equivalence is determined by whether different realizations lead to the same optimal action, as defined

by a utility function. This gives rise to the concept of *isoteleic volumes*, which are the continuous analogues of the pragmatic equivalence classes (fibers) introduced in Section II.

The coding theorems established in Sections VII–IX are developed for discrete, finite alphabets, where the syntactic, semantic, and pragmatic layers are characterized by quotient spaces of finite cardinality. For continuous alphabets, the pragmatic equivalence classes induced by the isoteleia mapping become uncountable sets, and the discrete entropy measures must be replaced by their differential counterparts. To bridge the discrete and continuous formulations, we adopt the standard quantization argument: partition the continuous space $\mathcal{W}$ into small cells of volume Δ , apply the discrete theory to the resulting finite alphabet, and then take the limit $\Delta \rightarrow 0$. Under this limiting procedure, the quotient-space structure of the isoteleia mapping converges to the continuous notion of *isoteleic volumes* Ω , which quantify the average measure of signal space that maps to the same optimal action. The entropy hierarchy $H_p \leq H_s \leq H$ is preserved in the limit, with the discrete pragmatic entropy giving way to the differential pragmatic entropy defined below. Consequently, all subsequent Gaussian channel and source characterizations are consistent with—and indeed inherit the operational meaning of—the discrete coding theorems proved earlier.

To formalize pragmatic information for continuous messages, we first define the isoteleia mapping for continuous variables and introduce the notion of isoteleic volumes.

Definition 44 (Isoteleic Volume): Let W be a continuous random variable with alphabet $\mathcal{W} \subseteq \mathbb{R}^d$ and probability density function $p(w)$. Let $\tilde{\mathcal{W}}$ be the associated semantic alphabet, and let $\underline{\mathcal{W}}$ be the pragmatic alphabet. The pragmatic variable $\underline{W}$ is induced by the isoteleia mapping $e : \underline{\mathcal{W}} \rightarrow 2^{\tilde{\mathcal{W}}}$, which, in turn, is derived from a utility function $U(s, a)$ defined on states and actions.

For each pragmatic symbol $\underline{w}_i \in \underline{\mathcal{W}}$, its *isoteleic volume* Ω_i is defined as the total measure of the set of all semantic (and hence syntactic) realizations that map to the same optimal action. Formally,

$$\Omega_i \triangleq |\{w \in \mathcal{W} : a^*(w) = \underline{w}_i\}|, \quad (242)$$

where $a^*(w)$ is the optimal action for the realization w , defined by

$$a^*(w) = \arg \max_{a \in \mathcal{A}} \mathbb{E}_{S|w}[U(S, a)]. \quad (243)$$

The collection $\{\Omega_i\}_{i=1}^{|\mathcal{W}|}$ forms a partition of the support of W , i.e., $\bigcup_i \Omega_i = \mathcal{W}$ and $\Omega_i \cap \Omega_j = \emptyset$ for $i \neq j$.

Analogous to the average synonymous length S in semantic information theory, we define the *average isoteleic volume* as the expectation of the isoteleic volume under the distribution of W .

Definition 45 (Average Isoteleic Volume): The average isoteleic volume Ω is defined as

$$\Omega \triangleq \mathbb{E}_{w \sim p(w)} [|\Omega_{a^*(w)}|] = \sum_{i=1}^{|\mathcal{W}|} \Pr(\underline{w}_i) \cdot |\Omega_i|, \quad (244)$$

where $\Pr(\underline{w}_i) = \int_{\Omega_i} p(w)dw$ is the probability mass of the pragmatic class $\underline{w}_i$.

The average isoteleic volume Ω quantifies the “compression” achieved by grouping different continuous signals into the same pragmatic class. A larger Ω indicates that more distinct signals lead to the same optimal action, thus reducing decision uncertainty.

A.1 Pragmatic Entropy for Continuous Messages

We now define the pragmatic entropy for a continuous random variable. The definition is derived by quantizing the continuous space and taking the limit, analogous to the derivation of differential entropy.

Definition 46 (Pragmatic Entropy for Continuous Variables): Let W be a continuous random variable with probability density function $p(w)$, supported on $\mathcal{W} \subseteq \mathbb{R}^d$. Let $\underline{W}$ be the associated pragmatic variable induced by the isoteleia mapping, with average isoteleic volume Ω . The pragmatic entropy of $\underline{W}$ is defined as

$$H_p(\underline{W}) \triangleq - \int_{\mathcal{W}} p(w) \log p(w) dw - \log \Omega, \quad (245)$$

where $\log \Omega = \sum_i \Pr(\underline{w}_i) \log |\Omega_i|$ is the logarithm of the average isoteleic volume.

Remark 17: The first term in Eq. (245) is the classical differential entropy $h(W) = - \int p(w) \log p(w) dw$.

The second term, $\log \Omega$, accounts for the reduction in uncertainty due to pragmatic abstraction. Unlike the discrete case, the pragmatic differential entropy $H_p(\underline{W})$ is not an absolute measure of information but a relative one, defined up to a reference measure. To maintain a meaningful non-negative pragmatic entropy (in prabits), the average isoteleic volume Ω must satisfy: $\log \Omega \leq h(W)$. When the isoteleia mapping is trivial (i.e., each signal leads to a unique action, $\Omega = 1$), the pragmatic entropy reduces to the differential entropy: $H_p(\underline{W}) = h(W)$.

Corollary 8 (Lower Bound on Pragmatic Entropy): The pragmatic entropy is lower bounded by

$$H_p(\underline{W}) \geq h(W) - \log \Omega. \quad (246)$$

Equality holds when the isoteleic volumes are chosen proportionally to the probability mass of the pragmatic classes, i.e., when the partition is optimal.

Example 5 (Uniform Distribution): Let W be uniformly distributed over an interval of length L , i.e., $p(w) = 1/L$ for $w \in [0, L]$. Suppose the isoteleia mapping partitions this interval into K equal isoteleic volumes, each of size $\Omega_i = L/K$. Then the average isoteleic volume is $\Omega = L/K$. The pragmatic entropy is

$$H_p(\underline{W}) = - \int_0^L \frac{1}{L} \log \frac{1}{L} dw - \log \frac{L}{K} = \log K \text{ prabits.} \quad (247)$$

This result is analogous to the semantic entropy for a uniformly distributed source, but here the interpretation is in terms of optimal actions rather than meanings.

Example 6 (Gaussian Distribution): Let $W \sim \mathcal{N}(0, \sigma^2)$ be a Gaussian random variable. Under an optimal isoteleia mapping with average isoteleic volume Ω , the pragmatic entropy is

$$\begin{aligned} H_p(\underline{W}) &= \mathbb{E}[-\log p(W)] - \log \Omega \\ &= \frac{1}{2} \log(2\pi e \sigma^2) - \log \Omega \\ &= \frac{1}{2} \log \left(\frac{2\pi e \sigma^2}{\Omega^2} \right) \text{ prabits.} \end{aligned} \quad (248)$$

This shows that the pragmatic entropy of a Gaussian source decreases logarithmically with the average isoteleic volume.

The definitions can be extended to multiple continuous variables. Let (W, V) be a pair of continuous random variables with joint density $p(w, v)$, and let $(\underline{W}, \underline{V})$ be their associated pragmatic variables, induced by a joint isoteleia mapping with average isoteleic volume Ω_{wv} .

Definition 47 (Joint Pragmatic Entropy for Continuous Variables): The joint pragmatic entropy of $(\underline{W}, \underline{V})$ is defined as

$$H_p(\underline{W}, \underline{V}) \triangleq - \int_{\mathcal{W} \times \mathcal{V}} p(w, v) \log p(w, v) dw dv - \log \Omega_{wv}, \quad (249)$$

where Ω_{wv} is the average joint isoteleic volume.

Similarly, the conditional pragmatic entropy of $\underline{V}$ given $\underline{W}$ is defined as

$$H_p(\underline{V}|\underline{W}) \triangleq - \int_{\mathcal{W} \times \mathcal{V}} p(w, v) \log p(v|w) dw dv - \log \Omega_v, \quad (250)$$

where Ω_v is the average isoteleic volume for V .

A.2 Pragmatic Mutual Information for Continuous Variables

We now define the pragmatic mutual information for continuous variables. As in the discrete case, there are two variants: the *up pragmatic mutual information* and the *down pragmatic mutual information*.

Definition 48 (Up Pragmatic Mutual Information for Continuous Variables): Let (W, V) be a pair of continuous random variables with joint density $p(w, v)$. Let $\underline{W}$ and $\underline{V}$ be their pragmatic counterparts, induced by a joint isoteleia mapping with average isoteleic volumes Ω_w and Ω_v . The up pragmatic mutual information is defined as

$$\begin{aligned} I^p(\underline{W}; \underline{V}) &\triangleq H(W) + H(V) - H_p(\underline{W}, \underline{V}) \\ &= - \int_{\mathcal{W} \times \mathcal{V}} p(w, v) \log \frac{p(w)p(v)}{p(w, v)} dw dv + \log(\Omega_w \Omega_v). \end{aligned} \quad (251)$$

Definition 49 (Down Pragmatic Mutual Information for Continuous Variables): The down pragmatic mutual information is defined as

$$\begin{aligned} I_p(\underline{W}; \underline{V}) &\triangleq H_p(\underline{W}) + H_p(\underline{V}) - H(W, V) \\ &= - \int_{\mathcal{W} \times \mathcal{V}} p(w, v) \log \frac{p(w)p(v)}{p(w, v)} dw dv - \log(\Omega_w \Omega_v). \end{aligned} \quad (252)$$

Remark 18: The first term in both (251) and (252) is the classical mutual information $I(W; V)$. The difference lies in the second term. If $\mathbb{E}[\log(\Omega_w \Omega_v)] \geq 0$, then the up pragmatic mutual information is greater than or equal to the classical mutual information, while the down pragmatic mutual information is less than or equal to it. In practice, the down pragmatic mutual information may be negative; in such cases, one can take its positive part, $(I_p(\underline{W}; \underline{V}))^+$.

The relationship between these measures is summarized by the following hierarchy, which mirrors the discrete case:

$$I_p(\underline{W}; \underline{V}) \leq I(W; V) \leq I^p(\underline{W}; \underline{V}). \quad (253)$$

Example 7 (Gaussian Variables): Let (W, V) be jointly Gaussian with correlation coefficient ρ and marginal variances σ_w^2 and σ_v^2 . The classical mutual information is $I(W; V) = -\frac{1}{2} \log(1 - \rho^2)$. Under a joint isoteleia mapping with average volumes Ω_w and Ω_v , the up and down pragmatic mutual informations are

$$I^p(\underline{W}; \underline{V}) = -\frac{1}{2} \log(1 - \rho^2) + \log(\Omega_w \Omega_v), \quad (254)$$

$$I_p(\underline{W}; \underline{V}) = -\frac{1}{2} \log(1 - \rho^2) - \log(\Omega_w \Omega_v). \quad (255)$$

This demonstrates how pragmatic abstraction can either enhance or reduce the measured information, depending on whether the coarsening is applied to the joint entropy or to the marginal entropies.

This completes the extension of pragmatic entropy and mutual information to continuous messages. In the following subsections, we will apply these measures to derive the pragmatic capacity of the Gaussian channel and the pragmatic rate-distortion function for Gaussian sources, along with their associated cost and value counterparts.

B. Pragmatic Capacity of Gaussian Channel and Pragmatic Cost of Information

We now investigate the pragmatic capacity of the additive white Gaussian noise (AWGN) channel and its dual—the pragmatic cost of information. The received signal at time k is modeled as

$$Y_k = X_k + Z_k, \quad (256)$$

with X_k being the channel input, $Z_k \sim \mathcal{N}(0, \sigma^2)$ i.i.d. Gaussian noise, and Y_k the output. The channel is subject to an average power constraint $\mathbb{E}[X^2] \leq P$.

Let $\underline{X}$ and $\underline{Y}$ be the pragmatic variables associated with the input and output, respectively, induced by the isoteleia mappings e_X and e_Y (and their compositions with the synonymous mappings) according to a given utility function. The pragmatic channel capacity is defined as the maximum achievable rate of pragmatic information transmission, i.e.,

$$C_p \triangleq \max_{p(x)} \max_{g_{XY}} I^p(\underline{X}; \underline{Y}), \quad (257)$$

where $g_{XY} = f_{XY} \circ e_{XY}$ is the joint reification mapping, and $I^p(\underline{X}; \underline{Y}) = H(X) + H(Y) - H_p(\underline{X}, \underline{Y})$ is the up pragmatic mutual information.

For the Gaussian channel, the input distribution that maximizes the classical mutual information is $X \sim \mathcal{N}(0, P)$. Under the pragmatic framework, we assume that the joint isoteleia mapping partitions the input-output space such that the average isoteleic volumes for X and Y are Ω_x and Ω_y , respectively. Without loss of generality, we consider a symmetric case where $\Omega_x = \Omega_y = \Omega$, with $\Omega \geq 1$ being the average isoteleic volume. This volume represents the average measure of the set of signals that lead to the same optimal terminal action.

Using the continuous version of the up pragmatic mutual information derived in (251), we have

$$\begin{aligned} I^p(\underline{X}; \underline{Y}) &= I(X; Y) + \log(\Omega_x \Omega_y) \\ &= \frac{1}{2} \log \left(1 + \frac{P}{\sigma^2} \right) + \log(\Omega^2) \\ &= \frac{1}{2} \log \left(\Omega^4 \left(1 + \frac{P}{\sigma^2} \right) \right). \end{aligned} \quad (258)$$

Thus, the pragmatic capacity of the Gaussian channel (per real dimension) is given by

$$C_p = \frac{1}{2} \log \left(\Omega^4 \left(1 + \frac{P}{\sigma^2} \right) \right) \text{ prabits per channel use.} \quad (259)$$

By the capacity-cost duality established in Section VI, the bilateral pragmatic cost of information is the inverse of the capacity function. From Eq. (259), solving for P as a function of R yields:

$$\text{CoI}_p(R) = \sigma^2 \left(\frac{2^{2R}}{\Omega^4} - 1 \right) \quad \text{for } R \geq 0. \quad (260)$$

This represents the minimum transmit power required to achieve a pragmatic rate R (in prabits per channel use) over the AWGN channel.

For asymmetric systems where only one side is coarsened, we have two additional cost measures, that is, perceptual/expressive pragmatic cost of information. For the perceptual path (observation Y carries information about pragmatic action $\underline{X}$), the prospective perceptual achievable rate from Section V-C is:

$$\begin{aligned} C_p^{\text{p-perc}}(P) &= \frac{1}{2} \log \left(1 + \frac{P}{\sigma^2} \right) + \log \Omega \\ &= \frac{1}{2} \log \left(\Omega^2 \left(1 + \frac{P}{\sigma^2} \right) \right). \end{aligned} \quad (261)$$

The inverse gives the perceptual pragmatic cost of information:

$$\text{CoI}_p^{\text{p-perc}}(R) = \sigma^2 \left(\frac{2^{2R}}{\Omega^2} - 1 \right). \quad (262)$$

This is the minimum sensing power required to extract R prabits of pragmatic information from the observation, when the observation itself is not coarsened.

For the expressive path (pragmatic label $\underline{Y}$ conveys information about source X), the prospective expressive achievable rate is:

$$C_p^{\text{p-expr}}(P) = \frac{1}{2} \log \left(\Omega^2 \left(1 + \frac{P}{\sigma^2} \right) \right), \quad (263)$$

which is identical to the perceptual case under symmetry. Thus,

$$\text{CoI}_p^{\text{p-expr}}(R) = \sigma^2 \left(\frac{2^{2R}}{\Omega^2} - 1 \right). \quad (264)$$

Remark 19 (Comparison of CoI Measures): The three CoI measures satisfy the hierarchy:

$$\text{CoI}_p(R) \leq \text{CoI}_p^{\text{p-perc}}(R) = \text{CoI}_p^{\text{p-expr}}(R), \quad (265)$$

since $\Omega^4 \geq \Omega^2$ for $\Omega \geq 1$. This confirms the general principle: bilateral pragmatic abstraction (coarsening both sides) is more efficient than single-sided abstraction, requiring less power to achieve the same pragmatic rate.

We now extend the result to the band-limited Gaussian channel. Consider a Gaussian channel with bandwidth B (Hz), two-sided power spectral density $N_0/2$, and signal power constraint P . The classical Shannon capacity for this channel is

$$C = B \log \left(1 + \frac{P}{N_0 B} \right) \quad \text{bits per second.} \quad (266)$$

By the Shannon-Nyquist sampling theorem, a band-limited signal of duration T can be represented by $2BT$ independent samples per real dimension. The total pragmatic capacity over the $2BT$ dimensions gives:

$$C_p = B \log \left(\Omega^4 \left(1 + \frac{P}{N_0 B} \right) \right) \quad \text{prabits per second.} \quad (267)$$

The corresponding bilateral CoI (power per second) is:

$$\text{CoI}_p(R) = N_0 B \left(\frac{2^{R/B}}{\Omega^4} - 1 \right) \quad (\text{watts}). \quad (268)$$

For the single-sided cases, the perceptual and expressive CoI become:

$$\text{CoI}_p^{\text{p-perc}}(R) = \text{CoI}_p^{\text{p-expr}}(R) = N_0 B \left(\frac{2^{R/B}}{\Omega^2} - 1 \right). \quad (269)$$

Example 8 (Numerical comparison of pragmatic cost measures): Let $\sigma^2 = 1$ for the discrete-time case, and for the band-limited case set $B = 1$ Hz, $N_0 = 1$. Take $\Omega = 2$. The table below compares the minimum required power (in units of σ^2 or watts) for various target rates R (in prabits per channel use or prabits/s) under bilateral and single-sided coarsening. Negative values indicate that zero power suffices (the rate lies below the zero-power capacity).

For the discrete bilateral case, the zero-power capacity is $C_p(0) = \frac{1}{2} \log(16) = 2$ prabits per channel use, so any $R \leq 2$ requires zero power. For the discrete perceptual path, $C_p^{\text{p-perc}}(0) = \frac{1}{2} \log(4) = 1$ prabit, thus $R = 1$ is free but $R = 2$ costs 3 units. In the band-limited case with $B = 1$, the bilateral zero-power capacity is $C_p(0) = \log(16) = 4$ prabits/s, hence $R = 3$ is still free; the perceptual zero-power capacity is $\log(4) = 2$ prabits/s, so $R = 2$ is free and $R = 3$ requires 1 W. These numerical results illustrate that bilateral abstraction (coarsening both source and observation) yields the greatest power savings for a given pragmatic rate.

Remark 20: The pragmatic capacity and CoI formulas reveal that the isoteleic volume Ω acts as a power amplification factor. In the band-limited case, Ω^4 effectively multiplies the signal-to-noise ratio, while Ω^2 multiplies it for single-sided cases. This provides a clear design guideline: to minimize transmit power for a given pragmatic rate, one should exploit pragmatic abstraction on both the source and the observation sides whenever possible.

C. Pragmatic Rate Distortion of Gaussian Source and Pragmatic Value of Information

As a dual to the channel capacity problem, we now investigate the pragmatic rate-distortion function for a Gaussian source and its corresponding pragmatic value of information. This function characterizes the minimum rate required to represent a continuous source such that the distortion, measured at the pragmatic level, does not exceed a given threshold.

Consider a Gaussian source $X \sim \mathcal{N}(0, P)$ with variance P . Let $\hat{X}$ be the reconstruction of X , and let $\underline{X}$ and $\hat{\underline{X}}$ be the corresponding pragmatic variables induced by the isoteleia mappings e_X and $e_{\hat{X}}$ (and their compositions with synonymous mappings). The distortion measure is defined at the pragmatic level as the mean squared error between the optimal actions corresponding to the source and its reconstruction:

$$d_p(\underline{x}, \hat{\underline{x}}) \triangleq (a^*(\underline{x}) - a^*(\hat{\underline{x}}))^2, \quad (270)$$

where $a^*(\underline{x})$ is the optimal action for the pragmatic class $\underline{x}$. For the Gaussian source, we adopt the standard mean squared error (MSE) distortion measure, which is

TABLE IX: CoI values for different scenarios and rates.

Scenario	CoI expression	$R = 1$	$R = 2$	$R = 3$
Discrete bilateral	$\sigma^2(2^{2R}/\Omega^4 - 1)$	-0.75 (0)	0 (0)	3
Discrete perceptual/expressive	$\sigma^2(2^{2R}/\Omega^2 - 1)$	0 (0)	3	15
Band-limited bilateral	$N_0 B(2^{R/B}/\Omega^4 - 1)$	-0.875 (0)	-0.75 (0)	-0.5 (0)
Band-limited perceptual/expressive	$N_0 B(2^{R/B}/\Omega^2 - 1)$	-0.5 (0)	0 (0)	1

widely used and has a well-known rate-distortion characterization. Specifically, we consider the distortion constraint $\mathbb{E}[(X - \hat{X})^2] \leq D$, which, under appropriate isoteleia mappings, translates to a pragmatic distortion constraint.

The pragmatic rate-distortion function is defined as

$$R_p(D) \triangleq \min_{\mathcal{S}_X, \mathcal{S}_{\hat{X}}} \min_{p(\hat{x}|x): \mathbb{E}[d_p(\underline{X}, \underline{\hat{X}})] \leq D} I_p(\underline{X}; \underline{\hat{X}}), \quad (271)$$

where $I_p(\underline{X}; \underline{\hat{X}}) = H_p(\underline{X}) + H_p(\underline{\hat{X}}) - H(X, \hat{X})$ is the down pragmatic mutual information.

For a Gaussian source with variance P , the classical rate-distortion function is $R(D) = \frac{1}{2} \log(P/D)$ for $0 \leq D \leq P$. Under the pragmatic framework, the isoteleia mapping groups signals that lead to the same optimal action, effectively reducing the “volume” of the source space that needs to be distinguished. Let Ω_x and $\Omega_{\hat{x}}$ be the average isoteleic volumes for the source and reconstruction, respectively. Assuming symmetry, we set $\Omega_x = \Omega_{\hat{x}} = \Omega$, where $\Omega \geq 1$ is the average isoteleic volume.

Using the continuous expression for the down pragmatic mutual information derived in Eq. (252), we have

$$\begin{aligned} I_p(\underline{X}; \underline{\hat{X}}) &= I(X; \hat{X}) - \log(\Omega_x \Omega_{\hat{x}}) \\ &\geq \frac{1}{2} \log\left(\frac{P}{D}\right) - \log(\Omega^2) \\ &= \frac{1}{2} \log\left(\frac{P}{\Omega^4 D}\right), \end{aligned} \quad (272)$$

where the inequality follows from the fact that for any test channel satisfying the MSE constraint $\mathbb{E}[(X - \hat{X})^2] \leq D$, the classical mutual information satisfies $I(X; \hat{X}) \geq \frac{1}{2} \log(P/D)$. The equality is achieved when the test channel is Gaussian and the isoteleia mapping is optimal (i.e., partitions the space proportionally to the probability mass of the pragmatic classes).

Thus, the pragmatic rate-distortion function for a Gaussian source is given by

$$R_p(D) = \begin{cases} \frac{1}{2} \log\left(\frac{P}{\Omega^4 D}\right), & 0 \leq D \leq \frac{P}{\Omega^4}, \\ 0, & D > \frac{P}{\Omega^4}. \end{cases} \quad (273)$$

By the rate-distortion-value duality established in Section VI, the bilateral pragmatic value of information is the Legendre-Fenchel dual of the rate-distortion function. For the Gaussian quadratic utility case, the value-rate function $\Phi_p(R)$ is obtained by solving the distortion-rate relationship.

From Eq. (273), the distortion as a function of rate R is:

$$D_p(R) = \frac{P}{\Omega^4} 2^{-2R}, \quad 0 \leq R < \infty. \quad (274)$$

For a quadratic utility $U(X, A) = -(X - A)^2$, the maximum expected utility at rate R is:

$$\Phi_p(R) = -D_p(R) = -\frac{P}{\Omega^4} 2^{-2R}. \quad (275)$$

The baseline utility (without information) is $U_0 = -P$. Therefore, the bilateral pragmatic value of information is:

$$\text{VoI}_p(R) = \Phi_p(R) - U_0 = P \left(1 - \frac{2^{-2R}}{\Omega^4}\right). \quad (276)$$

This is the maximum utility gain achievable when both the source and the reconstruction are coarsened pragmatically.

For asymmetric systems, we have two single-sided value measures corresponding to the perceptual and expressive rate-distortion functions from Section V-C. Due to symmetry, these two cases yield identical expressions. For either side, the rate-distortion function is

$$R_p^{\text{side}}(D) = \frac{1}{2} \log\left(\frac{P}{\Omega^2 D}\right), \quad 0 \leq D \leq \frac{P}{\Omega^2}, \quad (277)$$

TABLE X: VoI and distortion for Gaussian source with $P = 1$, $\Omega = 2$.

Scenario	Measure	$R = 0$	$R = 1$	$R = 2$
Bilateral	$\text{VoI}_p(R)$	0.9375	0.984375	0.996094
	$D_p(R) = P/\Omega^4 \cdot 2^{-2R}$	$1/16 = 0.0625$	$1/64 \approx 0.015625$	$1/256 \approx 0.003906$
Perceptual / Expressive	$\text{VoI}_p^{\text{perc/expr}}(R)$	0.75	0.9375	0.984375
	$D_p^{\text{perc/expr}}(R) = P/\Omega^2 \cdot 2^{-2R}$	$1/4 = 0.25$	$1/16 = 0.0625$	$1/64 \approx 0.015625$

with corresponding distortion-rate function $D_p^{\text{side}}(R) = \frac{P}{\Omega^2} 2^{-2R}$. The pragmatic value of information for both the perceptual and expressive paths is therefore

$$\text{VoI}_p^{\text{perc}}(R) = \text{VoI}_p^{\text{expr}}(R) = P \left(1 - \frac{2^{-2R}}{\Omega^2} \right). \quad (278)$$

We shall refer to either of these as the single-sided VoI.

Remark 21 (Comparison of VoI Measures): The three VoI measures satisfy the hierarchy:

$$\text{VoI}_p^{\text{perc}}(R) = \text{VoI}_p^{\text{expr}}(R) \leq \text{VoI}_p(R), \quad (279)$$

since $1/\Omega^2 \geq 1/\Omega^4$ for $\Omega \geq 1$. This is consistent with the general principle that bilateral pragmatic abstraction (coarsening both the source and the reconstruction) yields higher utility gain for a given rate than single-sided abstraction, because it discards more decision-irrelevant information, allowing the same rate to be used more efficiently for the task.

Example 9 (Numerical comparison of VoI and distortion for Gaussian source): Let $P = 1$ and $\Omega = 2$. The table below compares the pragmatic value of information $\text{VoI}(R)$ and the corresponding minimal distortion $D(R)$ for bilateral, perceptual, and expressive coarsening at selected rates R (in prabits). Recall that under symmetry, perceptual and expressive values are identical.

At $R = 1$, bilateral coarsening yields a VoI of 0.984 utils and distortion 0.0156, while single-sided coarsening gives 0.9375 utils and distortion 0.0625 — a four-fold reduction in distortion for the same rate. Equivalently, to achieve the single-sided distortion of 0.0625 with bilateral coarsening, only $R = 0$ is needed (since $D_p(0) = 0.0625$), saving 1 prabit. As $R \rightarrow \infty$, both VoI approach 1 (perfect action reconstruction) and distortion tends to 0. The classical case $\Omega = 1$ is recovered by setting $\Omega = 1$, which gives $\text{VoI}(R) = 1 - 2^{-2R}$, strictly smaller than any pragmatic VoI for finite R .

This example quantifies the benefit of coarsening both ends: bilateral pragmatic compression achieves lower distortion and higher utility gain than single-sided compression at the same rate, highlighting the efficiency of task-oriented abstraction.

Remark 22: The closed-form expressions for CoI and VoI derived above assume that the action space is continuous Euclidean, enabling the use of MSE as the distortion measure. For discrete action spaces (e.g., classification tasks), the corresponding CoI and VoI functions must be derived using discrete distortion measures such as Hamming distance or utility loss, which generally do not admit such simple closed forms. The general methodology, however, remains the same: CoI is the inverse of the achievable rate functions, and VoI is the Legendre-Fenchel dual of the rate-distortion functions, both extending naturally to the single-sided cases via the perceptual and expressive mutual informations.

The pragmatic cost of information and pragmatic value of information for Gaussian channels and sources complete the bilateral and single-sided duality spectrum in the continuous domain, providing a comprehensive framework for task-oriented communication system design with explicit resource-utility trade-offs.

XI. Joint Optimization of Pragmatic Information System

In this section, we present the pragmatic source-channel coding theorem, which extends the classical separation principle to the pragmatic domain and establishes that reliable transmission is possible iff the pragmatic entropy (or pragmatic rate-distortion function) does not exceed the pragmatic channel capacity. This condition is equivalent to the existence of a scheme that achieves the full pragmatic VoI while keeping the pragmatic CoI within the resource budget, thereby bridging information-theoretic limits with decision-theoretic objectives. We then formulate a unified optimization

over communication rate, control law, and pragmatic abstraction in closed-loop systems, and derive optimality conditions via the pragmatic Lagrangian for co-designing communication and control under resource constraints.

A. Pragmatic Source-Channel Coding

We now consider the joint source-channel coding problem in the pragmatic domain, where the goal is to transmit a syntactic source over a noisy channel such that the receiver can recover the pragmatic intention rather than the exact source symbols. This problem generalizes Shannon's classical source-channel coding theorem by relaxing the requirement of exact symbol reconstruction and allowing errors that do not affect the optimal action.

Consider a discrete memoryless syntactic source W with alphabet $\mathcal{W}$ and probability mass function $P(W)$. Let $\tilde{W}$ and $\underline{W}$ be the associated semantic and pragmatic variables induced by the synonymous mapping $f : \tilde{\mathcal{W}} \rightarrow 2^{\mathcal{W}}$ and the isoteleia mapping $e : \underline{\mathcal{W}} \rightarrow 2^{\tilde{\mathcal{W}}}$, respectively, with the reification mapping $g = f \circ e : \underline{\mathcal{W}} \rightarrow 2^{\mathcal{W}}$. The source generates an i.i.d. sequence W^n according to $P(W^n) = \prod_{k=1}^n P(W_k)$.

Let the channel be a discrete memoryless channel with input alphabet $\mathcal{X}$, output alphabet $\mathcal{Y}$, and transition probability $P(Y|X)$. The pragmatic channel capacity C_p is defined as in Section V-A : $C_p \triangleq \max_{p(x)} \max_{g_{XY}} I^p(\underline{X}; \underline{Y})$, where $I^p(\underline{X}; \underline{Y})$ is the up pragmatic mutual information, and g_{XY} is the joint reification mapping.

Figure 16 illustrates the block diagram of pragmatic source-channel coding. At the transmitter side, the syntactic source W^n first passes through the reification mapping $g^n = f^n \circ e^n$, which collapses semantically equivalent syntactic symbols into pragmatic classes $\underline{W}^n$ representing the optimal terminal actions. These pragmatic classes are then encoded into channel codewords X^n via a pragmatic channel encoder, which may exploit

the structure of the pragmatic equivalence classes to achieve higher rates than classical coding. The codewords are transmitted over the noisy channel, producing the received sequence Y^n . At the receiver side, the pragmatic channel decoder recovers an estimate of the pragmatic index $\hat{\underline{W}}^n$ from the channel output, ignoring syntactic and semantic variations that do not affect the terminal action. The demapping operation g^{-n} then maps the estimated pragmatic index to a representative syntactic sequence $\hat{W}^n$, which is passed to the destination. The entire system is evaluated by the error probability $P_e^{(n)} = \Pr\{g^n(\psi_n(Y^n)) \neq g^n(W^n)\}$, i.e., the probability that the recovered pragmatic intention differs from the true one. This architecture embodies the separation principle: the pragmatic source code (reification mapping) and the pragmatic channel code can be designed independently while preserving asymptotic optimality.

A pragmatic source-channel code of block length n consists of:

1) An encoding function $\phi_n : \mathcal{W}^n \rightarrow \mathcal{X}^n$ that maps each source sequence to a channel input sequence, possibly exploiting the reification mapping g^n to first map the source to its pragmatic class;

2) A decoding function $\psi_n : \mathcal{Y}^n \rightarrow \mathcal{W}^n$ that maps each channel output to an estimated source sequence (or directly to an estimated pragmatic index), followed by demapping g^{-n} to recover the pragmatic intention.

The error probability is defined as the probability that the decoded pragmatic intention differs from the true pragmatic intention:

$$P_e^{(n)} \triangleq \Pr\{g^n(\psi_n(Y^n)) \neq g^n(W^n)\}. \quad (280)$$

We now state the fundamental theorem of pragmatic source-channel coding.

Theorem 46 (Pragmatic Source-Channel Coding Theorem): Let W be a discrete memoryless syntactic source with associated pragmatic variable $\underline{W}$ having pragmatic entropy $H_p(\underline{W})$, and let $p(y|x)$ be a discrete memoryless channel with pragmatic capacity C_p .

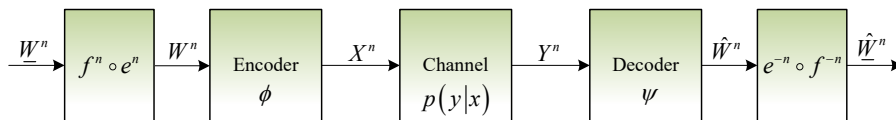


Fig. 16: Block diagram of pragmatic source channel coding.

1) **Lossless Case:** If $H_p(\underline{W}) < C_p$, then for any $\epsilon > 0$, there exists, for sufficiently large n , a sequence of pragmatic source-channel codes with error probability $P_e^{(n)} < \epsilon$. On the contrary, if $H_p(\underline{W}) > C_p$, then for any sequence of pragmatic source-channel codes, the error probability is bounded away from zero for sufficiently large n .

2) **Lossy Case:** For a given pragmatic distortion measure $d_p(\underline{w}, \hat{\underline{w}})$ and distortion constraint D , if $R_p(D) < C_p$, then there exists a sequence of pragmatic source-channel codes such that $\mathbb{E}[d_p(\underline{W}, \hat{\underline{W}})] \leq D$. On the contrary, if $R_p(D) > C_p$, then no such sequence exists.

Proof. We outline the proof following the standard separation principle, adapted to the pragmatic domain.

Achievability: Assume $H_p(\underline{W}) < C_p$. Choose $\epsilon > 0$ such that $H_p(\underline{W}) + \epsilon < C_p - \epsilon$. By the pragmatic lossless source coding theorem (Theorem 37), there exists a pragmatic source code that maps the source sequences to pragmatic indices at rate $R_s < H_p(\underline{W}) + \epsilon$ with error probability less than $\epsilon/2$. By the pragmatic channel coding theorem (Theorem 41), since $R_s < C_p$, there exists a pragmatic channel code that transmits these indices over the channel with error probability less than $\epsilon/2$. By the union bound, the overall error probability is less than ϵ . The decoder first recovers the pragmatic index from the channel output, then outputs a representative syntactic sequence from the corresponding reified typical set.

Converse: Suppose a sequence of codes exists with $P_e^{(n)} \rightarrow 0$. Following the standard Fano's inequality argument adapted to the pragmatic domain, we have

$$H_p(\underline{W}) \leq \frac{1}{n} I(\underline{W}^n; \underline{Y}^n) + \delta_n, \quad (281)$$

where $\delta_n \rightarrow 0$ as $P_e^{(n)} \rightarrow 0$. By the definition of pragmatic channel capacity and the data processing inequality,

$$\frac{1}{n} I(\underline{W}^n; \underline{Y}^n) \leq C_p. \quad (282)$$

Thus, $H_p(\underline{W}) \leq C_p$, which contradicts $H_p(\underline{W}) > C_p$. The proof for the lossy case follows analogously using the pragmatic rate-distortion function $R_p(D)$. $\square$

The pragmatic source-channel coding theorem has a direct interpretation in terms of the pragmatic value of information (VoI) and pragmatic cost of information (CoI) established in Section VI.

The condition $H_p(\underline{W}) < C_p$ (or $R_p(D) < C_p$) can be equivalently stated as the existence of a rate R such that:

$$\text{VoI}_p(R) = \text{VoI}_p^{\max} \quad \text{and} \quad \text{CoI}_p(R) \leq P_{\max}, \quad (283)$$

where $\text{VoI}_p^{\max}$ is the maximum achievable pragmatic value (the utility gain obtained when all pragmatic distinctions are preserved), and $P_{\max}$ is the available physical resource budget (e.g., maximum transmit power). This follows from the monotonicity of VoI and CoI:

- The pragmatic value of information $\text{VoI}_p(R)$ is non-decreasing in R and saturates at its maximum when $R \geq H_p(\underline{W})$ (for lossless transmission) or $R \geq R_p(D)$ (for lossy transmission).
- The pragmatic cost of information $\text{CoI}_p(R)$ is non-decreasing and convex in R , representing the minimum resource required to achieve rate R .

Therefore, the achievability condition $H_p(\underline{W}) < C_p$ guarantees that there exists a rate R such that:

$$\text{VoI}_p(R) = \text{VoI}_p^{\max} \quad \text{and} \quad \text{CoI}_p(R) \leq P_{\max}. \quad (284)$$

Conversely, if $H_p(\underline{W}) > C_p$, then even at the maximum achievable rate C_p , the system cannot transmit enough information to preserve all pragmatic distinctions, resulting in a loss of pragmatic value:

$$\text{VoI}_p(C_p) < \text{VoI}_p^{\max}. \quad (285)$$

This gap represents the *pragmatic value loss* due to the channel bottleneck.

Corollary 9 (Pragmatic Value-Cost Characterization):

A pragmatic communication system achieves the full pragmatic value of information if and only if the transmission rate R satisfies:

$$R \geq H_p(\underline{W}) \quad (\text{lossless}) \quad \text{or} \quad R \geq R_p(D) \quad (\text{lossy}), \quad (286)$$

and the available physical resources $P_{\max}$ satisfy:

$$\text{CoI}_p(R) \leq P_{\max}. \quad (287)$$

Remark 23: The pragmatic source-channel coding theorem provides a rigorous theoretical foundation for task-oriented communication. Unlike classical source-channel coding, which requires the syntactic entropy $H(W)$ to be less than the syntactic channel capacity C , the pragmatic version relaxes this condition to $H_p(\underline{W}) \leq C_p$. Since $H_p(\underline{W}) \leq H(W)$ and $C_p \geq C$, the pragmatic condition is strictly weaker than the classical one. This explains the performance gains observed in semantic and pragmatic communication systems: by tolerating syntactic and semantic errors that do not affect the optimal action, the system can operate reliably in regimes where classical communication would fail.

Example 10 (Autonomous Driving Revisited): Consider the two-vehicle autonomous driving scenario from Sections III and IV, where the syntactic source has entropy $H(W) \approx 1.9893$ bits per symbol, while the pragmatic entropy is $H_p(\underline{W}) \approx 0.9928$ prabits per symbol. Suppose the communication channel has syntactic capacity $C = 1.5$ bits per channel use. Classical source-channel coding would fail because $H(W) > C$. However, if the pragmatic capacity $C_p \geq 0.9928$ (which holds for any channel with $C_p \geq C$), the pragmatic source-channel coding theorem guarantees reliable transmission of the optimal action (Go/Stop) even though the exact traffic light symbols cannot be reliably transmitted. This demonstrates the power of pragmatic abstraction in overcoming physical communication bottlenecks.

A.1 Separation Principle and the Imperative of Joint Optimization

The pragmatic source-channel coding theorem (Theorem 46) is proved via a separation architecture, which builds on two independent results:

1) The pragmatic lossless source coding theorem (Section VII) guarantees that the pragmatic source can be compressed to rate $H_p(\underline{W})$ (or $R_p(D)$ in the lossy case) without losing any pragmatically relevant information.

2) The pragmatic channel coding theorem (Section VIII) guarantees that any rate up to C_p can be reliably transmitted over the noisy channel.

Concatenating the pragmatic source code and the pragmatic channel code yields an end-to-end system that achieves reliable pragmatic transmission *if and only if* the source rate does not exceed the channel capacity, i.e.,

$$H_p(\underline{W}) \leq C_p \quad (\text{lossless}) \quad \text{or} \quad R_p(D) \leq C_p \quad (\text{lossy}).$$

This constitutes the *pragmatic separation principle*—the asymptotic bedrock of the theory.

In the classical information-theoretic setting where the blocklength n tends to infinity and no constraints are imposed on encoding/decoding complexity, the separation principle is strictly optimal. It allows the system designer to optimize the pragmatic source code (e.g., feature extraction, action classification) and the pragmatic channel code (e.g., modulation, error correction) independently, while guaranteeing that the concatenated system achieves the fundamental limit. This modularity is a direct consequence of the asymptotic equipartition property and the joint typicality arguments used in the proofs, and it is fully consistent with the backward-compatibility requirement of Principle III in Section II.

The optimality of separation, however, is an asymptotic statement. In practical systems, where the blocklength n is finite, or when the encoder and decoder are subject to computational complexity, delay, or memory constraints, the separation architecture is *strictly suboptimal* with respect to the end-to-end task utility. In such regimes, the optimal code cannot be factored into independent source and channel components; instead, the encoder must jointly consider the source statistics, the channel characteristics, and the decision utility to minimize the overall distortion or maximize the expected reward. This fundamental gap between asymptotic theory and engineering reality is precisely the motivation for the joint optimization framework that follows.

This observation directly justifies the “joint optimization” framework developed in Section VI. While the separation principle identifies the ultimate performance bounds (capacities and rate-distortion functions), the *optimal allocation of finite resources*—such

as transmit power, latency, bandwidth, or model size—requires solving the coupled Lagrangian:

$$\mathcal{L}_{\text{global}}(R; \lambda) = \Phi_p(R) - \lambda \text{CoI}_p(R),$$

where the same rate R simultaneously determines the coarsening level of the isoteleia mapping (source-side abstraction) and the error-protection strength (channel-side coding). The cross-layer coupling therefore does not lie in the *code structure*, which can be separated asymptotically, but in the *rate-resource trade-off*, which must be jointly optimized to maximize the net benefit. Thus, the separation principle provides the theoretical bedrock, while the Lagrangian framework delivers the practical tool for engineering finite-dimensional, resource-constrained systems.

Remark 24: The pragmatic separation principle generalizes both the classical separation theorem (when the isoteleia and synonymous mappings are trivial) and the semantic separation theorem (when only the synonymous mapping is non-trivial). It unifies all previous results as special cases and offers new insights into the design of task-oriented communication systems, while the accompanying joint optimization framework addresses the real-world constraints that lie beyond the asymptotic ideal.

A.2 Gaussian Source over Gaussian Channel

To illustrate the joint optimization framework in a concrete setting, we now specialize to the canonical case of a Gaussian source transmitted over an additive white Gaussian noise (AWGN) channel. This example demonstrates how the pragmatic cost of information (CoI) and pragmatic value of information (VoI) jointly determine the optimal operating point of a task-oriented communication system.

Consider a Gaussian source $X \sim \mathcal{N}(0, P)$ with variance P , transmitted over an AWGN channel with noise variance σ^2 and power constraint $\mathbb{E}[|X|^2] \leq P_{\text{tx}}$. The channel output is $Y = X + Z$, where $Z \sim \mathcal{N}(0, \sigma^2)$. Let the isoteleic volume be $\Omega \geq 1$, representing the pragmatic abstraction that groups signals leading to the same optimal action.

From Sections B and C, we have the following closed-form expressions for the continuous Gaussian

case:

- **Pragmatic channel capacity** (per real dimension):

$$C_p(P_{\text{tx}}) = \frac{1}{2} \log \left(\Omega^4 \left(1 + \frac{P_{\text{tx}}}{\sigma^2} \right) \right) \quad \text{prabits per channel use.} \quad (288)$$

- **Bilateral pragmatic cost of information** (inverse of capacity):

$$\text{CoI}_p(R) = \sigma^2 \left(\frac{2^{2R}}{\Omega^4} - 1 \right) \quad \text{for } R \geq 0. \quad (289)$$

- **Pragmatic rate-distortion function:**

$$R_p(D) = \frac{1}{2} \log \left(\frac{P}{\Omega^4 D} \right), \quad 0 \leq D \leq \frac{P}{\Omega^4}. \quad (290)$$

- **Bilateral pragmatic value of information** (for quadratic utility $U(X, A) = -(X - A)^2$):

$$\text{VoI}_p(R) = P \left(1 - \frac{2^{-2R}}{\Omega^4} \right). \quad (291)$$

The pragmatic source-channel coding theorem (Theorem 46) states that reliable transmission of pragmatic information is possible if and only if

$$R_p(D) \leq C_p(P_{\text{tx}}). \quad (292)$$

Substituting the closed-form expressions from (288) and (290), the feasibility condition becomes

$$\frac{1}{2} \log \left(\frac{P}{\Omega^4 D} \right) \leq \frac{1}{2} \log \left(\Omega^4 \left(1 + \frac{P_{\text{tx}}}{\sigma^2} \right) \right). \quad (293)$$

Simplifying, we obtain the equivalent condition:

$$D \geq \frac{P}{\Omega^8 (1 + P_{\text{tx}}/\sigma^2)}. \quad (294)$$

Here the factor Ω^8 results from the product of the pragmatic gains at both ends: the source–reconstruction pair contributes Ω^4 through the rate–distortion function, and the input–output pair contributes another Ω^4 through the channel capacity, under the assumption of equal isoteleic volumes for both the source and the channel.

This inequality reveals the fundamental trade-off: for a given transmit power P_{tx} and pragmatic abstraction Ω , the minimum achievable pragmatic distortion is

$$D_{\min}(P_{\text{tx}}) = \frac{P}{\Omega^8 (1 + P_{\text{tx}}/\sigma^2)}. \quad (295)$$

Equivalently, for a target distortion D , the minimum required transmit power is

$$P_{\text{tx}}^{\min}(D) = \sigma^2 \left(\frac{P}{\Omega^8 D} - 1 \right). \quad (296)$$

The joint feasibility condition can also be expressed directly in terms of VoI and CoI. Recall that the full pragmatic value $\text{VoI}_p^{\max} = P$ is achieved when $R \geq R_p(D)$ (i.e., the rate is sufficient to meet the distortion requirement). The required rate to achieve a given VoI level is obtained by inverting (291):

$$R_{\text{VoI}}(V) = \frac{1}{2} \log \left(\frac{1}{\Omega^4 (1 - V/P)} \right), \quad 0 \leq V \leq P. \quad (297)$$

The system achieves the target VoI V if and only if the available power satisfies

$$\text{CoI}_p(R_{\text{VoI}}(V)) \leq P_{\text{tx}}. \quad (298)$$

Substituting the CoI expression (289), this becomes

$$\sigma^2 \left(\frac{2^{2R_{\text{VoI}}(V)}}{\Omega^4} - 1 \right) \leq P_{\text{tx}}. \quad (299)$$

Since $2^{2R_{\text{VoI}}(V)} = 1/(\Omega^4 (1 - V/P))$, we obtain

$$\sigma^2 \left(\frac{1}{\Omega^8 (1 - V/P)} - 1 \right) \leq P_{\text{tx}}. \quad (300)$$

Solving for V , the maximum achievable pragmatic value for a given transmit power is

$$\text{VoI}_p^{\max}(P_{\text{tx}}) = P \left(1 - \frac{1}{\Omega^8 (1 + P_{\text{tx}}/\sigma^2)} \right). \quad (301)$$

This expression directly quantifies the *value-cost trade-off*: the utility gain increases with transmit power but saturates at P as $P_{\text{tx}} \rightarrow \infty$. The pragmatic abstraction Ω amplifies the effective signal-to-noise ratio, reducing the power required to achieve a given utility level. For example, when $\Omega = 1$ (no pragmatic abstraction), (301) reduces to the classical value-cost relationship:

$$\text{VoI}^{\max}(P_{\text{tx}}) = P \left(1 - \frac{1}{1 + P_{\text{tx}}/\sigma^2} \right) = \frac{PP_{\text{tx}}}{P_{\text{tx}} + \sigma^2}, \quad (302)$$

which is the well-known SNR-dependent utility gain for a Gaussian channel with quadratic utility.

Example 11: Let $P = 1$, $\sigma^2 = 1$, and consider three values of the isoteleic volume: $\Omega = 1$ (classical), $\Omega = 1.5$, and $\Omega = 2$. For a target VoI of $V = 0.9$ (i.e., 90% of the maximum utility), the required transmit powers are:

$$\begin{aligned} \Omega = 1 : P_{\text{tx}} &= \sigma^2 \left(\frac{1}{\Omega^8 (1 - V/P)} - 1 \right) \\ &= \frac{1}{0.1} - 1 = 9 \text{ (units)}, \end{aligned} \quad (303)$$

$$\begin{aligned} \Omega = 1.5 : P_{\text{tx}} &= \frac{1}{1.5^8 \cdot 0.1} - 1 \\ &\approx \frac{1}{25.63 \cdot 0.1} - 1 \\ &\approx -0.61 \text{ (zero power suffices)}, \end{aligned} \quad (304)$$

$$\begin{aligned} \Omega = 2 : P_{\text{tx}} &= \frac{1}{2^8 \cdot 0.1} - 1 = \frac{1}{25.6} - 1 \\ &\approx -0.96 \text{ (zero power suffices)}. \end{aligned} \quad (305)$$

With $\Omega = 1.5$, the pragmatic abstraction alone ($P_{\text{tx}} = 0$) already achieves $\text{VoI} = P(1 - 1/\Omega^8) = 1 - (2/3)^8 \approx 0.961$, exceeding the target 0.9. This demonstrates that pragmatic abstraction can dramatically reduce or eliminate the need for transmit power in task-oriented communication.

For asymmetric systems where only one side is coarsened, the perceptual and expressive CoI/VoI functions from Section VI-D yield analogous joint feasibility conditions. For the perceptual path, the achievable VoI under power constraint is

$$\text{VoI}_p^{\text{perc}, \max}(P_{\text{tx}}) = P \left(1 - \frac{1}{\Omega^4 (1 + P_{\text{tx}}/\sigma^2)} \right), \quad (306)$$

which is strictly less than the bilateral case (301) for any finite $\Omega > 1$ and P_{tx} , since the effective SNR amplification is Ω^4 rather than Ω^8 . This quantifies the performance loss when only the source or only the reconstruction is coarsened, confirming the benefits of coarsening both sides of the communication link.

Remark 25: The Gaussian example illustrates the full power of the pragmatic information-theoretic framework: it provides a unified, analytically tractable model for jointly optimizing communication resources (power), decision performance (distortion/utility), and pragmatic abstraction (isoteleic volume). The closed-form expressions reveal the precise quantitative relationships among these quantities, enabling system designers to make principled trade-offs in task-oriented communication system design.

B. Communication-Control-Decision Joint Optimization

We now extend the Lagrangian dual framework from the static source-channel coding setting to the general closed-loop system where communication, control, and decision-making are jointly optimized. In many practical systems—such as autonomous vehicles, robotic control, and industrial automation—the communication rate, control policy, and pragmatic abstraction must be co-designed to maximize the overall task performance under resource constraints. We present two formal theorems that characterize the optimal trade-offs in both static (single-step) and dynamic (multi-step) settings, with explicit optimality conditions that bridge information-theoretic limits and decision-theoretic objectives.

B.1 Static Joint Optimization: Lagrangian Approach and Single-Sided Extensions

We first consider a single-step decision problem where the system observes a state $S \in \mathcal{S}$, takes an action $A \in \mathcal{A}$, and receives a utility $U(S, A)$. The observation is obtained through a communication channel that conveys a pragmatic message $\underline{Y}$ about the state, with a rate R that determines the quality of the pragmatic information. The receiver chooses a policy $\pi : \underline{Y} \rightarrow \mathcal{A}$ that maps the pragmatic observation to an action. The joint optimization problem is to maximize the expected utility subject to a communication cost constraint.

Theorem 47 (Static Joint Optimization): For a single-step decision problem with state S , action A , and utility $U(S, A)$, let the pragmatic observation $\underline{Y}$ be obtained at rate R satisfying $I_p(\underline{Y}; S) \leq R$, and let the communication cost be given by the pragmatic cost of information

$\text{CoI}_p(R)$. The Lagrangian relaxation of the constrained utility maximization problem

$$\begin{aligned} & \max_{\pi, P(\underline{Y}|S)} \mathbb{E}_{S, \underline{Y}}[U(S, \pi(\underline{Y}))] \quad \text{subject to} \\ & \text{CoI}_p(R) \leq P_{\max} \end{aligned} \quad (307)$$

yields the *global pragmatic Lagrangian*

$$\mathcal{L}_{\text{global}}(\pi, R; \lambda) = \mathbb{E}_{S, \underline{Y}}[U(S, \pi(\underline{Y}))] - \lambda \text{CoI}_p(R), \quad (308)$$

where $\lambda > 0$ is the Lagrange multiplier (shadow price of communication resources). The optimal policy π^* and optimal rate R^* satisfy the following necessary and sufficient conditions:

$$\pi^*(\underline{Y}) = \arg \max_a \mathbb{E}_{S|Y}[U(S, a)], \quad (309)$$

$$\lambda \text{CoI}'_p(R^*) = \frac{\partial}{\partial R} \mathbb{E}_{S, \underline{Y}}[U(S, \pi^*(\underline{Y}))], \quad (310)$$

where the derivative in (310) is understood in the sense of marginal utility gain with respect to the rate. Moreover, the value of the Lagrangian at the optimum equals the maximum net benefit, that is behavioral capacity:

$$\begin{aligned} \max_{\pi, R} \mathcal{L}_{\text{global}}(\pi, R; \lambda) &= \max_{P(\underline{Y}|S)} \{ \mathbb{E}_{S, \underline{Y}}[U(S, a^*(\underline{Y}))] - \\ & \quad \lambda \text{CoI}_p(R) \}, \end{aligned} \quad (311)$$

with $a^*(\underline{Y}) = \arg \max_a \mathbb{E}_{S|Y}[U(S, a)]$.

Sketch of Proof. The Lagrangian is concave in π and R under mild regularity conditions (e.g., U bounded, CoI_p convex). The first-order conditions follow from the concavity and the fact that the optimal policy is the Bayes decision rule. The marginal utility gain in (310) is obtained by differentiating the optimal value with respect to R , which equals the rate of improvement of the expected utility when the constraint on mutual information is relaxed. The global optimum is achieved when the marginal benefit of increasing the rate equals its marginal cost, scaled by the shadow price. $\square$

Corollary 10 (Single-Sided Lagrangian Extensions): For asymmetric systems where only one side of the communication link is coarsened, the corresponding single-sided Lagrangians are obtained by substituting the appropriate CoI functions:

- **Perceptual pragmatic Lagrangian:**
 $\mathcal{L}_{\text{perc}}(\pi, R; \lambda) = \mathbb{E}_{S, \underline{Y}}[U(S, \pi(\underline{Y}))] - \lambda \text{CoI}_p^{\text{perc}}(R)$, where $\text{CoI}_p^{\text{perc}}(R)$ is the perceptual cost (sensing power for rate R).
- **Expressive pragmatic Lagrangian:**
 $\mathcal{L}_{\text{expr}}(\pi, R; \lambda) = \mathbb{E}_{S, \underline{Y}}[U(S, \pi(\underline{Y}))] - \lambda \text{CoI}_p^{\text{expr}}(R)$, where $\text{CoI}_p^{\text{expr}}(R)$ is the expressive cost (transmit power for rate R).

The optimality conditions for each case are identical to (309) and (310) with the corresponding CoI function.

Example 12 (Linear Quadratic Gaussian (LQG) with Perceptual Cost): Consider a scalar state $S \sim \mathcal{N}(0, \sigma_s^2)$, quadratic utility $U(S, A) = -(S - A)^2$, and perceptual cost $\text{CoI}_p^{\text{perc}}(R) = \sigma^2(2^{2R}/\Omega^2 - 1)$ with isoteleic volume Ω . The Lagrangian $\mathcal{L}_{\text{perc}}(P_{\text{tx}}; \lambda) = -\sigma_s^2/(1 + \Omega^2 P_{\text{tx}}/\sigma^2) - \lambda P_{\text{tx}}$ yields the optimal power

$$P_{\text{tx}}^* = \max \left\{ \sigma^2 \left(\frac{\sqrt{\sigma_s^2}}{\Omega \sqrt{\lambda}} - 1 \right), 0 \right\}. \quad (312)$$

This closed-form solution illustrates the trade-off between control performance and communication cost: the optimal power decreases as the shadow price λ increases, and the pragmatic abstraction Ω amplifies the effective signal-to-noise ratio.

Example 13 (Gaussian Source over Gaussian Channel: Closed-Form Bilateral Pragmatic Lagrangian): We now instantiate Theorem 47 for the canonical case of a Gaussian source transmitted over an AWGN channel, where both the source and the observation are coarsened pragmatically (bilateral abstraction). This provides a closed-form verification of the static joint optimization conditions.

Let the state (source) be $X \sim \mathcal{N}(0, P)$ and the pragmatic observation be $\underline{Y}$ obtained through an AWGN channel with noise variance σ^2 and transmit power P_{tx} . The isoteleic volume $\Omega \geq 1$ captures the pragmatic equivalence. From Section XI-A.2, the bilateral pragmatic value and cost functions are (cf. (291) and (289)):

$$\Phi_p(R) = P \left(1 - \frac{2^{-2R}}{\Omega^4} \right), \quad (313)$$

$$\text{CoI}_p(R) = \sigma^2 \left(\frac{2^{2R}}{\Omega^4} - 1 \right), \quad (314)$$

where R is the pragmatic rate (prabits per channel use). Substituting into the global Lagrangian of Theorem 47 (after maximizing over the policy π , which yields $\Phi_p(R)$), we obtain the explicit Lagrangian:

$$\begin{aligned} \mathcal{L}(R) &= \Phi_p(R) - \lambda \text{CoI}_p(R) \\ &= P \left(1 - \frac{2^{-2R}}{\Omega^4} \right) - \lambda \sigma^2 \left(\frac{2^{2R}}{\Omega^4} - 1 \right). \end{aligned} \quad (315)$$

Equivalently, symmetrically:

$$\mathcal{L}(R) = (P + \lambda \sigma^2) - \frac{1}{\Omega^4} (P \cdot 2^{-2R} + \lambda \sigma^2 \cdot 2^{2R}). \quad (316)$$

Figure 17 illustrates the global Lagrangian structure. Subfigure (a) shows the pragmatic value, the weighted cost, and the resulting net benefit, with the optimal rate R^* marked where marginal value equals marginal cost. Subfigure (b) demonstrates that increasing Ω uniformly shifts the net benefit curve upward while leaving R^* unchanged, confirming that pragmatic abstraction amplifies the effective signal-to-noise ratio without altering the optimal communication rate.

Applying the first-order optimality condition from Theorem 47 ($\mathcal{L}'(R^*) = 0$):

$$\frac{d\mathcal{L}}{dR} = \frac{2 \ln 2}{\Omega^4} (P \cdot 2^{-2R^*} - \lambda \sigma^2 \cdot 2^{2R^*}) = 0, \quad (317)$$

we obtain the optimal rate:

$$R^* = \frac{1}{4} \log_2 \left(\frac{P}{\lambda \sigma^2} \right)^+, \quad (318)$$

where $(x)^+ = \max\{x, 0\}$ ensures non-negativity. The maximized Lagrangian (net benefit) is then:

$$\mathcal{L}^* = \mathcal{L}(R^*) = P + \lambda \sigma^2 - \frac{2\sqrt{P\lambda\sigma^2}}{\Omega^4}, \quad (319)$$

provided $P > \lambda \sigma^2$; otherwise $R^* = 0$ and $\mathcal{L}^* = 0$ (the system opts not to communicate).

This maximized net benefit is precisely the behavioral capacity $\mathcal{E}_p(\lambda)$ defined in Definition 24. Substituting $\lambda = (P/\sigma^2)2^{-4R^*}$ into (319) yields the behavioral capacity as a function of the optimal rate:

$$\mathcal{E}_p(R^*) = P \left(1 + 2^{-4R^*} - \frac{2 \cdot 2^{-2R^*}}{\Omega^4} \right), \quad R^* \geq 0. \quad (320)$$

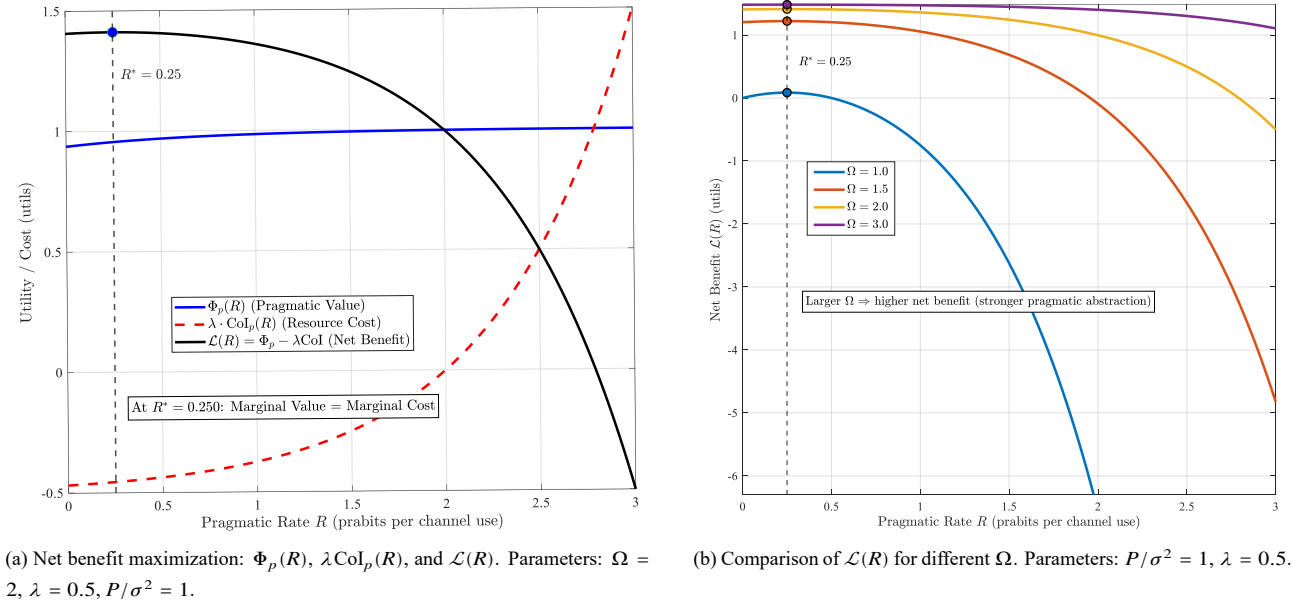


Fig. 17: Global pragmatic Lagrangian and the effect of pragmatic abstraction.

Equations (319) and (320) explicitly characterize the behavioral capacity in terms of the shadow price and the optimal communication rate, respectively.

Figure 18 further characterizes the behavioral capacity. Subfigure (a) shows how the capacity increases with the invested communication rate, with diminishing returns and saturation at the perfect-information utility P . A larger Ω shifts the curve upward, indicating that

the same rate yields higher net utility under stronger pragmatic abstraction. Subfigure (b) reveals the decay of capacity as resources become more expensive; the zero-capacity threshold marks the point where communication ceases. Larger Ω flattens the decline, demonstrating that abstraction provides robustness against rising resource costs.

This closed-form solution explicitly confirms the

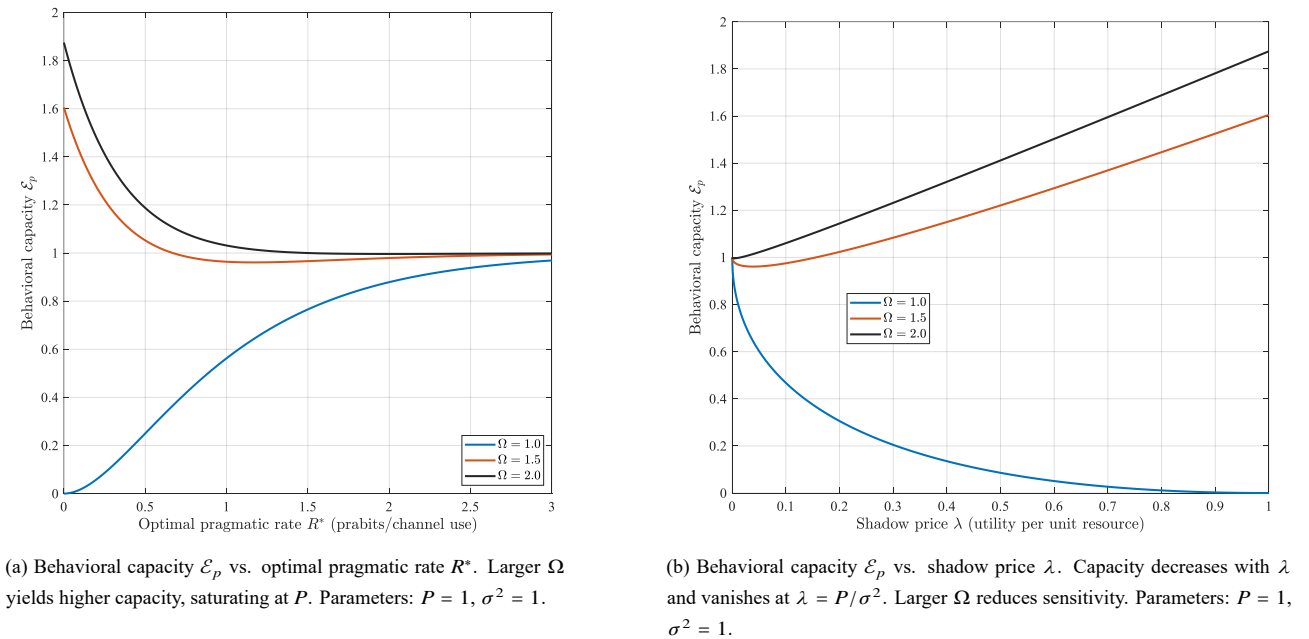


Fig. 18: Behavioral capacity as a function of optimal rate and resource shadow price.

two optimality conditions of Theorem 47:

1) The optimal policy π^* is the Bayes decision rule (implicit in the definition of $\Phi_p(R)$);

2) The marginal value equals the marginal cost: $\Phi'_p(R^*) = \lambda \text{CoI}'_p(R^*)$, which is precisely the balance expressed in (317).

Moreover, the expression (319) reveals the four-way trade-off among source power P , channel noise σ^2 , resource shadow price λ , and pragmatic abstraction Ω . When $\Omega = 1$, it recovers the classical Gaussian benchmark; when $\Omega > 1$, the effective SNR is amplified by Ω^4 on both sides, drastically reducing the power needed to achieve a given net benefit.

B.2 Dynamic Joint Optimization: Sequential Decision Making

In dynamic systems, the state evolves over time according to a controlled Markov process, and the decision-maker must balance immediate rewards against future consequences. The system is modeled as a Partially Observable Markov Decision Process (POMDP) with communication constraints. At each time step t , the system state S_t evolves as $S_{t+1} \sim P(\cdot|S_t, A_t)$, and the decision-maker receives a pragmatic observation $\underline{Y}_t$ about S_t at a rate R_t that incurs a cost $\text{CoI}_p(R_t)$. The goal is to find a policy $\pi = \{\pi_t\}_{t \geq 0}$ mapping the history of pragmatic observations to actions, maximizing the expected discounted sum of utilities subject to a total communication resource budget.

Theorem 48 (Dynamic Joint Optimization for POMDP): For an infinite-horizon discounted POMDP with state space $\mathcal{S}$, action space $\mathcal{A}$, observation space $\mathcal{Y}$, transition kernel $P(s'|s, a)$, observation kernel $P(y|s', R)$ (where the communication rate R affects the observation quality), utility $U(s, a)$, discount factor $\gamma \in (0, 1)$, and pragmatic communication cost $\text{CoI}_p(R)$ per step, the Lagrangian relaxation of the constrained expected discounted utility maximization

$$\begin{aligned} \max_{\pi, \{R_t\}} \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t U(S_t, A_t) \right] \quad \text{subject to} \\ \sum_{t=0}^{\infty} \gamma^t \text{CoI}_p(R_t) \leq C_{\text{total}} \end{aligned} \quad (321)$$

yields the *dynamic pragmatic Lagrangian*

$$\mathcal{L}_{\text{dyn}}(\pi, \{R_t\}; \lambda) = \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t (U(S_t, A_t) - \lambda \text{CoI}_p(R_t)) \right], \quad (322)$$

where $\lambda > 0$ is the Lagrange multiplier for the total resource constraint.

Let $b \in \Delta(\mathcal{S})$ denote the belief state (the posterior distribution over the state given the history of actions and observations). The optimal policy and rate schedule satisfy the Bellman equation [5] in belief-state form:

$$\begin{aligned} V(b) = \max_{a, R \geq 0} \left\{ \sum_s b(s) U(s, a) - \lambda \text{CoI}_p(R) + \right. \\ \left. \gamma \sum_{s, s', y} b(s) P(s'|s, a) P(y|s', R) V(b'_y) \right\}, \end{aligned} \quad (323)$$

where b'_y is the updated belief state after observing y (and taking action a with rate R):

$$b'_y(s') = \frac{P(y|s', R) \sum_s P(s'|s, a) b(s)}{\sum_{\tilde{s}} P(y|\tilde{s}, R) \sum_s P(\tilde{s}|s, a) b(s)}. \quad (324)$$

The optimal action a^* and optimal rate R^* at belief b are characterized by the following first-order conditions:

$$\begin{aligned} a^* = \arg \max_a \left\{ \sum_s b(s) U(s, a) + \right. \\ \left. \gamma \sum_s b(s) \sum_{s'} P(s'|s, a) \cdot \right. \\ \left. \sum_y P(y|s', R^*) V(b'_y) \right\}, \end{aligned} \quad (325)$$

$$\begin{aligned} \lambda \text{CoI}'_p(R^*) = \frac{\partial}{\partial R} \left(\gamma \sum_s b(s) \sum_{s'} P(s'|s, a^*) \cdot \right. \\ \left. \sum_y P(y|s', R) V(b'_y) \right) \Big|_{R=R^*}, \end{aligned} \quad (326)$$

where b'_y is given by (324) and depends on a^* and R . The derivative in (326) accounts for the improvement in future value due to better state estimation (sharper belief) resulting from a higher rate R .

Sketch of Proof. The proof follows from the principle of optimality applied to the belief-state MDP, which is equivalent to the original POMDP. Before applying the Bellman equation, we first establish its well-definedness. Define the Bellman operator $\mathcal{T}$ on the space of bounded continuous functions on $\Delta(\mathcal{S})$ by

$$(\mathcal{T}V)(b) \triangleq \max_{a, R \geq 0} \left\{ \sum_s b(s) U(s, a) - \lambda \text{CoI}_p(R) + \gamma \sum_{s, s', y} b(s) P(s'|s, a) P(y|s', R) V(b'_y) \right\}.$$

For any two bounded functions V_1, V_2 and any belief b , the transition kernels sum to 1, which gives

$$\begin{aligned} & |(\mathcal{T}V_1)(b) - (\mathcal{T}V_2)(b)| \\ & \leq \gamma \sup_{a, R} \sum_{s, s', y} b(s) P(s'|s, a) P(y|s', R) |V_1(b'_y) - V_2(b'_y)| \\ & \leq \gamma \|V_1 - V_2\|_\infty. \end{aligned}$$

Taking the supremum over b yields $\|\mathcal{T}V_1 - \mathcal{T}V_2\|_\infty \leq \gamma \|V_1 - V_2\|_\infty$. Since $\gamma < 1$, $\mathcal{T}$ is a contraction. By the Banach fixed-point theorem, there exists a unique fixed point V^* in the space of bounded continuous functions, ensuring that the belief-state Bellman equation (323) is well-defined. The dynamic isoteleia mapping e_{V^*} defined through V^* is then uniquely determined and recovers the static mapping e when $\gamma = 0$.

With this established, the augmented reward process $U(S_t, A_t) - \lambda \text{CoI}_p(R_t)$ leads to the belief-state Bellman equation (323), where the expectation over the next belief accounts for both state transition and observation likelihood. The optimality conditions (325) and (326) are obtained by differentiating the right-hand side of (323) with respect to a and R , respectively, under suitable regularity conditions (e.g., concavity of the value function and convexity of CoI_p). The coupling between action and rate appears through the belief update, as both influence the future belief distribution. $\square$

To avoid a conceptual gap between the static hierarchical framework (Section II–IX) and the dynamic POMDP formulation, we emphasize that the pragmatic variable $\underline{S}$ appearing in the Bellman equation (324) is induced by the *dynamic* isoteleia mapping e_V defined in Definition 3. The resulting pragmatic equivalence classes are therefore functions of the future value function V . In the steady-state case, e_V collapses to the static mapping e used in the coding theorems, ensuring that the hierarchical entropy inequalities remain valid as limiting bounds.

Remark 26 (Interpretation and Design Insights): The

dynamic Lagrangian framework provides several key insights for system design:

1) **Adaptive Resource Allocation:** The optimal rate R^* varies with the belief state, allocating more resources when the belief is diffuse (high uncertainty) or when future rewards are particularly sensitive to estimation accuracy. This yields a *value-of-information-guided* scheduling policy.

2) **Belief-Dependent Optimization:** Unlike the original erroneous formulation, the correct belief-state Bellman equation explicitly incorporates the observation model and the impact of R on the quality of future beliefs. The action and rate decisions are coupled through the belief update, reflecting the inherent trade-off between control and sensing/communication.

3) **Connection to CoI and VoI:** The term $\lambda \text{CoI}_p(R)$ represents the instantaneous communication cost, while the future value $V(b'_y)$ encodes the long-term utility gain from improved state information. The optimal policy balances immediate and future value, mirroring the economic principle of marginal benefit equals marginal cost.

Example 14 (Linear Quadratic Gaussian (LQG) with Linear Communication Cost): Consider a scalar linear system $S_{t+1} = aS_t + bA_t + W_t$ with $W_t \sim \mathcal{N}(0, \sigma_w^2)$, quadratic utility $U(S_t, A_t) = -qS_t^2 - rA_t^2$, and communication cost $\text{CoI}_p(R_t) = cR_t$ (linear in rate). Assume Gaussian observations with precision proportional to R_t , i.e., $Y_t = S_t + Z_t$, $Z_t \sim \mathcal{N}(0, \sigma^2/R_t)$. The belief state is Gaussian, characterized by its mean and variance. The optimal control law remains linear in the estimated state (separation principle holds for LQG with quadratic cost and Gaussian noise). The optimal rate R^* is constant over time when the system is stationary, and can be found by solving a Riccati equation augmented with the rate penalty. The steady-state rate decreases with the shadow price λ and depends on the system dynamics through the controllability and observability parameters. This example demonstrates how the dynamic Lagrangian framework provides a tractable method for co-designing communication and control in practical systems.

The dynamic joint optimization framework com-

pletes the pragmatic information-theoretic spectrum, providing a unified mathematical tool for designing task-oriented communication and control systems that optimally trade off information value, physical resource consumption, and long-term performance.

XII. Conclusions

A. Summary of Contributions

We have established a mathematical theory of pragmatic information that connects communication, control, and decision-making within a coherent framework. The theory is built upon the isoteleia mapping $e : \mathcal{W} \rightarrow 2^{\mathcal{V}}$, which formalizes equifinality—the principle that distinct semantic paths converging to the same optimal action are pragmatically equivalent. This induces a three-tier hierarchy of syntactic, semantic, and pragmatic information, with each successive abstraction discarding task-irrelevant distinctions.

The main contributions are threefold. First, we have developed a complete set of information-theoretic measures at the pragmatic level—entropy $H_p(\underline{W})$, up/down mutual information $I^p(\underline{X}; \underline{Y})$ and $I_p(\underline{X}; \underline{Y})$, channel capacity C_p , and rate-distortion function $R_p(D)$ —and established the fundamental hierarchies $H_p \leq H_s \leq H$, $C_p \geq C_s \geq C$, and $R_p \leq R_s \leq R$. Second, we have introduced pragmatic value of information (VoI) and pragmatic cost of information (CoI) as decision-theoretic and economic duals to rate-distortion and capacity, respectively, and formulated a Lagrangian dual framework for cross-layer optimization. The pragmatic efficiency functional $\mathcal{E}_p(\lambda) = \sup_R [\Phi_p(R) - \lambda \text{CoI}_p(R)]$ defines a behavioral capacity for resource-constrained intelligent systems under the stated assumptions. Third, we have proved three coding theorems—lossless source coding, channel coding, and rate-distortion coding—that generalize Shannon’s classical results and their semantic counterparts. We have further extended the framework to continuous messages, deriving closed-form Gaussian expressions, and formulated joint optimization of communication, control, and decision-making in both static and dynamic settings.

B. Engineering Implications and Deployment Pathways

The pragmatic information theory developed in this paper may have implications for a range of engineering disciplines and applications. By shifting the focus from symbol fidelity to the effectiveness of information in guiding actions, the framework offers a mathematical language for task-oriented communication, networked control, autonomous systems, and embodied AI, and suggests conceptual directions for human-machine interaction and social-cognitive systems. The Value of Information (VoI) and Cost of Information (CoI) offer a unified economic and decision-theoretic basis for resource allocation, enabling system designers to balance utility gains against physical resource consumption. The Lagrangian dual framework provides a principled approach for cross-layer optimization, bridging the gap between information-theoretic limits and engineering practice. The framework also offers new perspectives on emerging challenges such as semantic communication, edge AI, intelligent transportation, and human learning, where information must be evaluated by its effectiveness in achieving goals rather than its syntactic accuracy.

Looking toward engineering practice, a critical path for deploying the proposed theory lies in the realization of the isoteleia mapping—the core mathematical object that groups semantically distinct signals into pragmatically equivalent action classes. In real-world scenarios where the utility function U is not explicitly specified, the isoteleia mapping can be acquired implicitly through large-scale data-driven learning. Specifically, deep neural networks can serve as universal approximators of the optimal action function, effectively inducing the equivalence relation from data without requiring an analytic utility model. This paradigm aligns naturally with inverse reinforcement learning and reinforcement learning from human feedback (RLHF), where the network learns to map high-dimensional observations (semantic inputs) to optimal action classes (pragmatic outputs) by minimizing task-oriented loss functions.

Furthermore, in practical deployment, the isoteleia

mapping should not be a fixed, static structure. It must adapt dynamically to changing environmental contexts and evolving task objectives. For instance, in autonomous driving, the mapping from a visual scene to the action “brake” may depend critically on the current speed, road condition, and surrounding traffic. This context-dependent behavior suggests an extension of the isoteleia mapping to a conditional form depending on the contextual state or scene encoding. In terms of neural implementation, this can be achieved through context-gating mechanisms or conditional neural networks that modulate the mapping based on the current environment.

Equally important is the principle of determinism in decision-making. For safety-critical systems, the mapping from observation to action should be deterministic: a given input must produce a unique, consistent output. This ensures reproducibility, facilitates fault diagnosis, and aligns with the $\arg \max$ operation in the decision axiom. In practice, this can be realized by deploying deterministic policies at inference time, or by using techniques such as Gumbel-Softmax with temperature annealing to ensure that the learned mapping converges to a hard, deterministic decision boundary.

To concretize these ideas into a deployable pipeline, we envision an end-to-end pragmatic information system consisting of three functional stages: (i) an offline training phase, where a deep neural network is trained on large-scale task-specific data to approximate the conditional isoteleia mapping; (ii) a context-adaptive inference phase, where the network dynamically adjusts its decision boundaries based on real-time contextual inputs; and (iii) an online adaptation mechanism that continuously refines the mapping through reinforcement learning or online fine-tuning, enabling the system to respond to non-stationary environments. This pipeline directly addresses the practical challenges of implementing the theory in real-world applications, providing a bridge between the mathematical abstraction of the isoteleia mapping and the constraints of physical hardware and real-time computation.

C. Limitations and Future Directions

Several limitations of the current framework should be acknowledged. The theory assumes stationarity and ergodicity, which may not hold in social, economic, or biological systems where statistical properties evolve over time. The coding theorems are asymptotic in nature; finite-blocklength analyses and the corresponding dispersion bounds remain to be developed. The framework primarily addresses single-agent or centralized settings; multi-agent systems with conflicting or adversarial utilities require game-theoretic extensions. The dynamic optimization relies on belief-state Bellman equations that are computationally intractable in high dimensions, necessitating approximate solution methods. Finally, while deep neural networks offer a practical path for learning isoteleia mappings, they introduce approximation errors and generalization gaps not captured by the current theoretical guarantees.

Looking forward, several directions merit further investigation. The extension to non-stationary and non-ergodic processes calls for time-varying entropies and online adaptive coding strategies. Finite-blocklength analysis would provide dispersion bounds and practical guidelines for system design. Multi-agent and networked pragmatic information systems—including collaborative autonomous fleets, adversarial communication, and competitive economic interactions—require the development of game-theoretic equilibrium concepts and incentive-compatible protocols. The integration of deep learning with pragmatic coding demands provably efficient algorithms with finite-sample guarantees and convergence rates. Approximate dynamic programming techniques, such as deep reinforcement learning and function approximation, can be brought to bear on the high-dimensional belief-state optimization problem. Beyond these technical directions, the framework opens avenues for cross-disciplinary applications in cognitive science, linguistics, and economics, where information exchange is inherently goal-directed and context-dependent.

The mathematical theory of pragmatic information provides a unified language for goal-directed information processing under resource constraints. While its

full consequences will be revealed through future theoretical and empirical advances, the framework lays a solid foundation for this emerging direction in information theory.

References

- [1] Shannon C E. A mathematical theory of communication[J]. Bell System Technical Journal, 1948, 27(4): 623-656.
- [2] Shannon C E, Weaver W. The mathematical theory of communication[M]. Champaign: The University of Illinois Press, 1949.
- [3] Wiener N. Cybernetics or control and communication in the animal and the machine[M]. Cambridge: The Technology Press and New York: John Wiley & Sons, 1948.
- [4] Tsien H S. Engineering cybernetics[M]. New York: McGraw-Hill, 1954.
- [5] Bellman R. The theory of dynamic programming[J]. Bulletin of the American Mathematical Society, 1954, 60(6): 503-515.
- [6] Kalman R E. A new approach to linear filtering and prediction problems[J]. Journal of Basic Engineering, 1960, 82(1): 35-45.
- [7] Stratonovich R L. On value of information[J]. Izvestiya of USSR Academy of Sciences, Technical Cybernetics, 1965(5): 3-12.
- [8] Stratonovich R L. Theory of information and its value[M]. Cham: Springer International Publishing, 2020.
- [9] Weaver W. Recent contributions to the mathematical theory of communication[J]. ETC: A Review of General Semantics, 2017, 74(1/2): 136-157.
- [10] Carnap R, Bar-Hillel Y. An outline of a theory of semantic information[R]. Research Laboratory of Electronics, Massachusetts Institute of Technology, 1952.
- [11] Floridi L. Outline of a theory of strongly semantic information[J]. Minds and Machines, 2004, 14(2): 197-221.
- [12] Luca A D, Termini S. A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory[J]. Information and Control, 1972, 20(4): 301-312.
- [13] Luca A D, Termini S. Entropy of L-fuzzy sets[J]. Information and Control, 1974, 24(1): 55-73.
- [14] Wu W L. General source and general entropy[J]. Journal of Beijing University of Posts and Telecommunications, 1982, 5(1): 29-41.
- [15] Zhong Y X. Comprehensive measure of information[J]. Journal of Beijing University of Posts and Telecommunications, 1986, 9(2): 12-19.
- [16] Zhong Y X. Intelligence oriented comprehensive information theory — In memory of the 50th anniversary of Shannon information theory[J]. Journal of Beijing University of Posts and Telecommunications, 1998, 21(4): 1-6.
- [17] Lu C G. Coding significance of generalized entropy and generalized mutual information[J]. Journal on Communications, 1994, 15(6): 37-44.
- [18] Lu C G. A semantic generalization of Shannon's information theory and applications[J]. Entropy, 2025, 27(5): 461.
- [19] Bao J, Basu P, Dean M K, et al. Towards a theory of semantic communication[C]//Proceedings of the 2011 IEEE Network Science Workshop. Piscataway: IEEE Press, 2011: 110-117.
- [20] Liu J K, Zhang W Y, Poor H V. A rate-distortion framework for characterizing semantic information[C]//Proceedings of the 2021 IEEE International Symposium on Information Theory (ISIT). Piscataway: IEEE Press, 2021: 2894-2899.
- [21] Guo T, Wang Y Z, Han J, et al. Semantic compression with side information: a rate-distortion perspective[PP]. V1. (2022-08-12)[2026-09-01]. arXiv: arXiv.2208.06094.
- [22] Shao Y L, Cao Q, Gunduz D. A theory of semantic communication[PP]. V4. (2025-08-31)[2026-09-01]. arXiv: arXiv.2212.01485.
- [23] Tang J C, Yang Q Q, Zhang Z Y. Information-theoretic limits on compression of semantic information[PP]. V1. (2023-06-04)[2026-09-01]. arXiv: arXiv.2306.02305.
- [24] Niu K, Zhang P. A mathematical theory of semantic communication[J]. Journal on Communications, 2024, 45(6): 7-59.
- [25] Niu K, Zhang P. The mathematical theory of semantic communication[M]. Berlin: Springer, 2025.
- [26] Zhang P, Niu K, Liang Z J, et al. Beyond Shannon: semantic information theory and methodology[J]. IEEE Transactions on Network Science and Engineering, 2026, 13: 8062-8079.
- [27] Weizsäcker C F V, Weizsäcker E V. Wiederaufnahme der begrifflichen frage: was ist information?[J]. Nova Acta Leopoldina, 1972, 37(206): 535-555.
- [28] Gernert D. Pragmatic information: historical exposition and general overview[J]. Mind and Matter, 2006, 4(2): 141-167.
- [29] Weinberger E D. A theory of pragmatic information and its application to the quasi-species model of biological evolution[J]. Biosystems, 2002, 66(3): 105-119.
- [30] Weinberger E D. Towards a theory of pragmatic information[PP]. V8. (2026-08-27)[2026-09-01]. arXiv: arXiv.2403.12324.
- [31] Shi G M, Xiao Y, Li Y Y, et al. From semantic communication to semantic-aware networking: model, architecture, and open problems[J]. IEEE Communications Magazine, 2021, 59(8): 44-50.
- [32] Qin Z J, Tao X M, Lu J H, et al. Semantic communications: principles and challenges[PP]. V5. (2022-06-27)[2026-09-01]. arXiv: arXiv.2201.01389.
- [33] Xie H Q, Qin Z J, Tao X M, et al. Task-oriented multi-user semantic communications[J]. IEEE Journal on Selected Areas in Communications, 2022, 40(9): 2584-2597.
- [34] Gündüz D, Qin Z J, Aguerri I E, et al. Beyond transmitting bits: context, semantics, and task-oriented communications[J]. IEEE Journal on Selected Areas in Communications, 2023, 41(1): 5-41.
- [35] Zhang P, Xu W J, Gao H, et al. Toward wisdom-evolutionary and primitive-concise 6G: a new paradigm of semantic communication networks[J]. Engineering, 2022, 8: 60-73.
- [36] Niu K, Dai J C, Yao S S, et al. A paradigm shift toward semantic communications[J]. IEEE Communications Magazine, 2022, 60(11): 113-119.
- [37] Xiao L, Johansson M, Hindi H, et al. Joint optimization of communication rates and linear systems[J]. IEEE Transactions on Automatic Control, 2003, 48(1): 148-153.
- [38] Park P, Coleri Ergen S, Fischione C, et al. Wireless network design for control systems: a survey[J]. IEEE Communications Surveys & Tutorials, 2018, 20(2): 978-1013.
- [39] Wang Y, Wu S H, Lei C J, et al. A review on wireless networked control system: the communication perspective[J]. IEEE Internet of Things Journal, 2024, 11(5): 7499-7524.
- [40] Zhao G D, Ali Imran M, Pang Z B, et al. Toward real-time control in future wireless networks: communication-control co-design[J]. IEEE Communications Magazine, 2019, 57(2): 138-144.
- [41] Negi N, Chakraborty A. Optimal co-designs of communication and control in bandwidth-constrained cyber - physical systems[J]. Automatica, 2022, 142: 110288.
- [42] Lu Z B, Guo G. Control and communication scheduling co-design for networked control systems: a survey[J]. International Journal of Systems Science, 2023, 54(1): 189-203.
- [43] Tishby N, Polani D. Information theory of decisions and actions[M]//Cutsuridis V, Hussain A, Taylor J G. Perception-action cycle: models, architectures, and hardware. New York: Springer, 2011: 601-636.
- [44] Molchanov I. Theory of random sets[M]. 2nd edition. Berlin: Springer, 2017.
- [45] Cover T M, Thomas J A. Elements of information theory[M]. New York: Wiley, 1991.
- [46] Gamal A E, Kim Y H. Network information theory[M]. Cambridge: Cambridge University Press, 2011.

Appendix

A. Proof of Theorem 33

Proof. Property (1) follows directly from the definition of $\underline{A}_\epsilon^{(n)}$. By Theorem 28, the probability of the event $\underline{W}^n \in \underline{A}_\epsilon^{(n)}$ tends to 1 as $n \rightarrow \infty$. Thus for any $\epsilon > 0$, there exists n_0 such that for all $n \geq n_0$, $\Pr\{\underline{W}^n \in \underline{A}_\epsilon^{(n)}\} > 1 - \epsilon$, proving property (2). Summing over the set $\underline{A}_\epsilon^{(n)}$, using property (2) and the bounds from property (1), we obtain

$$(1 - \epsilon)2^{n(H_p(\underline{W}) - \epsilon)} \leq |\underline{A}_\epsilon^{(n)}| \leq 2^{n(H_p(\underline{W}) + \epsilon)}, \quad (327)$$

which proves property (3). $\square$

B. Proof of Theorem 34

Proof. Property (1) follows directly from the definition of $E_\epsilon^{(n)}(\underline{w}^n)$. To prove the left inequality of property (2), we write

$$\begin{aligned} P(\underline{w}^n) &= \sum_{\tilde{w}^n \in E_\epsilon^{(n)}(\underline{w}^n)} P(\tilde{w}^n) \\ &\leq \sum_{\tilde{w}^n \in E_\epsilon^{(n)}(\underline{w}^n)} P(\underline{w}^n) 2^{-n(H_s(\tilde{W}) - H_p(\underline{W}) - \epsilon)} \\ &= P(\underline{w}^n) |E_\epsilon^{(n)}(\underline{w}^n)| 2^{-n(H_s(\tilde{W}) - H_p(\underline{W}) - \epsilon)}. \end{aligned} \quad (328)$$

Dividing by $P(\underline{w}^n) > 0$ yields

$$|E_\epsilon^{(n)}(\underline{w}^n)| \geq 2^{n(H_s(\tilde{W}) - H_p(\underline{W}) - \epsilon)}. \quad (329)$$

For the right inequality, we have

$$\begin{aligned} P(\underline{w}^n) &= \sum_{\tilde{w}^n \in E_\epsilon^{(n)}(\underline{w}^n)} P(\tilde{w}^n) \\ &\geq \sum_{\tilde{w}^n \in E_\epsilon^{(n)}(\underline{w}^n)} P(\underline{w}^n) 2^{-n(H_s(\tilde{W}) - H_p(\underline{W}) + \epsilon)} \\ &= P(\underline{w}^n) |E_\epsilon^{(n)}(\underline{w}^n)| 2^{-n(H_s(\tilde{W}) - H_p(\underline{W}) + \epsilon)}, \end{aligned} \quad (330)$$

so

$$|E_\epsilon^{(n)}(\underline{w}^n)| \leq 2^{n(H_s(\tilde{W}) - H_p(\underline{W}) + \epsilon)}. \quad (331)$$

This completes the proof. $\square$

C. Proof of Theorem 36

Proof. Property (1) follows directly from the definition of $B_\epsilon^{(n)}(\underline{w}^n)$. To prove the left inequality of property (2), we write

$$\begin{aligned} P(\underline{w}^n) &= \sum_{w^n \in B_\epsilon^{(n)}(\underline{w}^n)} P(w^n) \\ &\leq \sum_{w^n \in B_\epsilon^{(n)}(\underline{w}^n)} P(\underline{w}^n) 2^{-n(H(W) - H_p(\underline{W}) - \epsilon)} \\ &= P(\underline{w}^n) |B_\epsilon^{(n)}(\underline{w}^n)| 2^{-n(H(W) - H_p(\underline{W}) - \epsilon)}. \end{aligned} \quad (332)$$

Dividing by $P(\underline{w}^n) > 0$ gives

$$|B_\epsilon^{(n)}(\underline{w}^n)| \geq 2^{n(H(W) - H_p(\underline{W}) - \epsilon)}. \quad (333)$$

For the right inequality,

$$\begin{aligned} P(\underline{w}^n) &= \sum_{w^n \in B_\epsilon^{(n)}(\underline{w}^n)} P(w^n) \\ &\geq \sum_{w^n \in B_\epsilon^{(n)}(\underline{w}^n)} P(\underline{w}^n) 2^{-n(H(W) - H_p(\underline{W}) + \epsilon)} \\ &= P(\underline{w}^n) |B_\epsilon^{(n)}(\underline{w}^n)| 2^{-n(H(W) - H_p(\underline{W}) + \epsilon)}, \end{aligned} \quad (334)$$

so

$$|B_\epsilon^{(n)}(\underline{w}^n)| \leq 2^{n(H(W) - H_p(\underline{W}) + \epsilon)}. \quad (335)$$

This completes the proof. $\square$

D. Proof of Theorem 39

Proof. Properties (1)-(3) follow directly from the definition of $\underline{A}_\epsilon^{(n)}$ and the weak law of large numbers applied to the i.i.d. pragmatic sequence pair, with the aggregated probabilities over the reified equivalence classes. For property (4), since $\dot{X}^n$ and $\dot{Y}^n$ are independent, we have

$$\Pr\{(\dot{\underline{X}}^n, \dot{\underline{Y}}^n) \in \underline{A}_\epsilon^{(n)}\} = \sum_{(\underline{x}^n, \underline{y}^n) \in \underline{A}_\epsilon^{(n)}} P(\underline{x}^n) P(\underline{y}^n).$$

Using the bounds on $P(\underline{x}^n)$ and $P(\underline{y}^n)$ from the entropy hierarchy $H_p(\underline{X}) \leq H_s(\tilde{X}) \leq H(X)$ and the size bound on $\underline{A}_\epsilon^{(n)}$, we obtain

$$\Pr\{(\dot{\underline{X}}^n, \dot{\underline{Y}}^n) \in \underline{A}_\epsilon^{(n)}\}$$

$$\begin{aligned} &\leq 2^{n(H_p(\underline{X}, \underline{Y}) + \epsilon)} 2^{-n(H(X) - \epsilon)} 2^{-n(H(Y) - \epsilon)} \\ &= 2^{-n(IP(\underline{X}; \underline{Y}) - 3\epsilon)}. \end{aligned}$$

Similarly,

$$\begin{aligned} &\Pr\{(\underline{\dot{X}}^n, \underline{\dot{Y}}^n) \in \underline{A}_\epsilon^{(n)}\} \\ &\geq (1 - \epsilon) 2^{n(H_p(\underline{X}, \underline{Y}) - \epsilon)} 2^{-n(H(X) + \epsilon)} 2^{-n(H(Y) + \epsilon)} \\ &= (1 - \epsilon) 2^{-n(IP(\underline{X}; \underline{Y}) + 3\epsilon)}. \end{aligned}$$

This proves property (4). $\square$

E. Proof of Theorem 40

Proof. Property (1) follows directly from the definition of $B_\epsilon^{(n)}(\underline{x}^n, \underline{y}^n)$. To prove the left inequality of property (2), we write

$$\begin{aligned} &P(\underline{x}^n, \underline{y}^n) \\ &= \sum_{(\underline{x}^n, \underline{y}^n) \in B_\epsilon^{(n)}(\underline{x}^n, \underline{y}^n)} P(\underline{x}^n, \underline{y}^n) \\ &\leq \sum_{(\underline{x}^n, \underline{y}^n) \in B_\epsilon^{(n)}(\underline{x}^n, \underline{y}^n)} P(\underline{x}^n, \underline{y}^n) 2^{-n(H(X, Y) - H_p(\underline{X}, \underline{Y}) - \epsilon)} \\ &= P(\underline{x}^n, \underline{y}^n) |B_\epsilon^{(n)}(\underline{x}^n, \underline{y}^n)| 2^{-n(H(X, Y) - H_p(\underline{X}, \underline{Y}) - \epsilon)}. \end{aligned}$$

Dividing by $P(\underline{x}^n, \underline{y}^n) > 0$ yields the lower bound. The right inequality follows similarly by reversing the inequality. $\square$

F. Proof of Theorem 43

Proof. Properties (1)-(3) follow directly from the definition of $\underline{A}_\epsilon^{(n)}$ and the weak law of large numbers applied to the i.i.d. pragmatic sequence pair, with the aggregated probabilities over the reified equivalence classes. For property (4), since $\dot{X}^n$ and $\hat{X}^n$ are independent, we have

$$\Pr\{(\underline{\dot{X}}^n, \underline{\hat{X}}^n) \in \underline{A}_\epsilon^{(n)}\} = \sum_{(\underline{x}^n, \underline{\hat{x}}^n) \in \underline{A}_\epsilon^{(n)}} P(\underline{x}^n) P(\underline{\hat{x}}^n).$$

Using the bounds on $P(\underline{x}^n)$ and $P(\underline{\hat{x}}^n)$ from the entropy hierarchy $H_p(\underline{X}) \leq H_s(\tilde{X}) \leq H(X)$ and the size bound

on $\underline{A}_\epsilon^{(n)}$, we obtain the desired exponential bounds. $\square$

ABOUT THE AUTHOR



Kai Niu received the B.S. degree in information engineering and the Ph.D. degree in signal and information processing from Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 1998 and 2003, respectively. He is currently a Professor of the School of Artificial Intelligence, BUPT. His research interests include information theory, coding theory, and semantic communication. He has published more than 200 academic papers, and proposed high performance construction algorithm of polar codes applied in 5G standard, and won the first prize of Natural Science of Science and Technology Award of the Chinese Institute of Electronics.



Ping Zhang is currently a Professor of the School of Information and Communication Engineering at Beijing University of Posts and Telecommunications, the director of State Key Laboratory of Networking and Switching Technology, a member of IMT-2020 (5G) Experts Panel, a member of Experts Panel for China's 6G development. He served as Chief Scientist of National Basic Research Program (973 Program), an expert in Information Technology Division of National High-tech R&D program (863 Program), and a member of Consultant Committee on International Cooperation of National Natural Science Foundation of China. His research interests mainly focus on wireless communication. He is an Academician of the Chinese Academy of Engineering (CAE).