Communicating Classification Results over Noisy Channels

Noel Teku Sudarshan Adiga Ravi Tandon
Dept. of Electrical and Computer Engineering
University of Arizona, Tucson, Arizona, USA
Email: {*nteku1*, *adiga*, *tandonr*}@arizona.edu

***Abstract*—In this work, the problem of communicating decisions of a classifier over a noisy channel is considered. With machine learning based models being used in variety of time-sensitive applications, transmission of these decisions in a reliable and timely manner is of significant importance. To this end, we study the scenario where a probability vector (representing the decisions of a classifier) at the transmitter, needs to be transmitted over a noisy channel. Under the assumption that the distortion between the original probability vector and the reconstructed one at the receiver is measured via f-divergence, we study the trade-off between transmission latency and the distortion. We completely analyze this trade-off for the setting when uniform quantization is used to encode the probability vector, and the latency incurred is obtained via results on finite-blocklength channel capacity. Our results show that there is an interesting interplay between source distortion (i.e., distortion for the probability vector measured via f-divergence) and the subsequent channel encoding/decoding parameters; and indicate that a *joint* design of these parameters is crucial to navigate the latency-distortion tradeoff.**

I. Introduction

In recent years, machine learning (ML) has been increasingly applied to time-sensitive applications, including Vehicle to Vehicle (V2V) and Vehicle to Infrastructure (V2I) communications. These applications require reliable and rapid data transmission for tasks such as trajectory prediction [1] and lane change detection [2]. Similarly, this need for reliable, fast communication extends to other domains like internet of things (IoT) and edge computing. Coinciding with the increasing use of ML in low-latency applications, there has also been a growing body of work on context-dependent low-latency communications; which includes semantic communications [3]–[6], ultra-reliable low latency communications (URLLC) [7], [8], and joint source channel coding (JSCC) [9]–[12].

Semantic communication generally focuses on sending *context dependent* features/decisions dependent on the data to the receiver (rather than the entire raw message) [6]. In doing so, the amount of bits required for transmission is often reduced [5]. For example, in [13], a transformer-based network was used to learn/transmit semantic features of sentences and decode the received features to ensure that the original meaning of the sentences were preserved. In [14], an approach to modeling the length of a semantic message and its distortion based on noise due to the model and the channel is presented along with masking strategies that can be applied before transmission. [15], [16] present rate-distortion approaches for semantic communications for general block-wise distortion functions. The focus of URLLC is to design protocols in order to transmit transmitting low-data rate (short packets) with high reliability (low probability of error) within a small latency [8]. A rate-distortion analysis is also performed in [17] for short control packets, assuming transmissions are being made to a remote agent, where the distortion measures considered are quantization error and the freshness of the data (age of information); however, this analysis is done under the assumption of noiseless channels.

Overview and Main Contributions: In this paper, we focus on the following problem: a transmitter wishes to send a probability vector (e.g., representing the decisions of a ML based classifer) to a receiver over a noisy channel. Transmitting the results of a classification task incurs lower latency/overhead compared to sending a compressed form of the data required for classification at the receiver. It also enables the receiver to quickly execute tasks that depend on knowing the classification results, which applies to goal-oriented communications [5]. Additionally, this problem falls under the umbrella of JSCC as its objective is to attain low-latency transmissions by operating in the finite blocklength regime [11]. The main new elements herein are two fold: we measure utility of the reconstruction of the probability vector in terms of statistical divergence measures; and secondly, we simultaneously want to minimize the transmission latency over the noisy channel. We note that there has been prior work on quantizing probability distributions, including [18]–[22]. In particular, [21], [22] investigated quantizing probability distributions in order to minimize Kullback-Leibler (KL)-divergence (with [22] providing additional analysis for L_1 & L_2^2 distances) by performing a non-linear operation and then using uniform quantization. However, the existing works did not study the scenario when a probability vector has to be transmitted through a noisy channel, and what would be the right quantization strategy/parameters if the goal is to minimize latency. By considering the distortion introduced by the channel and quantization, we aim to analyze the trade off between the end-to-end distortion of the system and the incurred transmission latency. Our main contributions are as follows: (a) Characterizing the end-to-end distortion between received and transmitted vectors using the statistical

This work was supported by NSF grants CAREER 1651492, CCF-2100013, CNS-2209951, CNS-1822071, CNS-2317192.

ICC 2024 - IEEE International Conference on Communications | 978-1-7281-9054-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICC51166.2024.10622771

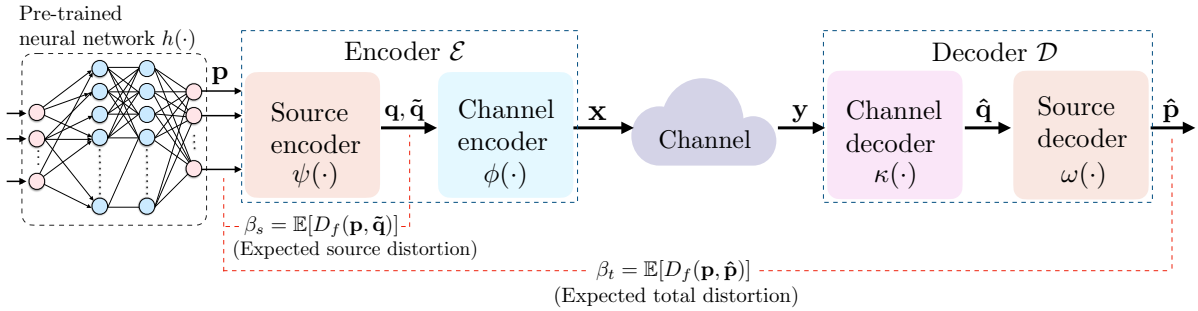


Fig. 1: End-to-End block diagram for communicating classifier decisions (probability vector) efficiently over a noisy channel.

f-divergence measure, especially when classification probabilities are uniformly quantized; (b) Determining the bit budget required to meet specific latency and distortion constraints, considering both source and channel distortion; (c) We show the interplay between source distortion and encoder/decoder parameters; and leveraging this relationship, along with insights from finite block length analysis, to explore a trade-off between latency and end-to-end distortion.

II. System Model

We consider the scenario illustrated in Fig. 1: a pre-trained classifier (e.g., a neural network), denoted as $h(\cdot)$, is used for a k-class classification problem and is situated at a transmitter. The output classification probabilities are represented as $\mathbf{p} = [\mathbf{p}[1], \mathbf{p}[2], \cdots, \mathbf{p}[k]]^\top$, where $\mathbf{p} \in \mathbb{R}^{k\times 1}$. Let $\hat{\mathbf{p}} = [\hat{\mathbf{p}}[1], \hat{\mathbf{p}}[2], \cdots, \hat{\mathbf{p}}[k]]^\top$ denote the estimated classifier output at the receiver. In this paper, we measure the distortion between $\mathbf{p}$ and $\hat{\mathbf{p}}$ via f-divergence, defined as

$$D_{\mathrm{f}}(\mathbf{p}, \hat{\mathbf{p}}) = \sum_{i=1}^{k} f\left(\frac{\mathbf{p}[i]}{\hat{\mathbf{p}}[i]}\right) \hat{\mathbf{p}}[i]. \tag{1}$$

The transmitter's goal is to communicate the probability vector $\mathbf{p}$ within a latency budget of $T_{\max}$ with minimum total expected distortion β_t, i.e., $\mathbb{E}(D_{\mathrm{f}}(\mathbf{p}, \hat{\mathbf{p}})) \leq \beta_t$, where the expectation is over the noisy channel realizations. We next describe the main components (source/channel encoder/decoder(s)): a source encoder $\psi(\cdot)$ quantizes the probability vector $\mathbf{p}$, such that $\mathbf{q} = \psi(\mathbf{p})$. The lossy compression caused by quantization results in source distortion, denoted by β_s. The total number of bits required by $\mathbf{q}$, given the source distortion, is represented as $J(\beta_s)$, where $J(\cdot)$ is a function of β_s. We note that $\mathbf{q}$ may not necessarily be a probability vector. We normalize the values in $\mathbf{q}$ to obtain the corresponding probability vector $\tilde{\mathbf{q}}$ after source encoding, where $\tilde{\mathbf{q}}[i] = \mathbf{q}[i]/\sum_{i=1}^{k} \mathbf{q}[i]$ and $\tilde{\mathbf{q}} \in \mathbb{R}^{k\times 1}$. The source distortion β_s is quantified as $\beta_s = D_{\mathrm{f}}(\mathbf{p}, \tilde{\mathbf{q}})$. We use the channel encoder $\phi(\cdot)$ to generate the n-length channel input $\mathbf{x} = \phi(\mathbf{q})$, where $\mathbf{x} = [\mathbf{x}[1], \mathbf{x}[2], \cdots, \mathbf{x}[n]]^\top$ and $\mathbf{x} \in \mathcal{X}^n$. Let $\mathcal{E}$ denote the source and channel encoder pair. We consider the scenario of a bandwidth constrained AWGN channel, where the channel output is given by $\mathbf{y}[i] = \mathbf{x}[i] + \mathbf{z}[i]$, for all $i \in [n]$; where $\mathbf{y} \in \mathbb{R}^{n\times 1}$ and the AWGN noise vector is given by $\mathbf{z} = [\mathbf{z}[1], \mathbf{z}[2], \cdots, \mathbf{z}[n]]^\top$ with $\mathbf{z} \in \mathbb{R}^{n\times 1}$. The signal-to-noise ratio (SNR) of the channel for a bandwidth B_0 Hz, is defined as $\gamma_0 = \frac{P}{N_0}$, where P denotes the signal power and N_0 denotes the noise power. To simulate using the same transmit powers at different bandwidths, as done in [8], we define the operational SNR for a channel of bandwidth B Hz as $\gamma = \frac{\gamma_0 B_0}{B}$ where $\frac{B_0}{B}$ acts as a scaling factor for relating different channel conditions. We denote the decoding error probability by $\epsilon^*(n)$, where $\epsilon^*(n) \in [0, 1]$. At the receiver, we consider a channel decoder, denoted by $\kappa(\cdot)$, such that $\hat{\mathbf{q}} = \kappa(\mathbf{y})$. Subsequently, we consider the source decoder $\omega(\cdot)$ and a normalization operation to obtain an estimate of the classifier probabilities, given by $\hat{\mathbf{p}} = \omega(\hat{\mathbf{q}})$. Let $\mathcal{D}$ denote the source and channel decoder pair.

The channel noise, in addition to the source distortion, contributes to the total end-to-end distortion. Given a specific SNR, it is possible to vary the source distortion β_{s} to achieve a maximum total expected distortion of β_{t}. In other words, we have $\beta_s \in [0, \beta_{\mathrm{t}}]$. This choice will also affect the incurred transmission latency; given a bandwidth of B Hz, the time required to transmit an n-length vector $\mathbf{x}$ is calculated as:

$$T(\mathcal{E}, \mathcal{D}) = \frac{n}{2B}. \tag{2}$$

In this paper, we focus on understanding the tradeoff between latency and distortion for the task of communicating probability distributions. Specifically, given the channel statistics (e.g., bandwidth, SNR) and desired maximum latency $T_{\max}$, the optimal distortion can be defined as follows:

$$D^*(T_{\max}) \triangleq \min_{(\mathcal{E},\mathcal{D})} \quad \beta_t(\mathcal{E}, \mathcal{D}), \quad \text{s.t.} \quad T(\mathcal{E}, \mathcal{D}) \leq T_{\max}. \tag{3}$$

Alternatively, we can fix the maximum permissible distortion $\beta_{\max}$, and minimize the total latency T over encoder-decoder pairs as

$$T^*(\beta_{\max}) \triangleq \min_{(\mathcal{E},\mathcal{D})} \quad T(\mathcal{E}, \mathcal{D}), \quad \text{s.t.} \quad \beta_t(\mathcal{E}, \mathcal{D}) \leq \beta_{\max}. \tag{4}$$

In the lemma stated next, we show that the optimal latency $T^*(\beta_{\max})$ is a convex non-increasing function of the total distortion $\beta_{\max}$; and likewise, we show that the minimal distortion $D^*(T_{\max})$ is a convex non-increasing function of $T_{\max}$.

Lemma 1. *$T^*(\beta_{max})$ is convex non-increasing function of β_{max}. $D^*(T_{max})$ is convex non-increasing function of T_{max}.*

Proof. The minimal achievable latency $T^*(\beta)$ is a non-increasing function of β. This is clear from the fact that

any decoder which satisfies a distortion constraint of β also satisfies the distortion constraint of β' for $\beta' \geq \beta$.

We next show that $T^*(\beta)$ is a convex function of β. Let $T^*(\beta_{t_1})$ ($T^*(\beta_{t_2})$, respectively) represent the minimum latencies obtained using encoder-decoder pair $(\mathcal{E}_1^*, \mathcal{D}_1^*)$ ($(\mathcal{E}_2^*, \mathcal{D}_2^*)$, respectively) that satisfy $D_{\mathrm{f}}(\mathbf{p}, \hat{\mathbf{p}}_1) \leq \beta_{t_1}$ ($D_{\mathrm{f}}(\mathbf{p}, \hat{\mathbf{p}}_2) \leq \beta_{t_2}$, respectively). We define a new encoder-decoder pair $(\mathcal{E}_3, \mathcal{D}_3)$ such that,

$$(\mathcal{E}_3, \mathcal{D}_3) = \begin{cases} (\mathcal{E}_1^*, \mathcal{D}_1^*) & \text{with probability } \alpha \\ (\mathcal{E}_2^*, \mathcal{D}_2^*) & \text{with probability } 1-\alpha. \end{cases}$$

The expected latency using $(\mathcal{E}_3, \mathcal{D}_3)$ is $\alpha T^*(\beta_{t_1}) + (1-\alpha)T^*(\beta_{t_2})$. The total distortion using $(\mathcal{E}_3, \mathcal{D}_3)$ is $D_{\mathrm{f}}(\mathbf{p}, \alpha\hat{\mathbf{p}}_1 + (1-\alpha)\hat{\mathbf{p}}_2)$, which can be upper bounded as,

$$\begin{aligned} D_{\mathrm{f}}(\mathbf{p}, \alpha\hat{\mathbf{p}}_1 + (1-\alpha)\hat{\mathbf{p}}_2) &\overset{(a)}{\leq} \alpha D_{\mathrm{f}}(\mathbf{p}, \hat{\mathbf{p}}_1) + (1-\alpha) D_{\mathrm{f}}(\mathbf{p}, \hat{\mathbf{p}}_2) \\ &\overset{(b)}{\leq} \alpha\beta_{t_1} + (1-\alpha)\beta_{t_2}, \end{aligned} \tag{5}$$

where (a) follows from convexity of f-divergence, and (b) follows from the bounds on end-to-end distortion for the two individual decoders. Let us now consider the optimal encoder-decoder pair $(\mathcal{E}^*, \mathcal{D}^*)$ that satisfies the distortion constraint $D_{\mathrm{f}}(\mathbf{p}, \hat{\mathbf{p}}) \leq \alpha\beta_{t_1} + (1-\alpha)\beta_{t_2}$. The minimum latency using $(\mathcal{E}^*, \mathcal{D}^*)$ is then $T^*(\alpha\beta_{t_1} + (1-\alpha)\beta_{t_2})$. Recall, $(\mathcal{E}^*, \mathcal{D}^*)$ is optimal encoder-decoder pair, and therefore, the corresponding latency must be always less than the latencies obtained using any $(\mathcal{E}_3, \mathcal{D}_3)$ pair. That is,

$$T^*(\alpha\beta_{t_1} + (1-\alpha)\beta_{t_2}) \leq \alpha T^*(\beta_{t_1}) + (1-\alpha)T^*(\beta_{t_2}).$$

This proves that $T^*(\beta)$ is convex. $D^*(T)$ can be shown to be convex in a similar manner by leveraging the convexity of f-divergence. □

III. Main Results & Discussion

In this section, we present the framework for analyzing the latency-distortion tradeoff. We begin by assuming a noiseless channel and uniform quantization as the source encoder (i.e., transforming $\mathbf{p}$ to $\mathbf{q}$) and analyze the corresponding source distortion (Lemma 2). We note that other non-uniform quantization techniques can be readily substituted in this framework. We then incorporate and analyze the impact of channel noise on the end-to-end distortion (Lemma 3). Subsequently, we use results on finite-blocklength capacity, which allow us to connect latency with the overall distortion. This, in turn, also leads to an explicit optimization (Theorem 1), which can be solved to trade latency with distortion.

A. Quantizing Classifier Probabilities

Suppose we have a total budget of J bits to quantize the k-dimensional probability vector $\mathbf{p}$. Under uniform quantization, we use $j = \lfloor J/k \rfloor$ bits to quantize each element $\mathbf{p}[i], i = 1, 2, \ldots, k$. We denote $\mathbf{q}[i]$ as the resulting quantized output. Note that $\mathbf{q}$ may not necessarily be a probability vector. We can however, normalize it as $\tilde{\mathbf{q}}[i] = \frac{\mathbf{q}[i]}{\sum_{\ell=1}^{k} \mathbf{q}[\ell]}$, for $i = 1, \ldots, k$. Recall that our objective is to minimize the f-divergence between $\mathbf{p}$ & $\hat{\mathbf{p}}$; in the noiseless scenario, this would be equivalent to minimizing $D_{\mathrm{f}}(\mathbf{p}, \tilde{\mathbf{q}})$, as β_s would be the only distortion present. When $f(x) = \frac{1}{2}|x-1|$, f-divergence results in total variation (TV): $D_{\mathrm{f}}(\mathbf{p}, \mathbf{q}) = D_{\mathrm{TV}}(\mathbf{p}, \mathbf{q}) = \frac{1}{2}\sum_{i=1}^{K} |\mathbf{p}[i] - \mathbf{q}[i]|$ [23]. For the remainder of this paper, we use TV as the divergence metric. The next lemma shows a sufficient condition on the quantization budget to achieve a source distortion of β_s.

Lemma 2. *For a k-class classification problem, if the total quantization budget satisfies*

$$J \geq 2k \cdot \log_2\left(\frac{k}{\beta_s}\right), \tag{6}$$

then $D_{TV}(\mathbf{p}, \tilde{\mathbf{q}}) \leq \beta_s$.

Proof. We consider uniform quantization with bins of width $\frac{1}{2^j}$, where $j = \lfloor J/k \rfloor$. For $r \in \{1, 2, \cdots, 2^j - 1\}$, we define the values for the r^{th} bin in the range $\left[\frac{r}{2^j}, \frac{r+1}{2^j}\right)$. In other words, for any $i \in [k]$, $\mathbf{q}[i]$ is obtained by mapping the value of $\mathbf{p}[i]$ in the range $\left[\frac{r_i}{2^j}, \frac{r_i+1}{2^j}\right)$ to $\frac{r_i+\frac{1}{2}}{2^j}$. We define the probability vector $\tilde{\mathbf{q}}$, by normalizing the values in $\mathbf{q}$, to ensure that the resulting values in the quantized vector sum to 1. Therefore, for any $i \in [k]$, we can write $\mathbf{p}[i]$ and $\tilde{\mathbf{q}}[i]$ as follows:

$$\mathbf{p}[i] = \frac{r_i + \delta_i}{2^j}, \qquad \tilde{\mathbf{q}}[i] = \frac{r_i + \frac{1}{2}}{S \cdot 2^j}, \tag{7}$$

where $\delta_i \in [0, 1)$ and $S = \sum_{i=1}^{k} \frac{r_i + \frac{1}{2}}{2^j}$. Returning to our goal, recall that we wish to pick J such that $D_{\mathrm{TV}}(\mathbf{p}, \tilde{\mathbf{q}}) \leq \beta_s$. To this end, we first bound

$$\begin{aligned} D_{\mathrm{TV}}(\mathbf{p}, \tilde{\mathbf{q}}) &= \frac{1}{2}\sum_{i=1}^{k} |\mathbf{p}[i] - \tilde{\mathbf{q}}[i]| \\ &= \frac{1}{2}\sum_{i=1}^{k} \left| \frac{r_i + \delta_i}{2^j} - \frac{r_i + \frac{1}{2}}{S \cdot 2^j} \right| \\ &= \frac{1}{2^{(j+1)}}\sum_{i=1}^{k} \left| r_i\left(1 - \frac{1}{S}\right) + \delta_i - \frac{1}{2S} \right| \\ &\overset{(a)}{\leq} \frac{1}{2^{(j+1)}}\sum_{i=1}^{k} \left(\left| r_i\left(1 - \frac{1}{S}\right) \right| + \left| \delta_i - \frac{1}{2S} \right| \right) \\ &\overset{(b)}{\leq} \frac{1}{2^{(j+1)}}\sum_{i=1}^{k} \left(\left| 2^j\left(1 - \frac{1}{S}\right) \right| + \left| \delta_i - \frac{1}{2S} \right| \right), \end{aligned} \tag{8}$$

where (a) follows from triangle inequality, and (b) follows from the fact that $r_i < 2^j$. We also know that, $\sum_i^k \mathbf{p}[i] = 1$ and $\sum_i^k \tilde{\mathbf{q}}[i] = S$. Consider the difference,

$$\left| \sum_{i=1}^{k} \frac{r_i + \delta_i}{2^j} - \sum_{i=1}^{k} \frac{r_i + \frac{1}{2}}{2^j} \right| \leq \sum_{i=1}^{k} \frac{|\delta_i - \frac{1}{2}|}{2^j} \leq \frac{k}{2^{j+1}}. \tag{9}$$

Therefore, we have that,

$$|1 - S| \leq \frac{k}{2^{j+1}}. \tag{10}$$

Suppose that j is given as

$$j = \log_2\left(\frac{k}{2\alpha}\right) \tag{11}$$

for some $\alpha \in (0, 0.5]$. Therefore, from (10) we have that,

$$1 - \alpha \leq S \leq 1 + \alpha. \tag{12}$$

Using (11) and (12) in (8), we can further bound $D_{\text{TV}}(\mathbf{p}, \tilde{\mathbf{q}})$ as,

$$\begin{aligned} D_{\text{TV}}(\mathbf{p}, \tilde{\mathbf{q}}) &\leq \left[\frac{\alpha k}{2(1-\alpha)} + \max\left\{\frac{\alpha}{2(1-\alpha)}, \alpha - \frac{\alpha}{2(1+\alpha)}\right\}\right] \\ &\overset{(a)}{\leq} \frac{\alpha k}{2(1-\alpha)} + \frac{\alpha}{2(1-\alpha)} = \frac{\alpha}{1-\alpha}\left(\frac{k+1}{2}\right), \end{aligned} \tag{13}$$

where (a) holds for all $\alpha \in (0, 0.5]$. Now, since we require $D_{\text{TV}}(\mathbf{p}, \tilde{\mathbf{q}}) \leq \beta_s$, we can pick α such that $\frac{\alpha}{1-\alpha}\left(\frac{k+1}{2}\right) \leq \beta_s$. We can pick α to satisfy this constraint with equality, i.e.,

$$\alpha^* = \frac{2\beta_{\text{s}}}{k + 1 + \beta_{\text{s}}}. \tag{14}$$

Next, substituting (14) in (11), we can then claim that as long as the total bit budget $J = kj \geq k\log_2(k/2\alpha^*)$, then $D_{\text{TV}}(\mathbf{p}, \tilde{\mathbf{q}}) \leq \beta_s$. Now, using the fact that $k \geq 2$ and $\beta_s \in [0, 1]$, it can be readily verified that $2k \cdot \log_2\left(\frac{k}{\beta_s}\right) \geq k\log_2(k/2\alpha^*)$, completing the proof of Lemma 2. □

Remark 1. *(Impact of normalization) The proof of Lemma 2 is non-trivial due to the nature of the vectors involved. To apply the statistical f-divergence measure, we normalize the entries of $\mathbf{q}$ by their sum, i.e., $S = \sum_i \mathbf{q}[i]$. However, this normalization operation makes the analysis of bounding the f-divergence challenging. The proof above overcomes this issue, by first assuming that the number of bits j is of the form $j = \log(k/2\alpha)$, and then we are able to bound the sum S as $S \in [1-\alpha, 1+\alpha]$. This allows us to determine the number of bits required to achieve a desired source distortion β_s.*

Impact of Channel Noise & Decoding Error: We now incorporate the effects of channel noise and decoding errors in the analysis of the end-to-end distortion. The next lemma shows a bound on the overall expected distortion (expectation is over the channel noise realizations) if the source distortion is bounded by β_s, and the decoding error probability is given by $\epsilon^*(n)$.

Lemma 3. *For a given source distortion β_s and decoding error probability $\epsilon^*(n)$, the overall expected distortion is upper bounded as follows:*

$$\mathbb{E}[D_{TV}(\mathbf{p}, \hat{\mathbf{p}}(\kappa(\mathbf{y})))] \leq (1 - \epsilon^*(n))\beta_s + \epsilon^*(n). \tag{15}$$

Proof. We can bound the end-to-end expected distortion as

$$\begin{aligned} \mathbb{E}[D_{\text{TV}}(\mathbf{p}, \hat{\mathbf{p}}(\kappa(\mathbf{y}))] &\overset{(a)}{=} P(\psi(\mathbf{p}) = \kappa(\mathbf{y}))D_1 + P(\psi(\mathbf{p}) \neq \kappa(\mathbf{y}))D_2 \\ &\overset{(b)}{\leq} (1 - \epsilon^*(n))D_1 + \epsilon^*(n)D_2 \\ &\overset{(c)}{\leq} (1 - \epsilon^*(n))\beta_s + \epsilon^*(n), \end{aligned} \tag{16}$$

where (a) follows from the total probability theorem, and $D_1 = \mathbb{E}[D_{\text{f}}(\mathbf{p}, \psi(\mathbf{p}))|\psi(\mathbf{p}) = \kappa(\mathbf{y})]$ ($D_2 = \mathbb{E}[D_{\text{f}}(\mathbf{p}, \hat{\mathbf{p}})|\psi(\mathbf{p}) \neq \kappa(\mathbf{y})]$, respectively) is the expected distortion when quantized probability vector is exactly constructed (not exactly reconstructed, respectively) at the receiver; (b) follows from considering a bound on the decoding error probability such that, $\mathbb{P}(\psi(\mathbf{p}) \neq \kappa(\mathbf{y})) \leq \epsilon^*(n)$. (c) follows using the source distortion constraint $D_1 \leq \beta_s$, and from the fact that the total distortion is quantified using TV-divergence which allows us to bound $D_2 \leq 1$. □

Remark 2. *An observation from Lemma 3 is that there is no explicit dependence on the specific quantization technique. The bound on overall distortion is only dependent on the source distortion β_s and decoding error probability introduced via $\epsilon^*(n)$. This indicates that this framework can work generally for different quantization techniques (for instance, one could replace uniform quantization with some other sophisticated non-uniform quantizer) and the bound will only be a function of the corresponding source distortion.*

B. Tradeoff Between Latency & End-to-End Distortion

In this Section, we show how the lemmas shown up to this point can be used to devise a framework for analyzing the tradeoff between latency and distortion. In Lemma 3, we showed that the overall expected distortion $\mathbb{E}[D_{\text{TV}}(\mathbf{p}, \hat{\mathbf{p}}(\kappa(\mathbf{y}))]$ can be upper bounded by $(1 - \epsilon^*(n))\beta_s + \epsilon^*(n)$. For brevity, we refer to $\mathbb{E}[D_{\text{TV}}(\mathbf{p}, \hat{\mathbf{p}}(\kappa(\mathbf{y}))]$ as β_t (i.e., the total expected distortion). To derive a relationship between the overall distortion β_t and latency T, we first recall the finite blocklength result from [8], which states that the decoding error probability $\epsilon^*(n)$ that can be assured for sending J bits through an AWGN channel is given by [8] [24]:

$$\epsilon^*(n, \gamma, J) = Q\left(\frac{nC(\gamma) - J + \frac{1}{2}\log_2 n}{\sqrt{nV(\gamma)}}\right), \tag{17}$$

where n represents the blocklength, γ represents the SNR, $C(\gamma)$ represents the capacity defined by $\frac{1}{2}\log_2(1+\gamma)$ and $V(\gamma)$ denotes the channel dispersion defined by $\frac{\gamma(\gamma+2)}{2(\gamma+1)^2}(\log_2(e))^2$.

Let us now return to the problem of transmitting a probability vector $\mathbf{p}$ over an AWGN channel. Observe that the number of bits one can use to represent $\mathbf{p}$ can be chosen as a function of the source distortion β_s (via Lemma 2, i.e., $J(\beta_s)$). However, the choice of β_s, and therefore $J(\beta_s)$ also directly impact the decoding error probability $\epsilon^*(n, \gamma, J(\beta_s))$ as given in (17). Thus, the resulting overall distortion from Lemma 3 can then be bounded by $(1-\epsilon^*(n, \gamma, J(\beta_s)))\beta_s + \epsilon^*(n, \gamma, J(\beta_s))$. Hence, if we are given a target total expected distortion of β_t, one can then optimize β_s to minimize latency while satisfying the total distortion budget. This is the core idea behind our approach and is formalized in the following Theorem.

Theorem 1. *Given a total distortion budget β_t, using uniform quantization, we can achieve the following latency:*

$$T_{Unif}(\beta_t) = \min_{0 \leq \beta_s \leq \beta_t} \frac{n(\beta_s)}{2B} \tag{18}$$

where

$$\sqrt{n(\beta_s)} = \frac{r + \sqrt{r^2 + 4C(\gamma)J(\beta_s)}}{2C(\gamma)}, \tag{19}$$

$J(\beta_s) = 2k \log_2\left(\frac{k}{\beta_s}\right)$, *and* $r = \sqrt{V(\gamma)}Q^{-1}\left(\frac{\beta_t - \beta_s}{1-\beta_s}\right)$.

Proof. First, the number of bits to quantize $\mathbf{p}$ can be obtained based on the choice of quantization technique (we pick $J(\beta_s) = 2k \log_2\left(\frac{k}{\beta_s}\right)$ for uniform quantization based on Lemma 2). We can then rearrange Lemma 3 to solve for the desired decoding error probability $\epsilon^*(n, \gamma, J(\beta_s)) = \frac{\beta_t - \beta_s}{1-\beta_s}$ in terms of β_t & β_s. Hence, the next step is to find the minimum number of channel uses (n) that can support the desired decoding error probability of $\frac{\beta_t - \beta_s}{1-\beta_s}$ by using (17). Specifically, we wish to solve for the smallest non-negative integer n satisfying:

$$\left(\frac{\beta_t - \beta_s}{1 - \beta_s}\right) \leq Q\left(\frac{nC(\gamma) - J(\beta_s) + \frac{1}{2}\log_2 n}{\sqrt{nV(\gamma)}}\right). \tag{20}$$

As the $Q(\cdot)$ function (complementary CDF of standard Gaussian) is monotonically decreasing, this means that for any $n > 1$, we can bound the r.h.s. of (20) as:

$$Q\left(\frac{nC(\gamma) - J(\beta_s) + \frac{1}{2}\log_2 n}{\sqrt{nV(\gamma)}}\right) \leq Q\left(\frac{nC(\gamma) - J(\beta_s)}{\sqrt{nV(\gamma)}}\right). \tag{21}$$

Thus, we can find n by instead solving for the simpler equation $\left(\frac{\beta_t - \beta_s}{1-\beta_s}\right) = Q\left(\frac{nC(\gamma) - J(\beta_s)}{\sqrt{nV(\gamma)}}\right)$. Applying $Q^{-1}(\cdot)$ on both sides, we arrive at the following:

$$nC(\gamma) - \sqrt{nV(\gamma)}Q^{-1}((\beta_t - \beta_s)/(1-\beta_s)) - J(\beta_s) = 0. \tag{22}$$

This equation can be viewed as a quadratic by setting $\tilde{n} = \sqrt{n}$. Solving for n, we arrive at the latency expression (setting $T = n/2B$). One can then optimize the latency by minimizing over all $\beta_s \in [0, \beta_t]$, thus completing the proof of Theorem 1. □

IV. Experimental Results

In this section, results are presented investigating the tradeoff between the latency and overall distortion for sending a probability vector to a receiver over the AWGN channel. Unless otherwise stated, we assume that $k = 10$, $B = 25$ kHz, $B_0 = 10$ kHz, and $\gamma_0 = 5$ dB, which yields the SNR of $\gamma \approx 1.021$ dB. We also assume $0 < \epsilon^*(n) < 0.5$.

Minimum latency for a fixed total distortion (β_t): Figure 2a shows the lower convex hull of the minimum latencies that can be attained at different β_t under uniform quantization. These results were obtained by solving the optimization problem presented in Theorem 1. The figure indicates an inverse relationship between latency and β_t, where a smaller latency can be attained at the cost of a higher end-to-end distortion and vice versa. Because $T^*(\cdot)$ is convex, as proven in Lemma 1, the figure also indicates that there exists an encoder-decoder pair, whose convex combination can attain smaller latencies while still satisfying the constraint on β_t.

Interplay between source distortion (β_s) and latency: Figure 2b shows the latencies attained for different values of β_s at a fixed total distortion β_t. Three observations can be made. First, as β_t increases, it is noted that the optimum source distortion β_s also increases, which is in agreement with the theoretical result in Lemma 3. Secondly, higher values of β_t (which correspond to higher source distortions β_s) allow for lower achievable latencies. Third, surges in the latency occur as β_s approaches β_t. This intuitively makes sense because as β_s approaches β_t, this implies that no compensation for the distortion is being performed by the source encoder/decoder. This means that the channel encoder/decoder are responsible for satisfying the requirement on β_t and can only do so by using higher blocklengths.

Latency as a function of channel bandwidth: Figure 2c displays the latencies achieved for different bandwidths and values of β_t. It is important to recall that the Signal-to-Noise Ratio (SNR) is related to the bandwidth B as $\gamma = \frac{\gamma_0 B_0}{B}$, indicating that SNR decreases as B increases. Firstly, it is observed that as the total distortion β_t increases, the latencies decrease. Secondly, for a fixed total distortion β_t, the latency decreases with an increase in bandwidth B.

Impact of number of classes, k (dimensionality of $\mathbf{p}$): Figure 3 presents the plot of latencies against total distortion β_t for a varying number of classes k. In line with our theoretical result in (6), it is noted that for larger values of k, the minimum bit-budget required to achieve a given distortion constraint is higher, as observed in Fig. 3.

V. Conclusion

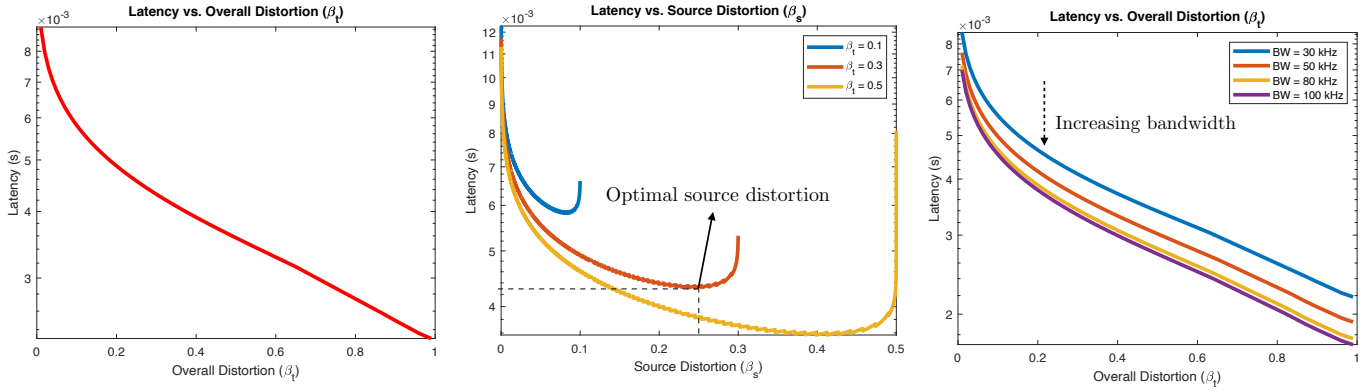
In this work, we have investigated a framework where results from a classification task are sent from a transmitter to a receiver. Specifically, we analyze the tradeoff between the latency associated with transmitting this result against the distortion incurred with quantizing the result and the impact of channel noise on the transmission. Our results show that there is an interesting interplay between source distortion (i.e., distortion for the probability vector measured via f-divergence) and the subsequent channel encoding/decoding parameters; and indicate that a *joint* design of these parameters is crucial to navigate the latency-distortion tradeoff. There are several directions for future work, including investigating non-uniform quantization, fading channels, assuming different divergence functions, as well as developing converse results (for instance, obtaining lower bounds on the distortion as a function of a target latency, and vice-versa).

References

[1] D. Choi, J. Yim, M. Baek, S. Lee, "Machine Learning-Based Vehicle Trajectory Prediction Using V2V Communications and On-Board Sensors", in Electronics 2021, vol. 10, no. 4: 420.

[2] A. H. Sakr, G. Bansal, V. Vladimerou and M. Johnson, "Lane Change Detection Using V2V Safety Messages," 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, USA, pp. 3967-3973, 2018.

[3] D. Gündüz et al., "Beyond Transmitting Bits: Context, Semantics, and Task-Oriented Communications," in IEEE Journal on Selected Areas in Communications, vol. 41, no. 1, pp. 5-41, Jan. 2023.



(a) (b) (c)

Fig. 2: (a) Lower convex hull of latencies for different β_t for the case of uniform quantization (obtained from Theorem 1). (b) This figure shows the latency as a function of the source distortion for different values of $\beta_t = 0.1, 0.3, 0.5$. We can clearly observe that for every value of β_t, there is an optimum choice of source distortion β_s which yields the smallest latency. (c) Lower convex hull of observed latencies as a function of β_t for different values of bandwdith ($B = 30, 50, 80, 100$ kHz).

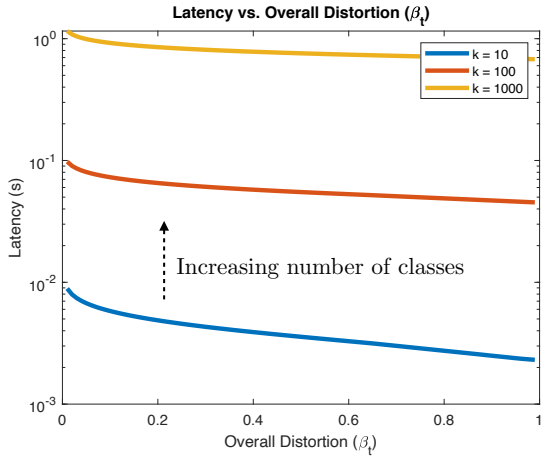


Fig. 3: Impact of the number of classes k (dimensionality of $\mathbf{p}$) on the latency-distortion tradeoff.

[4] S. Ma, et al. "A Theory for Semantic Communications." arXiv preprint arXiv:2303.05181 (2023).

[5] W. Yang et al., "Semantic Communications for Future Internet: Fundamentals, Applications, and Challenges," in IEEE Communications Surveys & Tutorials, vol. 25, no. 1, pp. 213-250, First quarter 2023.

[6] X. Luo, H. -H. Chen and Q. Guo, "Semantic Communications: Overview, Open Issues, and Future Research Directions," in IEEE Wireless Communications, vol. 29, no. 1, pp. 210-219, February 2022.

[7] G. Durisi, T. Koch and P. Popovski, "Toward Massive, Ultrareliable, and Low-Latency Wireless Communication With Short Packets," in Proceedings of the IEEE, vol. 104, no. 9, pp. 1711-1726, Sept. 2016.

[8] P. Popovski et al., "Wireless Access in Ultra-Reliable Low-Latency Communication (URLLC)," in IEEE Transactions on Communications, vol. 67, no. 8, pp. 5783-5801, Aug. 2019.

[9] M. Fresia, F. Peréz-Cruz, H. V. Poor and S. Verdú, "Joint Source and Channel Coding," in IEEE Signal Processing Magazine, vol. 27, no. 6, pp. 104-113, Nov. 2010.

[10] D. B. Kurka, and D. Gündüz. "Bandwidth-agile image transmission with deep joint source-channel coding.", in IEEE Transactions on Wireless Communications, vol. 20, no.12, pp. 8081-8095, Dec. 2021.

[11] N. Farsad, R. Milind, and A. Goldsmith. "Deep learning for joint source-channel coding of text.", 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2018.

[12] T. Tung, and D. Gündüz. "Deep joint source-channel and encryption coding: Secure semantic communications.", ICC 2023-IEEE International Conference on Communications. IEEE, 2023.

[13] H. Xie, Z. Qin, G. Y. Li and B. -H. Juang, "Deep Learning Enabled Semantic Communication Systems," in IEEE Transactions on Signal Processing, vol. 69, pp. 2663-2675, 2021.

[14] H. Xie, Z. Qin and G. Y. Li, "Semantic Communication With Memory," in IEEE Journal on Selected Areas in Communications, vol. 41, no. 8, pp. 2658-2669, Aug. 2023.

[15] Y. Xiao, X. Zhang, Y. Li, G. Shi and T. Başar, "Rate-Distortion Theory for Strategic Semantic Communication," 2022 IEEE Information Theory Workshop (ITW), Mumbai, India, pp. 279-284, 2022.

[16] J. Liu, W. Zhang and H. V. Poor, "A Rate-Distortion Framework for Characterizing Semantic Information," 2021 IEEE International Symposium on Information Theory (ISIT), Melbourne, Australia, pp. 2894-2899, 2021.

[17] S. Enayati and H. Pishro-Nik, "Quantization Rate and AoI-Induced Distortion Trade-off Analysis with Application to Remote Agents," 2022 IEEE Wireless Communications and Networking Conference (WCNC), Austin, TX, USA, pp. 782-787, 2022.

[18] Y. A. Reznik, "An Algorithm for Quantization of Discrete Probability Distributions," 2011 Data Compression Conference, Snowbird, UT, USA, pp. 333-342, 2011.

[19] R. Cabasag, S. Huq, E. Mendoza, & M. K. Roychowdhury, "Optimal quantization for discrete distributions", 2020. arXiv preprint arXiv:2008.03255.

[20] S. Matsuura and K. Hiroshi, "Statistical Estimation of Quantization for Probability Distributions: Best Equivariant Estimator of Principal Points." International Conference on Machine Learning, Optimization, and Data Science. Cham: Springer International Publishing, 2021.

[21] A. Adler, J. Tang and Y. Polyanskiy, "Quantization of Random Distributions under KL Divergence,", 2021 IEEE International Symposium on Information Theory (ISIT), Melbourne, Australia, pp. 2762-2767, 2021.

[22] A. Adler, J. Tang and Y. Polyanskiy, "Efficient Representation of Large-Alphabet Probability Distributions," in IEEE Journal on Selected Areas in Information Theory, vol. 3, no. 4, pp. 651-663, Dec. 2022.

[23] A. Rényi. "On measures of entropy and information." Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics. Vol. 4. University of California Press, 1961.

[24] Y. Polyanskiy, H. V. Poor and S. Verdu, "Channel Coding Rate in the Finite Blocklength Regime," in IEEE Transactions on Information Theory, vol. 56, no. 5, pp. 2307-2359, May 2010.