# Importance Matching Lemma for Lossy Compression with Side Information

**Buu Phan**[*]
University of Toronto

**Ashish Khisti**[*]
University of Toronto
Qualcomm AI Research[†]

**Christos Louizos**
Qualcomm AI Research[†]

## Abstract

We propose two extensions to existing importance sampling based methods for lossy compression. First, we introduce an importance sampling based compression scheme that is a variant of ordered random coding (Theis and Ahmed, 2022) and is amenable to direct evaluation of the achievable compression rate for a finite number of samples. Our second and major contribution is the *importance matching lemma*, which is a finite proposal counterpart of the recently introduced Poisson matching lemma (Li and Anantharam, 2021). By integrating with deep learning, we provide a new coding scheme for distributed lossy compression with side information at the decoder. We demonstrate the effectiveness of the proposed scheme through experiments involving synthetic Gaussian sources, distributed image compression with MNIST and vertical federated learning with CIFAR-10.

## 1 INTRODUCTION

Lossy compression has become increasingly important in the field of machine learning, fueled by the expanding scale of data (Deng et al., 2009), models (Ramesh et al., 2021; OpenAI, 2023), and infrastructure(Tang et al., 2018). Furthermore, with the growing demand for decentralized and distributed learning, compression techniques that operate in multi-terminal settings must be considered. We focus on a class of lossy compression techniques based on channel simulation, targeting one-shot or finite block length settings. The sender, upon observing $X \sim p_X(\cdot)$, communicates a noisy sample $Y \sim p_{Y|X}(.)$ to the decoder at a rate of $R$ bits/sample. In practice, the distribution $p_{Y|X}(.|x)$ is selected to satisfy a variety of constraints e.g., fidelity constraint or distribution constraints. In a recent work, Li and El Gamal (2018) propose a method based on the *Poisson functional representation lemma* (PFRL) with near optimal compression rate:

$$R \leq I(X;Y) + \log(I(X;Y)+1) + 5. \tag{1}$$

Here $I(X;Y)$, which denotes the mutual information between $X$ and $Y$ is well known to be a lower bound on the compression rate (Cuff (2013); Bennett et al. (2002)). However PFRL, requires an *infinite* number of samples to be generated between the encoder and decoder. More practical approaches for channel simulation have been developed using importance sampling (Chatterjee and Diaconis, 2018) for a variety of applications e.g., neural compression (Flamich et al., 2020; Theis et al., 2022), federated learning (Isik et al., 2023; Triastcyn et al., 2021), differential privacy (Shah et al., 2022), and model compression (Havasi et al., 2019). We will refer to these approaches as importance sampling based compression (ISC). In these methods, the output samples follow a proxy distribution $\tilde{p}_{Y|X}(\cdot)$ whose divergence w.r.t the target distribution $p_{Y|X}(\cdot)$ can be made arbitrarily small, provided that the number of samples is sufficiently large. Ordered random coding (ORC, Theis and Ahmed (2022)) is a recently proposed method in this family that also achieves near-optimal compression rate in (1). The analysis of compression rate in ORC is based on one-to-one comparison of the selected sample index with PFRL.

To our knowledge ISC methods till date have not considered distributed source coding (DSC) which enables higher compression rates by exploiting the correlations between multiple sources (El Gamal and Kim, 2011). We note that this is particularly relevant in many machine learning setups (Castiglia et al., 2022; Mital et al., 2022). While the information-theoretic limits of DSC have been well studied in classical settings,

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. 

[*] Equal Contribution.

[†] Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc. and/or its subsidiaries.

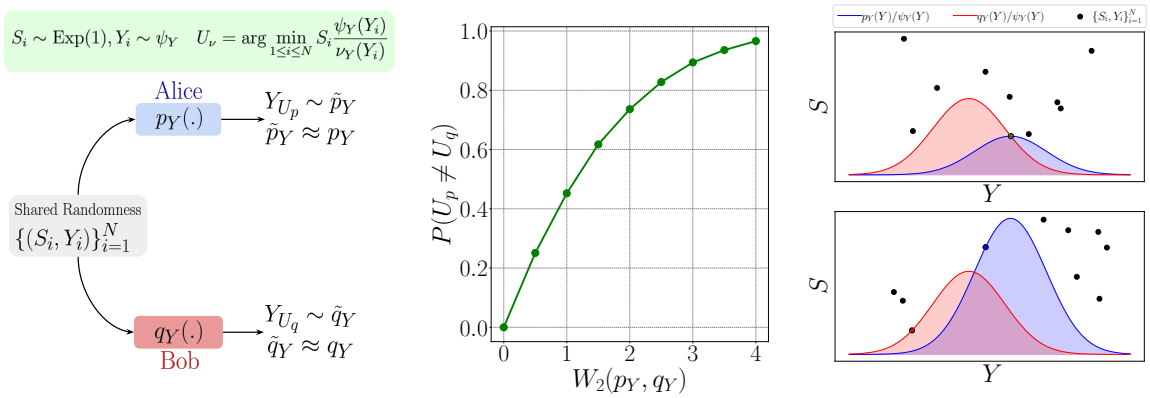



Figure 1: (Left) Overview of IML: Alice and Bob independently sample $Y_{U_p}, Y_{U_q}$ by applying the Gumbel-max trick on the shared randomness. (Middle) The empirical mismatch probability with respect to the Wasserstein-2 distance $W_2(p_Y, q_Y)$, where $p_Y = \mathcal{N}(m, 1)$, $q_Y = \mathcal{N}(-m, 1)$ and $m \in [0, \infty)$. (Right) Mechanism of IML: each party scales their respective importance weights function $\frac{p_i(y)}{\psi_i(y)}$ and $\frac{q_i(y)}{\psi_i(y)}$ until one point $\{S_i, Y_i\}$ falls on the curve. Top and bottom figures show the matching and mismatching case respectively.

practical implementations remain challenging. First, the joint source distributions are often unknown and need to be learned using data. Second quantization based approaches for DSC are challenging to implement in higher dimensions (Zamir and Shamai, 1998) and are mostly studied in one-dimensional case (Liu et al., 2006; Chen and Tuncel, 2010; Domanovitz et al., 2022). The extension of PFRL to DSC settings has been recently proposed in (Li and Anantharam, 2021) through the introduction of a new analysis technique called the Poisson matching lemma (PML). It however requires an infinite number of samples to be generated and can be challenging to implement in practice. In this work, as our main contribution, we introduce a new theoretical tool called the *importance matching lemma*, which enables us to extend ISC to DSC settings with provable guarantees. Our main contributions are:

- We propose an ISC scheme, which we call *communication efficient importance sampling based compression* (CE-ISC). Our scheme is amenable to direct evaluation of the achievable compression rate. We also discuss a potential extension to the case of multiple importance sampling.
- We introduce a new analysis tool called the *importance matching lemma* (IML), which is the counterpart of PML to importance sampling. This enables us to significantly expand the scope of ISC. We discuss in detail the application of ISC to DSC using IML.
- We conduct experimental studies on synthetic Gaussian sources, distributed compression involving the MNIST dataset, and a vertical federated learning setting with CIFAR-10 to demonstrate the effectiveness of our approach. We propose a data-driven approach to implement the decoding rule, and make use of a feedback link from the decoder to encoder to improve the rate-distortion performance.

A core technical challenge in our work involves analysis of self-normalized importance sampling, where bounds on standard quantities (e.g., bias and variance) are considerably challenging to characterize (Agapiou et al., 2017). In our work, we are required to perform novel analysis of such methods for characterizing the compression rate in ISC and error probabilities associated with IML.

**Related Work.**

*Channel Simulation.* Our work falls in the class of ISC schemes, which include ORC (Theis and Ahmed, 2022) and minimum random coding (MRC) (Havasi et al., 2019). Our proposed scheme is different from these techniques, is amenable to direct evaluation of the achievable rate and appears compatible with multiple importance sampling (Elvira et al., 2019). While ISC schemes are approximate sampling techniques with an upper bound on the total number of proposal samples, other methods such as PFRL based sampling (Li and El Gamal, 2018), A* sampling (Maddison et al., 2014; Flamich et al., 2022) and rejection sampling (Harsha et al., 2007; Flamich and Theis, 2023) are exact sampling techniques and may require arbitrarily large number of proposal samples in the worst case.

*Distributed Source Coding (DSC).* To our knowledge the only channel simulation technique that extends to DSC is from Li and Anantharam (2021) discussed previously. Deep learning based DSC has been considered in some recent works (Mital et al. (2022); Whang et al. (2021); Ozyilkan et al. (2023)), which provide empirical evidence that neural networks could learn the binning. In contrast, our approach is motivated by

theoretical analysis of IML and integrates deep learning to incorporate complex joint distributions. Traditional information theoretic approaches for one-shot DSC (Verdú, 2012; Liu et al., 2015; Song et al., 2016) do not appear amenable to practical implementations. Finally quantization based approaches for DSC have several limitations as are already discussed before.

# 2 COMMUNICATION-EFFICIENT IMPORTANCE SAMPLING

We introduce our communication efficient ISC scheme, provide analysis of the achievable compression rate and discuss an extension to multiple importance sampling.

## 2.1 Problem Setup

We begin by introducing the setup of approximate channel simulation under communication constraints. Let $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ be a pair of random variables distributed according to $p_{X,Y}$, with marginal distributions $p_X(\cdot)$ and $p_Y(\cdot)$, respectively. The encoder observes a realization $x$ of $X \sim p_X$. Additionally, there is a shared source of randomness denoted by $W \in \mathcal{W}$, which is available to both the encoder and the decoder. We define the mappings $f$ and $g$ for the encoder and decoder respectively:

$$f\colon \mathcal{X} \times \mathcal{W} \to \mathcal{M}, \tag{2}$$

$$g\colon \mathcal{M} \times \mathcal{W} \to \mathcal{Y}, \tag{3}$$

where $\mathcal{M} \in \{0,1\}^\star$ denotes the (variable length) message that the encoder sends to the decoder and $\ell(M)$ denotes the length of message $M$. A rate-divergence tuple $(R, \tilde{\epsilon})$ is defined to be one-shot achievable if there exists $f, g$ satisfying:

$$E[\ell(M)] \leq R, \tag{4}$$

$$D_{\mathrm{TV}}(\tilde{p}_{Y|X}(.|x), p_{Y|X}(.|x)) \leq \tilde{\epsilon}, \forall x \in \mathcal{X} \tag{5}$$

where $\tilde{p}_{Y|X}(.|x)$ denotes the proxy distribution realized by our choice of $f$ and $g$, while $p_{Y|X}(.|x)$ is the desired target distribution.

## 2.2 Coding Scheme

The encoder and decoder generate shared randomness $W = \{S_i, Y_i\}_{i=1}^N$ where $Y_i \sim p_Y(\cdot)$ and $S_i \sim \mathrm{Exp}(1)$ are sampled in an i.i.d. fashion. The encoder, upon observing $X = x$ must select an index $i \in \{1, ..., N\}$ and communicate it such that both (4) and (5) are satisfied. Specifically, the encoding function is described in the following steps:

1. *Index Selection.* Upon observing $X{=}x$, the encoder selects index $U$ by:

$$U = \arg\min_{1 \leq i \leq N} S_i \frac{p_Y(Y_i)}{p_{Y|X}(Y_i|x)} \tag{6}$$

which is also known as Gumbel-max trick (Maddison et al., 2014).
2. *Index Reordering.* Instead of directly sending $U$ to the decoder, which will cost $\log(N)$ bits, the encoder will send its corresponding position $K$ in the sorted list: $S_{\pi(1)} \leq S_{\pi(2)} \leq \ldots \leq S_{\pi(N)}$, such that $\pi(K){=}U$.
3. *Entropy Coding.* Finally, the encoder convert $K$ into a bit string $M$ by entropy coding with a Zipf distribution (Li and El Gamal, 2018, Section 2).

The decoding step $g$ is straightforward. As the sorted list is both known to the encoder and decoder, the decoder can recover $K$ (and by so, $U$) losslessly from $M$, and outputs $Y_U$.

Note that our index selection in step 1 is the same as in importance sampling (Havasi et al., 2019). In particular, given a common sequence of samples $Y_1, \ldots, Y_N$, our approach selects the sample $Y_i$ with probability:

$$\lambda_i = \left\{ \frac{p_{Y|X}(Y_i|x)}{p_Y(Y_i)} \right\} \Big/ \left\{ \sum_{i=1}^{N} \frac{p_{Y|X}(Y_i|x)}{p_Y(Y_i)} \right\}. \tag{7}$$

Conditioned on the common sequence $Y_1^N$, our setting reduces to the exponential functional representation lemma for discrete alphabets in (Li, 2017, Def. 4.1). Following their approach, we transmit the position of the selected index in the sorted list of $\{S_i\}$ in steps 2 & 3. Our construction is different from ORC (Theis and Ahmed, 2022, Sec. 3.4) that reorders (sorts) the exponential random variables before the index selection operation in (6). Interestingly both schemes lead to the same output distribution when the samples are generated in an i.i.d. fashion.

Note that the proxy distribution generated by our scheme is:

$$\tilde{p}_{Y|X}(y|x) = E_{Y_1,\ldots,Y_N}\left[ \sum_{i=1}^{N} \lambda_i \cdot \delta(y - Y_i) \right] \tag{8}$$

where $\lambda_i$ are defined in (7). Following prior works (Havasi et al., 2019; Theis and Ahmed, 2022; Chatterjee and Diaconis, 2018) the output distribution can be close to the target distribution $p_{Y|X}(\cdot)$ for sufficiently many samples. In particular, under the standard assumption:

$$\frac{p_{Y|X}(y|x)}{p_Y(y)} \leq \omega, \forall x, y, \tag{9}$$

we can construct a $N_0(\epsilon)$ such that for $N \geq N_0(\epsilon)$, we have that $D_{\mathrm{TV}}(\tilde{p}_{Y|X}(.|x), p_{Y|X}(.|x)) \leq \epsilon$. A characterization of $N_0(\epsilon)$ is provided in the Section 7 of the supplementary material for sake of completeness. We next provide analysis of the achievable rate.

**Theorem 1.** *Given $(X, Y) \sim p_{X,Y}$, and $N, K$ as in the scheme in Sec. 2.2, we have that:*

$$E[\log K | X = x] \leq E_{Y_1^N}\left[ D(\lambda || \mathbf{u}) \right] + \delta \tag{10}$$

*where $\lambda = (\lambda_1, \ldots, \lambda_N)$ is defined via (7), $D(.||.)$ is the KL divergence, $\mathbf{u} = (1/N, \ldots, 1/N)$ is associated with the uniform distribution and $\delta = 1 + \log e/e$ is a constant. Furthermore,*

$$H[K] \leq I(X;Y) + \frac{\Delta}{N} + \log\left(I(X;Y) + \frac{\Delta}{N} + 1\right) + 4 \quad (11)$$

*where $\Delta := \Delta(p_{X,Y})$ is a constant defined in the supplementary material via (31) and (32) that does not depend on $N$.* □

We provide the proof of Theorem 1 in the Section 8 in the supplementary material. While the upper bound in (10) follows through connection with (Li, 2017, Chapter 4), the derivation of (11) is complicated by the normalizing term in the denominator in (7). Theorem 1 demonstrates that the proposed scheme also achieves a near optimal compression rate and an additive penalty of at most $\Theta(1/N)$. We also provide an alternative bound to (11) in the supplementary material in section 9, which is simpler to derive but involves a multiplicative penalty term.

### 2.3 Beyond i.i.d. samples

Our discussion so far has assumed that the proposal samples $\{Y_i\}$ are generated in an i.i.d. fashion from a distribution $p_Y(\cdot)$. However in variance reduction schemes, such as multiple importance sampling (Elvira et al., 2019), it is required that different samples be generated from different distributions. As a simple example, suppose that $Y_1, \ldots, Y_{\bar{N}}$ are sampled i.i.d. from $p_Y^{(1)}(\cdot)$ and $Y_{\bar{N}+1}, \ldots, Y_N$ are sampled i.i.d. from $p_Y^{(2)}(\cdot)$ where $\bar{N} = N/2$ and $p_Y^{(1)}(\cdot)$ and $p_Y^{(2)}(\cdot)$ are selected to satisfy $p_Y(y) = \frac{1}{2}p_Y^{(1)}(y) + \frac{1}{2}p_Y^{(2)}(y)$. Given $X = x$, the probability that the output index $K = i$ in Multiple Importance Sampling (MIS) is proportional to $\lambda_i$ (see scheme N3 in Elvira et al. (2019)) in (7). As a result our proposed coding scheme in Section 2.2 can be immediately used as stated. In the rate analysis, note that the upper bound in (10) in Theorem 1 also applies with the difference that $Y_1^N$ are not i.i.d. but distributed according to either $p_Y^{(1)}(\cdot)$ or $p_Y^{(2)}(\cdot)$. We discuss further analysis of this specific case in Section 10 in the supplementary material. We show that under a simplifying assumption that the denominator in (7) equals its expectation, the distribution of the output samples equals the target distribution and the associated compression rate matches ORC. We also study a numerical example involving a Gaussian mixture model and demonstrate that MIS with our proposed compression scheme can achieve significantly lower bias and variance than the ORC scheme for a given number of samples.

## 3 IMPORTANCE MATCHING LEMMAS

### 3.1 Importance Matching Lemma

The Poisson Matching Lemma (PML) (Li and Anantharam, 2021) enables the application of Poisson Functional Representation lemma (PFRL) to a broad class of problems in multi-terminal source and channel coding with provable guarantees. In this section, we introduce the Importance Matching Lemma (IML), which enables application of importance sampling to such settings. We demonstrate the application of IML to a specific setting of source coding with side information in the next section.

Let $Y_1, \ldots, Y_N$ be sampled i.i.d. from distribution $p_Y(\cdot)$ and let $p_{Y|X}(\cdot|X = x)$ and $q_{Y|X}(\cdot|X = x)$ be two conditional distributions. We note that such $p_Y(\cdot)$ can be replaced by any distribution over $\mathcal{Y}$ such that (9) is satisfied as our proof does not require any relation between $p_Y(\cdot)$ and $p_{Y|X}(\cdot)$ to hold. We generate two indices $U_p$ and $U_q$ as follows:

$$U_p = \arg\min_{1 \leq i \leq N} \frac{S_i}{\lambda_i^p}, \qquad U_q = \arg\min_{1 \leq i \leq N} \frac{S_i}{\lambda_i^q} \quad (12)$$

Where $S_1, \ldots, S_N$ are sampled i.i.d. Exp(1) and $\lambda_i^p$ and $\lambda_i^q$ are the importance weight counterparts of (7):

$$\lambda_i^p = \frac{\frac{p_{Y|X}(Y_i|X=x)}{p_Y(Y_i)}}{\sum_{j=1}^N \frac{p_{Y|X}(Y_j|X=x)}{p_Y(Y_j)}}, \lambda_i^q = \frac{\frac{q_{Y|X}(Y_i|X=x)}{p_Y(Y_i)}}{\sum_{j=1}^N \frac{q_{Y|X}(Y_j|X=x)}{p_Y(Y_j)}}. \quad (13)$$

The indices $U_p$ and $U_q$ selected via importance sampling have the same proposal distribution $p_Y(\cdot)$ but different target distributions $p_{Y|X}$ and $q_{Y|X}$ respectively. We bound the error event that $\{U_p \neq U_q\}$.

**Proposition 1.** *Letting $\Omega = \{y_1, \ldots, y_N\}$ denote the sequence of samples:*

$$\Pr(U_p \neq U_q | \Omega, X = x, U_p = k) \leq$$
$$1 - \left(1 + \frac{p_{Y|X}(y_k|x)}{q_{Y|X}(y_k|x)} \frac{\left(\frac{1}{N}\sum_{j=1}^N \frac{q_{Y|X}(y_j|x)}{p_Y(y_j)}\right)}{\left(\frac{1}{N}\sum_{j=1}^N \frac{p_{Y|X}(y_j|x)}{p_Y(y_j)}\right)}\right)^{-1}. \quad (14)$$

The proof of Prop. 1 is Section 11 in the supplementary material. It exploits the fact that conditioned on the samples $Y_1^N$, the operation in (12) can be viewed as sampling over a discrete alphabet with probabilities given by $\lambda_i^p$ and $\lambda_i^q$ respectively and arguments as in (Li and Anantharam, 2021) are applicable. Our main result in this section is the following where the conditioning on all $Y_i$, except $Y_k$ is removed.

**Theorem 2.** *Define $\bar{N} = N - 1$, we have:*

$$\Pr(U_p \neq U_q | Y_k = y_k, U_p = k, X = x) \leq 1 - \left(1 + \frac{p_{Y|X}(y_k|x)}{q_{Y|X}(y_k|x)} \mu_{y_k}(\bar{N})\right)^{-1} \quad (15)$$

*and $\mu_{y_k}(\bar{N})$ is defined via (79)-(81) in Section 12 in the supplementary material. Note that $\mu_{y_k}(\bar{N})$ scales as $\Theta(1)$ as $\bar{N} \to \infty$ under some mild assumptions on the distributions (see Remark 12 in Supplementary Material).* □

The proof of Theorem 2 is in Section 12 in the supplementary material. The main challenge is associated with the normalizing terms in (13). We also provide an alternate bound in Section 13 in the supplementary material that has a shorter derivation, but requires stronger assumptions on the distribution.

Intuitively, IML bounds the mismatch between the sampled indices when different target distributions are used in importance sampling. This is further illustrated in Fig. 1. As is the case with PML, it turns out that in practice we need a conditional version of IML.

### 3.2 Conditional Importance Matching Lemma.

Suppose that $(S_i, Y_i)_{i=1}^N$ be sampled i.i.d. as in Section 3.1. Let $(X, Y, Z)$ be a triplet of random variables with a joint distribution $p_{X,Y,Z}(\cdot)$. Let $Q_{Y|Z}(\cdot)$ be an arbitrary conditional distribution satisfying $\frac{Q_{Y|Z}(y|z)}{P_Y(y)} \leq \omega$ for all $(y, z)$. Given $X=x$ sampled independently of $(S_i, Y_i)_{i=1}^N$, suppose that we sample $Y = Y_{U_P}$ using importance sampling i.e., we select

$$U_P = \arg\min_{1 \leq i \leq N} \frac{S_i}{\lambda_i^P}, \quad \lambda_i^P = \frac{\frac{p_{Y|X}(Y_i|X=x)}{p_Y(Y_i)}}{\sum_{j=1}^N \frac{p_{Y|X}(Y_j|X=x)}{p_Y(Y_j)}}. \quad (16)$$

Next, given $X = x$ and $Y = y$ we generate a sample $Z \sim p_{Z|X,Y}(\cdot|X = x, Y = y)$ and note that

$$Z \to (X, Y) \to (S_i, Y_i)_{i=1}^N \quad (17)$$

is satisfied by construction. Given a realization $Z = z$ we sample

$$U_Q = \arg\min_{1 \leq i \leq N} \frac{S_i}{\lambda_i^Q}, \ \lambda_i^Q = \frac{\frac{Q_{Y|Z}(Y_i|Z=z)}{p_Y(Y_i)}}{\sum_{j=1}^N \frac{Q_{Y|Z}(Y_j|Z=z)}{p_Y(Y_j)}}. \quad (18)$$

**Theorem 3.** *The error probability satisfies:*

$$\Pr\left(U_P \neq U_Q | U_P = k, X = x, Z = z, \Omega\right) \leq 1 - \left(1 + \frac{p_{Y|X}(y_k|x)}{Q_{Y|Z}(y_k|z)} \frac{\left(\frac{1}{N}\sum_{j=1}^N \frac{Q_{Y|Z}(y_j|z)}{p_Y(y_j)}\right)}{\left(\frac{1}{N}\sum_{j=1}^N \frac{p_{Y|X}(y_j|x)}{p_Y(y_j)}\right)}\right)^{-1}, \quad (19)$$

*where $\Omega = \{y_1, \ldots, y_N\}$, and furthermore,*

$$\Pr(U_p \neq U_q | Y_k = y_k, U_p = k, X = x, Z = z) \leq 1 - \left(1 + \mu_{y_k}(\bar{N}) \frac{p_{Y|X}(y_k|x)}{Q_{Y|Z}(y_k|z)}\right)^{-1}. \quad (20)$$

*where $\mu_{y_k}(\bar{N})$ defined via (109)-(111) in the supplementary material, scales as $\Theta(1)$ when $\bar{N} \to \infty$ under some mild assumptions (c.f. Remark 14).* □

The proof of Thm. 3 is in Section 14 in the supplementary material. As discussed in Remark 14 there, under mild conditions on the distributions, we can exhibit an $N_1(\epsilon)$ such that

$$\mu_y(N) \leq 1 + \epsilon, \quad \forall N \geq N_1(\epsilon) \quad (21)$$

In the special case $Z = X$, (19) reduces to (14) in Prop. 1. Likewise (20) reduces to (15) in Theorem 2. The value of Theorem 3 is that it extends IML to any $Z$ that satisfies (17). In other words, the conditional version of IML is an extension of Theorem 2 where the decoder is revealed an observation $Z$ regarding the sample $Y=Y_{U_P}$ selected by the encoder, and improves the decoding rule (cf. (18)) by making use of the posterior distribution $Q_{Y|Z}(\cdot)$. We demonstrate an application of this result in the next section.

## 4 LOSSY COMPRESSION WITH SIDE-INFORMATION

### 4.1 Problem Setup

Just as the Poisson Matching Lemma (PML) has broad applications in multi-terminal source and channel coding settings, IML developed in Section 3.1 can have analogous applications using importance sampling. We demonstrate the application of IML to the classical problem of lossy compression with side information at the decoder (Wyner and Ziv, 1976). As illustrated in Figure 2(left), a source sample $V \sim p_V(\cdot)$ observed at the encoder, must be lossily compressed into a bit sequence of average rate $R$ bits/sample and sent to the decoder. Besides the information from the encoder, the decoder also has access to the side information $T \sim p_{T|V}(.|V = v)$, both of which are employed to output $W$ that is required to be approximately sampled from a conditional distribution $p_{W|V}(\cdot|v)$. In practice this conditional distribution can be selected to satisfy an average distortion constraint $E[d(V, \hat{V})]$, where $\hat{V}$ is the final reconstruction expressed as a function of $W$ and $T$, i.e. $\hat{V} = \tilde{g}(W, T)$.

We present a scheme based on importance sampling and present the error analysis by making use of IML in the previous section.

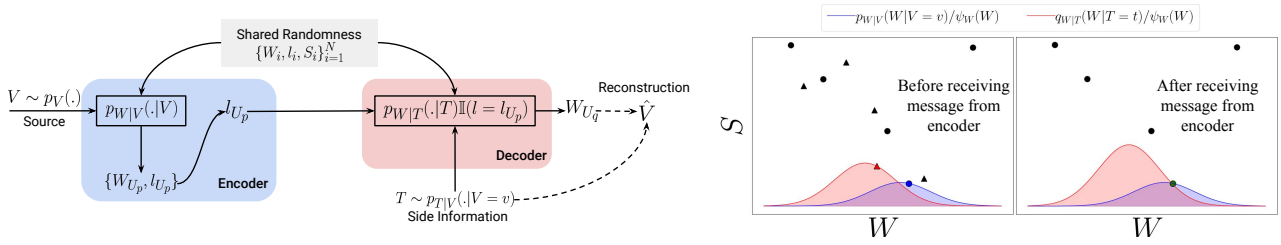



Figure 2: (Left) Source coding with side information at the decoder with conditional IML. (Right) Decoding mechanism: the encoder scales $\frac{p_{W|V}(w|v)}{\psi_W(w)}$ and selects $W_{U_p}$(blue circle). Left sub-figure: the decoder selects incorrect indices by purely scaling $\frac{p_{W|T}(w|t)}{\psi_W(w)}$ (equivalently, rate $R$=0). Right sub-figure: we generate extra one-bit information $l_i$ for each codeword index by randomly marking it either a triangle ($l_i$=0) or a circle ($l_i$=1). Upon receiving $l_{U_P}$=1 from the encoder, the decoder eliminates all indices marked by triangle and correctly decode the index among the circles.

### 4.2 Coding Scheme

Let $p_W(\cdot)$ be the marginal distribution of $W$ and $p_{W|T}(w|t) = \sum_v p_{W|V}(w|v)p_{V|T}(v|t)$ be the conditional of $W$ given $T$. Following the construction in Section 2.2, we sample $W_1, \ldots, W_N$ i.i.d. from the distribution $p_W(\cdot)$. In addition, let $L > 0$ be an integer and let $p_{\mathsf{l}}(\cdot)$ be uniform over the set of integers $\{1, 2, \ldots, L\}$. We generate $l_1, \ldots, l_N$ i.i.d. from $p_{\mathsf{l}}(\cdot)$. Let us define $Y = (W, l)$ with $p_Y(y) = p_W(w)p_{\mathsf{l}}(l)$ and note that $Y_i = (W_i, l_i)$ is sampled i.i.d. from $p_Y(\cdot)$. Further let $X = V$ and $p_{Y|X}(w, l|v) = p_{W|V}(w|v)p_{\mathsf{l}}(l)$ be the target distribution used at the encoder with the knowledge of $v$. Finally we let $S_1, \ldots, S_N$ be a sequence of i.i.d. exponential random variables Exp(1) known to both the encoder and the decoder. Following (12), the encoder selects an index $U_p$ given by:

$$U_p = \arg\min_{1\le i\le N} \frac{S_i}{\frac{p_{Y|X}(Y_i|v)}{p_Y(Y_i)}} = \arg\min_{1\le i\le N} \frac{S_i}{\frac{p_{W|V}(W_i|v)}{p_W(W_i)}}. \tag{22}$$

The encoder, in turn, transmits $l_{U_p} \in \{1, 2, \ldots, L\}$ to the decoder using $\log L$ bits. In defining the decoding rule, we let $Z = (T, l_{U_p})$. and note that $Z \to (X, Y_{U_p}) \to (S_i, Y_i)_{i=1}^N$ is satisfied. Let

$$Q_{Y|Z}(y|z) = Q_{(W,l)|(T,l_{U_p})}(w, l|t, l_{U_p}) \tag{23}$$

$$= p_{W|T}(w|t)\mathbb{I}(l = l_{U_p}) \tag{24}$$

be the distribution used at the decoder. Following (16), the decoder outputs an index $U_q$ given by:

$$U_q = \arg\min_{1\le i\le N} \frac{S_i}{\frac{Q_{Y|Z}(Y_i|t,l_{U_p})}{p_Y(Y_i)}} = \arg\min_{1\le i\le N} \frac{S_i}{\frac{p_{W|T}(W_i|t)\mathbb{I}(l_i=l_{U_p})}{p_W(W_i)p_{\mathsf{l}}(l_i)}} \tag{25}$$

The decoder finally outputs $\hat{W}$=$W_{U_q}$ as the sample.

Since the encoder selects the sample $W_{U_p}$ using importance sampling (22), it follows from the discussion in Section 2.2 that for sufficiently large $N$ the distribution $p_{W_{U_p}|V}(\cdot)$ can be arbitrarily close to the target distribution $p_{W|V}(\cdot)$. The error probability can be bounded as below:

**Proposition 2.** *For sufficiently large $N$,*

$$\Pr(U_p \neq U_q) \le E_{V,W,T}\left[1 - \left(1 + (1+\epsilon)L^{-1}2^{i(W;V|T)}\right)^{-1}\right] \tag{26}$$

*where $i_{W,V|T}(w; v|t) = \log \frac{p_{W|V}(w|v)}{p_{W|T}(w|t)}$ is the conditional information density and recall $T \to V \to W$.* □

The proof of Prop. 2 is in Section 15 in the supplementary material. Note that (26) provides a tradeoff between the compression rate $R$= $\log L$ and the error probability. The larger the value of $R$, the smaller will be the error probability. We illustrate how this decoding mechanism reduces the error in Figure 2 (right) for the case $R$=0 and $R$=1. Similar to results in Li and Anantharam (2021), we can extend this result to bound the excess distortion probability $P_e$= $\Pr(d(V, \hat{V})>D)$ where $\hat{V}$=$\tilde{g}(W_{U_q}, T)$ is the reconstruction output by the decoder and $d(.,.)$ is the distortion (Section 16 in the Supplementary).

**Remark 1.** *Note that the error probability directly depends on the compression rate $R = \log L$ and the conditional information density $i(W; V|T)$. Provided that $N$ is sufficiently large (which is necessary to track the target distribution in ISC) it only exhibits a second-order dependence on $N$ via the $\epsilon$ factor.*

**Remark 2.** *In some of our experiments, we will consider compressing $k$ i.i.d. samples together. The error probability can be recovered by setting $L = 2^{k\cdot R}$ and $i(W^k; V^k|T^k) = \sum_{i=1}^k i(W_i; V_i|T_i)$.*

### 4.3 Decision Feedback Based Scheme

In practice, a decoding error from (26) can result in high average reconstruction distortion. This motivates

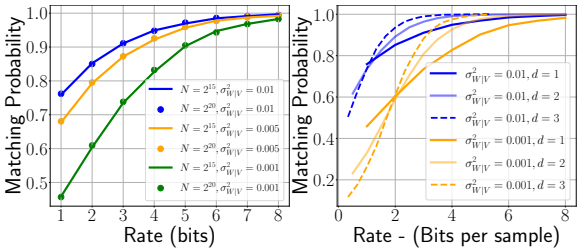
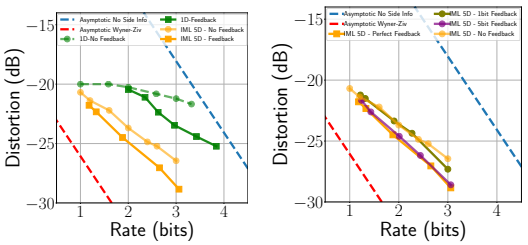



Figure 3: Left: Empirical matching probabilities with different target distribution and number of proposals. Right: effects of compressing multiple samples jointly on the matching probability (Best view in screen).




(a) Rate-Distortion Performance. (b) Comparison of Different Feedback Rates.

Figure 4: Analysis and rate-distortion performance of different IML schemes. (Best view in screen)

us to use feedback communication to correct the errors and improve the rate-distortion performance as follows:

1. *Index Selection.* The encoder communicates $\log_2(L)$ least significant bits (LSB) of the selected index $U_p$, see (22), to the decoder.
2. *Decoding and Feedback.* The decoder outputs $U_q$ using (25) and send $\log_2(N/L)$ most significant bits (MSB) to the encoder.
3. *Re-tranmission.* From the feedback MSB, if the index is correct, the encoder responds with an acknowledgment bit. Otherwise, it sends the MSB of its selection to the decoder.

We verified in our experiments that the use of LSB instead of random bits in step 1 did not have any noticeable difference. While the feedback rate of $\log_2(N/L)$ bits guarantees that the encoder can perfectly locate the decoded index, we experimentally observe that we can reduce it by sending a hashed value of MSB, while tolerating a slight increase in distortion. Note that this small hash of the decoded index requires significantly fewer bits than full side information (e.g., by $> 500$ times in Sec 5.2).

Our rate-distortion analysis uses the total length of the messages in both index selection and re-transmission (including any acknowledgement messages) for computing the rate. We do not include the rate of the feedback message, however. This can be justified if there is an asymmetric cost in communication in the forward and reverse directions, e.g., wireless channels. For details about the rate distortion analysis of this scheme, see Section 17 in the supplementary material.

# 5 EXPERIMENTS

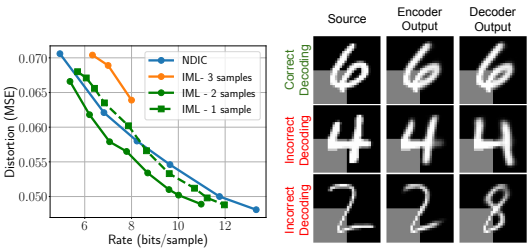
We experimentally study ISC schemes for the setup in Section 4 with different datasets as discussed below. For real-world datasets, we note that the decoding step requires computing $\log \frac{p_W(W_i)}{p_{W|T}(W_i|t)}$, which can be learned by training a neural estimator (Hermans et al., 2020). Further details are available in Section 24 in the supplementary.

## 5.1 Synthetic Gaussian Source

For the setup in Section 4.1 we assume that the source $V \sim \mathcal{N}(0, \sigma_V^2{=}1.0)$ and the side information $T = V + \zeta$ where $\zeta \sim \mathcal{N}(0, 0.01)$, i.e $p_{T|V}(.|v){=}\mathcal{N}(v, \sigma^2_{T|V} = 0.01)$. Furthermore, the encoder and decoder have access to the shared randomness $(S_i, Y_i, \ell_i)_i^N$ as described previously. The decoder must ideally output $W \sim p_{W|V}$, where $p_{W|V}(\cdot|v){=}\mathcal{N}(v, \sigma^2_{W|V})$. The encoder follows (22) to select the index $U_P$ and transmit $l_{U_p}$ to the decoder, using $\log_2(L)$ bits. Upon receiving $l_{U_p}$, the decoder selects its index following (25) and outputs $\hat{W}{=}W_{U_q}$. Finally, note that in this scenario, a closed-form solution for $p_{W|T}$ exists, and a refined reconstruction $\hat{V}{=}\tilde{g}(\hat{W}, T)$ can be generated with inverse variance weighting (see Section 24 in the supplementary).

Figure 3 illustrates the empirical matching probability: $p_m = \Pr(U_p = U_q)$ in our setup. In the left figure, we plot $p_m$ as a function of rate for different choices of $\sigma^2_{W|V}$ and different number of proposal candidates $N$. We note that provided $N$ is sufficiently large (which is required for guaranteeing $\tilde{p}_{W|V} \approx p_{W|V}$ in all ISC schemes), it has a negligible effect on $p_m$ as noted in Remark 12. We also note that consistent with the theoretical analysis (1) increasing the rate for a fixed $\sigma^2_{W|V}$ increases $p_m$ and (2) increasing $\sigma^2_{W|V}$ with a fixed rate increases $p_m$. Finally in the right sub-figure in Fig. 3, we also demonstrate the behaviour of $p_m$ when compressing multiple (say $k$) independent source samples. In the regime where $p_m$ (and correspondingly $R$) is large, which is of practical interest, we observe that compressing $k > 1$ independent samples improves $p_m$. This effect is also reflected in our theoretical analysis in Remark 2. In particular if we approximate $\sum_{i=1}^k i(W_i; V_i|T_i)$ by the expectation, $kI(W; V|T)$, then the error probability is decreasing in $k$ provided $R > I(W; V|T)$.

Figure 4a presents the rate-distortion (RD) trade-off with and without feedback when compressing either a single sample or 5 samples together (which we will refer to as 1D and 5D respectively). First, note that there is




(a) Rate-Distortion Performance. (b) Decoded Examples (No Feedback).

Figure 5: Distributed Image Compression with MNIST. In (a), the orange curve is IML without feedback and the performance is restricted to 8 bits due to limited computing resources. In (b), the gray area denotes the side-information sent to the decoder, which is 0 at the encoder side.

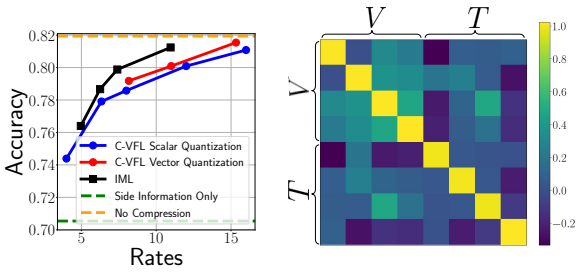



Figure 6: VFL with CIFAR-10. Left: Rate-Accuracy performance. Right: Covariance matrix between the source and side information's embeddings ($V$ and $T$ respectively).

a substantial improvement in the RD trade-off in the 5D case. It is worth mentioning that each operating point in the figure is such that compressing multiple samples is beneficial (see Fig. 3). Secondly, when considering the 1D Gaussian case, feedback demonstrates a significant enhancement in performance, albeit at the cost of an additional acknowledgment bit required in the re-transmission step. The overhead of the acknowledgment bit gets amortized over 5 samples in the 5D case, resulting in the overhead of 0.2 bits/sample. As such when comparing the 1D and 5D compression schemes with feedback, the consistent improvement in the latter can be attributed to both the reduced overhead in the acknowledgment bit, as well as improved matching probability. Note that the feedback communication here is perfect as described previously in Section 4.3.

We show the effects of different feedback rates on the RD tradeoffs in Figure 4b for the 5D case. Note that 1 and 5 bits are not sufficient to recover the index since we use $N{=}2^{27}$ and the maximum value of $L$ is $2^{15}$. Nevertheless, they can correct most of the decoding error as the result shows. With 1 bit of feedback, we can offset the penalty of the overhead in retransmission and with 5 bit feedback, the performance remains close to that achieved with full feedback (at least 12 bits). This demonstrates that limited number of feedback bits may also be sufficient in practice.

### 5.2 Distributed Image Compression

We validate the efficacy of our method through a distributed image compression setting (Whang et al., 2021; Mital et al., 2022). We consider the MNIST dataset where the side information is the cropped bottom-left quadrant of the image, see Figure 5b, while the source image to be reconstructed is the remaining.

Directly sending the noisy source image as in the Gaussian case will incur high complexity. Instead, we rely on the learned compression approach, where a $\beta$-VAE (Higgins et al., 2016) is trained to first project the source image to the embedding vector of size 4. This vector, together with the side information, is then fed into a neural network to reconstruct the source image, where the whole process is trained end-to-end. Once the $\beta$-VAE converges, we train a neural estimator (details in Section 23 in the supplementary) for the decoding process at test time, where the proposal distribution of the 4D embedding is the prior distribution used in $\beta$-VAE. At test time, we deploy the feedback process described in Section 4.3. Note that the variance of the target distribution is dependent on the input and we vary the $\beta$ factor to obtain different rate-distortion tradeoffs. Also, we use a fixed transmission size of 5 bits per feedback, irrespective of the number of samples to be compressed.

We compare our approach with NDIC (Mital et al., 2022), a method that enhances compression rates in this task by modeling the common information where we also set its bottleneck dimension to 4. Figure 5a shows that our approach achieves comparable performance with a single sample (4D vector) while consistently outperforming NDIC when jointly compressing two samples (8D vector), as in the Gaussian case. It is also worth highlighting that due to the high dimensionality of the side information ($14 \times 14$), employing classical 1D binning is not straightforward. Unlike NDIC which heavily relies on the learning capability of neural networks for encoding and decoding, our scheme explicitly exploits the statistical correlation between the source and side information and gives stronger RD performance. Figure 5b shows examples of when the decoder selects the correct/incorrect index, showing that the neural estimator selects messages that are semantically related to the side information. Finally, we provide additional experiment with different feedback rates in Section 24 in the supplementary material.

### 5.3 Vertical Federated Learning

#### 5.3.1 CIFAR-10 Dataset.

We demonstrate the applicability of our method to the compressed-vertical federated learning setting proposed by Castiglia et al. (2022), whose work we will refer to as C-VFL. Specifically, given a pre-trained model

(e.g. the neural networks at the server and each party), we want to exploit the statistical correlation between the features or learned embeddings to efficiently compress and therefore save communication costs from each party to the server during inference, while also minimizing the accuracy drop. We adapt the setup and network architecture in C-VFL for CIFAR-10 to the two-party scenario, where each party is assigned a non-overlapping quadrant of an input image. Each party then transforms their given quadrant into an embedding of dimension 4. Assuming perfect transmission, one party's embedding is compressed losslessly (32 bits per dimension) and treated as side information, while the other party's embedding is compressed in a lossy manner. Note that we exclude the possibility of splitting images into half since the side information alone in this case achieves near-optimal accuracy, rendering the source information unnecessary, and vice versa. This lets us evaluate the effectiveness of different compression methods when one party's information alone is insufficient for optimal accuracy.

Following our CE-IS approach, we compress the embedding by communicating its noisy version. Here, each of the 4 dimensions is perturbed with independent Gaussian noise with zero mean and the same variance[1], whose value is varied to obtain different rate-accuracy tradeoffs. To further the efficiency, similar to the MNIST experiment, we exploit the correlation from the side information by training a neural estimator (see Supplementary, Section 23) and use the feedback scheme to communicate the perturbed embedding. Note that we employ 4-bit feedback, which is sufficient for locating the index in this experiment.

We compare our method with scalar and vector quantization baselines, proposed by Castiglia et al. (2022). In scalar quantization, each dimension of the 4D embeddings is discretized, while in vector quantization, a 2D lattice is constructed for every 2 dimensions. In Figure 6 (left), we observe that our method outperforms the baselines, achieving near-optimal accuracy of 81.24% with ∼11 bits, while both vector and scalar quantization requires up to ∼15 bits for similar accuracy. This improvement can be attributed to the utilization of the correlation structure between the source and side information, as depicted in the covariance matrix in the right figure. These results further support the effectiveness of our method, even when the objective of the task considered is not related to source reconstruction.

### 5.3.2 UCI Breast Cancer Dataset.

We compare our IML method with scalar quantization on the Breast Cancer dataset (Wolberg and Street, 1995). For this task, we use the following features at the sender: "mean texture", "mean area", "mean smoothness", "mean concavity", and the following as side information, "mean symmetry", "mean fractal dimension", "texture error", "area error". Here, we directly compress the features at the sender and the combined input features (size of 8) will be fed into a neural network consisting of 2 hidden layers of size 8 (with ReLU activation) and an output layer of size 1. The results are shown in Figure 7, which shows that IML consistently outperforms scalar quantization. On the other hand, our scheme with the current hyperparameter tuning did not outperform vector quantization in this experiment. Results are averaged over 10 runs.

[1]For each dimension, before compressing, we shift and scale their values to zero-mean and unit variance. This helps us avoid searching the variance for each dimension.

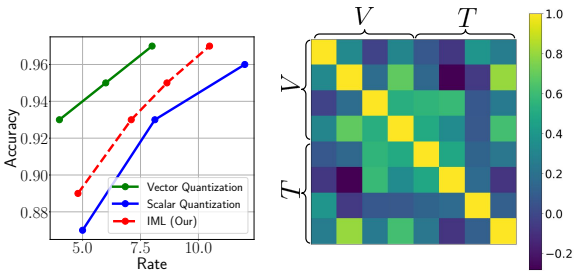



Figure 7: VFL with Breast Cancer Dataset. Left: Rate-Accuracy performance. Right: Covariance matrix between the source and side information's features.

## 6 CONCLUSIONS

We introduce a new one-shot ISC scheme with theoretical guarantees for lossy compression with side information at the decoder. Different from the previous work by (Li and Anantharam, 2021), we introduce the importance matching lemma that quantifies the influence of the number of proposals $N$ on the mismatch probability. We also present a detailed study of synthetic Gaussian sources to validate our theoretical results. On the practical side, we present an algorithm that uses neural networks to enable the extension of IML to complex probability distributions. We then demonstrated its effectiveness in the task of distributed image compression with MNIST and vertical federated learning with CIFAR-10.

For future research, an important direction is to further scale and extend our approach to other DSC and machine learning settings. We note that it may be possible to extend this method to higher dimensional source models, by employing similar techniques proposed by Havasi et al. (2019) for model compression. Specifically, one can split source vectors into $k$ smaller parts and transmit them separately. This can reduce the proposals by roughly a factor of $2^{O(k)}$ but increase the decoding error probability. Finally, another avenue for exploration is the elimination of feedback for latency reduction.

## References


Agapiou, S., Papaspiliopoulos, O., Sanz-Alonso, D., and Stuart, A. M. (2017). Importance sampling: Intrinsic dimension and computational cost. *Statistical Science*, pages 405–431.

Bennett, C. H., Shor, P. W., Smolin, J. A., and Thapliyal, A. V. (2002). Entanglement-assisted capacity of a quantum channel and the reverse shannon theorem. *IEEE transactions on Information Theory*, 48(10):2637–2655.

Castiglia, T. J., Das, A., Wang, S., and Patterson, S. (2022). Compressed-vfl: Communication-efficient learning with vertically partitioned data. In *International Conference on Machine Learning*, pages 2738–2766. PMLR.

Chatterjee, S. and Diaconis, P. (2018). The sample size required in importance sampling. *The Annals of Applied Probability*, 28(2):1099–1135.

Chen, X. and Tuncel, E. (2010). Low-delay prediction- and transform-based wyner–ziv coding. *IEEE transactions on signal processing*, 59(2):653–666.

Cranmer, K., Pavez, J., and Louppe, G. (2015). Approximating likelihood ratios with calibrated discriminative classifiers. *arXiv preprint arXiv:1506.02169*.

Cuff, P. (2013). Distributed channel synthesis. *IEEE Transactions on Information Theory*, 59(11):7071–7096.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Domanovitz, E., Severo, D., Khisti, A., and Yu, W. (2022). Data-driven optimization for zero-delay lossy source coding with side information. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5203–5207. IEEE.

El Gamal, A. and Kim, Y.-H. (2011). *Network information theory*. Cambridge university press.

Elvira, V., Martino, L., Luengo, D., and Bugallo, M. F. (2019). Generalized multiple importance sampling.

Flamich, G., Havasi, M., and Hernández-Lobato, J. M. (2020). Compressing images by encoding their latent representations with relative entropy coding. *Advances in Neural Information Processing Systems*, 33:16131–16141.

Flamich, G., Markou, S., and Hernández-Lobato, J. M. (2022). Fast relative entropy coding with a* coding. In *International Conference on Machine Learning*, pages 6548–6577. PMLR.

Flamich, G. and Theis, L. (2023). Adaptive greedy rejection sampling. *arXiv preprint arXiv:2304.10407*.

Graybill, F. A. and Deal, R. (1959). Combining unbiased estimators. *Biometrics*, 15(4):543–550.

Harsha, P., Jain, R., McAllester, D., and Radhakrishnan, J. (2007). The communication complexity of correlation. In *Twenty-Second Annual IEEE Conference on Computational Complexity (CCC'07)*, pages 10–23. IEEE.

Havasi, M., Peharz, R., and Hernández-Lobato, J. M. (2019). Minimal random code learning: Getting bits back from compressed model parameters. In *7th International Conference on Learning Representations, ICLR 2019*.

Hermans, J., Begy, V., and Louppe, G. (2020). Likelihood-free mcmc with amortized approximate ratio estimators. In *International conference on machine learning*, pages 4239–4248. PMLR.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2016). beta-vae: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*.

Isik, B., Pase, F., Gunduz, D., Koyejo, S., Weissman, T., and Zorzi, M. (2023). Communication-efficient federated learning through importance sampling. *arXiv preprint arXiv:2306.12625*.

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Li, C. T. (2017). *Information-Theoretic Limits of Distributed Randomness Generation*. PhD thesis, Stanford University.

Li, C. T. and Anantharam, V. (2021). A unified framework for one-shot achievability via the poisson matching lemma. *IEEE Transactions on Information Theory*, 67(5):2624–2651.

Li, C. T. and El Gamal, A. (2018). Strong functional representation lemma and applications to coding theorems. *IEEE Transactions on Information Theory*, 64(11):6967–6978.

Liu, J., Cuff, P., and Verdú, S. (2015). One-shot mutual covering lemma and marton's inner bound with a common message. In *2015 IEEE International Symposium on Information Theory (ISIT)*, pages 1457–1461. IEEE.

Liu, Z., Cheng, S., Liveris, A. D., and Xiong, Z. (2006). Slepian-wolf coded nested lattice quantization for wyner-ziv coding: High-rate performance analysis and code design. *IEEE Transactions on Information Theory*, 52(10):4358–4379.

Maddison, C. J., Tarlow, D., and Minka, T. (2014). A* sampling. *Advances in neural information processing systems*, 27.

Mital, N., Özyılkan, E., Garjani, A., and Gündüz, D. (2022). Neural distributed image compression using common information. In *2022 Data Compression Conference (DCC)*, pages 182–191. IEEE.

OpenAI (2023). Gpt-3.5 chatgpt. Accessed: Date.

Ozyilkan, E., Ballé, J., and Erkip, E. (2023). Learned wyner-ziv compressors recover binning. *arXiv preprint arXiv:2305.04380*.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. (2021). Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

Shah, A., Chen, W.-N., Balle, J., Kairouz, P., and Theis, L. (2022). Optimal compression of locally differentially private mechanisms. In *International Conference on Artificial Intelligence and Statistics*, pages 7680–7723. PMLR.

Song, E. C., Cuff, P., and Poor, H. V. (2016). The likelihood encoder for lossy compression. *IEEE Transactions on Information Theory*, 62(4):1836–1849.

Tang, H., Gan, S., Zhang, C., Zhang, T., and Liu, J. (2018). Communication compression for decentralized training. *Advances in Neural Information Processing Systems*, 31.

Theis, L. and Ahmed, N. Y. (2022). Algorithms for the communication of samples. In *International Conference on Machine Learning*, pages 21308–21328. PMLR.

Theis, L., Salimans, T., Hoffman, M. D., and Mentzer, F. (2022). Lossy compression with gaussian diffusion. *arXiv preprint arXiv:2206.08889*.

Triastcyn, A., Reisser, M., and Louizos, C. (2021). Dp-rec: Private & communication-efficient federated learning. *arXiv preprint arXiv:2111.05454*.

Verdú, S. (2012). Non-asymptotic achievability bounds in multiuser information theory. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1–8. IEEE.

Whang, J., Acharya, A., Kim, H., and Dimakis, A. G. (2021). Neural distributed source coding. *arXiv preprint arXiv:2106.02797*.

Wolberg, William, M. O. S. N. and Street, W. (1995). Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5DW2B.

Wyner, A. and Ziv, J. (1976). The rate-distortion function for source coding with side information at the decoder. *IEEE Transactions on information Theory*, 22(1):1–10.

Zamir, R. and Shamai, S. (1998). Nested linear/lattice codes for wyner-ziv encoding. In *1998 Information Theory Workshop (Cat. No. 98EX131)*, pages 92–93. IEEE.

## Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won't result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. [Yes]

   (b) The license information of the assets, if applicable. [Not Applicable]

   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

# Supplementary Material: Importance Matching Lemma for Lossy Compression with Side Information

## 7 Output Distribution of Importance Sampling

The result in this section has already been shown in prior works (Havasi et al., 2019; Theis and Ahmed, 2022; Chatterjee and Diaconis, 2018). In particular, following (Theis and Ahmed, 2022, Corollary 3.2), we have that, for each $x$, if we set:

$$N = 2^{[D_{\mathrm{KL}}(p_{Y|X}(\cdot|x)\,||\,p_Y(\cdot))+t]} \tag{27}$$

for any $t \geq 2\log(e)/(e)$, then $D_{\mathrm{TV}}(\tilde{p}_{Y|X}(.|x), p_{Y|X}(.|x)) \leq 4\epsilon$, where recall that $\tilde{p}_{Y|X}(.|x)$ is the output distribution of the importance sampling scheme as defined in (8) in the main paper. Here $\epsilon$ is given by:

$$\epsilon = 2^{-t/8} + \sqrt{2}\exp\left(-\frac{1}{4B^2}\left(t/2 - \frac{\log e}{e}\right)^2\right) \tag{28}$$

and $B = \log\omega$, where $\omega$ is defined in (9) in the main paper. Thus for any given $\epsilon > 0$, we can construct a $t(\epsilon)$ such that selecting $N \geq N(x,\epsilon)$, where

$$N(x,\epsilon) = 2^{[D_{\mathrm{KL}}(p_{Y|X}(\cdot|x)\,||\,p_Y(\cdot))+t(\epsilon)]}$$

guarantees that $D_{\mathrm{TV}}(\tilde{p}_{Y|X}(.|x), p_{Y|X}(.|x)) \leq 4\epsilon$. Finally since this bound must hold for every $x$ it suffices to take $N_0(\epsilon) = \max_x N(x,\epsilon)$.

## 8 Proof of Theorem 1

For convenience we re-state the Theorem below:

**Theorem** (Restatement of Theorem 1 in main paper)**.** *Given $(X,Y) \sim p_{X,Y}$, and $N, K$ are defined as in the scheme in Sec. 2.2 in the main paper, then we have that:*

$$E[\log K|X=x] \leq E_{Y_1^N}\left[D(\lambda||\mathbf{u})\right] + \delta \tag{29}$$

*where $\lambda = (\lambda_1,\ldots,\lambda_N)$ is defined in (7) (in the main paper), $\mathbf{u} = (1/N,\ldots,1/N)$ is associated with the uniform distribution and $\delta = 1 + \log e/e$ is a constant. Furthermore,*

$$H[K] \leq I(X;Y) + \frac{\Delta}{N} + \log(I(X;Y) + \frac{\Delta}{N} + 1) + 4, \tag{30}$$

*where $\Delta := \Delta(p_{X,Y})$ is a constant defined via (31) and (32) (below) that depends on the distributions $p_{Y|X}(\cdot|x)$, $p_Y(\cdot)$ and $\omega$ in (9), but not on $N$.*

Here we introduce:

$$\Delta = 6(\omega-1)\log\omega + E_{X\sim p_X(\cdot)}\left[\alpha(p_Y(\cdot), p_{Y|X}(\cdot|X=x))\right] \tag{31}$$

and

$$\begin{aligned}\alpha(p_Y(\cdot), p_{Y|X}(\cdot|x)) = 2(\omega-1) + 2\sqrt{\omega-1}(d_3(p_Y(\cdot)||p_{Y|X}(\cdot|x)) - d_2^2(p_Y(\cdot)||p_{Y|X}(\cdot|x)))^{\frac{1}{2}} \\ + 4\omega\cdot d_2(p_Y(\cdot)||p_{Y|X}(\cdot|x)),\end{aligned} \tag{32}$$

and finally

$$d_{N+1}(p_Y(\cdot), p_{Y|X}(\cdot|X=x)) = E_{Y\sim p_Y(\cdot)}\left[\frac{p_Y^N(\cdot)}{p_{Y|X}^N(\cdot|X=x)}\right], \tag{33}$$

for each $N \geq 1$.

### 8.1 Proof of (29) (Eq. (10) in main paper)

We start with the proof of (29). Following the description of the coding scheme in Section 2.2 in the main paper, we note the following: conditioned on $Y_1^n = y_1^n$, our construction is equivalent to channel simulation over a discrete alphabet

$$\Omega = \{y_1, \ldots, y_N\} \tag{34}$$

where the probability of selecting $y_i$ must equal:

$$\lambda_i = \left\{\frac{p_{Y|X}(y_i|x)}{p_Y(y_i)}\right\} / \left\{\sum_{i=1}^{N} \frac{p_{Y|X}(y_i|x)}{p_Y(y_i)}\right\}. \tag{35}$$

The index selection rule i.e., (6) in Section 2.2 in the main text and the associated compression scheme is equivalent to the construction of the Exponential Function Representation Lemma (Li, 2017, Chapter 4) which is summarized below.

1. **Input**: $\Omega = \{y_1, \ldots, y_N\}$ is a discrete alphabet, $\mu(\cdot)$ is a proposal distribution over $\Omega$ known to both the encoder and the decoder while $\nu(\cdot)$ is a target distribution only known to the encoder.
2. **Step 1**: The encoder and decoder sample $S_1, \ldots, S_N$ i.i.d. from $\mathrm{E}xp(1)$ distribution using shared randomness.
3. **Step 2**: The encoder and decoder compute $\phi_i = \frac{S_i}{\mu(y_i)}$ and sort them so that:
$$\phi_{\pi_1} \leq \phi_{\pi_2} \ldots \leq \phi_{\pi_N}.$$
4. **Step 3**: Given $\nu(\cdot)$, the encoder computes
$$I = \arg\min_{1\leq i\leq N} \frac{S_i}{\nu(y_i)}$$
and transmits index $K$ such that $I = \pi_K$.
5. **Analysis**: Following the analysis in (Li, 2017, Chapter 4) we can show that with the selected index $I$ we have $y_I \sim \nu(\cdot)$ and furthermore
$$E[\log K] \leq D(\nu(\cdot)||\mu(\cdot)) + \underbrace{\frac{\log e}{e} + 1}_{=\delta}. \tag{36}$$

We provide a proof of (36) for completeness in Section 18 in this document.

Note that our proposed scheme is equivalent to the exponential functional representation lemma over the discrete alphabet $\Omega$ where the proposal distribution $\mu(\cdot) = \mathbf{u}$ is the uniform distribution over all $N$ samples and the target distribution $\nu(\cdot) = (\lambda_1, \ldots, \lambda_N)$ is based on (35). It follows that the transmitted index $K$ satisfies:

$$E[\log K|Y_1^N = y_1^n, X = x] \leq D(\lambda||\mathbf{u}) + \delta \tag{37}$$

where $\lambda = (\lambda_1, \ldots, \lambda_N)$, with $\lambda_i$ defined in (35) and $\mathbf{u} = (1/N, 1/N, \ldots, 1/N)$ is the uniform distribution. Taking expectation w.r.t. $Y_1^N$ completes the proof.

### 8.2 Proof of (30) above (Eq. (11) in main paper)

We will provide some intuition behind the proof under certain heuristic assumptions. First note that:

$$E_{Y_1^N}\left[\sum_{i=1}^{N}\lambda_i \log\frac{\lambda_i}{u_i}\right] \tag{38}$$

$$= E_{Y_1^N}\left[\sum_{i=1}^{N}\lambda_i \log(N\lambda_i)\right] \tag{39}$$

$$= \sum_{i=1}^{N} E_{Y_1^N}\left[\lambda_i \log(N\lambda_i)\right] \tag{40}$$

$$= NE_{Y_1^N}[\lambda_1 \log(N\lambda_1)] = E_{Y_1^N}[N\lambda_1 \log(N\lambda_1)] \tag{41}$$

The last step follows from symmetry since $Y_1, \dots Y_N$ are sampled i.i.d. from $p_Y(\cdot)$.
Next observe that for each $i = 1, 2, \dots N$ we have that

$$E_{Y_i\sim P_Y(\cdot)}\left[\frac{P_{Y|X}(Y_i|x)}{P_Y(Y_i)}\right] = \int_y P_Y(y)\frac{P_{Y|X}(y|x)}{P_Y(y)}dy = \int_y P_{Y|X}(y|x) = 1.$$

Also since $Y_1, \dots Y_N$ are sampled i.i.d. it follows by law of large numbers that $\frac{1}{N}\sum_{i=1}^{N}\frac{P_{Y|X}(Y_i|x)}{P_Y(Y_i)} \to 1$ as $N \to \infty$.
Note the followings heuristic approximation:

$$N\lambda_1 = \frac{\frac{p_{Y|X}(Y_1|x)}{p_Y(Y_1)}}{\frac{1}{N}\sum_{i=1}^{N}\frac{p_{Y|X}(Y_i|x)}{p_Y(Y_i)}} \approx \frac{\frac{p_{Y|X}(Y_1|x)}{p_Y(Y_1)}}{\frac{1}{N}\sum_{i=2}^{N}\frac{p_{Y|X}(Y_i|x)}{p_Y(Y_i)}} \approx \frac{N}{N-1}\frac{p_{Y|X}(Y_i|x)}{p_Y(Y_i)} \tag{42}$$

In turn, for large $N$ by assuming $\frac{N}{N-1} \approx 1$ we have that

$$E_{Y_1^N}[N\lambda_1 \log(N\lambda_1)] \approx E_{Y_1}\left[\frac{p_{Y|X}(Y_1|x)}{p_Y(Y_1)}\log\frac{p_{Y|X}(Y_1|x)}{p_Y(Y_1)}\right] = D(P_{Y|X}(\cdot|x)||P_Y(\cdot)) \tag{43}$$

Thus it shows that under the above approximations $E[\log K]$ is upper bounded by $E_X[D(P_{Y|X}(\cdot|x)||P_Y(\cdot))] = I(X;Y) + \delta$. Finally as in (Li and El Gamal, 2018) the upper bound on $E[\log K]$ can be converted into an upper bound on $H(K)$ using the maximum entropy theorem:

$$H(K) \le E[\log K] + \log(E[\log K] + 1) + 1, \tag{44}$$

which will complete the proof.
In establishing (30) we formalize the heuristic argument by adding a penalty term that scales as $O(1/N)$. In particular we will show that:

**Proposition 3.** *For any $N \ge 1$ we have that:*

$$E_{Y_1^N}\left[\sum_{i=1}^{N}\lambda_i \log\frac{\lambda_i}{u_i}\right] \le D(p_{Y|X}(\cdot|x)||p_Y(\cdot)) + \frac{6(\omega-1)\log\omega}{N} + \frac{\alpha(p_Y(\cdot), p_{Y|X}(\cdot))}{N} \tag{45}$$

□

Note that upon taking expectation with respect to $X$ on both sides in (45) and using (37) we have that:

$$E[\log K] \le I(X;Y) + \frac{\Delta}{N} + \delta \tag{46}$$

where

$$\Delta = 6(\omega-1)\log\omega + E_{X\sim p_X(\cdot)}\left[\alpha(p_Y(\cdot), p_{Y|X}(\cdot|X=x))\right] \tag{47}$$

Finally using the maximum entropy theorem as in (Li and El Gamal, 2018) we can show that:

$$H(K) \leq I(X;Y) + \frac{\Delta}{N} + \log\left(I(X;Y) + \frac{\Delta}{N} + 1\right) + 4, \tag{48}$$

which completes the proof of (30).

It thus remains to provide a proof of Prop. 3. The proof is rather long and the main challenge is to handle the normalizing term in the expression for $\lambda_i$ carefully. It is presented in Section 19 in this document.

## 9 Alternative Bound for Eq. (11) in Theorem 1 in the main paper

We establish the following alternate upper bound on $H(K)$, which is the counterpart of (30). For any $\epsilon > 0$, we have:

$$H(K) \leq \alpha_N(\epsilon) I(X;Y) + \beta_N(\epsilon) + \log\left(\alpha_N(\epsilon) I(X;Y) + \beta_N(\epsilon) + 1\right) + 4 \tag{49}$$

where

$$\alpha_N(\epsilon) = \frac{N}{(N-1)(1-\epsilon)} \tag{50}$$

and

$$\beta_N(\epsilon) = \frac{N}{(N-1)(1-\epsilon)} \log \frac{N}{(N-1)(1-\epsilon)} + N \log N \exp\left(-2(N-1)\epsilon^2/\omega^2\right) \tag{51}$$

Note that the upper bound in (49) involves a multiplicative constant for $I(X;Y)$ and appears weaker than the bound in (11) in main paper. However by setting $\epsilon \to 0$ and $N \to \infty$ such that $N\epsilon^2 \to \infty$ we can have $\alpha_N(\epsilon) \to 1$ and $\beta_N(\epsilon) \to 0$, so that we can also attain the same rate as in Theorem 1 when $N \to \infty$.

The key step in the following proposition:

**Proposition 4.** *We have that for any $\epsilon > 0$*

$$E_{Y_1^N}\left[\sum_{i=1}^{N} \lambda_i \log \frac{\lambda_i}{u_i}\right] \leq \alpha_N(\epsilon) D(p_{Y|X}(\cdot|x)||p_Y(\cdot)) + \beta_N(\epsilon). \tag{52}$$

The proof of Prop. 4 is relegated to Section 20 in this document. Note it follows from Prop. 4,

$$E[\log K | X = x] \leq \alpha_N(\epsilon) D(p_{Y|X}(\cdot|x)||p_Y(\cdot)) + \beta_N(\epsilon), \tag{53}$$

and thus we have:

$$E[\log K] \leq \alpha_N(\epsilon) I(X;Y) + \beta_N(\epsilon). \tag{54}$$

The upper bound in (54) leads to an upper bound on $H(K)$ as in the previous section. That argument is similar and will not be repeated.

## 10 Multiple Importance Sampling

### 10.1 Analysis

We consider the setting discussed in Section 2.3 in the main text. We generate our samples as follows:

- $Y_1, \ldots, Y_{\bar{N}}$ are sampled i.i.d. from $p_Y^{(1)}(\cdot)$
- $Y_{\bar{N}+1}, \ldots, Y_N$ are sampled i.i.d. from $p_Y^{(2)}(\cdot)$

where we select $p_Y^{(1)}(\cdot)$ and $p_Y^{(2)}(\cdot)$ to satisfy: $p_Y(y) = \frac{1}{2}p_Y^{(1)}(y) + \frac{1}{2}p_Y^{(2)}(y)$.
Given $X = x$, the index $K$ in Multiple Importance Sampling (MIS) is selected Elvira et al. (2019) using the following probability distribution:

$$\Pr(K = i) = \lambda_i = \frac{\frac{p_{Y|X}(Y_i|X=x)}{p_Y(Y_i)}}{\sum_{j=1}^{N} \frac{p_{Y|X}(Y_j|X=x)}{p_Y(Y_j)}} \tag{55}$$

We perform approximate analysis assuming that $N$ is sufficiently large and that we can approximate

$$\sum_{j=1}^{N} \frac{p_{Y|X}(Y_j|X=x)}{p_Y(Y_j)} \approx N, \tag{56}$$

so that

$$\lambda_i \approx \frac{1}{N} \frac{p_{Y|X}(Y_i|X=x)}{p_Y(Y_i)}. \tag{57}$$

The above approximation is justified by noting that for $i = 1, 2, \ldots, \bar{N}$ we have that:

$$\begin{aligned} &\frac{1}{N} \sum_{i=1}^{\bar{N}} E_{Y_i, Y_{i+\bar{N}}} \left[ \frac{p_{Y|X}(Y_i|X=x)}{p_Y(Y_i)} + \frac{p_{Y|X}(Y_{i+\bar{N}}|X=x)}{p_Y(Y_{i+\bar{N}})} \right] \\ &= \frac{1}{2} \int_y \frac{p_{Y|X}(y|x)}{p_Y(y)} \left( p^{(1)}(y) + p^{(2)}(y) \right) dy = 1. \end{aligned} \tag{58}$$

We argue that under the simplifying assumption (57), the proxy distribution is close to the target distribution. Indeed, note that from (8) in the main text

$$\tilde{p}_{Y|X}(y|x) = E_{Y_1,\ldots,Y_N} \left[ \sum_{i=1}^{N} \lambda_i \cdot \delta(y - Y_i) \right] \tag{59}$$

$$\approx \frac{1}{N} E_{Y_1,\ldots,Y_N} \left[ \sum_{i=1}^{N} \frac{p_{Y|X}(Y_i|X=x)}{p_Y(Y_i)} \cdot \delta(y - Y_i) \right] \tag{60}$$

$$= \frac{1}{N} \sum_{i=1}^{N} \int_{y_i} \frac{p_{Y|X}(y_i|X=x)}{p_Y(y_i)} p_Y(y_i) \delta(y - y_i) \tag{61}$$

$$= p_{Y|X}(y|X=x). \tag{62}$$

Thus the distribution of the output samples will be close to the target distribution. For the approximate rate analysis, we note that the proof in Section 8.1 in this document still applies and we have:

$$E[\log K|X=x] \leq E_{Y_1,\ldots,Y_N}[D(\lambda||\mathbf{u})] + \delta \tag{63}$$

where again we have

$$\lambda_i = \frac{\frac{p_{Y|X}(Y_i|X=x)}{p_Y(Y_i)}}{\sum_{j=1}^{N} \frac{p_{Y|X}(Y_j|X=x)}{p_Y(Y_j)}}, \qquad i = 1, 2, \ldots, N$$

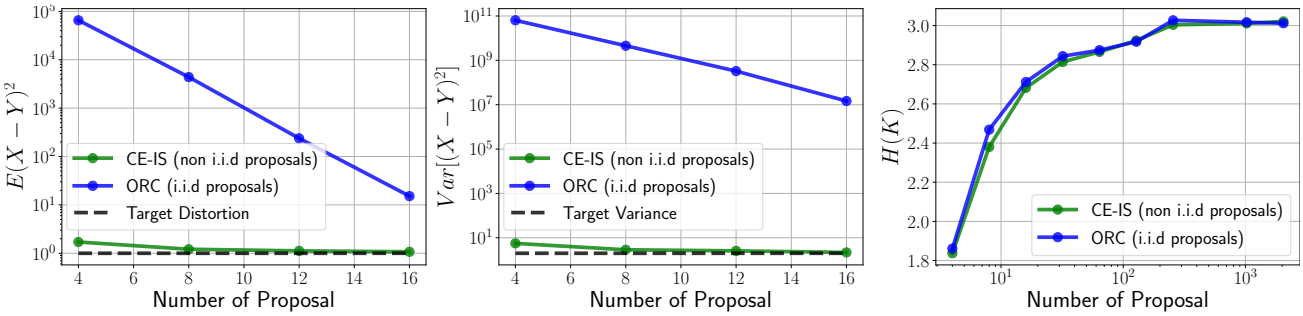



Figure 8: Multiple Importance Sampling. From left to right: expected distortion, distortion variance, and compression rate as a function of the number of proposals, $N$. MIS exhibits faster convergence to the target levels while maintaining a comparable compression rate to ORC. We set $m = 512$ and $D = 1$.

and $\mathbf{u}$ denotes the uniform distribution. Now note that:

$$E_{Y_1,\dots,Y_N}[D(\lambda||\mathbf{u})] = E_{Y_1,\dots,Y_N}\left[\sum_{i=1}^{\bar{N}} \left(\lambda_i \log N\lambda_i + \lambda_{i+\bar{N}} \log N\lambda_{i+\bar{N}}\right)\right] \tag{64}$$

$$\approx E_{Y_i,Y_{i+\bar{N}}}\left[\frac{1}{2}\frac{p_{Y|X}(Y_i|X=x)}{p_Y(Y_i)} \log \frac{p_{Y|X}(Y_i|X=x)}{p_Y(Y_i)} + \frac{1}{2}\frac{p_{Y|X}(Y_{i+\bar{N}}|X=x)}{p_Y(Y_{i+\bar{N}})} \log \frac{p_{Y|X}(Y_{i+\bar{N}}|X=x)}{p_Y(Y_{i+\bar{N}})}\right] \tag{65}$$

$$= D(p_{Y|X}(\cdot|X=x)||p_Y(\cdot)) \tag{66}$$

Thus to the first order approximation in $N$, the MIS scheme achieves the same compression rate as the IS scheme.

### 10.2 Numerical Example

We demonstrate the advantage of our proposed over ORC through a numerical example. We assume that the source sample $X$ is a mixture of two Gaussians: $p_X(x) = \frac{1}{2}p_X^{(1)}(x)+\frac{1}{2}p_X^{(2)}(x)$, where we have $p_X^{(1)}(x) = \mathcal{N}(m,1)$ and $p_X^{(2)}(x) = \mathcal{N}(-m,1)$. The conditional distribution of $Y$ given $X$ is given by $Y = X+\zeta$, where $\zeta \sim \mathcal{N}(0,D)$. Note that we can also express the marginal distribution of $Y$ as $p_Y(y) = \frac{1}{2}p_Y^{(1)}(y)+\frac{1}{2}p_Y^{(2)}(y)$ where $p_Y^{(1)}(y) = \mathcal{N}(m,1+D)$ and $p_Y^{(2)}(y) = \mathcal{N}(-m,1+D)$. In this example, we set $m = 512$, $D = 1$ and average our results over $2^{20}$ simulations. In Figure 8, given a source sample $X$ at the encoder and output $Y$ at the decoder, we compute $E[(Y-X)^2]$, $\mathrm{Var}((Y-X)^2)$ as well as the compression rate for our proposed scheme (computed from the index histogram) and ORC (with i.i.d. samples from $p_Y(\cdot)$) for a different number of candidate proposals. Here, $E[(Y-X)^2]$ is the expected distortion, and $\mathrm{Var}((Y-X)^2)$ is the distortion variance and is equal to $D$ and $2D^2$ respectively when $\tilde{p}_{Y|X}(.|x) = p_{Y|X}(.|x)$ for all $x$, which we refer to as target distortion and target variance. We observe that while both schemes achieve a similar compression rate, our proposed scheme outperforms ORC in other metrics indicating that it more closely approximates the target distribution $p_{Y|X}(.|x)$. This occurs because when a small number of i.i.d proposals $N$ is considered, there is a high probability that all the proposals are sampled from different modes of the source variable X. This probability is $2^{-(N+1)}$ and can result a significant distortion in the output. The extent of this distortion is primarily determined by the distance between the two modes $p_X^{(1)}(.)$ and $p_X^{(2)}(.)$ of $p_X(.)$, which is $4m^2$ in this example. On the other hand, such probability is 0 in our MIS scheme, enabling us to achieve a better convergence rate in this example.

In Figure 9, we show that when the proposals $\{Y_i\}_1^N$ are non-i.i.d, ORC is unable to simulate the target distribution $p_{Y|X}(.|x)$ properly. In this example, for each figure, we fix a source sample $X = x$ and plot the histogram of the obtained samples $Y \sim \tilde{p}_{Y|X}(.|x)$, from multiple sets of non-i.i.d proposals. We set $N = 512$ in this case and refer to $p_X^{(1)}(.) = \mathcal{N}(m,1)$ and $p_X^{(2)}(.) = \mathcal{N}(-m,1)$ as the positive and negative mode respectively. Figure 9 plots and compares the simulated histogram of our method CE-IS and ORC in the case where $X$ is from the

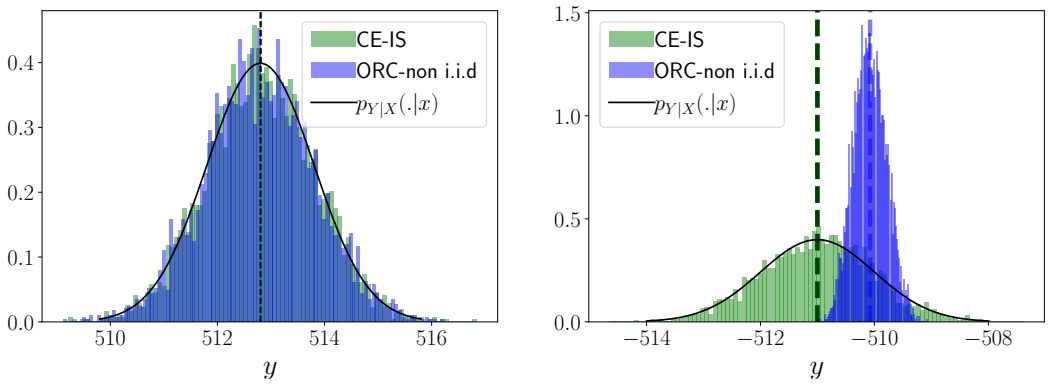



Figure 9: ORC does not simulate the target distribution properly in the case of non-i.i.d proposals. (Left) When $X = x$ is from the positive mode, i.e. $x = 512.88$ in this case, ORC can simulate the $\tilde{p}_{Y|X}(.|x)$ approximately close to the target $p_{Y|X}(.|x)$. (Right), $X = x$ is from the negative mode, i.e. $x = -509.93$, when ORC is unable to simulate $\tilde{p}_{Y|X}(.|x)$ accurately. Our method CE-IS can simulate accurately in both cases. We set $m = 512$ and $D = 1$.

postive mode (left) and negative mode (right). Our first observation is that our proposed scheme CE-IS is able to simulate the distribution accurately in both cases. On the other hand, while ORC seems to simulate accurately the distribution when $x > 0$, or from the positive mode (left figure), its simulated distribution when $x < 0$ (negative mode) is very different from the target distribution $p_{Y|X}(.|x)$ (right figure). This is because ORC requires the exponential variables $S_1, ..., S_{2N}$ to be sorted before performing index selection, which only works when the proposals are i.i.d. In the right figure, when the proposals are non-i.i.d and $x < 0$, ORC will ignore the first half of the proposals (since $\log \frac{p_Y(Y_i)}{p_{Y|X}(Y_i|x)}$ is extremely large due to large $m$) to select samples in the second half of the proposals $Y_{N+1}...Y_{2N}$. As a result, it loses the exponential race property to simulate proper distribution. When $x > 0$, on the other hand, ORC does not need to ignore the first half and as a result, can simulate approximately accurate $\tilde{p}_{Y|X}(.|x)$.

## 11 Proof of Prop. 1 in the Main Paper

We restate the result for convenience.

**Proposition.** *Let $\Omega = \{y_1, \ldots, y_N\}$ denote the sequence of samples.*

$$\Pr(U_p \neq U_q | \Omega, X = x, U_p = k) \leq 1 - \left(1 + \frac{p_{Y|X}(y_k|x)}{q_{Y|X}(y_k|x)} \frac{\left(\frac{1}{N}\sum_{j=1}^{N} \frac{q_{Y|X}(y_j|x)}{p_Y(y_j)}\right)}{\left(\frac{1}{N}\sum_{j=1}^{N} \frac{p_{Y|X}(y_j|x)}{p_Y(y_j)}\right)}\right)^{-1}. \tag{67}$$

The proof follows by observing that given $Y_1^N = y_1^N$, the sampling procedures in (12) in the main paper, is equivalent to the Poisson Matching Lemma applied to a discrete alphabet $\Omega = \{y_1, \ldots, y_N\}$ with target distributions given by

$$\lambda_i^p = \frac{\frac{p_{Y|X}(y_i|X=x)}{p_Y(y_i)}}{\sum_{j=1}^{N} \frac{p_{Y|X}(y_j|X=x)}{p_Y(y_j)}}, \quad \lambda_i^q = \frac{\frac{q_{Y|X}(y_i|X=x)}{p_Y(y_i)}}{\sum_{j=1}^{N} \frac{q_{Y|X}(y_j|X=x)}{p_Y(y_j)}}. \tag{68}$$

We provide the analysis for completeness below. Let us define:

$$\tilde{S}_p = \min_{1 \leq i \leq N} \frac{S_i}{\lambda_i^p} \sim \text{Exp}(\sum_i \lambda_i^p = 1). \tag{69}$$

We further condition on $U_p = k$ and $\tilde{S}_p = s$. It follows that for each $j \neq k$ we have that:

$$\frac{S_j}{\lambda_j^p} \geq s, \tag{70}$$

and thus $S_j - \lambda_j^p \cdot s$ is an Exponential(1) random variable. Since this holds for each $s$, it follows tht $S_j - \lambda_j^p \cdot \tilde{S}_p$ is also an Exponential(1) and is independent of the variable $\tilde{S}_p$.
We consider the following events:

$$\Pr(U_p \neq U_q|\Omega, U_p = k) = \Pr\left(\min_{j\neq k} \frac{S_j}{\lambda_j^q} \leq \frac{S_k}{\lambda_k^q} \;\middle|\; \Omega, U_p = k\right) \tag{71}$$

$$= \Pr\left(\min_{j\neq k} \frac{S_j}{\lambda_j^q} \leq \tilde{S}_p \frac{\lambda_k^p}{\lambda_k^q} \;\middle|\; \Omega, U_p = k\right) \tag{72}$$

$$\leq \Pr\left(\min_{j\neq k} \frac{S_j - \lambda_j^p \cdot \tilde{S}_p}{\lambda_j^q} \leq \tilde{S}_p \frac{\lambda_k^p}{\lambda_k^q} \;\middle|\; \Omega, U_p = k\right) \tag{73}$$

$$= \frac{\sum_{j\neq k} \lambda_j^q}{\sum_{j\neq k} \lambda_j^q + \frac{\lambda_k^q}{\lambda_k^p}} \tag{74}$$

$$\leq \frac{1}{1 + \frac{\lambda_k^q}{\lambda_k^p}} = 1 - \left(1 + \frac{\lambda_k^p}{\lambda_k^q}\right)^{-1} \tag{75}$$

$$= 1 - \left(1 + \frac{\frac{p_{Y|X}(y_k|x)}{p_Y(y_k)}}{\frac{q_{Y|X}(y_k|x)}{p_Y(y_k)}} \frac{\left(\sum_{j=1}^N \frac{q_{Y|X}(y_j|x)}{p_Y(y_j)}\right)}{\left(\sum_{j=1}^N \frac{p_{Y|X}(y_j|x)}{p_Y(y_j)}\right)}\right)^{-1} \tag{76}$$

$$= 1 - \left(1 + \frac{p_{Y|X}(y_k|x)}{q_{Y|X}(y_k|x)} \frac{\left(\frac{1}{N}\sum_{j=1}^N \frac{q_{Y|X}(y_j|x)}{p_Y(y_j)}\right)}{\left(\frac{1}{N}\sum_{j=1}^N \frac{p_{Y|X}(y_j|x)}{p_Y(y_j)}\right)}\right)^{-1} \tag{77}$$

## 12 Proof of Theorem 2 in Main Paper

In what follows, we will set $k = 1$ without loss of generality. The argument can be easily extended to any $k$. We rewrite the Theorem below for sake of convenience.

**Theorem** (Expanded version of Theorem 2 in main paper.)**.** *Define $\bar{N} = N - 1$, we have:*

$$\Pr(U_p \neq U_q|Y_1 = y_1, U_p = 1) \leq 1 - \left(1 + \frac{\lambda(y_1)}{\beta(y_1)}\mu_{y_1}(\bar{N})\right)^{-1}, \tag{78}$$

*where*

$$\mu_{y_1}(\bar{N}) = \left(\frac{\frac{\beta(y_1)}{\bar{N}} + 1}{\frac{\lambda(y_1)}{\bar{N}} + 1}\right) + \frac{1}{\bar{N}}\left(1 + \frac{\lambda(y_1)}{\bar{N}}\right)K(\bar{N}) + \frac{2\omega}{\bar{N}}\left(1 + \frac{\lambda(y_1)}{\bar{N}}\right)L(\bar{N}), \tag{79}$$

$$K(\bar{N}) = 4\frac{(\omega-1)}{(1 + \frac{\lambda(y_1)}{\bar{N}})^2}\left(1 + \frac{N+1}{\bar{N}}\omega\right)\sqrt{2 + 4\left(\frac{1 + \frac{\beta(y_1)}{\bar{N}}}{1 + \frac{2\lambda(y_1)}{\bar{N}}}\right)^2\left\{\left(1 + \frac{N+1}{\bar{N}}\omega\right)^2 + \frac{(\omega-1)}{\bar{N}}\right\}} \tag{80}$$

*and*

$$L(\bar{N}) = \sqrt{\omega - 1}\sqrt{(d_5(p_Y(\cdot)||p_{Y|X}(\cdot|X = x)) - d_3^2(p_Y(\cdot)||p_{Y|X}(\cdot|x)))} + (\omega - 1)d_3(p_Y(\cdot)||p_{Y|X}(\cdot|x)) \tag{81}$$

*in* (80) *and* (81) *respectively scale as* $\Theta(1)$ *as* $\bar{N} \to \infty$*. Also, we use* $\lambda(y_1) = \frac{p_{Y|X}(y_1|x)}{p_Y(y_1)}$*, and* $\beta(y_1) = \frac{q_{Y|X}(y_1|x)}{p_Y(y_1)}$*.*

Upon examining (80) and (81), note that as $\bar{N} \to \infty$, the dominating term in $K(\bar{N})$ scales as $\omega^3$, while $L(\bar{N})$ is upper-bounded by $\omega \cdot d_5(p_Y(\cdot)||p_{Y|X}(\cdot|x))$, regardless of $N$. Under the assumption that $d_5(p_Y(\cdot)||p_{Y|X}(\cdot|x)) < \infty$, see (33), we observe that for any $\epsilon > 0$ there is a sufficiently large $N_1(\epsilon)$, we will have that $\mu_{y_1}(N) \leq 1+\epsilon$ for any $N \geq N_1(\epsilon)$. In turn (78) recovers the bound in the Poisson Matching Lemma in (Li and Anantharam, 2021).

To proceed with the proof, observe that:

$$\Pr(U_p \neq U_q | Y_1 = y_1, U_p = 1) \tag{82}$$

$$= 1 - E_{Y_2^N}\left[\left(1 + \frac{p_{Y|X}(y_1|x)}{q_{Y|x}(y_1|x)} \frac{\left(\frac{1}{N}\sum_{j=1}^N \frac{q_{Y|X}(Y_j|x)}{p_Y(Y_j)}\right)}{\left(\frac{1}{N}\sum_{j=1}^N \frac{p_{Y|X}(Y_j|x)}{p_Y(Y_j)}\right)}\right)^{-1} \middle| Y_1 = y_1, U_p = 1\right] \tag{83}$$

$$\leq 1 - \left(1 + \frac{p_{Y|X}(y_1|x)}{q_{Y|X}(y_1|x)} E_{Y_2^N}\left[\frac{\left(\frac{1}{N}\sum_{j=1}^N \frac{q_{Y|X}(Y_j|x)}{p_Y(Y_j)}\right)}{\left(\frac{1}{N}\sum_{j=1}^N \frac{p_{Y|X}(Y_j|x)}{p_Y(Y_j)}\right)} \middle| Y_1 = y_1, U_p = 1\right]\right)^{-1}, \tag{84}$$

where the last step is a consequence of Jensen's inequality since the function $f(t) = 1/t$ is a convex function so that $E[f(t)] \geq f(E[t])$ holds. We thus need to upper bound the expectation above.

$$E_{Y_2^N}\left[\frac{\left(\sum_{j=1}^N \frac{q_{Y|X}(Y_j|x)}{p_Y(Y_j)}\right)}{\left(\sum_{j=1}^N \frac{p_{Y|X}(Y_j|x)}{p_Y(Y_j)}\right)} \middle| Y_1 = y_1, U_p = 1\right] \tag{85}$$

$$= \int_{y_2^N} \frac{\left(\sum_{j=1}^N \frac{q_{Y|X}(y_j|x)}{p_Y(y_j)}\right)}{\left(\sum_{j=1}^N \frac{p_{Y|X}(y_j|x)}{p_Y(y_j)}\right)} p_{Y_2^N|\{Y_1,U_p\}}(y_2,\ldots,y_N|Y_1 = y_1, U_p = 1) dy_2 \ldots dy_N \tag{86}$$

$$= \int_{y_2^N} \frac{\sum_{j=1}^N \beta(y_j)}{\sum_{j=1}^N \lambda(y_j)} p_{Y_2^N|\{Y_1,U_p\}}(y_2,\ldots,y_N|Y_1 = y_1, U_p = 1) dy_2 \ldots dy_N \tag{87}$$

$$\tag{88}$$

where we define

$$\beta(y_j) = \frac{q_{Y|X}(y_j|x)}{p_Y(y_j)} \qquad \lambda(y_j) = \frac{p_{Y|X}(y_j|x)}{p_Y(y_j)} \tag{89}$$

Now consider the joint density function $p_{Y_2^N|\{Y_1,U_p\}}(y_2,\ldots,y_N|Y_1 = y_1, U_p = 1)$:

$$p_{Y_2^N|\{Y_1,U_p\}}(y_2,\ldots,y_N|Y_1 = y_1, U_p = 1) \tag{90}$$

$$= \frac{\Pr(U_p = 1|Y_1 = y_1,\ldots,Y_N = y_N) p_{Y_2^N|Y_1}(y_2^N|Y_1 = y_1)}{\Pr(U_p = 1|Y_1 = y_1)} \tag{91}$$

$$= \frac{\Pr(U_p = 1|Y_1 = y_1,\ldots,Y_N = y_N) \prod_{j=2}^N p_{Y_j}(y_j)}{\Pr(U_p = 1|Y_1 = y_1)} \tag{92}$$

$$= \frac{\lambda(y_1)}{\sum_{j=1}^N \lambda(y_j)} \frac{\prod_{j=2}^N p_{Y_j}(y_j)}{\Pr(U_p = 1|Y_1 = y_1)} \tag{93}$$

Next we consider the probability $\Pr(U_p = 1|Y_1 = y_1)$.

$$\Pr(U_p = 1|Y_1 = y_1) = E_{Y_2^n}[\Pr(U_p = 1|Y_1 = y_1, Y_2 = y_2, \ldots, Y_N = y_N)] \tag{94}$$

$$= E_{Y_2^N}\left[\frac{\lambda(y_1)}{\lambda(y_1) + \sum_{j=2}^N \lambda(y_j)}\right] \tag{95}$$

$$\geq \frac{\lambda(y_1)}{\lambda(y_1) + E_{Y_2^N}[\sum_{j=2}^N \lambda(y_j)]} \tag{96}$$

$$= \frac{\lambda(y_1)}{\lambda(y_1) + (N-1)} \tag{97}$$

where we have again used the convexity of $f(t) = 1/t$ and applied Jensen's inequality.
Thus we can express the following:

$$E_{Y_2^N}\left[\frac{\left(\sum_{j=1}^N \frac{q_{Y|X}(Y_j|x)}{p_Y(Y_j)}\right)}{\left(\sum_{j=1}^N \frac{p_{Y|X}(Y_j|x)}{p_Y(Y_j)}\right)}\middle| Y_1 = y_1, U_p = 1\right] \tag{98}$$

$$\leq (N - 1 + \lambda(y_1)) \int_{y_2^N} \frac{\sum_{j=1}^N \beta(y_j)}{(\sum_{j=1}^N \lambda(y_j))^2} \prod_{j=2}^n p_Y(y_j) dy_2 \ldots dy_N \tag{99}$$

$$= (N - 1 + \lambda(y_1)) E_{Y_2^N}\left[\frac{\sum_{j=1}^N \beta(Y_j)}{(\sum_{j=1}^N \lambda(Y_j))^2}\middle| Y_1 = y_1\right] \tag{100}$$

We now establish the following bound:

**Proposition 5.** *For any $N \geq 1$ with $\bar{N} = N - 1$, we have that:*

$$\bar{N} E_{Y_2^N}\left[\frac{\sum_{i=1}^N \beta(Y_i)}{\left(\sum_{i=1}^N \lambda(Y_i)\right)^2}\middle| Y_1 = y_1\right] \leq \frac{\frac{\beta(y_1)}{\bar{N}} + 1}{\left(1 + \frac{\lambda(y_1)}{\bar{N}}\right)^2} + \frac{1}{\bar{N}} K(\bar{N}) + \frac{2\omega}{\bar{N}} L(\bar{N})$$

*and in turn using* (100)*, we have*

$$E_{Y_2^N}\left[\frac{\left(\sum_{j=1}^N \frac{q_{Y|X}(Y_j|x)}{p_Y(Y_j)}\right)}{\left(\sum_{j=1}^N \frac{p_{Y|X}(Y_j|x)}{p_Y(Y_j)}\right)}\middle| Y_1 = y_1, U_p = 1\right] \leq \frac{\beta(y_1) + \bar{N}}{\lambda(y_1) + \bar{N}} + \frac{1 + \frac{\lambda(y_1)}{\bar{N}}}{\bar{N}} K_1(\bar{N}) + \frac{2\omega(1 + \frac{\lambda(y_1)}{\bar{N}})}{\bar{N}} L(\bar{N}). \tag{101}$$

*where $K(\bar{N})$ is given in* (80) *and $L(\bar{N})$ is given in* (81)*.*

The proof of Prop. 5 is in rather long and relegated to Section 21. Note that substituting (101) into (84)

$$\Pr(U_p \neq U_q|Y_1 = y_1, U_p = 1) \leq 1 - \left(1 + \frac{\lambda(y_1)}{\beta(y_1)}\left(\frac{\frac{\beta(y_1)}{\bar{N}} + 1}{\frac{\lambda(y_1)}{\bar{N}} + 1} + \frac{1 + \frac{\lambda(y_1)}{\bar{N}}}{\bar{N}} K(\bar{N}) + \frac{2\omega(1 + \frac{\lambda(y_1)}{\bar{N}})}{\bar{N}} L(\bar{N})\right)\right)^{-1}, \tag{102}$$

which completes the proof.

## 13 Alternative Upper Bound in (15) in Theorem 2 in Main Paper

We establish the following upper bound:

$$\Pr(U_p \neq U_q|Y_1 = y_1, U_p = 1) \leq 1 - \left(1 + \frac{\lambda(y_1)}{\beta(y_1)} \mu'_{y_1}(N)\right)^{-1}. \tag{103}$$

where $\lambda(y_1)$ and $\beta(y_1)$ are defined in (89) and we have

$$\mu'_{y_1}(N) = (N-1+\lambda(y_1))\left(\frac{\beta(y_1)+(N-1)(1+\epsilon)}{(\lambda(y_1)+(N-1)(1-\epsilon))^2}+\frac{N\omega}{\lambda(y_1)^2}2e^{-(N-1)\epsilon^2/\omega^2}\right) \tag{104}$$

Note that this upper bound requires that $\lambda(y_1) > 0$, which is a stronger condition than the condition in Theorem 2. It can also be seen that for sufficiently large $N$, we can choose $\epsilon$ to be arbitrarily small and recover PML.

In order to establish (103), we will show the following:

**Proposition 6.** *For any $0 < \epsilon < 1$ we have that:*

$$E_{Y_2^N}\left[\frac{\sum_{j=1}^N \beta(Y_j)}{(\sum_{j=1}^N \lambda(Y_j))^2}\middle| Y_1 = y_1\right] \leq \frac{\beta(y_1)+(N-1)(1+\epsilon)}{(\lambda(y_1)+(N-1)(1-\epsilon))^2}+\frac{N\omega}{\lambda(y_1)^2}2e^{-(N-1)\epsilon^2/\omega^2} \tag{105}$$

*and in turn using* (100)*, we have*

$$E_{Y_2^N}\left[\frac{\left(\sum_{j=1}^N \frac{q_{Y|X}(Y_j|x)}{p_Y(Y_j)}\right)}{\left(\sum_{j=1}^N \frac{p_{Y|X}(Y_j|x)}{p_Y(Y_j)}\right)}\middle| Y_1 = y_1, U_p = 1\right] \leq \mu'_{y_1}(N) \tag{106}$$

□

Note that the bound in (103) follows directly by substituting (106) into (84). We relegate the proof of Prop. 6 to Section 22 in this document.

## 14 Proof of Theorem 3 in Main Paper

We state an expanded version of Theorem 3 in the main paper. We assume that $k = 1$ without loss of genrality.

**Theorem.** *Let $\Omega = \{y_1, \ldots, y_N\}$. The error probability satisfies:*

$$Pr\left(U_P \neq U_Q | U_P = 1, X = x, Z = z, \Omega\right) \leq 1 - \left(1 + \frac{p_{Y|X}(y_1|x)}{Q_{Y|Z}(y_1|z)}\frac{\left(\frac{1}{N}\sum_{j=1}^N \frac{Q_{Y|Z}(y_j|z)}{p_Y(y_j)}\right)}{\left(\frac{1}{N}\sum_{j=1}^N \frac{p_{Y|X}(y_j|x)}{p_Y(y_j)}\right)}\right)^{-1}, \tag{107}$$

*and furthermore,*

$$Pr(U_p \neq U_q | Y_1 = y_1, U_p = k, X = x, Z = z) \leq 1 - \left(1 + \mu_{y_1}(\bar{N})\frac{p_{Y|X}(y_1|x)}{Q_{Y|Z}(y_1|z)}\right)^{-1}. \tag{108}$$

*where*

$$\mu_{y_1}(N) = \left(\frac{\frac{\beta(y_1)}{\bar{N}}+1}{\frac{\lambda(y_1)}{\bar{N}}+1}\right)+\frac{1}{\bar{N}}\left(1+\frac{\lambda(y_1)}{\bar{N}}\right)K(\bar{N})+\frac{2\omega}{\bar{N}}\left(1+\frac{\lambda(y_1)}{\bar{N}}\right)L(\bar{N})), \tag{109}$$

*where $\bar{N} = N-1$, $\beta(y_1) = \frac{Q_{Y|Z}(y_1|z)}{p_Y(y_1)}$ and $\lambda(y_1) = \frac{p_{Y|X}(y_1|x)}{p_Y(y_1)}$ and*

$$K(\bar{N}) = 4\frac{(\omega-1)}{(1+\frac{\lambda(y_1)}{\bar{N}})^2}\left(1+\frac{N+1}{\bar{N}}\omega\right)\sqrt{2+4\left(\frac{1+\frac{\beta(y_1)}{\bar{N}}}{1+\frac{2\lambda(y_1)}{\bar{N}}}\right)^2\left\{\left(1+\frac{N+1}{\bar{N}}\omega\right)^2+\frac{(\omega-1)}{\bar{N}}\right\}}, \tag{110}$$

$$L(\bar{N}) = \sqrt{\omega-1}\sqrt{(d_5(p_Y(\cdot)||p_{Y|X}(\cdot|X=x)) - d_3^2(p_Y(\cdot)||p_{Y|X}(\cdot|x)))}+(\omega-1)d_3(p_Y(\cdot)||p_{Y|X}(\cdot|x)). \tag{111}$$

Upon examining (109)-(111), and assuming that $d_5(p_Y(\cdot)||p_{Y|X}(\cdot|x)) < \infty$, see (33), it is clear that there exists an $N_1(\epsilon)$ such that

$$\mu(N) \leq 1 + \epsilon, \quad \forall N \geq N_1(\epsilon) \tag{112}$$

We define $\Omega = \{y_1^N\}$ and consider:

$$\begin{aligned}
&\Pr\left(U_P \neq U_Q | U_P = k, X = x, Z = z, \Omega\right) && (113)\\
&= \Pr\left(\min_{j\neq k} \frac{S_j}{\lambda_j^q} \leq \frac{S_k}{\lambda_k^q} \;\middle|\; \Omega, U_p = k, X = x, Z = z\right) && (114)\\
&= \Pr\left(\min_{j\neq k} \frac{S_j}{\lambda_j^q} \leq \tilde{S}_P \frac{\lambda_k^p}{\lambda_k^q} \;\middle|\; \Omega, U_p = k, X = x, Z = z\right) && (115)\\
&\leq \Pr\left(\min_{j\neq k} \frac{S_j - \lambda_j^p \tilde{S}_P}{\lambda_j^q} \leq \tilde{S}_P \frac{\lambda_k^p}{\lambda_k^q} \;\middle|\; \Omega, U_p = k, X = x, Z = z\right) && (116)\\
&= \Pr\left(\min_{j\neq k} \frac{S_j - \lambda_j^p \tilde{S}_P}{\lambda_j^q} \leq \tilde{S}_P \frac{\lambda_k^p}{\lambda_k^q} \;\middle|\; \Omega, U_p = k, X = x\right) && (117)\\
&\leq \frac{1}{1 + \frac{\lambda_k^q}{\lambda_k^p}} = 1 - \left(1 + \frac{\lambda_k^p}{\lambda_k^q}\right)^{-1} && (118)\\
&= 1 - \left(1 + \frac{\frac{p_{Y|X}(y_k|x)}{p_Y(y_k)}}{\frac{Q_{Y|Z}(y_k|z)}{p_Y(y_k)}} \frac{\left(\sum_{j=1}^N \frac{Q_{Y|Z}(y_j|z)}{p_Y(y_j)}\right)}{\left(\sum_{j=1}^N \frac{p_{Y|X}(y_j|x)}{p_Y(y_j)}\right)}\right)^{-1} && (119)\\
&= 1 - \left(1 + \frac{p_{Y|X}(y_k|x)}{Q_{Y|Z}(y_k|z)} \frac{\left(\frac{1}{N}\sum_{j=1}^N \frac{Q_{Y|Z}(y_j|z)}{p_Y(y_j)}\right)}{\left(\frac{1}{N}\sum_{j=1}^N \frac{p_{Y|X}(y_j|x)}{p_Y(y_j)}\right)}\right)^{-1} && (120)
\end{aligned}$$

where (117) follows from the Markov condition (17) and the subsequent stepsilon follows the the analysis done previously.

We will assume that $k = 1$ without loss of generality. Taking expectation with respect to $\Omega_2 = \{Y_2^N\}$ we have that:

$$\begin{aligned}
&\Pr\left(U_P \neq U_Q | U_P = 1, X = x, Z = z, Y = y\right) && (121)\\
&= E_{Y_2^N}\left[1 - \left(1 + \frac{p_{Y|X}(Y_1|x)}{Q_{Y|Z}(Y_1|z)} \frac{\left(\frac{1}{N}\sum_{j=1}^N \frac{Q_{Y|Z}(Y_j|z)}{p_Y(Y_j)}\right)}{\left(\frac{1}{N}\sum_{j=1}^N \frac{p_{Y|X}(Y_j|x)}{p_Y(Y_j)}\right)}\right)^{-1} \middle| U_P = 1, X = x, Z = z, Y = y\right] && (122)\\
&\leq 1 - \left(1 + \frac{p_{Y|X}(Y_1|x)}{Q_{Y|Z}(Y_1|z)} \cdot E_{Y_2^N|\{U_p,X,Y,Z\}}\left[\frac{\left(\frac{1}{N}\sum_{j=1}^N \frac{Q_{Y|Z}(Y_j|z)}{p_Y(Y_j)}\right)}{\left(\frac{1}{N}\sum_{j=1}^N \frac{p_{Y|X}(Y_j|x)}{p_Y(Y_j)}\right)} \middle| U_P = 1, X = x, Z = z, Y = y\right]\right)^{-1} && (123)\\
&= 1 - \left(1 + \frac{p_{Y|X}(Y_1|x)}{Q_{Y|Z}(Y_1|z)} \cdot E_{Y_2^N|\{U_p,X,Y\}}\left[\frac{\left(\frac{1}{N}\sum_{j=1}^N \frac{Q_{Y|Z}(Y_j|z)}{p_Y(Y_j)}\right)}{\left(\frac{1}{N}\sum_{j=1}^N \frac{p_{Y|X}(Y_j|x)}{p_Y(Y_j)}\right)} \middle| U_P = 1, X = x, Y = y\right]\right)^{-1} && (124)
\end{aligned}$$

By defining $\beta(Y_j) = \frac{Q_{Y|Z}(Y_j|z)}{p_Y(Y_j)}$ and $\lambda(Y_j) = \frac{p_{Y|X}(Y_j|x)}{p_Y(Y_j)}$ and leading to the same sequence of stepsilon that lead to (101) in Prop. 5 in this document, we can complete the proof.

## 15 Proof of Prop. 2 in the Main document

**Proposition.**

$$\Pr(U_p \neq U_q) \leq E_{V,W,T}\left[1-\left(1+(1+\epsilon)L^{-1}2^{i(W;V|T)}\right)^{-1}\right] \tag{125}$$

*where* $i_{W,V|T}(w;v|t)=\log\frac{p_{W|V}(w|v)}{p_{W|T}(w|t)}$ *is the conditional information density.*

By application of the conditional IML in Theorem 3 in the main document, and assuming that $N$ is sufficiently large, as stated in (21), it follows that:

$$\begin{aligned}
&\Pr(U_p \neq U_q|U_p=k, X=v, Y_k=(w,l), Z=(t,l)) \\
&\leq 1-\left(1+(1+\epsilon)\frac{p_{Y|X}(w,l|v)}{Q_{Y|Z}(w,l|t,l)}\right)^{-1} &(126)\\
&= 1-\left(1+(1+\epsilon)\frac{p_{W|V}(w|v)p_{\mathsf{l}}(l)}{p_{W|T}(w|t)}\right)^{-1} &(127)\\
&= 1-\left(1+(1+\epsilon)L^{-1}2^{i_{W,V|T}(w;v|t)}\right)^{-1} &(128)
\end{aligned}$$

where $i_{W,V|T}(w;v|t)=\log\frac{p_{W|V}(w|v)}{p_{W|T}(w|t)}$ is the conditional information density. It thus follows that

$$\Pr(U_p \neq U_q) \leq E_{V,W,T}\left[1-\left(1+(1+\epsilon)L^{-1}2^{i(W;T|V)}\right)^{-1}\right]. \tag{129}$$

## 16 Bound on the Probability of Excess Distortion

We introduce and prove the following bound on the probability of excess distortion.

**Proposition.** *For a large enough* $N$*, the probability of excess distortion* $\Pr(d(V,\hat{V})>D)$*, where* $\hat{V}=\tilde{g}(W_{U_q},T)$ *is the reconstruction output by the decoder, can be bound as follow.*

$$P_e=\Pr(d(V,\hat{V})>D) \leq E_{W,V,T}\left\{1-\mathbb{I}(d(V,\tilde{g}(W,T))\leq D)\left(1+(1+\epsilon)L^{-1}2^{i(W;T|V)}\right)^{-1}\right\} \tag{130}$$

*where* $i_{W,T|V}(w;t|v)=\log\frac{p_{W|V}(w|v)}{p_{W|T}(w|t)}$ *is the conditional information density,* $d(.,.)$ *is the distortion measure.*

Proof: We recall that for sufficiently large $N$, $p_{W_{U_p}|V}$ can be arbitrarily close to $p_{W|V}$. As such:

$$\begin{aligned}
P_e &= 1-\Pr(d(V,\hat{V})\leq D) &(131)\\
&\overset{(a)}{\leq} 1-\Pr(d(V,\hat{V})\leq D, W_{U_p}=W_{U_q}) &(132)\\
&= 1-\Pr(d(V,\tilde{g}(W_{U_q},T))\leq D, W_{U_p}=W_{U_q}) &(133)\\
&\overset{(b)}{=} 1-\Pr(d(V,\tilde{g}(W_{U_p},T))\leq D, W_{U_p}=W_{U_q}) &(134)\\
&\overset{(c)}{=} 1-E_{W,V,T}\{\Pr(d(V,\tilde{g}(W_{U_p},T))\leq D, W_{U_p}=W_{U_q}|W_{U_p}=W,V,T)\} &(135)\\
&\overset{(d)}{=} 1-E_{W,V,T}\{\Pr(d(V,\tilde{g}(W,T))\leq D|W_{U_p}=W_{U_q}=W,V,T)\Pr\{W_{U_p}=W_{U_q}|W_{U_p}=W,V,T)\}\} &(136)\\
&\overset{(e)}{=} 1-E_{W,V,T}\{\mathbb{I}(d(V,\tilde{g}(W,T))\leq D)\Pr\{W_{U_p}=W_{U_q}|W_{U_p}=W,V,T)\}\} &(137)\\
&\overset{(f)}{\leq} E_{W,V,T}\{1-\mathbb{I}(d(V,\tilde{g}(W,T))\leq D)\Pr\{U_p=U_q|W_{U_p}=W,V,T\}\} &(138)\\
&\overset{(g)}{\leq} E_{W,V,T}\left\{1-\mathbb{I}(d(V,\tilde{g}(W,T))\leq D)\left(1+(1+\epsilon)L^{-1}2^{i(W;V|T)}\right)^{-1}\right\} &(139)
\end{aligned}$$

where (a) is by marginalization, (b) is by $W_{U_p} = W_{U_q}$, (c) is by the law of iterated expectation , (d) is by chain rule for joint probability, (e) the event $d(V, \tilde{g}(W,T)) \leq D$ is a function of the conditioned random variables and therefore the probability $\Pr(d(V, \tilde{g}(W,T)) \leq D | W_{U_p}{=}W_{U_q}{=}W, V, T)$ becomes the indicator function $\mathbb{I}(d(V, \tilde{g}(W,T)){\leq}D)$, (f) the event $\{W_{U_p} = W_{U_q}\} = \{U_p = U_q\} \cup \{U_p \neq U_q \text{ and } W_{U_p} = W_{U_q}\}$. For (g), we note that:

$$\begin{aligned}
\Pr\{U_p{=}U_q | W_{U_p}{=}w, v, t\} &= E_{\mathsf{k},\mathsf{l}}[\Pr\{U_p{=}U_q | W_{U_p}{=}w, v, t, U_p = \mathsf{k}, l_{U_p} = \mathsf{l}\}] && (140)\\
&= E_{\mathsf{k},\mathsf{l}}[\Pr\{U_p{=}U_q | U_p = \mathsf{k}, Y_\mathsf{k} = (W_{U_p}, l_{U_p}) = (w, \mathsf{l}), Z = (t, \mathsf{l})\}] && (141)\\
&= E_{\mathsf{k},\mathsf{l}}[1 - \Pr\{U_p{\neq}U_q | U_p = \mathsf{k}, Y_\mathsf{k} = (W_{U_p}, l_{U_p}) = (w, \mathsf{l}), Z = (t, \mathsf{l})\}] && (142)\\
&\geq E_{\mathsf{k},\mathsf{l}}\left[\left(1 + (1+\epsilon)L^{-1}2^{i(w;v|t)}\right)^{-1}\right] && (143)\\
&= \left(1 + (1+\epsilon)L^{-1}2^{i(w;v|t)}\right)^{-1} && (144)
\end{aligned}$$

where the first line (equality) is by the law of iterated expectation over $(U_p = \mathsf{k}, l_{U_p} = \mathsf{l})$; in the second line (equality) we rewrite it in the form of Thm. 3 (conditional importance matching lemma); the third line (equality) is because the two event $\{U_p{=}U_q\}$ and $\{U_p{\neq}U_q\}$ are complementary; the fourth line (inequality) is by applying Prop. 2; the final line (equality) is because the term inside the bracket does not depend on $\mathsf{k}$ and $\mathsf{l}$. Following this, we arrive at (g) above.

## 17 Rate Distortion Analysis for Feedback Scheme

We present the rate-distortion analysis for lossy compression with side information with feedback. Recall that in the feedback scheme, after sending the LSB of size $\log_2(L)$ of $U_p$ to the decoder, the encoder will outputs an acknowledgement bit of 1 if the feedback signal indicates that the decoder outputs the same index, i.e. $U_q = U_p$. On the other hand, if the signal indicates $U_q \neq U_p$, the encoder outputs the MSB of its selection $U_p$ to the decoder. This means that the encoder message will be $\log_2(L) + 1$ if the decoder outputs the correct index in the first try and $\log_2(N)$ otherwise.

Assuming perfect feedback, the output distribution between the encoder and decoder is the same, i.e. $\tilde{p}_{W|V}$, since the decoder always recover the correct $W_{U_p}$. Here, we note that the output distribution $\tilde{p}_{W|V}$ correspond to the number of samples $N$, and achieve the expected distortion $D = E[d(V, \hat{V})]$. Since each $L$ yields different matching probability, the rate $R(D)$ in this case is:

$$R(D) = \min_L R(D, L) \tag{145}$$

where:

$$\begin{aligned}
R(D, L) &= [\log_2(L) + 1][1 - \Pr(U_p \neq U_q)] + \log_2(N) \Pr(U_p \neq U_q) && (146)\\
&= \log_2(L) + 1 + (\log_2 N - \log_2 L - 1) \Pr(U_p \neq U_q) && (147)\\
&\leq \log_2(2L) + (\log_2 \frac{N}{2L}) E_{V,W,T}\left[1 - \left(1 + (1+\epsilon)L^{-1}2^{i(W;V|T)}\right)^{-1}\right] && (148)
\end{aligned}$$

## 18 Proof of Equation (36) in Section 8.1 in this document

We will show that:

$$E[\log K] \leq D(\nu(\cdot)||\mu(\cdot)) + \underbrace{\frac{\log e}{e} + 1}_{=\delta}. \tag{149}$$

Our proof directly follows (Li, 2017, Chapter 4) with a change in notation. We note that in Li (2017), $\nu(\cdot) = p_{Y|X}(\cdot|x)$ and $\mu = p_Y(\cdot)$. However $p_{Y|X}(\cdot|x)$ and $p_Y(\cdot)$ are treated as arbitrary distributions in the proof and

fact that $p_Y(\cdot)$ is related to $p_{Y|X}(\cdot|x)$ through marginalization is not required. We thus provide the proof using the notation in the present paper.

Without loss of generality we assume $\Omega = \{1, 2, \ldots, N\}$.

We introduce:

$$I = \arg\min_{y\in\Omega} \frac{S_y}{\nu(y)} \qquad \Theta = \min_{y\in\Omega} \frac{S_y}{\nu(y)} \tag{150}$$

It follows from the exponential-race property (Maddison et al., 2014) that (1) $Y = I \sim \nu(\cdot)$ (2) $\Theta \sim \text{Exp}(1)$ and (3) (1) $Y$ and $\Theta$ are mutually independent of each other.

Note note with $\phi_y = \frac{S_y}{\mu(y)}$, and since $K$ is the index of $\phi_Y$ in $\{\phi_y\}_{y\in\mathcal{Y}}$ sorted in ascending order:

$$K = \{y' : \phi_{y'} < \phi_y\} + 1 \tag{151}$$

$$E[\log K] = \sum_{y\in\Omega} \nu(y) E[\log K|Y = y] \tag{152}$$

$$= \sum_{y\in\Omega} \nu(y) E_\Theta[\log K|Y = y, \Theta = \theta] \tag{153}$$

$$= \sum_{y\in\Omega} \nu(y) \int_{\theta=0}^{\infty} e^{-\theta} E[\log K|Y = y, \Theta = \theta] d\theta. \tag{154}$$

Now consider:

$$E[\log K|Y = y, \Theta = \theta] = E\left[\log\left|y' \neq y : \phi_{y'} < \phi_y\right| + 1|Y = y, \Theta = \theta\right] \tag{155}$$

$$\leq \log E\left[\left|y' \neq y : \phi_{y'} < \phi_y\right| + 1|Y = y, \Theta = \theta\right] \tag{156}$$

$$= \log E\left[\left|y' \neq y : \phi_{y'} < \theta\frac{\nu(y)}{\mu(y)}\right| + 1\middle|Y = y, \Theta = \theta\right] \tag{157}$$

$$= \log E\left[\left|y' \neq y : \phi_{y'} < \theta\frac{\nu(y)}{\mu(y)}\right| + 1\middle|Y = y, \frac{S_{y'}}{\nu(y')} \geq \theta, \forall y' \in \Omega\right] \tag{158}$$

$$= \log E\left[\sum_{y'\neq y} \mathbb{I}\left\{\phi_{y'} < \theta\frac{\nu(y)}{\mu(y)}\right\} + 1\middle|Y = y, \frac{S_{y'}}{\nu(y')} \geq \theta, \forall y' \in \Omega\right] \tag{159}$$

$$= \log\left(\sum_{y'\neq y} \Pr\left(\phi_{y'} \leq \theta\frac{\nu(y)}{\mu(y)}\middle|Y = y, \frac{S_{y'}}{\nu(y')} \geq \theta, \forall y' \in \Omega\right) + 1\right) \tag{160}$$

$$= \log\left(\sum_{y'\neq y} \Pr\left(S_{y'} \leq \theta\frac{\nu(y)}{\mu(y)}\mu(y')\middle|Y = y, \frac{S_{y'}}{\nu(y')} \geq \theta, \forall y' \in \Omega\right) + 1\right) \tag{161}$$

$$= \log\left(\sum_{y'\neq y} \Pr\left(S_{y'} \leq \theta\frac{\nu(y)}{\mu(y)}\mu(y')\middle|\frac{S_{y'}}{\nu(y')} \geq \theta\right) + 1\right) \tag{162}$$

where (157) uses the fact that $\phi_y\mu(y) = \theta\nu(y) = S_y$ and (158) follows form the fact that the event $\{Y = y, \Theta = \theta\}$ is equivalent to the event $\{Y = y, \frac{S_{y'}}{\mu(y')} \geq \theta, \forall y' \in \Omega\}$ by definition of $\Theta$. Eq. (162) follows from the fact that the distribution of $S_{y'}$ only depends on the event $\{\frac{S_{y'}}{\nu(y')} \geq \theta\}$ given the conditioning in (161).

Next we define $r(y) = \nu(y)/\mu(y)$ for each $y \in \Omega$ and use the fact that $S_{y'} \sim \text{Exp}(1)$ so that:

$$\Pr\left(S_{y'} \leq \theta\frac{\nu(y)}{\mu(y)}\mu(y')\middle|\frac{S_{y'}}{\nu(y')} \geq \theta\right) \leq \mathbb{I}(r(y') \leq r(y))\left(1 - \exp(-\theta\mu(y')(r(y) - r(y')))\right) \tag{163}$$

$$\leq \mathbb{I}(r(y') \leq r(y))\theta\mu(y')(r(y) - r(y')) \tag{164}$$

$$\leq \theta\mu(y')r(y). \tag{165}$$

Thus using (162), we have:

$$\log\left(\sum_{y'\neq y}\Pr\left(S_{y'}\leq\theta\frac{\nu(y)}{\mu(y)}\mu(y')\Big|\frac{S_{y'}}{\nu(y')}\geq\theta\right)+1\right)\leq\log(\theta r(y)\sum_{y'}\mu(y')+1) \tag{166}$$

$$=\log(\theta r(y)+1) \tag{167}$$

Using (154) we have that:

$$E[\log K]\leq\sum_{y\in\Omega}\nu(y)\int_{\theta\geq 0}e^{-\theta}\log(\theta r(y)+1)d\theta \tag{168}$$

$$\leq\sum_{y\in\Omega}\nu(y)\log(r(y)+1) \tag{169}$$

$$=\sum_{y\in\Omega:r(y)\geq 1}\nu(y)\log(r(y)+1)+\sum_{y\in\Omega:r(y)\leq 1}\nu(y)\log(r(y)+1) \tag{170}$$

$$\leq\sum_{y\in\Omega:r(y)\geq 1}\nu(y)\log(r(y)+1)+\sum_{y\in\Omega:r(y)\leq 1}\nu(y) \tag{171}$$

$$\leq\sum_{y\in\Omega:r(y)\geq 1}\nu(y)(\log r(y)+1)+\sum_{y\in\Omega:r(y)\leq 1}\nu(y) \tag{172}$$

$$=\sum_{y\in\Omega:r(y)\geq 1}\nu(y)\log r(y)+1 \tag{173}$$

$$=\sum_{y\in\Omega}\nu(y)\log r(y)-\sum_{y\in\Omega:r(y)<1}\nu(y)\log r(y)+1 \tag{174}$$

$$=D(\nu(\cdot)||\mu(\cdot))-\sum_{y\in\Omega:r(y)<1}\nu(y)\log r(y)+1 \tag{175}$$

$$\leq D(\nu(\cdot)||\mu(\cdot))+\frac{\log e}{e}+1 \tag{176}$$

where we use Jensen's inequality in (169) and the following inequality in (Harsha et al., 2007, Appendix A): For any two distributions $P(\cdot)$ and $Q(\cdot)$ on $\mathcal{X}$ and any $\mathcal{X}'\subset\mathcal{X}$

$$-\sum_{x\in\mathcal{X}'}P(x)\log\frac{P(x)}{Q(x)}\leq\frac{\log e}{e}. \tag{177}$$

This completes the proof.

## 19 Proof of Prop. 3 in Section 8.2 in this document

For simplicity in notation we will use $p_i=p_Y(Y_i)$ and $q_i=p_Y(Y_i|X=x)$. Thus the objective we need to simplify reduces to

$$E_{Y_1^N}\left[\sum_{i=1}^{N}\frac{\frac{q_i}{p_i}\log\left(N\frac{\frac{q_i}{p_i}}{\sum_{j=1}^{N}\frac{q_j}{p_j}}\right)}{\sum_{j=1}^{N}\frac{q_j}{p_j}}\right]$$

$$=E_{Y_1^N}\left[\sum_{i=1}^{N}\frac{\frac{q_i}{p_i}\log\left(\frac{N}{\sum_{j=1}^{N}\frac{q_j}{p_j}}\right)}{\sum_{j=1}^{N}\frac{q_j}{p_j}}\right]+E_{Y_1^N}\left[\sum_{i=1}^{N}\frac{\frac{q_i}{p_i}\log\left(\frac{q_i}{p_i}\right)}{\sum_{j=1}^{N}\frac{q_j}{p_j}}\right] \tag{178}$$

$$=E_{Y_1^N}\left[\log\left(\frac{N}{\sum_{j=1}^{N}\frac{q_j}{p_j}}\right)\right]+E_{Y_1^N}\left[\sum_{i=1}^{N}\frac{\frac{q_i}{p_i}\log\left(\frac{q_i}{p_i}\right)}{\sum_{j=1}^{N}\frac{q_j}{p_j}}\right] \tag{179}$$

We will establish an upper bound on each of the two terms in (179) separately. Consider the first term:

$$E_{Y_1^N}\left[\log\left(\frac{N}{\sum_{j=1}^N \frac{q_j}{p_j}}\right)\right] \leq \log\left(E_{Y_1^N}\left[\frac{N}{\sum_{j=1}^N \frac{q_j}{p_j}}\right]\right) \tag{180}$$

which follows from Jensen's inequality. Now consider:

$$\begin{aligned}
E_{Y_1^N}\left[\frac{N}{\sum_{i=1}^N \frac{q_i}{p_i}}\right] - 1 &= E_{Y_1^N}\left[\frac{1}{\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}} - 1\right] && (181)\\
&= E_{Y_1^N}\left[\frac{1}{\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}}\left(1 - \frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}\right)\right] && (182)\\
&= E_{Y_1^N}\left[\left(\frac{1}{\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}} - 1\right)\left(1 - \frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}\right)\right] && (183)\\
&= E_{Y_1^N}\left[\left(\frac{1}{\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}}\right)\left(1 - \frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}\right)^2\right] && (184)\\
& && (185)
\end{aligned}$$

where (183) follows from the fact that $E_{Y_1^N}\left[1 - \frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}\right] = 0$. Next observe that:

$$\begin{aligned}
&E_{Y_1^N}\left[\left(\frac{1}{\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}}\right)\left(1 - \frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}\right)^2\right]\\
&= E_{Y_1^N}\left[\frac{1}{\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}}\left(1 - \frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}\right)^2 \mathbb{I}\left(\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i} \geq \frac{1}{2}\right)\right]\\
&+ E_{Y_1^N}\left[\frac{1}{\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}}\left(1 - \frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}\right)^2 \mathbb{I}\left(\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i} < \frac{1}{2}\right)\right] && (186)
\end{aligned}$$

and by using the triangular inequality, we have that:

$$\begin{aligned}
&\left|E_{Y_1^N}\left[\frac{N}{\sum_{i=1}^N \frac{q_i}{p_i}}\right] - 1\right|\\
&\leq \left|E_{Y_1^N}\left[\frac{1}{\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}}\left(1 - \frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}\right)^2 \mathbb{I}\left(\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i} \geq \frac{1}{2}\right)\right]\right|\\
&+ \left|E_{Y_1^N}\left[\frac{1}{\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}}\left(1 - \frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}\right)^2 \mathbb{I}\left(\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i} < \frac{1}{2}\right)\right]\right| && (187)
\end{aligned}$$

We consider each of the two terms in (187) separately. For the first term:

$$E_{Y_1^N}\left[\frac{1}{\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}}\left(1-\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}\right)^2 \mathbb{I}\left(\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}\geq \frac{1}{2}\right)\right]$$

$$\leq 2E_{Y_1^N}\left[\left(1-\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}\right)^2 \mathbb{I}\left(\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}\geq \frac{1}{2}\right)\right] \tag{188}$$

$$\leq 2E_{Y_1^N}\left[\left(1-\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}\right)^2\right] \tag{189}$$

$$= \frac{2}{N}\left(E_{Y\sim p(Y)}\left[\frac{q^2(Y)}{p^2(Y)}\right]-1\right) = \frac{2}{N}\left(d_2(q||p)-1\right) \tag{190}$$

where $d_2(q||p) = E_q\left[\frac{q}{p}\right]$. We next consider the second term in (187).

$$E_{Y_1^N}\left[\frac{1}{\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}}\left(1-\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}\right)^2 \mathbb{I}\left(\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}< \frac{1}{2}\right)\right]$$

$$\leq E_{Y_1^N}\left[\frac{1}{\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}}\mathbb{I}\left(\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}< \frac{1}{2}\right)\right] \tag{191}$$

We will next use the following key inequality established at the end of this section:

$$\frac{1}{\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}} \leq \frac{1}{N}\sum_{i=1}^N \frac{p_i}{q_i} \tag{192}$$

Also using $E_p\left[\frac{p}{q}\right] = d_2(p||q)$, we have

$$E_{Y_1^N}\left[\frac{1}{\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}}\mathbb{I}\left(\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}< \frac{1}{2}\right)\right] \leq E_{Y_1^N}\left[\left(\frac{1}{N}\sum_{i=1}^N \frac{p_i}{q_i}\right)\mathbb{I}\left(\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}< \frac{1}{2}\right)\right] \tag{193}$$

$$= E_{Y_1^N}\left[\left(\frac{1}{N}\sum_{i=1}^N \frac{p_i}{q_i} - d_2(p||q)\right)\mathbb{I}\left(\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}< \frac{1}{2}\right)\right] + d_2(p||q)\Pr\left(\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}< \frac{1}{2}\right) \tag{194}$$

$$\leq \sqrt{E_{Y_1^N}\left[\left(\frac{1}{N}\sum_{i=1}^N \frac{p_i}{q_i} - d_2(p||q)\right)^2\right]}\sqrt{E_{Y_1^N}\left[\mathbb{I}\left(\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}< \frac{1}{2}\right)\right]} + d_2(p||q)\Pr\left(\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}< \frac{1}{2}\right) \tag{195}$$

$$= \sqrt{E_{Y_1^N}\left[\left(\frac{1}{N}\sum_{i=1}^N \frac{p_i}{q_i} - d_2(p||q)\right)^2\right]}\sqrt{\Pr\left(\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}< \frac{1}{2}\right)} + d_2(p||q)\Pr\left(\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}< \frac{1}{2}\right) \tag{196}$$

where (195) follows from the Cauchy-Schwartz inequality: $E[X\cdot Y] \leq \sqrt{E[X^2]E[Y^2]}$. Using Chebyshev's Inequality, we have the following:

$$\Pr\left(\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}< \frac{1}{2}\right) \leq \Pr\left(\left|\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i} - 1\right| \geq \frac{1}{2}\right) \tag{197}$$

$$\leq \frac{\frac{1}{N}E_p\left[\left(\frac{q}{p}-1\right)^2\right]}{1/4} \tag{198}$$

$$= \frac{4}{N}\left(d_2(q||p)-1\right). \tag{199}$$

Finally note that:

$$E_{Y_1^N}\left[\left(\frac{1}{N}\sum_{i=1}^{N}\frac{p_i}{q_i}-d_2(p||q)\right)^2\right]=\frac{1}{N}\left(E_p\left[\frac{p^2}{q^2}\right]-d_2^2(p||q)\right)=\frac{1}{N}\left(d_3(p||q)-d_2^2(p||q)\right) \tag{200}$$

where $d_3(p||q)=E_p\left[\frac{p^2}{q^2}\right]$ as in (33). It follows that:

$$\begin{aligned}&E_{Y_1^N}\left[\frac{1}{\frac{1}{N}\sum_{i=1}^{N}\frac{q_i}{p_i}}\mathbb{I}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{q_i}{p_i}<\frac{1}{2}\right)\right]\\&\leq\frac{2}{N}\left((d_3(p||q)-d_2^2(p||q))(d_2(q||p)-1)\right)^{\frac{1}{2}}+\frac{4}{N}\left(d_2(q||p)-1\right)d_2(p||q)\end{aligned} \tag{201}$$

Collecting all the terms we have that:

$$\begin{aligned}E_{Y_1^N}\left[\frac{N}{\sum_{i=1}^{N}\frac{q_i}{p_i}}\right]&\leq 1+\frac{2}{N}\left(d_2(q||p)-1\right)\\&+\frac{2}{N}\left((d_3(p||q)-d_2^2(p||q))(d_2(q||p)-1)\right)^{\frac{1}{2}}+\frac{4}{N}\left(d_2(q||p)-1\right)d_2(p||q)\end{aligned} \tag{202}$$

Finally by using the fact that $d_2(q||p)\leq\omega$ and defining

$$\alpha(p,q)=2(\omega-1)+2\sqrt{\omega-1}(d_3(p||q)-d_2^2(p||q))^{\frac{1}{2}}+4\omega\cdot d_2(p||q) \tag{203}$$

we have that:

$$\log E_{Y_1^N}\left[\frac{N}{\sum_{i=1}^{N}\frac{q_i}{p_i}}\right]\leq\log\left(1+\frac{\alpha(p,q)}{N}\right)\leq\frac{\alpha(p,q)}{N} \tag{204}$$

We will now consider the second term in (179). Consider the following:

$$E_{Y_1^N}\left[\sum_{i=1}^{N}\frac{\frac{q_i}{p_i}\log\left(\frac{q_i}{p_i}\right)}{\sum_{j=1}^{N}\frac{q_j}{p_j}}-\frac{\sum_{i=1}^{N}\frac{q_i}{p_i}\log\frac{q_i}{p_i}}{N}\right]=E_{Y_1^N}\left[\frac{\frac{1}{N}\sum_{i=1}^{N}\frac{q_i}{p_i}\log\frac{q_i}{p_i}}{\frac{1}{N}\sum_{i=1}^{N}\frac{q_i}{p_i}}\left(1-\frac{\sum_{i=1}^{N}\frac{q_i}{p_i}}{N}\right)\right] \tag{205}$$

Now we consider the following:

$$\begin{aligned}&E\left[\left(\frac{\frac{1}{N}\sum_{i=1}^{N}\frac{q_i}{p_i}\log\frac{q_i}{p_i}}{\frac{1}{N}\sum_{i=1}^{N}\frac{q_i}{p_i}}\right)\left(1-\frac{\sum_{i=1}^{N}\frac{q_i}{p_i}}{N}\right)\right]\\&=E\left[\left(\frac{\frac{1}{N}\sum_{i=1}^{N}\frac{q_i}{p_i}\log\frac{q_i}{p_i}}{\frac{1}{N}\sum_{i=1}^{N}\frac{q_i}{p_i}}-D(q||p)\right)\left(1-\frac{\sum_{i=1}^{N}\frac{q_i}{p_i}}{N}\right)\right] &(206)\\&=E\left[\left(\frac{\frac{1}{N}\sum_{i=1}^{N}\frac{q_i}{p_i}\log\frac{q_i}{p_i}-D(q||p)\frac{1}{N}\sum_{i=1}^{N}\frac{q_i}{p_i}}{\frac{1}{N}\sum_{i=1}^{N}\frac{q_i}{p_i}}\right)\left(1-\frac{\sum_{i=1}^{N}\frac{q_i}{p_i}}{N}\right)\right] &(207)\\&=E\left[\left(\frac{\frac{1}{N}\sum_{i=1}^{N}\frac{q_i}{p_i}\log\frac{q_i}{p_i}-D(q||p)\frac{1}{N}\sum_{i=1}^{N}\frac{q_i}{p_i}}{\frac{1}{N}\sum_{i=1}^{N}\frac{q_i}{p_i}}\right)\left(1-\frac{\sum_{i=1}^{N}\frac{q_i}{p_i}}{N}\right)\mathbb{I}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{q_i}{p_i}\geq\frac{1}{2}\right)\right]\\&\quad+E\left[\left(\frac{\frac{1}{N}\sum_{i=1}^{N}\frac{q_i}{p_i}\log\frac{q_i}{p_i}-D(q||p)\frac{1}{N}\sum_{i=1}^{N}\frac{q_i}{p_i}}{\frac{1}{N}\sum_{i=1}^{N}\frac{q_i}{p_i}}\right)\left(1-\frac{\sum_{i=1}^{N}\frac{q_i}{p_i}}{N}\right)\mathbb{I}\left(\frac{1}{N}\sum_{i=1}^{N}\frac{q_i}{p_i}<\frac{1}{2}\right)\right] &(208)\end{aligned}$$

Here (206) follows from the fact that $E\left(1-\frac{\sum_{i=1}^N \frac{q_i}{p_i}}{N}\right)=0$. Using triangular inequality we have that:

$$\begin{aligned}
&\left|E_{Y_1^N}\left[\sum_{i=1}^N \frac{\frac{q_i}{p_i}\log\left(\frac{q_i}{p_i}\right)}{\sum_{j=1}^N \frac{q_j}{p_j}} - \frac{\sum_{i=1}^N \frac{q_i}{p_i}\log\frac{q_i}{p_i}}{N}\right]\right| \leq \\
&\quad\left|E\left[\left(\frac{\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}\log\frac{q_i}{p_i} - D(q||p)\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}}{\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}}\right)\left(1-\frac{\sum_{i=1}^N \frac{q_i}{p_i}}{N}\right)\mathbb{I}\left(\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}\geq\frac{1}{2}\right)\right]\right| \\
&\qquad+\left|E\left[\left(\frac{\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}\log\frac{q_i}{p_i} - D(q||p)\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}}{\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}}\right)\left(1-\frac{\sum_{i=1}^N \frac{q_i}{p_i}}{N}\right)\mathbb{I}\left(\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}<\frac{1}{2}\right)\right]\right| \qquad (209)
\end{aligned}$$

We will bound the two terms above separately. Now note that:

$$\begin{aligned}
&\left|E\left[\left(\frac{\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}\log\frac{q_i}{p_i} - D(q||p)\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}}{\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}}\right)\left(1-\frac{\sum_{i=1}^N \frac{q_i}{p_i}}{N}\right)\mathbb{I}\left(\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}\geq\frac{1}{2}\right)\right]\right| \\
&\leq 2E\left[\left|\left(\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}\log\frac{q_i}{p_i} - D(q||p)\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}\right)\right|\left|\left(1-\frac{\sum_{i=1}^N \frac{q_i}{p_i}}{N}\right)\right|\right] \qquad (210)\\
&\leq 2\sqrt{E\left[\left(\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}\log\frac{q_i}{p_i} - D(q||p)\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}\right)^2\right]}\sqrt{E\left[\left(1-\frac{\sum_{i=1}^N \frac{q_i}{p_i}}{N}\right)^2\right]} \qquad (211)
\end{aligned}$$

Here the last step follows from Chebyshev's inequality. Next note that:

$$E\left[\left(1-\frac{\sum_{i=1}^N \frac{q_i}{p_i}}{N}\right)^2\right] = \frac{1}{N}\left(d_2(q||p)-1\right) \qquad (212)$$

and

$$\begin{aligned}
&E\left[\left(\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}\log\frac{q_i}{p_i} - D(q||p)\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}\right)^2\right] \\
&= E\left[\left(\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}\log\frac{q_i}{p_i} - D(q||p) + D(q||p) - D(q||p)\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}\right)^2\right] \qquad (213)\\
&\leq 2E\left[\left(\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}\log\frac{q_i}{p_i} - D(q||p)\right)^2\right] + 2E\left[\left(D(q||p)\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i} - D(q||p)\right)^2\right] \qquad (214)\\
&= \frac{2}{N}\left(E_p\left[\left(\frac{q}{p}\log\frac{q}{p}\right)^2\right] - D^2(q||p)\right) + \frac{2}{N}D^2(q||p)(d_2(q||p)-1) \qquad (215)
\end{aligned}$$

Thus we have that:

$$\begin{aligned}
&\left|E\left[\left(\frac{\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}\log\frac{q_i}{p_i} - D(q||p)\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}}{\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}}\right)\left(1-\frac{\sum_{i=1}^N \frac{q_i}{p_i}}{N}\right)\mathbb{I}\left(\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}\geq\frac{1}{2}\right)\right]\right| \\
&\leq \frac{2}{N}\sqrt{\left(2\left(E_p\left[\left(\frac{q}{p}\log\frac{q}{p}\right)^2\right] - D^2(q||p)\right) + 2D^2(q||p)(d_2(q||p)-1)\right)(d_2(q||p)-1)} \qquad (216)\\
&= \frac{2}{N}g(p,q) \qquad (217)
\end{aligned}$$

where

$$g(p,q) = \sqrt{\left(2\left(E_p\left[\left(\frac{q}{p}\log\frac{q}{p}\right)^2\right] - D^2(q||p)\right) + 2D^2(q||p)(d_2(q||p)-1)\right)(d_2(q||p)-1)} \tag{218}$$

Furthermore using the fact that $q(y)/p(y) \leq \omega$ we can show that:

$$g(p,q) \leq 2(\omega-1)\log\omega. \tag{219}$$

In a similar manner consider:

$$\left|E\left[\left(\frac{\frac{1}{N}\sum_{i=1}^N \frac{q_i}{p_i}\log\frac{q_i}{p_i} - D(q||p)\frac{1}{N}\sum_{i=1}^N\frac{q_i}{p_i}}{\frac{1}{N}\sum_{i=1}^N\frac{q_i}{p_i}}\right)\left(1-\frac{\sum_{i=1}^N\frac{q_i}{p_i}}{N}\right)\mathbb{I}\left(\frac{1}{N}\sum_{i=1}^N\frac{q_i}{p_i}<\frac{1}{2}\right)\right]\right|$$

$$\leq (\log\omega)\, E\left[\mathbb{I}\left(\frac{1}{N}\sum_{i=1}^N\frac{q_i}{p_i}<\frac{1}{2}\right)\right] \tag{220}$$

$$\leq \frac{\log\omega}{N}\left(4\left(d_2(q||p)-1\right)\right) = \frac{h(p,q)}{N} \tag{221}$$

where we used the bound in (199) and define

$$\begin{aligned} h(p,q) &= \log\omega\left(4\left(d_2(q||p)-1\right)\right) \\ &\leq 4(\omega-1)\log\omega. \end{aligned} \tag{222}$$

We have established that:

$$E_{Y_1^N}\left[\sum_{i=1}^N \frac{\frac{q_i}{p_i}\log\left(\frac{q_i}{p_i}\right)}{\sum_{j=1}^N\frac{q_j}{p_j}}\right] \leq E_{Y_1^N}\left[\frac{\sum_{i=1}^N\frac{q_i}{p_i}\log\frac{q_i}{p_i}}{N}\right] + \frac{6(\omega-1)\log\omega}{N} \tag{223}$$

$$= E_Y\left[\frac{p_{Y|X}(Y|x)}{p_Y(Y)}\log\frac{p_{Y|X}(Y|x)}{p_Y(Y)}\right] + \frac{6(\omega-1)\log\omega}{N} \tag{224}$$

$$= D(p_{Y|X}(\cdot|X=x)||p_Y(\cdot)) + \frac{6(\omega-1)\log\omega}{N} \tag{225}$$

Thus collecting (179), (204) and (225) we have:

$$E_{Y_1^N}\left[\sum_{i=1}^N \frac{\frac{q_i}{p_i}\log\left(N\frac{\frac{q_i}{p_i}}{\sum_{j=1}^N\frac{q_j}{p_j}}\right)}{\sum_{j=1}^N\frac{q_j}{p_j}}\right] \leq D(p_{Y|X}(\cdot|X=x)||p_Y(\cdot)) + \frac{6(\omega-1)\log\omega}{N} + \frac{\alpha(p_Y(\cdot),p_{Y|X}(\cdot))}{N} \tag{226}$$

where $\alpha(\cdot)$ is defined in (203). It only remains to establish the following inequality stated in (192):

$$\frac{1}{\frac{1}{N}\sum_{i=1}^N\frac{q_i}{p_i}} \leq \frac{1}{N}\sum_{i=1}^N\frac{p_i}{q_i}. \tag{227}$$

Now define $\beta_i = \frac{q_i}{p_i}$ as before. And observe that:

$$\frac{1}{\frac{1}{N}\sum_{i=1}^N\frac{q_i}{p_i}} = \left(\frac{\sum_{i=1}^N\beta_i}{N}\right)\left(\frac{\sum_{i=1}^N\beta_i\frac{p_i}{q_i}}{\sum_{i=1}^N\beta_i}\right)\left(\frac{\sum_{i=1}^N\beta_i\frac{p_i}{q_i}}{\sum_{i=1}^N\beta_i}\right) \tag{228}$$

$$= \left(\frac{\sum_{i=1}^N\beta_i}{N}\right)E_\beta\left[\frac{p}{q}\right]E_\beta\left[\frac{p}{q}\right] \tag{229}$$

$$\leq \frac{\sum_{i=1}^N\beta_i}{N}E_\beta\left[\left(\frac{p}{q}\right)^2\right] \tag{230}$$

$$= \frac{1}{N}\sum_{i=1}^N\frac{p_i}{q_i} \tag{231}$$

where (230) is a consequence of Cauchy-Schwartz inequality.

## 20 Proof of Prop. 4 in Section 9 in this document.

Applying Hoeffding's Inequality we get that:

$$\Pr\left(\frac{1}{N-1}\sum_{j=2}^{N}\frac{p_{Y|X}(Y_j|x)}{p_Y(Y_j)} \leq 1-\epsilon\right) \leq \exp\left(-2(N-1)\epsilon^2/\omega^2\right). \tag{232}$$

Equivalently, if

$$\Omega = \left\{ y_2^N : \frac{1}{N-1}\sum_{j=2}^{N}\frac{p_{Y|X}(y_j|x)}{p_Y(y_j)} \leq 1-\epsilon \right\}$$

then $\Pr(\Omega) \leq \exp\left(-2(N-1)\epsilon^2/\omega^2\right)$.
Now revisiting (41) we have

$$E_{Y_1^N}[N\lambda_1 \log(N\lambda_1)] = E_{Y_1^N}[N\lambda_1 \log(N\lambda_1)|\Omega]\Pr(\Omega) + E_{Y_1^N}[N\lambda_1 \log(N\lambda_1)|\Omega^c]\Pr(\Omega^c) \tag{233}$$

$$\leq E_{Y_1^N}[N\lambda_1 \log(N\lambda_1)|\Omega]\exp\left(-2(N-1)\epsilon^2/\omega^2\right) + E_{Y_1^N}[N\lambda_1 \log(N\lambda_1)|\Omega^c] \tag{234}$$

We bound each of the two terms in (234). First consider

$$E_{Y_1^N}[N\lambda_1 \log(N\lambda_1)|\Omega] = E\left[ N\frac{\frac{p_{Y|X}(Y_1|x)}{p_Y(Y_1)}}{\sum_{j=1}^{N}\frac{p_{Y|X}(Y_j|x)}{p_Y(Y_j)}} \log\left( N\frac{\frac{p_{Y|X}(Y_1|x)}{p_Y(Y_1)}}{\sum_{j=1}^{N}\frac{p_{Y|X}(Y_j|x)}{p_Y(Y_j)}} \right) \middle| \Omega \right] \tag{235}$$

$$\leq N\log N, \tag{236}$$

where the second step uses the fact that $\frac{P_{Y|X}(y|x)}{P_Y(y)} \geq 0$. Furthermore we have that

$$E_{Y_1^N}[N\lambda_1 \log(N\lambda_1)|\Omega^c] = E\left[ N\frac{\frac{p_{Y|X}(Y_1|x)}{p_Y(Y_1)}}{\sum_{j=1}^{N}\frac{p_{Y|X}(Y_j|x)}{p_Y(Y_j)}} \log\left( N\frac{\frac{p_{Y|X}(Y_1|x)}{p_Y(Y_1)}}{\sum_{j=1}^{N}\frac{p_{Y|X}(Y_j|x)}{p_Y(Y_j)}} \right) \middle| \Omega^c \right] \tag{237}$$

$$\leq E_{Y_1\sim P_Y(\cdot)}\left[ \frac{N}{(N-1)(1-\epsilon)}\frac{p_{Y|X}(Y_1|x)}{p_Y(Y_1)} \log \frac{N}{(N-1)(1-\epsilon)}\frac{p_{Y|X}(Y_1|x)}{p_Y(Y_1)} \right] \tag{238}$$

$$= \frac{N}{(N-1)(1-\epsilon)} D(p_{Y|X}(\cdot|x)||p_Y(\cdot)) + \frac{N}{(N-1)(1-\epsilon)} \log \frac{N}{(N-1)(1-\epsilon)} \tag{239}$$

## 21 Proof of Prop. 5 in this document

For simplicity we will denote $\lambda_i = \lambda(Y_i) = \frac{p_{Y|X}(Y_i|X)}{p_Y(Y_i)}$ and $\beta_i = \beta(Y_i) = \frac{q_{Y|X}(Y_i|X=x)}{q_Y(Y_i)}$. For simplicity, we keep the conditioning on $Y_1 = y_1$ implicit. For convenience, we define $\bar{N} = N-1$ and consider the following normalization:

$$E_{Y_2^N}\left[ \frac{\frac{1}{\bar{N}}\sum_{i=1}^{N}\beta_i}{\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right)^2} \right] \tag{240}$$

and let us define:

$$\Omega_1 = E_{Y_2^N}\left[ \frac{1}{\bar{N}}\sum_{i=1}^{N}\beta_i \right] = \frac{\beta_1}{\bar{N}} + 1 \tag{241}$$

as well as

$$\Omega_2 = E_{Y_2^N}\left[\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right)^2\right] \tag{242}$$

$$= \frac{1}{\bar{N}^2}E_{Y_2^N}\left[\left(\lambda_1^2 + \sum_{i=2}^{N}\lambda_i^2 + 2\lambda_1\sum_{j=2}^{N}\lambda_j + 2\sum_{i=2}^{N}\sum_{j=i+1}^{N}\lambda_i\lambda_j\right)\right] \tag{243}$$

$$= \frac{\lambda_1^2}{\bar{N}^2} + \frac{d_2(q||p)}{\bar{N}} + \frac{2\lambda_1}{\bar{N}} + \frac{\bar{N}-1}{\bar{N}} \tag{244}$$

$$= \left(1 + \frac{\lambda_1}{\bar{N}}\right)^2 + \frac{d_2(q||p)-1}{\bar{N}} \tag{245}$$

$$= 1 + \frac{\gamma}{\bar{N}} \tag{246}$$

where we use the fact that $E_{Y_i}[\lambda_i] = 1$ and $E_{Y_i}[\lambda_i^2] = d_2(q||p)$ and define

$$\gamma = 2\lambda_1 + d_2(q||p) - 1 + \frac{\lambda_1^2}{\bar{N}}. \tag{247}$$

Next we consider the following:

$$E_{Y_2^N}\left[\frac{\frac{1}{\bar{N}}\sum_{i=1}^{N}\beta_i}{\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right)^2}\right] - \frac{\Omega_1}{\Omega_2} \tag{248}$$

$$= E_{Y_2^N}\left[\frac{1}{\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right)^2}\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\beta_i - \frac{\Omega_1}{\Omega_2}\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right)^2\right)\right] \tag{249}$$

Now note that:

$$E_{Y_2^N}\left[\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\beta_i - \frac{\Omega_1}{\Omega_2}\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right)^2\right)\right] = 0. \tag{250}$$

Thus (249) is equivalent to:

$$E_{Y_2^N}\left[\left(\frac{1}{\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right)^2} - \frac{1}{c^2}\right)\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\beta_i - \frac{\Omega_1}{\Omega_2}\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right)^2\right)\right] \tag{251}$$

for any constant $c \neq 0$. We select

$$c = E_{Y_2^N}\left[\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right] = \frac{\lambda_1}{\bar{N}} + 1 \tag{252}$$

We can express (251) as follows:

$$E_{Y_2^N}\left[\frac{1}{\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right)^2}\left(1 - \frac{\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right)^2}{c^2}\right)\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\beta_i - \frac{\Omega_1}{\Omega_2}\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right)^2\right)\right]$$

$$= E_{Y_2^N}\left[\frac{1}{\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right)^2}\left(1 - \frac{\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right)^2}{c^2}\right)\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\beta_i - \frac{\Omega_1}{\Omega_2}\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right)^2\right)\mathbb{I}\left(\frac{1}{\bar{N}}\sum_{i=2}^{\bar{N}}\lambda_i \geq \frac{1}{2}\right)\right]$$

$$+ E_{Y_2^N}\left[\frac{1}{\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right)^2}\left(1 - \frac{\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right)^2}{c^2}\right)\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\beta_i - \frac{\Omega_1}{\Omega_2}\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right)^2\right)\mathbb{I}\left(\frac{1}{\bar{N}}\sum_{i=2}^{\bar{N}}\lambda_i < \frac{1}{2}\right)\right] \tag{253}$$

Now consider make use of the triangular inequality:

$$\begin{aligned}
&\left| E_{Y_2^N}\left[\frac{\frac{1}{\bar{N}}\sum_{i=1}^N \beta_i}{\left(\frac{1}{\bar{N}}\sum_{i=1}^N \lambda_i\right)^2}\right] - \frac{\Omega_1}{\Omega_2}\right| \\
&\leq \left| E_{Y_2^N}\left[\frac{1}{\left(\frac{1}{\bar{N}}\sum_{i=1}^N \lambda_i\right)^2}\left(1 - \frac{\left(\frac{1}{\bar{N}}\sum_{i=1}^N \lambda_i\right)^2}{c^2}\right)\left(\frac{1}{\bar{N}}\sum_{i=1}^N \beta_i - \frac{\Omega_1}{\Omega_2}\left(\frac{1}{\bar{N}}\sum_{i=1}^N \lambda_i\right)^2\right)\mathbb{I}\left(\frac{1}{\bar{N}}\sum_{i=2}^{\bar{N}} \lambda_i \geq \frac{1}{2}\right)\right]\right| \\
&+ \left| E_{Y_2^N}\left[\frac{1}{\left(\frac{1}{\bar{N}}\sum_{i=1}^N \lambda_i\right)^2}\left(1 - \frac{\left(\frac{1}{\bar{N}}\sum_{i=1}^N \lambda_i\right)^2}{c^2}\right)\left(\frac{1}{\bar{N}}\sum_{i=1}^N \beta_i - \frac{\Omega_1}{\Omega_2}\left(\frac{1}{\bar{N}}\sum_{i=1}^N \lambda_i\right)^2\right)\mathbb{I}\left(\frac{1}{\bar{N}}\sum_{i=2}^{\bar{N}} \lambda_i < \frac{1}{2}\right)\right]\right| \quad (254)
\end{aligned}$$

We now consider each of the two terms in (254) separately:

$$\begin{aligned}
&\left| E_{Y_2^N}\left[\frac{1}{\left(\frac{1}{\bar{N}}\sum_{i=1}^N \lambda_i\right)^2}\left(1 - \frac{\left(\frac{1}{\bar{N}}\sum_{i=1}^N \lambda_i\right)^2}{c^2}\right)\left(\frac{1}{\bar{N}}\sum_{i=1}^N \beta_i - \frac{\Omega_1}{\Omega_2}\left(\frac{1}{\bar{N}}\sum_{i=1}^N \lambda_i\right)^2\right)\mathbb{I}\left(\frac{1}{\bar{N}}\sum_{i=2}^{\bar{N}} \lambda_i \geq \frac{1}{2}\right)\right]\right| \\
&\leq \frac{4}{c^2}\left| E_{Y_2^N}\left[\left(c^2 - \left(\frac{1}{\bar{N}}\sum_{i=1}^N \lambda_i\right)^2\right)\left(\frac{1}{\bar{N}}\sum_{i=1}^N \beta_i - \frac{\Omega_1}{\Omega_2}\left(\frac{1}{\bar{N}}\sum_{i=1}^N \lambda_i\right)^2\right)\mathbb{I}\left(\frac{1}{\bar{N}}\sum_{i=2}^{\bar{N}} \lambda_i \geq \frac{1}{2}\right)\right]\right| && (255) \\
&\leq \frac{4}{c^2} E_{Y_2^N}\left[\left|\left(c^2 - \left(\frac{1}{\bar{N}}\sum_{i=1}^N \lambda_i\right)^2\right)\right|\left|\left(\frac{1}{\bar{N}}\sum_{i=1}^N \beta_i - \frac{\Omega_1}{\Omega_2}\left(\frac{1}{\bar{N}}\sum_{i=1}^N \lambda_i\right)^2\right)\right|\mathbb{I}\left(\frac{1}{\bar{N}}\sum_{i=2}^{\bar{N}} \lambda_i \geq \frac{1}{2}\right)\right] && (256) \\
&\leq \frac{4}{c^2} E_{Y_2^N}\left[\left|\left(c^2 - \left(\frac{1}{\bar{N}}\sum_{i=1}^N \lambda_i\right)^2\right)\right|\left|\left(\frac{1}{\bar{N}}\sum_{i=1}^N \beta_i - \frac{\Omega_1}{\Omega_2}\left(\frac{1}{\bar{N}}\sum_{i=1}^N \lambda_i\right)^2\right)\right|\right] && (257) \\
&\leq \frac{4}{c^2}\sqrt{E_{Y_2^N}\left[\left(c^2 - \left(\frac{1}{\bar{N}}\sum_{i=1}^N \lambda_i\right)^2\right)^2\right]}\sqrt{E_{Y_2^N}\left[\left(\frac{1}{\bar{N}}\sum_{i=1}^N \beta_i - \frac{\Omega_1}{\Omega_2}\left(\frac{1}{\bar{N}}\sum_{i=1}^N \lambda_i\right)^2\right)^2\right]} && (258)
\end{aligned}$$

We now consider each of the two terms above separately.

$$\begin{aligned}
E_{Y_2^N}\left[\left(c^2 - \left(\frac{1}{\bar{N}}\sum_{i=1}^N \lambda_i\right)^2\right)^2\right] &= E_{Y_2^N}\left[\left(c - \left(\frac{1}{\bar{N}}\sum_{i=1}^N \lambda_i\right)\right)^2\left(c + \left(\frac{1}{\bar{N}}\sum_{i=1}^N \lambda_i\right)\right)^2\right] && (259) \\
&= E_{Y_2^N}\left[\left(1 - \left(\frac{1}{\bar{N}}\sum_{i=2}^N \lambda_i\right)\right)^2\left(c + \left(\frac{1}{\bar{N}}\sum_{i=1}^N \lambda_i\right)\right)^2\right] && (260) \\
&\leq \left(1 + \frac{N+1}{\bar{N}}\omega\right)^2 E_{Y_2^N}\left[\left(1 - \left(\frac{1}{\bar{N}}\sum_{i=2}^N \lambda_i\right)\right)^2\right] && (261) \\
&= \left(1 + \frac{N+1}{\bar{N}}\omega\right)^2 \frac{1}{\bar{N}} E_p\left[\frac{q^2}{p^2} - 1\right] && (262) \\
&= \frac{1}{\bar{N}}\left(1 + \frac{N+1}{\bar{N}}\omega\right)^2 \left(d_2(q||p) - 1\right). && (263)
\end{aligned}$$

We now consider the second term in (258).

$$E_{Y_2^N}\left[\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\beta_i - \frac{\Omega_1}{\Omega_2}\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right)^2\right)^2\right]$$

$$= E_{Y_2^N}\left[\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\beta_i - \Omega_1 + \Omega_1 - \frac{\Omega_1}{\Omega_2}\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right)^2\right)^2\right] \tag{264}$$

$$\leq 2E_{Y_2^N}\left[\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\beta_i - \Omega_1\right)^2\right] + 2\left(\frac{\Omega_1}{\Omega_2}\right)^2 E_{Y_2^N}\left[\left(\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right)^2 - \Omega_2\right)^2\right] \tag{265}$$

We consider each term in (265) separately.

$$E_{Y_2^N}\left[\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\beta_i - \Omega_1\right)^2\right] = E_{Y_2^N}\left[\left(\frac{1}{\bar{N}}\sum_{i=2}^{N}\beta_i - 1\right)^2\right] \tag{266}$$

$$= \frac{1}{\bar{N}}E_p\left[\left(\frac{r}{p}\right)^2 - 1\right] \tag{267}$$

$$= \frac{1}{\bar{N}}\left(d_2(r||p) - 1\right) \tag{268}$$

Next consider

$$E_{Y_2^N}\left[\left(\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right)^2 - \Omega_2\right)^2\right] = E_{Y_2^N}\left[\left(\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right)^2 - \left(1+\frac{\lambda_1}{\bar{N}}\right)^2 - \frac{d_2(q||p)-1}{\bar{N}}\right)^2\right] \tag{269}$$

$$\leq 2E_{Y_2^N}\left[\left(\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right)^2 - \left(1+\frac{\lambda_1}{\bar{N}}\right)^2\right)^2\right] + 2\left(\frac{d_2(q||p)-1}{\bar{N}}\right)^2 \tag{270}$$

Now we can upper bound the first term as follows:

$$E_{Y_2^N}\left[\left(\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right)^2 - \left(1+\frac{\lambda_1}{\bar{N}}\right)^2\right)^2\right] \tag{271}$$

$$= E_{Y_2^N}\left[\left(\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right) - \left(1+\frac{\lambda_1}{\bar{N}}\right)\right)^2\left(\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right) + \left(1+\frac{\lambda_1}{\bar{N}}\right)\right)^2\right] \tag{272}$$

$$\leq \left(1+\frac{N+1}{\bar{N}}\omega\right)^2 E_{Y_2^N}\left[\left(1 - \left(\frac{1}{\bar{N}}\sum_{i=2}^{N}\lambda_i\right)\right)^2\right] \tag{273}$$

$$= \frac{1}{\bar{N}}\left(1+\frac{N+1}{\bar{N}}\omega\right)^2\left(d_2(q||p)-1\right) \tag{274}$$

As a result, using (265), (268) and (274) we have:

$$E_{Y_2^N}\left[\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\beta_i - \frac{\Omega_1}{\Omega_2}\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right)^2\right)^2\right] \tag{275}$$

$$\leq \frac{2}{\bar{N}}\left(d_2(r||p)-1\right) + \frac{4}{\bar{N}}\left(\frac{\Omega_1}{\Omega_2}\right)^2\left\{\left(1+\frac{N+1}{\bar{N}}\omega\right)^2\left(d_2(q||p)-1\right) + \frac{\left(d_2(q||p)-1\right)^2}{\bar{N}}\right\} \tag{276}$$

Consequently using (258), (263) and (276), we can show that:

$$\left| E_{Y_2^N}\left[ \frac{1}{\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right)^2}\left(1-\frac{\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right)^2}{c^2}\right)\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\beta_i-\frac{\Omega_1}{\Omega_2}\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right)^2\right)\mathbb{I}\left(\frac{1}{\bar{N}}\sum_{i=2}^{\bar{N}}\lambda_i\geq\frac{1}{2}\right)\right]\right|$$

$$\leq \frac{4}{\bar{N}c^2}\sqrt{\left(1+\frac{N+1}{\bar{N}}\omega\right)^2(d_2(q||p)-1)}$$

$$\times\sqrt{2\left(d_2(p||r)-1\right)+4\left(\frac{\Omega_1}{\Omega_2}\right)^2\left\{\left(1+\frac{N+1}{\bar{N}}\omega\right)^2(d_2(p||q)-1)+\frac{(d_2(q||p)-1)^2}{\bar{N}}\right\}} \tag{277}$$

$$\leq 4\frac{(\omega-1)}{\bar{N}c^2}\left(1+\frac{N+1}{\bar{N}}\omega\right)\sqrt{2+4\left(\frac{\Omega_1}{\Omega_2}\right)^2\left\{\left(1+\frac{N+1}{\bar{N}}\omega\right)^2+\frac{(\omega-1)}{\bar{N}}\right\}} \tag{278}$$

$$=\frac{1}{\bar{N}}K_1(\bar{N}) \tag{279}$$

where we repeatedly use the fact that $d_2(q||p)\leq\omega$ and $d_2(r||p)\leq\omega$. Here we have introduced:

$$K_1(\bar{N})=4\frac{(\omega-1)}{(1+\frac{\lambda_1}{\bar{N}})^2}\left(1+\frac{N+1}{\bar{N}}\omega\right)\sqrt{2+4\left(\frac{1+\frac{\beta_1}{\bar{N}}}{1+\frac{\gamma_1}{\bar{N}}}\right)^2\left\{\left(1+\frac{N+1}{\bar{N}}\omega\right)^2+\frac{(\omega-1)}{\bar{N}}\right\}} \tag{280}$$

$$\leq 4\frac{(\omega-1)}{(1+\frac{\lambda_1}{\bar{N}})^2}\left(1+\frac{N+1}{\bar{N}}\omega\right)\sqrt{2+4\left(\frac{1+\frac{\beta_1}{\bar{N}}}{1+\frac{2\lambda_1}{\bar{N}}}\right)^2\left\{\left(1+\frac{N+1}{\bar{N}}\omega\right)^2+\frac{(\omega-1)}{\bar{N}}\right\}}, \tag{281}$$

where we use the fact that $\gamma_1\geq 2\lambda_1$ following (247). Note that $K_1(\bar{N})=\Theta(1)$.
Now consider the other term in (254):

$$\left| E_{Y_2^N}\left[ \frac{1}{\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right)^2}\left(1-\frac{\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right)^2}{c^2}\right)\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\beta_i-\frac{\Omega_1}{\Omega_2}\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right)^2\right)\mathbb{I}\left(\frac{1}{\bar{N}}\sum_{i=2}^{\bar{N}}\lambda_i<\frac{1}{2}\right)\right]\right| \tag{282}$$

$$\leq \left| E_{Y_2^N}\left[ \frac{1}{\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right)^2}\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\beta_i-\frac{\Omega_1}{\Omega_2}\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right)^2\right)\mathbb{I}\left(\frac{1}{\bar{N}}\sum_{i=2}^{\bar{N}}\lambda_i<\frac{1}{2}\right)\right]\right| \tag{283}$$

$$\leq \left| \omega E_{Y_2^N}\left[ \frac{1}{\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right)^2}\mathbb{I}\left(\frac{1}{\bar{N}}\sum_{i=2}^{\bar{N}}\lambda_i<\frac{1}{2}\right)\right]\right| \tag{284}$$

Next, we can show that:

$$\frac{1}{\left(\frac{1}{\bar{N}}\sum_{i=1}^{N}\lambda_i\right)^2}\leq\frac{1}{\left(\frac{1}{\bar{N}}\sum_{i=2}^{N}\lambda_i\right)^2} \tag{285}$$

$$\leq\frac{1}{\bar{N}}\sum_{i=2}^{N}\left(\frac{p_i}{q_i}\right)^2 \tag{286}$$

The proof of (286) will be shown at the end of this section. Furthermore note that $E_p[p^2/q^2]=d_3(p||q)$. Thus

we can upper-bound (284) as:

$$\leq \left| \omega E_{Y_2^N} \left[ \left( \frac{1}{\bar{N}} \sum_{i=2}^{\bar{N}} \frac{p_i^2}{q_i^2} \right) \mathbb{I}\left( \frac{1}{\bar{N}} \sum_{i=2}^{\bar{N}} \lambda_i < \frac{1}{2} \right) \right] \right| \tag{287}$$

$$= \omega \left| E_{Y_2^N} \left[ \left( \frac{1}{\bar{N}} \sum_{i=2}^{\bar{N}} \frac{p_i^2}{q_i^2} - d_3(p||q) + d_3(p||q) \right) \mathbb{I}\left( \frac{1}{\bar{N}} \sum_{i=2}^{\bar{N}} \lambda_i < \frac{1}{2} \right) \right] \right| \tag{288}$$

$$\leq \omega \left| E_{Y_2^N} \left[ \left( \frac{1}{\bar{N}} \sum_{i=2}^{\bar{N}} \frac{p_i^2}{q_i^2} - d_3(p||q) \right) \mathbb{I}\left( \frac{1}{\bar{N}} \sum_{i=2}^{\bar{N}} \lambda_i < \frac{1}{2} \right) \right] \right| + \omega \left( d_3(p||q) \right) E\left[ \mathbb{I}\left( \frac{1}{\bar{N}} \sum_{i=2}^{\bar{N}} \lambda_i < \frac{1}{2} \right) \right] \tag{289}$$

$$\leq \omega \left| E_{Y_2^N} \left[ \left( \frac{1}{\bar{N}} \sum_{i=2}^{\bar{N}} \frac{p_i^2}{q_i^2} - d_3(p||q) \right) \mathbb{I}\left( \frac{1}{\bar{N}} \sum_{i=2}^{\bar{N}} \lambda_i < \frac{1}{2} \right) \right] \right| + \omega \left( d_3(p||q) \right) \frac{4}{\bar{N}} (d_2(q||p) - 1) \tag{290}$$

The first term above can be upper bounded using Cauchy-schwartz as follows:

$$\left| E_{Y_2^N} \left[ \left( \frac{1}{\bar{N}} \sum_{i=1}^{\bar{N}} \frac{p_i^2}{q_i^2} - d_3(p||q) \right) \mathbb{I}\left( \frac{1}{\bar{N}} \sum_{i=2}^{\bar{N}} \lambda_i < \frac{1}{2} \right) \right] \right| \tag{291}$$

$$\leq \sqrt{ E_{Y_2^N} \left[ \left( \frac{1}{\bar{N}} \sum_{i=1}^{\bar{N}} \frac{p_i^2}{q_i^2} - d_3(p||q) \right)^2 \right] } \sqrt{ E\left[ \mathbb{I}\left( \frac{1}{\bar{N}} \sum_{i=2}^{\bar{N}} \lambda_i < \frac{1}{2} \right) \right] } \tag{292}$$

$$\leq \sqrt{ \frac{1}{\bar{N}} (d_5(p||q) - d_3(p||q)^2) } \sqrt{ \frac{4}{\bar{N}} (d_2(q||p) - 1) } \tag{293}$$

It thus follows that we can express:

$$\left| E_{Y_2^N} \left[ \frac{1}{\left( \frac{1}{\bar{N}} \sum_{i=1}^{N} \lambda_i \right)^2} \left( 1 - \frac{\left( \frac{1}{\bar{N}} \sum_{i=1}^{N} \lambda_i \right)^2}{c^2} \right) \left( \frac{1}{\bar{N}} \sum_{i=1}^{N} \beta_i - \frac{\Omega_1}{\Omega_2} \left( \frac{1}{\bar{N}} \sum_{i=1}^{N} \lambda_i \right)^2 \right) \mathbb{I}\left( \frac{1}{\bar{N}} \sum_{i=2}^{\bar{N}} \lambda_i < \frac{1}{2} \right) \right] \right|$$
$$\leq \frac{2\omega}{\bar{N}} L(\bar{N}) \tag{294}$$

where

$$L(\bar{N}) = \sqrt{ (d_5(p||q) - d_3(p||q)^2) } \sqrt{ d_2(q||p) - 1 } + \left( d_3(p||q) \right) \left( d_2(q||p) - 1 \right) \tag{295}$$

$$\leq \sqrt{\omega - 1} \sqrt{ (d_5(p||q) - d_3(p||q)^2) } + (\omega - 1) d_3(p||q) \tag{296}$$

Thus using (254), (281) and (296) it follows that:

$$\bar{N} E_{Y_2^N} \left[ \frac{\sum_{i=1}^{N} \beta_i}{\left( \sum_{i=1}^{N} \lambda_i \right)^2} \right] \leq \frac{\frac{\beta_1}{\bar{N}} + 1}{\left( 1 + \frac{\lambda_1}{\bar{N}} \right)^2} + \frac{1}{\bar{N}} K_1(\bar{N}) + \frac{2\omega}{\bar{N}} L(p, q)$$

It remains to show

$$\frac{1}{\left( \frac{1}{\bar{N}} \sum_{i=2}^{N} \frac{q_i}{p_i} \right)^2} \leq \frac{1}{\bar{N}} \sum_{i=2}^{N} \left( \frac{p_i}{q_i} \right)^2 \tag{297}$$

Note that:

$$\frac{1}{\left(\frac{1}{\bar{N}}\sum_{i=2}^{N}\frac{q_i}{p_i}\right)^2} = \frac{\sum_{i=2}^{\bar{N}}\frac{q_i}{p_i}}{\bar{N}}\left(\frac{\sum_{i=2}^{\bar{N}}\frac{q_i}{p_i}\frac{p_i}{q_i}}{\sum_{i=2}^{\bar{N}}\frac{q_i}{p_i}}\right)^3 \tag{298}$$

$$= \frac{\sum_{i=2}^{\bar{N}}\frac{q_i}{p_i}}{\bar{N}}\left(E_\lambda\left[\frac{p}{q}\right]\right)^3 \tag{299}$$

$$\leq \frac{\sum_{i=2}^{\bar{N}}\frac{q_i}{p_i}}{\bar{N}}E_\lambda\left[\frac{p^3}{q^3}\right] \tag{300}$$

$$= \frac{\sum_{i=2}^{\bar{N}}\frac{q_i}{p_i}}{\bar{N}}\left(\frac{\sum_{i=2}^{\bar{N}}\frac{q_i}{p_i}\frac{p_i^3}{q_i^3}}{\sum_{i=2}^{\bar{N}}\frac{q_i}{p_i}}\right) \tag{301}$$

$$= \frac{1}{\bar{N}}\sum_{i=2}^{\bar{N}}\frac{p_i^2}{q_i^2}. \tag{302}$$

In (299) we define $\lambda$ to be probability vector which select index $i$ with probability proportional to $q_i/p_i$. We use Jensen's inequality in (300) since the function $f(x) = x^3$ is convex on $x \geq 0$.

## 22 Proof of Prop. 6 in this document

For $0 < \epsilon < 1$, define

$$\mathcal{E} = \left\{(Y_j)_{j=2}^N : \frac{1}{N-1}\sum_{j=2}^{N}\lambda(Y_j) \geq 1-\epsilon, \frac{1}{N-1}\sum_{j=2}^{N}\beta(Y_j) \leq 1+\epsilon\right\}, \tag{303}$$

then using Hoeffding's inequality and the union bound, we have that:

$$\Pr\left(\mathcal{E}^c\right) \leq 2\exp\left(-2(N-1)\epsilon^2/\omega^2\right). \tag{304}$$

Now observe that:

$$E_{Y_2^N}\left[\frac{\sum_{j=1}^{N}\beta(Y_j)}{(\sum_{j=1}^{N}\lambda(Y_j))^2}\middle| Y_1 = y_1\right] \tag{305}$$

$$= E_{Y_2^N}\left[\frac{\sum_{j=1}^{N}\beta(Y_j)}{(\sum_{j=1}^{N}\lambda(Y_j))^2}\middle| Y_1 = y_1, \mathcal{E}\right] \tag{306}$$

$$+ E_{Y_2^N}\left[\frac{\sum_{j=1}^{N}\beta(Y_j)}{(\sum_{j=1}^{N}\lambda(Y_j))^2}\middle| Y_1 = y_1, \mathcal{E}^c\right]\Pr(\mathcal{E}^c) \tag{307}$$

$$\leq \frac{\beta(y_1)+(N-1)(1+\epsilon)}{(\lambda(y_1)+(N-1)(1-\epsilon))^2} + \frac{N\omega}{\lambda(y_1)^2}2e^{-(N-1)\epsilon^2/\omega^2} \tag{308}$$

Collecting all the terms we have that:

$$E_{Y_2^N}\left[\frac{\left(\sum_{j=1}^{N}\frac{q_{Y|X}(Y_j|x)}{p_Y(Y_j)}\right)}{\left(\sum_{j=1}^{N}\frac{p_{Y|X}(Y_j|x)}{p_Y(Y_j)}\right)}\middle| Y_1 = y_1, U_p = 1\right]$$

$$\leq (N-1+\lambda(y_1))\left(\frac{\beta(y_1)+(N-1)(1+\epsilon)}{(\lambda(y_1)+(N-1)(1-\epsilon))^2} + \frac{N\omega}{\lambda(y_1)^2}2e^{-(N-1)\epsilon^2/\omega^2}\right) \tag{309}$$

## 23 Decoding with Neural Estimator

We recall that in the problem of lossy compression with side information at the decoder, the random variables $V, W$ and $T$ follow the Markov chain $T - V - W$. Following the setup and algorithm in Section 2 in the main paper, the encoding step is relatively straightforward, given that $p_{W|V}(.|.)$ and $p_W(.)$ can be predefined. On the other hand, during the decoding step, the decoder needs to compute the following quantity:

$$\begin{aligned} U_q &= \arg \min_{1 \leq i \leq N} \frac{S_i}{\frac{Q_{Y|Z}(Y_i|t, l_{U_p})}{p_Y(Y_i)}} \\ &= \arg \min_{1 \leq i \leq N} \frac{S_i}{\frac{p_{W|T}(W_i|t)\mathbb{I}(l_i = l_{U_p})}{p_W(W_i)p_l(l_i)}}, \end{aligned}$$

which can be hard to compute due to the presence of $p_{W|T}(W_i|t)$, especially when the distribution is unknown and complicated. As a result, the quantity $\log \frac{p_W(W_i)}{p_{W|T}(W_i|t)}$ has to be learned from the training dataset. We note that while techniques like Markov Chain Monte Carlo (MCMC) or variational inference can be employed, their usage may lead to significant time complexity or sub optimal performance due to the inherent limitations in expressing complex distributions accurately.

Instead, we construct and train a neural network $\Gamma \colon \mathcal{W} \times \mathcal{T} \to [0, 1]$ to directly estimate the above ratio (Hermans et al., 2020; Cranmer et al., 2015), by classifying whether $W, T \sim p_{W,T}(.)$ (positive samples) or $W, T \sim p_W(.)p_T(.)$ (negative samples). Following the Markov chain $T - V - W$, one can construct a positive sample by first sampling from the training set a pair of $\{T, V\}$ and then get $W \sim p_{W|V}(.)$ where $p_{W|V}(.)$ is predefined. On the other hand, to obtain negative samples, we sample $\{T, V\}$ from the training set and $W \sim p_W(.)$. Note that ratio between positive and negative samples should be 1. Furthermore, we define $\Gamma(W, T) = \sigma(h_\gamma(W, T))$ where $h_\gamma$ is a neural networks with parameters $\gamma$ and $\sigma$ is the sigmoid activation:

$$\sigma(x) = \frac{1}{1 + \exp(-x)}$$

Hermans et al. (2020) shows that the logit values of the optimal classifier can be then used as a log-likelihood estimator, that is:

$$h_{\gamma^*}(W, T) \approx -\log \frac{p_W(W)p_T(T)}{p_{W,T}(W, T)} = -\log \frac{p_W(W)}{p_{W|T}(W|T)}. \tag{310}$$

which is the quantity we would like to estimate. We train our classifier using the standard cross-entropy loss with Adam optimizer. For details about the neural network architecture of each experiment (MNIST and CIFAR-10), refer Section 24 below.

## 24 ADDITIONAL EXPERIMENT RESULTS

### 24.1 Synthetic Gaussian Case

We provide details of how we compute the conditional distribution $p_{W|T}(.)$, inverse variance weighting and additional experimental results

**Calculating $p_{W|T}$.** We recall the setup we are following. Assume that the source $V \sim \mathcal{N}(0, \sigma_V^2 = 1.0)$ and the side information $T = V + \zeta_{T|V}$ where $\zeta_{T|V} \sim \mathcal{N}(0, 0.01)$, i.e $p_{T|V}(.|v) = \mathcal{N}(v, \sigma_{T|V}^2 = 0.01)$. Furthermore, the encoder and decoder have access to the shared randomness $(S_i, Y_i, \ell_i)_i^N$ as described previously. The decoder must ideally output $W \sim p_{W|V}$, where $p_{W|V}(\cdot|v) = \mathcal{N}(v, \sigma_{W|V}^2)$. We start with the joint distribution of $V$ and $T$, which can be expressed as:

$$\begin{pmatrix} V \\ T \end{pmatrix} \sim \mathcal{N}\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_V^2 & \sigma_V^2 \\ \sigma_V^2 & \sigma_T^2 \end{bmatrix} \right),$$

where $\sigma_T^2 = \sigma_V^2 + \sigma_{T|V}^2$. Following this, we have the conditional probability of $V$ given $T$ as:

$$p_{V|T}(.|T=t) = \mathcal{N}\left(\frac{\sigma_V^2}{\sigma_T^2}t, \left(1 - \frac{\sigma_V^2}{\sigma_T^2}\right)\sigma_V^2\right),$$

Using the Markov chain $T - V - W$, we have:

$$\begin{aligned} p_{W|T}(w|t) &= \int_{-\infty}^{\infty} p_{W|V}(w|v,t)p_{V|T}(v|t)dv \\ &= \int_{-\infty}^{\infty} p_{W|V}(w|v)p_{V|T}(v|t)dv \end{aligned}$$

As $p_{W|V}(.)$ and $p_{V|T}(.)$ are two Gaussians with fix variance, we obtain:

$$p_{W|T}(.|t) = \mathcal{N}\left(\frac{\sigma_V^2}{\sigma_T^2}t, \sigma_W^2 - \frac{\sigma_V^4}{\sigma_T^2}\right)$$

where $\sigma_W^2 = \sigma_V^2 + \sigma_{W|V}^2$. We then use this quantity to compute the decoder index as explained in the main paper.

**Inverse Variance Weighting.** We combine the decoder output $W_{U_q}$ with the side information $T \sim p_{T|V}$ to obtain a lower variance estimator $\hat{V}$ of $V$ by applying the inverse variance weighting fusion method proposed by (Graybill and Deal, 1959), which we show its effectiveness in Figure 10. We note that in the case without feedback, the decoder's output $W_{U_q}$ might not closely follow the target Gaussian distribution due to mismatching error, and applying inverse variance weighting might yield suboptimal results, i.e. its distortion is higher than that of using side-information alone, which is also demonstrated in Figure 10. As such, in the main paper, when this situation happens, we simply ignore the information from the source and only consider the side information for reconstruction. In the case where feedback is used, our inverse variance weighting estimator performs consistently well since $W_{U_q}$ now follows more closely to the target Gaussian distribution.

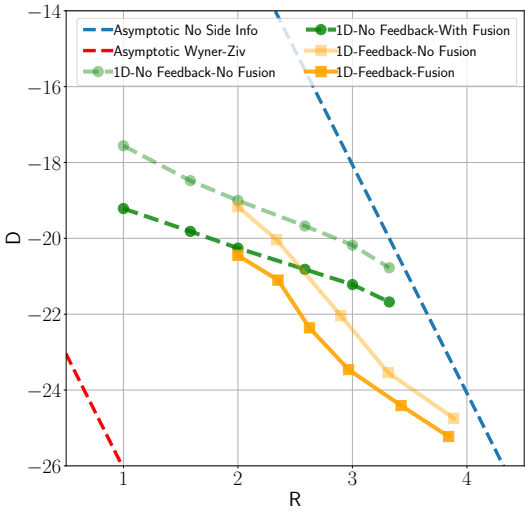



Figure 10: Effects of inverse variance weighting (fusion) on improving the estimator accuracy.

**Simulation Parameters.** For all the experiments where we only compress 1 sample, we set $N = 2^{15}$ and use grid-search on $L \in \{2, 4, 6, 8, 10\}$, $\sigma_{W|V}^2 \in \{0.01, 0.008, 0.006, 0.005, 0.003, 0.002, 0.001\}$. For the 5D case, we set $N = 2^{27}$, use grid-search on $L \in \{2^5, 2^6, 2^7, 2^8, 2^9, 2^{10}, 2^{11}, 2^{12}, 2^{13}, 2^{14}\}$. We note that if the encoder detects a mismatch between the two indices, it can send either the rest or a part of the MSB of its index, which we refer to as $L_2$. For reference, we provide some optimal parameters we found in Table 1. Results in the main paper are averaged over 10 runs.

**Additional Results.**

*Matching Probability and Side Information Quality.* We show in Figure 11 the matching probability as a function of side-information quality $\Delta$ for different $\sigma_{W|V}^2$ and $L$, where we measure the side-information quality by its

Table 1: Rate-Distortion Parameters for 5D Gaussian Compression with Perfect Feedback.

| Dimension | $L$ | $L_2$ | $\sigma^2_{W\|V}$ | Rate | Distortion (dB) |
|---|---|---|---|---|---|
| 1 | 2 | 3 | 0.01 | 2.133 | -20.61 |
| 1 | 2 | 6 | 0.008 | 2.35 | -21.10 |
| 1 | 2 | 8 | 0.005 | 2.625 | -22.36 |
| 1 | 2 | 12 | 0.003 | 2.966 | -23.46 |
| 1 | 2 | 16 | 0.001 | 3.425 | -24.41 |
| 5 | $2^5$ | $2^2$ | 0.01 | 1.18 | -21.78 |
| 5 | $2^8$ | $2^5$ | 0.008 | 1.33 | -22.33 |
| 5 | $2^5$ | $2^{14}$ | 0.005 | 1.88 | -24.49 |
| 5 | $2^{12}$ | $2^8$ | 0.003 | 2.65 | -27.04 |
| 5 | $2^{14}$ | $2^6$ | 0.001 | 3.06 | -28.84 |

associated distance to the source, i.e $\Delta = |t - v|$. To obtain the matching probability, for each $\Delta$, we sample $V \sim p_V(.)$, send the side information $T = V \pm \Delta$ to the decoder and compare $U_q$ and $U_p$. This process is simulated and averaged over 1000 runs to obtain the matching probability. We observe that the matching probability decrease with $\Delta$ and increasing $\sigma^2_{W|V}$ and $L$ consistently improves the matching rate for all $\Delta$. Finally, we note that the matching probability is not 1.0 for $\Delta = 0$ due to the distribution mismatch between $p_{W|V}(.|v)$ and $p_{W|T}(.|t)$.

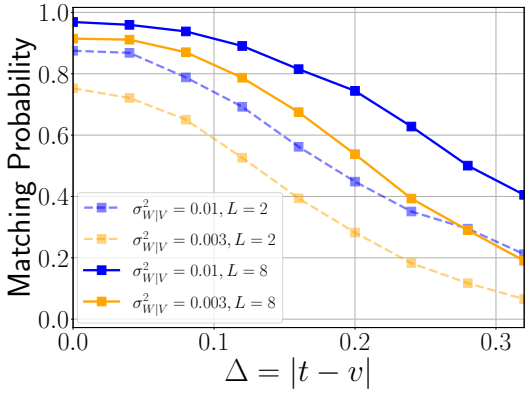



Figure 11: Matching Probability w.r.t side information quality. We use the distance $\Delta = |t - v|$ to quantify the side information quality, lower $\Delta$ correspond to better quality.

*On Feedback Error.* We provide the feedback error (collision) due to hashing during the feedback step in Table 2. Specifically, we define feedback error as the probability that the encoder index and decoder index are different but have the same hash value. Also, when their hash values are different, the two indices must be different. In general, we observe that this error depends on multiple factors, such as the number of samples we are compressing, the matching probability in the first round (which depends on $L$), the number of proposals, and $\sigma^2_{W|V}$. As a simple illustrative example, consider the case when we compress 5 samples jointly with the number of proposals $N = 2^{25}$, then using 4 bits feedback is sufficient for the encoder to obtain the exact position when $L = 2^{21}$ but it is not the case when $L = 2^{10}$. Overall, we found that we can adjust the parameters to obtain low feedback error in most of the cases.

### 24.2 Distributed Image Compression

We provide details on the network architecture and additional results for the distributed lossy image compression experiment on MNIST dataset (Lecun et al., 1998).

Table 2: Feedback error with different parameters. The first column represents the feedback rate for the encoder to recover the full index. The second column represents the (hashed) non-ideal feedback rate. We measure feedback rate by bits.

| Feedback Rate (Ideal) | Feedback Rate (non-ideal) | Dimension | $L$ | N | $\sigma^2_{W\|V}$ | Error Probability |
|---|---|---|---|---|---|---|
| 11 | 1 | 1 | $2^1$ | $2^{12}$ | 0.01 | 13.1% |
| 11 | 1 | 1 | $2^1$ | $2^{12}$ | 0.008 | 16.1% |
| 9 | 1 | 1 | $2^3$ | $2^{12}$ | 0.004 | 7.21% |
| 22 | 1 | 5 | $2^5$ | $2^{27}$ | 0.01 | 12.36% |
| 20 | 1 | 5 | $2^7$ | $2^{27}$ | 0.01 | 4.14% |
| 19 | 1 | 5 | $2^8$ | $2^{27}$ | 0.005 | 4.88% |
| 17 | 1 | 5 | $2^{10}$ | $2^{27}$ | 0.005 | 2.48% |
| 13 | 1 | 5 | $2^{14}$ | $2^{27}$ | 0.001 | 1.39% |
| 22 | 5 | 5 | $2^5$ | $2^{27}$ | 0.01 | 1.09% |
| 17 | 5 | 5 | $2^{10}$ | $2^{27}$ | 0.005 | 0.09% |
| 13 | 5 | 5 | $2^{14}$ | $2^{27}$ | 0.001 | 0.03% |

### 24.2.1 Network Architecture

**Encoder-Decoder Network.** We show the architecture our $\beta-$VAE network, including the encoder network $f_e(v)$, the projection network $h(t)$, and the decoder $g(w,t)$ in the Table.3. Convolutional and transposed convolutional layers are denoted as "conv" and "upconv" respectively, which are accompanied by a number of filters, kernel size, stride, and padding. For "upconv", we have an additional parameter which is the output padding at the end. The encoder network maps an image into 2 vectors of size 4 (total 8D output), where the first vector represents the output mean $f_e(v)^{(1)}$ and the second vector $f_e(v)^{(2)}$ represents the output variance. Specifically, we define $p_{W|V}(.) = \mathcal{N}(f_e(v)^{(1)}, f_e(v)^{(2)})$ and use the prior distribution $p_W(.) = \mathcal{N}(0,1)$.

At the decoder side, a projection network $h(t)$ first maps the side information image $T$ to a vector of size 128, which is then combined with a vector of size 4 from the encoder, resulting in a 132D vector. This $132D$ vector is then fed into a decoder network $g(w,t)$ that outputs a reconstruction of size $28 \times 28$, which we denote as $\hat{V}$.

**Loss Function** We train our $\beta-$VAE network by optimize the following rate-distortion loss:

$$\mathcal{L} = \beta(V - \hat{V})^2 - E_V[D_{\mathrm{KL}}(p_{W|V}(.|v)||p_W(.))] \tag{311}$$

where we vary $\beta$ for different rate-distortion tradeoff. We train each model for 30 epochs on an NVIDIA-RTX A4500, which takes 30 minutes per model.

Table 3: Encoder, project network, and decoder for MNIST distributed image compression.

**(a)**Encoder $f_e(v)$

| |
|---|
| Input $28 \times 28 \times 1$ |
| conv (128:3:1:1), ReLU |
| conv (128:3:2:1), ReLU |
| conv (128:3:2:1), ReLU |
| Flatten |
| Linear (6272, 512), ReLU |
| Linear (512, 8) |

**(b)**Projection Network $h(t)$

| |
|---|
| Input $14 \times 14 \times 1$ |
| conv (32:3:1:1), ReLU |
| conv (64:3:2:1), ReLU |
| conv (128:3:2:1), ReLU |
| Flatten |
| Linear (2048, 512), ReLU |
| Linear (512, 128) |

**(c)**Decoder Network $g(w,t)$

| |
|---|
| Input-(4+128) |
| Linear-$(132, 512)$, ReLU |
| upconv (64:3:2:1:1), ReLU |
| upconv (32:3:2:1:1), ReLU |
| upconv (1:3:1:1), Tanh |

**Neural Estimator Network.** The neural estimator network in this case consists of two subnetworks. The first subnetwork will project the side-information into an embedding of size 128 and the second subnetwork combines that 128D embedding with the 4D embedding, either from $p_{W|V}$ or $p_W$, and outputs the probability of whether $T, W$ are from the joint or from the marginal distributions. We note that the projection network architecture is the same as the one in our $\beta$-VAE network. Finally, we note that this model is trained with 100 epochs.

**Simulation Parameters.** To train our network, we varies $\beta \in \{0.15, 0.25, 0.35, 0.45, 0.55, 0.65, 0.75, 0.85, 0.95\}$. For one sample compression, we perform grid search on $N \in \{2^7, 2^8, 2^9, 2^{10}, 2^{11}\}$, $L \in \{2^4, 2^5, 2^6, 2^7, 2^8\}$. For two sample compression, we use $N \in \{2^{20}, 2^{21}, 2^{22}, 2^{23}, 2^{24}, 2^{25}\}$ and $L \in \{2^{15}, 2^{16}, 2^{17}, 2^{18}, 2^{19}, 2^{20}\}$. In both

Table 4: Neural Estimator Networks for Distributed Image Compression.

| **(a)**Projection Network |
|---|
| Input $14 \times 14 \times 1$ |
| conv (32:3:1:1), ReLU |
| conv (64:3:2:1), ReLU |
| conv (128:3:2:1), ReLU |
| Flatten |
| Linear (2048, 512), ReLU |
| Linear (512, 128) |

| **(b)** Combine and Classify |
|---|
| Input $128 + 4$ |
| Linear (132, 128), l-ReLU |
| Linear (128,128), l-ReLU |
| Linear (128,128), l-ReLU |
| Linear (128, 1) |

Table 5: Rate-Distortion Parameters for MNIST Compression with Feedback.

| Number of Samples | $L$ | $N$ | $\beta$ | Rate | Distortion (MSE) |
|---|---|---|---|---|---|
| 1 | $2^{10}$ | $2^{15}$ | 0.95 | 11.96 | 0.0488 |
| 1 | $2^{7}$ | $2^{12}$ | 0.75 | 8.64 | 0.0566 |
| 1 | $2^{3}$ | $2^{8}$ | 0.25 | 6.865 | 0.0635 |
| 2 | $2^{20}$ | $2^{25}$ | 0.95 | 11.01 | 0.0489 |
| 2 | $2^{15}$ | $2^{20}$ | 0.75 | 7.79 | 0.0565 |
| 2 | $2^{10}$ | $2^{15}$ | 0.35 | 6.2 | 0.0618 |

cases, we send the full MSB index in the second transmission. We provide some optimal values shown in Table 5. Results in the main paper are averaged over 10 runs.

### 24.2.2 Additional Examples

We provide additional examples where the decoder outputs correct/incorrect reconstructions during the first transmission. This again confirms that our neural estimator selects a semantically meaningful message from the encoder's output.

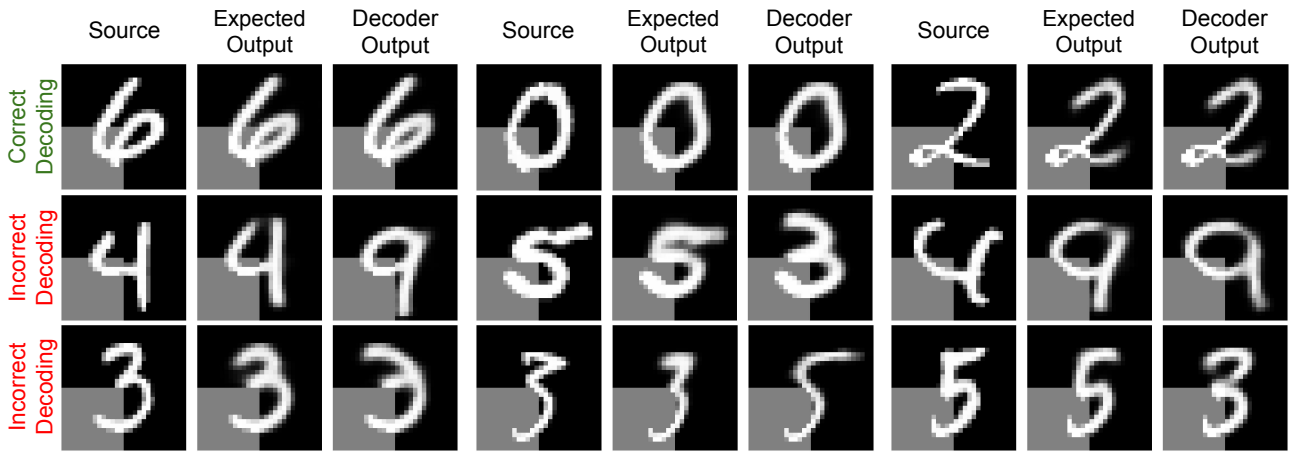



Figure 12: Distributed Image Compression without Feedback. Additional Results.

### 24.2.3 Compression with limited Feedback

We provide a new experiment R-D plot for our feedback-free scheme (compressing 3 samples together) for the MNIST experiment in Sec. 5.2 in Fig. 13. We also include the case where the feedback signal is imperfect. Although slightly worse than the NDIC baseline, the latter requires engineering neural networks, involving complex loss function with several hyper-parameters.

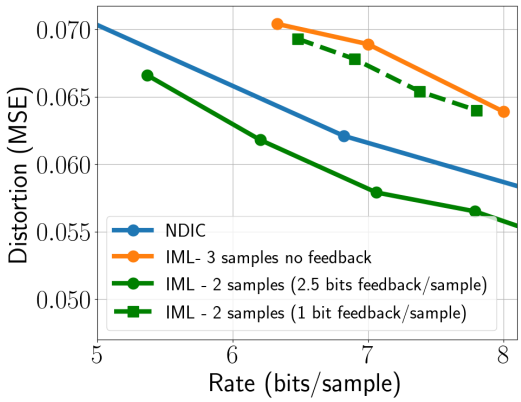



Figure 13: MNIST Distributed Compression with Different Feedback Rates. Here, 2.5 bits feedback/sample is sufficient to recover the index when two samples are jointly compressed.

### 24.3 Vertical Federated Learning - CIFAR 10

We provide details on network architecture and simulation parameters in the vertical federated learning experiment with CIFAR-10 (Krizhevsky et al., 2009).

**Network Architecture** We present our networks, including the model at each party, the server model, and the neural estimator module in the Table 6. We use the "residual block" which is shown in Figure 14. Each party model in this case will project its quadrant to a 4D embedding and send them to the server model, which will output the prediction. We train the model for 100 epochs on an NVIDIA-RTX A4500, which converges after 2 hours training.

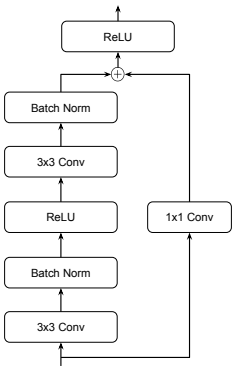



Figure 14: Residual Block. In our description, each "residual block" is described by number of filters and stride of the 3x3 convolution operator, which we set the value of padding to 1.

**Loss Function.** We train our network end-to-end with a standard cross-entropy loss. We augment the dataset by applying horizontal flip and random cropping to the original image, before cropping the two quadrants (bottom-left and top-right), that would be then distributed to both parties.

**Simulation Parameters.** We perform grid-search and show the optimal parameters in Table 7. Results in the main paper are averaged over 10 runs.

### 24.4 Breast Cancer Dataset

The parameters for IML is shown in Table 8. We note that in this experiment, there is no neural network at the encoder side and we aim to lossily transmit the features to the decoder.

Table 6: Party Model, Server Model, and Neural Estimator.

**(a)**Party Model

| Input |
|---|
| conv (64:3:1:1), BatchNorm2D, l-ReLU |
| residual block (64:1) |
| residual block (128:2) |
| residual block (256:2) |
| residual block (512:2) |
| Linear (2048, 4) |

**(b)**Server Module

| Input $4 \times 2$ |
|---|
| Linear (8, 128) |
| Linear (128, 10) |

**(c)**Neural Estimator

| Input-$(4 \times 2)$ |
|---|
| Linear (8, 128), l-ReLU |
| Linear (128, 128), l-ReLU |
| Linear (128, 128), l-ReLU |
| Linear (128, 128), l-ReLU |
| Linear (128, 1) |

Table 7: Parameters for C-VFL experiments with CIFAR-10.

| $L$ | $N$ | $\sigma^2_{W\|V}$ | Rate | Accuracy |
|---|---|---|---|---|
| $2^3$ | $2^6$ | 0.07 | 4.96 | 0.764 |
| $2^4$ | $2^8$ | 0.06 | 6.5 | 0.789 |
| $2^5$ | $2^9$ | 0.04 | 7.4 | 0.798 |
| $2^9$ | $2^{13}$ | 0.005 | 11.41 | 0.811 |

Table 8: Parameters for C-VFL experiments with Breast Cancer Dataset (Our method).

| $L$ | $N$ | $\sigma^2_{W\|V}$ | Rate | Accuracy |
|---|---|---|---|---|
| $2^3$ | $2^6$ | 0.07 | 4.8 | 0.9 |
| $2^5$ | $2^9$ | 0.04 | 7.3 | 0.93 |
| $2^6$ | $2^{10}$ | 0.01 | 8.6 | 0.95 |
| $2^9$ | $2^{12}$ | 0.005 | 10.48 | 0.97 |