A Theoretical Results

A.1 Gaussian Case

Recall that for $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$,

$$R(D) = \begin{cases} \frac{1}{2}\log(\sigma_X^2/D), & 0 \le D \le \sigma_X^2, \\ 0, & D > \sigma_X^2, \end{cases} \tag{15}$$

and in the first case the function is attained by some $p_{\hat{X}|X}$ with marginal $\hat{X} \sim \mathcal{N}(\mu_X, \sigma_X^2 - D)$[7].

Theorem 1. For $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$, the rate-distortion-perception function under squared error distortion and squared W_2 distance is achieved by some $\hat{X}$ jointly Gaussian with X and is given by

$$R(D,P) = \begin{cases} \frac{1}{2}\log \frac{\sigma_X^2(\sigma_X - \sqrt{P})^2}{\sigma_X^2(\sigma_X-\sqrt{P})^2 - (\frac{\sigma_X^2 + (\sigma_X - \sqrt{P})^2 - D}{2})^2} \\ \qquad\qquad \text{if } \sqrt{P} < \sigma_X - \sqrt{|\sigma_X^2 - D|}, \\ \max\{\frac{1}{2}\log\frac{\sigma_X^2}{D}, 0\} \qquad \text{if } \sqrt{P} \ge \sigma_X - \sqrt{|\sigma_X^2 - D|}. \end{cases}$$

We will first need a Lemma from estimation theory. Let $\hat{X}$ be a random variable with $\mathbb{E}[\hat{X}] = \mu_{\hat{X}}$, $\mathrm{Var}(\hat{X}) = \sigma_{\hat{X}}^2$ and $\mathrm{Cov}(X, \hat{X}) = \theta$. Let $\hat{X}_G$ be a random variable jointly Gaussian with X with the same first and second order statistics as $\hat{X}$.

Lemma 1. Given $\mu_{\hat{X}}$, $\sigma_{\hat{X}}^2$, and θ, we have that

$$\mathbb{E}[(X - \mathbb{E}[X|\hat{X}_G])^2] \ge \mathbb{E}[(X - \mathbb{E}[X|\hat{X}])^2].$$

The proof of this result can be found in a standard estimation theory reference, e.g. Chapter 3, page 134 of the 6.432 notes by Willsky & Wornell [38].

Proof of Theorem 1. We shall show that there is no loss of optimality in assuming that $\hat{X}$ is jointly Gaussian with X. It is clear that $\mathbb{E}[(X - \hat{X})^2] = \mathbb{E}[(X - \hat{X}_G)^2]$, as the first and second order statistics are all given. Note that by expanding out $W_2(p_X, p_{\hat{X}})$, one can see that the optimal coupling is identified only through the cross-term between X and $\hat{X}$; since every coupling of p_X and $p_{\hat{X}}$ induces a Gaussian coupling of p_X and $p_{\hat{X}_G}$ with the same covariance, it follows that

$$W_2^2(p_X, p_{\hat{X}}) \ge W_2^2(p_X, p_{\hat{X}_G}). \tag{16}$$

Finally, we have

$$\begin{aligned} I(X;\hat{X}) &= h(X) - h(X|\hat{X}) \\ &\ge h(X) - h(X - \mathbb{E}[X|\hat{X}]) \\ &\overset{(a)}{\ge} h(X) - \frac{1}{2}\log(2\pi e \mathbb{E}[(X - \mathbb{E}[X|\hat{X}])^2]) \\ &\overset{(b)}{\ge} h(X) - \frac{1}{2}\log(2\pi e \mathbb{E}[(X - \mathbb{E}[X|\hat{X}_G])^2]) \\ &= h(X) - h(X - \mathbb{E}[X|\hat{X}_G])) \\ &\overset{(c)}{=} h(X) - h(X|\hat{X}_G) \\ &= I(X;\hat{X}_G), \end{aligned} \tag{17}$$

where (a) is because the Gaussian distribution maximizes differential entropy for a given variance, (b) follows from Lemma 1 and (c) is because the estimation error is independent of $\hat{X}_G$. Thus, it suffices to solve the problem

$$\begin{aligned} R(D,P) = \min_{p_{\hat{X}_G|X}} \quad & I(X;\hat{X}_G) \\ \text{s.t.} \quad & \mathbb{E}[(X - \hat{X}_G)^2] \le D, \quad W_2^2(p_X, p_{\hat{X}_G}) \le P. \end{aligned} \tag{18}$$

Note that we can write

$$\mathbb{E}[(X - \hat{X}_G)^2] = (\mu_X - \mu_{\hat{X}})^2 + \sigma_X^2 + \sigma_{\hat{X}}^2 - 2\theta, \tag{19}$$

and we have from standard results (e.g. minimizing (19), or more generally [8]) that

$$W_2^2(p_X, p_{\hat{X}_G}) = (\mu_X - \mu_{\hat{X}})^2 + (\sigma_X - \sigma_{\hat{X}})^2. \tag{20}$$

Finally, recall that the mutual information between the two Gaussian distributions is given by

$$I(X; \hat{X}_G) = \frac{1}{2} \log \frac{\sigma_X^2 \sigma_{\hat{X}}^2}{\sigma_X^2 \sigma_{\hat{X}}^2 - \theta^2}, \tag{21}$$

so there is no loss of optimality in assuming $\mu_{\hat{X}} = \mu_X$ and $\theta \geq 0$. Now we consider when each constraint is active. Suppose that P was active and D was inactive. Then

$$\begin{aligned} D &> \sigma_X^2 + \sigma_{\hat{X}}^2 - 2\theta \\ &= \sigma_X^2 + (\sigma_X - \sqrt{P})^2 - 2\theta. \end{aligned} \tag{22}$$

Hence, we can decrease θ to reduce the mutual information until either D is active or the rate is zero.

If D is active, then the perception constraint is satisfied automatically when $(\sigma_{\hat{X}} - \sigma_X)^2 \leq P$, or $\sqrt{P} \geq \sigma_X - \sqrt{|\sigma_X^2 - D|}$ (here we have used the solution to $R(D)$ from (15)). When $\sqrt{P} < \sigma_X - \sqrt{|\sigma_X^2 - D|}$, both P and D are active, and consequently we have $\sigma_{\hat{X}}^2 = (\sigma_X - \sqrt{P})^2$ and $\theta = \frac{\sigma_X^2 + \sigma_{\hat{X}}^2 - D}{2}$. Noting that the other case is simply the solution to $R(D)$, this concludes the proof. □

Alternatively, we may express the minimum achievable distortion in terms of P and R as

$$D(P, R) = \begin{cases} \sigma_X^2 + (\sigma_X - \sqrt{P})^2 - 2\sigma_X(\sigma_X - \sqrt{P})\sqrt{1 - 2^{-2R}}, & P < (\sigma_X - \sqrt{\sigma_X^2 - \sigma_X^2 2^{-2R}})^2, \\ \sigma_X^2 2^{-2R}, & P \geq (\sigma_X - \sqrt{\sigma_X^2 - \sigma_X^2 2^{-2R}})^2. \end{cases}$$

For any fixed R, as P increases from 0 to $(\sigma_X - \sqrt{\sigma_X^2 - \sigma_X^2 2^{-2R}})^2$, $D(P, R)$ decreases from $2\sigma_X^2 - 2\sigma_X^2\sqrt{1 - 2^{-2R}}$ to $\sigma_X^2 2^{-2R}$; further increasing P does not affect $D(P, R)$ anymore.

Moreover, the proof of Theorem 1 can be modified to handle to the case $d(p_X, p_{\hat{X}}) = \mathrm{KL}(p_X, p_{\hat{X}})$, where $\mathrm{KL}(p_X, p_{\hat{X}}) = \int p_{\hat{X}}(x) \log \frac{p_{\hat{X}}(x)}{p_X(x)} \mathrm{d}x$ is the KL-divergence between p_X and $p_{\hat{X}}$. Given $(\mu_{\hat{X}}, \sigma_{\hat{X}}^2)$, $\mathrm{KL}(p_X, p_{\hat{X}})$ is minimized when $p_{\hat{X}}$ is a Gaussian distribution. We have that

$$\begin{aligned} \mathrm{KL}(p_X, p_{\hat{X}_G}) &= \frac{\sigma_{\hat{X}}^2 - \sigma_X^2}{2\sigma_X^2} + \frac{1}{2} \log \frac{\sigma_X^2}{\sigma_{\hat{X}}^2}, \\ W_2^2(p_X, p_{\hat{X}_G}) &= (\sigma_X - \sigma_{\hat{X}})^2. \end{aligned}$$

When $\sigma_{\hat{X}} \leq \sigma_X$, both functions are monotonically decreasing in $\sigma_{\hat{X}}$. This implies that the rate-distortion-perception functions under $\mathrm{KL}(p_X, \cdot)$ and $W_2^2(p_X, \cdot)$ also share a one-to-one correspondence in P.

A.2 Achievability of Universal Representations

Before moving on to the achievability of universal representations, we first discuss the functional representations lemmas which play an integral part in the proof. The functional representation lemma states that for jointly distributed random variables X and Y, there exists a random variable U independent of X, and function ϕ such that $Y = \phi(X, U)$. Here, U is not necessarily unique. The strong functional representation lemma [19] states further that there exists a U which is informative of Y in the sense that

$$H(Y|U) \leq I(X; Y) + \log(I(X; Y) + 1) + 4.$$

Note that X and Y may be continuous random variables, and the entropy is still well-defined as long as $Y|U = u$ is discrete for each u. The construction given in [19] satisfies this property.

Theorem 2.

(a) $R^*(\Theta) \le R(\Theta) + \log(R(\Theta)+1) + 5$.

(b) $R^*(\Theta) \ge R(\Theta)$.

Proof of Theorem 2. (a) Let Z be jointly distributed with X such that for any $(D,P) \in \Theta$, there exists $p_{\hat{X}_{D,P}|Z}$ satisfying $\mathbb{E}[\Delta(X,\hat{X}_{D,P})] \le D$ and $d(p_X, p_{\hat{X}_{D,P}}) \le P$. It follows by the strong functional representation lemma that there exist a random variable V, independent of X, and a deterministic function ϕ such that $Z = \phi(X,V)$ and $H(\phi(X,V)|V) \le I(X;Z) + \log(I(X;Z)+1) + 4$. So with V available at both the encoder and the decoder, we can use a class of prefix-free binary codes indexed by V with the expected codeword length no greater than $I(X;Z) + \log(I(X;Z)+1) + 5$ to lossless represent Z. Now it suffices for the decoder to simulate $p_{\hat{X}_{D,P}|Z}$. Specifically, it follows by the functional representation lemma that there exists a random variable $V_{D,P}$, independent of (X,V), and a deterministic function $\psi_{D,P}$ such that $\hat{X}_{D,P} = \psi_{D,P}(Z, V_{D,P})$. Note that V and $V_{D,P}$ can be extracted from random seed U.

(b) For any random variable U, encoding function $f_U : \mathcal{X} \to \mathcal{C}_U$, and decoding functions $g_{U,D,P} : \mathcal{C}_U \to \hat{\mathcal{X}}$, $(D,P) \in \Theta$ satisfying $\mathbb{E}[\Delta(X,\hat{X}_{D,P})] \le D$ and $d(p_X, p_{\hat{X}_{D,P}}) \le P$, we have

$$\begin{aligned}
\mathbb{E}[\ell(f_U(X))] &\ge H(f_U(X)|U) \\
&= I(X; f_U(X)|U) \\
&= I(X; f_U(X), U) \\
&\ge R(\Theta),
\end{aligned}$$

where the last inequality follows by defining $(f_U(X), U)$ as Z, which satisfies the conditions in the definition of $R(\Theta)$. □

Theorem 3. Let $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ be a scalar Gaussian source and assume MSE and $W_2^2(\cdot,\cdot)$ losses. Let Θ be any non-empty set of (D,P) pairs. Then

$$A(\Theta) = 0. \tag{23}$$

Moreover, for any representation Z jointly Gaussian with X such that

$$I(X;Z) = \sup_{(D,P)\in\Theta} R(D,P), \tag{24}$$

we have

$$\Theta \subseteq \Omega(p_{Z|X}) = \Omega(I(X;Z)). \tag{25}$$

Proof of Theorem 3. Let $R = \sup_{(D,P)\in\Theta} R(D,P)$. It is clear that $\Theta \subseteq \Omega(R)$. The distortion-perception tradeoff with respect to R, i.e., the lower boundary of $\Omega(R)$, is given by

$$D = \sigma_X^2 + (\sigma_X - \sqrt{P})^2 - 2\sigma_X(\sigma_X - \sqrt{P})\sqrt{1-2^{-2R}}, \quad P \in [0, (\sigma_X - \sqrt{\sigma_X^2 - \sigma_X^2 2^{-2R}})^2].$$

Every point in $\Omega(R)$ is dominated in a component-wise manner by some (D,P) on this tradeoff. Let Z be jointly Gaussian with X such that $I(X;Z) = R$. Note that $I(X;Z) = R$ implies $\rho_{XZ}^2 = 1 - 2^{-2R}$, where $\rho_{XZ} = \frac{\mathbb{E}[(X-\mu_X)(Z-\mu_Z)]}{\sigma_X\sigma_Z}$. For any (D,P) on the tradeoff, define $\hat{X}_{D,P} = \text{sign}(\rho_{XZ})\frac{\sigma_X - \sqrt{P}}{\sigma_Z}(Z-\mu_Z) + \mu_X$, where $\text{sign}(\rho_{XZ}) = 1$ if $\rho_{XZ} \ge 0$ and $\text{sign}(\rho_{XZ}) = -1$ otherwise. One may verify by direct substitution that

$$\begin{aligned}
W_2^2(p_X, p_{\hat{X}_{D,P}}) &= (\sigma_X - \sigma_{\hat{X}_{D,P}})^2 = P, \\
\mathbb{E}[(X - \hat{X}_{D,P})^2] &= \sigma_X^2 + \sigma_{\hat{X}_{D,P}}^2 - 2\sigma_X(\sigma_X - \sqrt{P})|\rho_{XZ}| \\
&= \sigma_X^2 + (\sigma_X - \sqrt{P})^2 - 2\sigma_X(\sigma_X - \sqrt{P})\sqrt{1-2^{-2R}} \\
&= D.
\end{aligned}$$

This shows that $\Omega(p_{Z|X}) = \Omega(R)$, which further implies $A(\Theta) = 0$. □

Proposition 1 (Equivalence of zero rate penalty and full distortion-perception region)**.** Suppose the following regularity conditions hold:

1) $\sup_{(D,P)\in\Omega(R)} R(D,P) = R'$,

2) the infimum in the definition of $R(\Omega(R'))$ is attainable.

Then the equality $A(\Omega(R')) = 0$ holds if and only if there exists some representation Z with $I(X;Z) = R'$ such that $\Omega(p_{Z|X}) = \Omega(I(X;Z))$.

Proof of Proposition 1. If there exists some representation Z with $I(X;Z) = R'$ such that $\Omega(p_{Z|X}) = \Omega(I(X;Z))$, then $R(\Omega(R')) \leq R'$. Now under condition 1), we must have $A(\Omega(R')) \leq 0$, which implies $A(\Omega(R')) = 0$ as $A(\Omega(R'))$ must be nonnegative.

Under condition 2), there exists some representation Z with $I(X;Z) = R(\Omega(R'))$ such that $\Omega(p_{Z|X}) \supseteq \Omega(R')$. If $A(\Omega(R')) = 0$, then $R(\Omega(R')) = \sup_{(D,P)\in\Omega(R)} R(D,P)$, which together with condition 1) yields $R(\Omega(R')) = R$. Note that $I(X;Z) = R'$ implies $\Omega(p_{Z|X}) \subseteq \Omega(R')$, and consequently we must have $\Omega(p_{Z|X}) = \Omega(R')$. □

Theorem 4. Assume MSE loss and any perception measure $d(\cdot,\cdot)$. Let Z be any arbitrary representation of X. Then

$$\Omega(p_{Z|X}) \subseteq \left\{(D,P) : D \geq \mathbb{E}[\|X-\tilde{X}\|^2] + \inf_{p_{\hat{X}}:d(p_X,p_{\hat{X}})\leq P} W_2^2(p_{\tilde{X}},p_{\hat{X}})\right\} \subseteq \text{cl}(\Omega(p_{Z|X})),$$

where $\tilde{X} = \mathbb{E}[X|Z]$ is the reconstruction minimizing squared error distortion with X under the representation Z and $cl(\cdot)$ denotes set closure. In particular, the two extreme points $(D^{(a)},P^{(a)}) = (\mathbb{E}[\|X-\tilde{X}\|^2], d(p_X,p_{\tilde{X}}))$ and $(D^{(b)},P^{(b)}) = (\mathbb{E}[\|X-\tilde{X}\|^2] + W_2^2(p_{\tilde{X}},p_X), 0)$ are contained in $\text{cl}(\Omega(p_{Z|X}))$.

Proof of Theorem 4. For any $(D,P) \in \Omega(p_{Z|X})$, there exists some $\hat{X}_{D,P}$ jointly distributed with (X,Z) such that $X \leftrightarrow Z \leftrightarrow \hat{X}_{D,P}$ form a Markov chain, $\mathbb{E}[\Delta(X,\hat{X}_{D,P})] \leq D$, and $d(p_X,p_{\hat{X}_{D,P}}) \leq P$. Note that

$$\begin{aligned}
D &\geq \mathbb{E}[\|X-\hat{X}_{D,P}\|^2] \\
&= \mathbb{E}[\|X-\tilde{X}\|^2] + \mathbb{E}[\|\tilde{X}-\hat{X}_{D,P}\|^2] \\
&\geq \mathbb{E}[\|X-\tilde{X}\|^2] + W_2^2(p_{\tilde{X}},p_{\hat{X}_{D,P}}) \\
&\geq \mathbb{E}[\|X-\tilde{X}\|^2] + \inf_{p_{\hat{X}}:d(p_X,p_{\hat{X}})\leq P} W_2^2(p_{\tilde{X}},p_{\hat{X}}).
\end{aligned}$$

Therefore, we have $\Omega(p_{Z|X}) \subseteq \{(D,P) : D \geq \mathbb{E}[\|X-\tilde{X}\|^2] + \inf_{p_{\hat{X}}:d(p_X,p_{\hat{X}})\leq P} W_2^2(p_{\tilde{X}},p_{\hat{X}})\}$.

On the other hand, given

$$(D',P') \in \{(D,P) : D \geq \mathbb{E}[\|X-\tilde{X}\|^2] + \inf_{p_{\hat{X}}:d(p_X,p_{\hat{X}})\leq P} W_2^2(p_{\tilde{X}},p_{\hat{X}})\},$$

for any $\epsilon > 0$, we can find some $p_{\hat{X}'}$ such that $d(p_X,p_{\hat{X}'}) \leq P'$ and $D' + \epsilon \geq \mathbb{E}[\|X-\tilde{X}\|^2] + W_2^2(p_{\tilde{X}},p_{\hat{X}'})$. Let $\hat{X}'$ be jointly distributed with (X,Z) such that $X \leftrightarrow Z \leftrightarrow \hat{X}'$ form a Markov chain and $\mathbb{E}[\|\tilde{X}-\hat{X}'\|^2] \leq W_2^2(p_{\tilde{X}},p_{\hat{X}'}) + \epsilon$. It is possible to find such $\hat{X}'$ by the Markov condition. Note that

$$\begin{aligned}
\mathbb{E}[\|X-\hat{X}'\|^2] &= \mathbb{E}[\|X-\tilde{X}\|^2] + \mathbb{E}[\|\tilde{X}-\hat{X}'\|^2] \\
&\leq \mathbb{E}[\|X-\tilde{X}\|^2] + W_2^2(p_{\tilde{X}},p_{\hat{X}'}) + \epsilon \\
&\leq D' + 2\epsilon.
\end{aligned}$$

Therefore, we have

$$\{(D,P) : D \geq \mathbb{E}[\|X-\tilde{X}\|^2] + \inf_{p_{\hat{X}}:d(p_X,p_{\hat{X}})\leq P} W_2^2(p_{\tilde{X}},p_{\hat{X}})\} \subseteq \text{cl}(\Omega(p_{Z|X})).$$

Choosing $p_{\hat{X}} = p_{\tilde{X}}$ and $p_{\hat{X}} = p_X$ shows respectively that $(D^{(a)}, P^{(a)})$ and $(D^{(b)}, P^{(b)})$ are contained in $\{(D,P) : D \geq \mathbb{E}[\|X - \tilde{X}\|^2] + \inf_{p_{\hat{X}}: d(p_X, p_{\hat{X}}) \leq P} W_2^2(p_{\tilde{X}}, p_{\hat{X}})\}$. □

Quantitative results for the additive and multiplicative gaps. Since $P_3 = 0$ (i.e., $p_{\hat{X}_{D_3,P_3}} = p_X$), it follows that

$$D_3 = \mathbb{E}[\|X - \hat{X}_{D_3,P_3}\|^2] = 2\sigma_X^2 - 2\mathbb{E}[(X - \mu_X)^T(\hat{X}_{D_3,P_3} - \mu_X)]. \tag{26}$$

Note that $I(X; \mathbb{E}[X|\hat{X}_{D_3,P_3}]) \leq I(X; \hat{X}_{D_3,P_3}) = R(D_1, \infty)$, which implies $\mathbb{E}[\|X - \mathbb{E}[X|\hat{X}_{D_3,P_3}]\|^2] \geq D_1$. Let $c = \frac{2\sigma_X^2 - D_3}{2\sigma_X^2}$. We have

$$\begin{aligned} D_1 &\leq \mathbb{E}[\|X - \mathbb{E}[X|\hat{X}_{D_3,P_3}]\|^2] \\ &\leq \mathbb{E}[\|X - \mu_X - c(\hat{X}_{D_3,P_3} - \mu_X)\|^2] \\ &= (1 + c^2)\sigma_X^2 - 2c\mathbb{E}[(X - \mu_X)^T(\hat{X}_{D_3,P_3} - \mu_X)] \\ &\overset{\text{(a)}}{=} \frac{4\sigma_X^2 D_3 - D_3^2}{4\sigma_X^2}, \end{aligned}$$

where (a) is due to (26). So

$$D_3 \geq 2\sigma_X^2 - 2\sigma_X\sqrt{\sigma_X^2 - D_1},$$

which together with the fact that $D^{(b)} \leq 2D_1$ implies

$$\begin{aligned} D^{(b)} - D_3 &\leq 2D_1 - 2\sigma_X^2 + 2\sigma_X\sqrt{\sigma_X^2 - D_1}, \\ \frac{D^{(b)}}{D_3} &\leq \frac{D_1}{\sigma_X^2 - \sigma_X\sqrt{\sigma_X^2 - D_1}}. \end{aligned}$$

It is easy to verify that

$$\begin{aligned} \frac{1}{2}\sigma_X^2 &\geq 2D_1 - 2\sigma_X^2 + 2\sigma_X\sqrt{\sigma_X^2 - D_1} \overset{D_1 \approx 0 \text{ or } \sigma_X^2}{\approx} 0, \\ 2 &\geq \frac{D_1}{\sigma_X^2 - \sigma_X\sqrt{\sigma_X^2 - D_1}} \overset{D_1 \approx \sigma_X^2}{\approx} 1. \end{aligned}$$

A similar argument can be used to bound the gap between (D_1, P_1) and the upper-left extreme point $(\tilde{D}^{(a)}, \tilde{P}^{(a)})$ of blue curve. Note that

$$\tilde{D}^{(a)} = \mathbb{E}[\|X - \mathbb{E}[X|\hat{X}_{D_3,P_3}]\|^2] \leq \frac{4\sigma_X^2 D_3 - D_3^2}{4\sigma_X^2},$$

which together with the fact that $D_1 \geq \frac{1}{2}D_3$ implies

$$\begin{aligned} \tilde{D}^{(a)} - D_1 &\leq \frac{1}{2}D_3 - \frac{D_3^2}{4\sigma_X^2}, \\ \frac{\tilde{D}^{(a)}}{D_1} &\leq 2 - \frac{D_3}{2\sigma_X^2}. \end{aligned}$$

Finally, this implies

$$\frac{1}{4}\sigma_X^2 \geq \frac{1}{2}D_3 - \frac{D_3^2}{4\sigma_X^2} \overset{D_3 \approx 0 \text{ or } 2\sigma_X^2}{\approx} 0, \tag{27}$$

$$2 \geq 2 - \frac{D_3}{2\sigma_X^2} \overset{D_3 \approx 2\sigma_X^2}{\approx} 1. \tag{28}$$

We have previously dealt with the one-shot setting. Now we consider the case where we jointly encode an i.i.d. sequence X^n where each symbol has marginal distribution p_X. Here we assume that $d(\cdot, \cdot)$ is convex in its second argument.

Definition 4. Let Θ be an arbitrary set of (D, P) pairs. A Θ-universal representation of asymptotic rate R is said to exist if we can find a sequence of random variables $U^{(n)}$, encoding functions $f^{(n)}_{U^{(n)}} : \mathcal{X}^n \to \mathcal{C}^{(n)}_U$ and decoding functions $g^{(n)}_{U^{(n)},D,P} : \mathcal{C}^{(n)}_{U^{(n)}} \to \hat{\mathcal{X}}^n$, $(D, P) \in \Theta$, satisfying

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\Delta(X(i), \hat{X}_{D,P}(i))] \le D, \tag{29}$$

$$d\left(p_X, \frac{1}{n}\sum_{i=1}^{n} p_{\hat{X}_{D,P}(i)}\right) \le P \tag{30}$$

such that

$$\limsup_{n\to\infty} \frac{1}{n}\mathbb{E}[\ell(f^{(n)}_{U^{(n)}}(X^n))] \le R,$$

where $\hat{X}^n_{D,P} \triangleq g^{(n)}_{U^{(n)},D,P}(f^{(n)}_{U^{(n)}}(X^n))$. The minimum of such R with respect to Θ is denoted as $R^{(\infty)}(\Theta)$.

Theorem 5. $R^{(\infty)}(\Theta) = R(\Theta)$.

Remark 1. The same conclusion holds if constraint (29) is replaced with

$$\mathbb{E}[\Delta(X(i), \hat{X}_{D,P}(i))] \le D, \quad i = 1, \cdots, n, \tag{31}$$

and/or constraint (30) is replaced with

$$d(p_X, p_{\hat{X}_{D,P}(i)}) \le P, \quad i = 1, \cdots, n. \tag{32}$$

Note that (31) and (32) are more restrictive than (29) and (30), respectively, as

$$(31) \Rightarrow (29),$$

$$(32) \Rightarrow \frac{1}{n}\sum_{i=1}^{n} d(p_X, p_{\hat{X}_{D,P}(i)}) \le P \Rightarrow (30).$$

Moreover, it is easy to verify that under constraints (31) and (32), Theorem 5 holds without the convexity assumption on $d(\cdot, \cdot)$.

Proof of Theorem 5. Let Z be jointly distributed with X such that for any $(D, P) \in \Theta$, there exists $p_{\hat{X}_{D,P}|Z}$ satisfying $\mathbb{E}[\Delta(X, \hat{X}_{D,P})] \le D$ and $d(p_X, p_{\hat{X}_{D,P}}) \le P$. Construct

$$p_{Z^n|X^n} \triangleq \prod_{i=1}^{n} p_{Z(i)|X(i)}$$

with $p_{Z(i)|X(i)} = p_{Z|X}$, $i = 1, \cdots, n$. It follows by the strong functional representation lemma that there exists a random variable $V^{(n)}$, independent of X^n, and a deterministic function $\phi^{(n)}$ such that $Z^n = \phi(X^n, V^{(n)})$ and $H(\phi(X^n, V^{(n)})|V^{(n)}) \le I(X^n; Z^n) + \log(I(X^n; Z^n) + 1) + 4$. So with $V^{(n)}$ available at both the encoder and the decoder, we can use a class of prefix-free binary codes indexed by $V^{(n)}$ with the expected codeword length no greater than $I(X^n; Z^n) + \log(I(X^n; Z^n) + 1) + 5$ to lossless represent Z^n. Moreover, by the functional representation lemma, there exist a random variable $V_{D,P}$, independent of $(X^n; V^{(n)})$, and a deterministic function $\psi_{D,P}$ such that $p_{\psi_{D,P}(Z(i),V_{D,P})|Z(i)} = p_{\hat{X}_{D,P}|Z}$. Note that $V^{(n)}$ and $V_{D,P}$ can be extracted from random seed $U^{(n)}$. Define $\hat{X}_{D,P}(i) = \psi_{D,P}(Z(i), V_{D,P})$, $i = 1, \cdots, n$. It is easy to verify that

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\Delta(X(i), \hat{X}_{D,P}(i))] = \mathbb{E}[\Delta(X, \hat{X}_{D,P})] \le D,$$

$$d\left(p_X, \frac{1}{n}\sum_{i=1}^{n} p_{\hat{X}_{D,P}(i)}\right) = d(p_X, p_{\hat{X}_{D,P}}) \le P.$$

Moreover, notice that

$$\begin{aligned}
&\frac{1}{n}I(X^n;Z^n)+\frac{1}{n}\log(I(X^n;Z^n)+1)+\frac{5}{n}\\
&=I(X;Z)+\frac{1}{n}\log(nI(X;Z)+1)+\frac{5}{n}\\
&\overset{n\to\infty}{\to} I(X;Z).
\end{aligned}$$

This proves that $R^{(\infty)}(\Theta)\le R(\Theta)$.

For any random variable $U^{(n)}$, encoding function $f^{(n)}_{U^{(n)}}:\mathcal{X}^n\to\mathcal{C}^{(n)}_{U^{(n)}}$ and decoding function $g^{(n)}_{U^{(n)},D,P}:\mathcal{C}^{(n)}_{U^{(n)}}\to\hat{\mathcal{X}}^n$, $(D,P)\in\Theta$ satisfying (29) and (30), we have

$$\begin{aligned}
\frac{1}{n}\mathbb{E}[\ell(f^{(n)}_{U^{(n)}}(X^n))] &\ge \frac{1}{n}H(f^{(n)}_{U^{(n)}}(X^n)|U^{(n)})\\
&=\frac{1}{n}I(X^n;f^{(n)}_{U^{(n)}}(X^n)|U^{(n)})\\
&=\frac{1}{n}\sum_{i=1}^{n}I(X(i);f^{(n)}_{U^{(n)}}(X^n)|U^{(n)},X^{i-1})\\
&=\frac{1}{n}\sum_{i=1}^{n}I(X(i);f^{(n)}_{U^{(n)}}(X^n),U^{(n)},X^{i-1})\\
&\ge\frac{1}{n}\sum_{i=1}^{n}I(X(i);f^{(n)}_{U^{(n)}}(X^n),U^{(n)})\\
&=I(X(T);f^{(n)}_{U^{(n)}}(X^n),U^{(n)}|T)\\
&=I(X(T);f^{(n)}_{U^{(n)}}(X^n),U^{(n)},T),
\end{aligned}$$

where T is uniformly distributed over $\{1,\cdots,n\}$ and is independent of X^n and $U^{(n)}$. Note that $\hat{X}_{D,P}(T)$ is a function of $(f^{(n)}_{U^{(n)}}(X^n),U^{(n)},T)$ for any $(D,P)\in\Theta$. Since

$$\begin{aligned}
&p_{X(T)}=p_X,\\
&\mathbb{E}[\Delta(X(T),\hat{X}_{D,P}(T))]=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\Delta(X(i),\hat{X}_{D,P}(i))]\le D,\\
&d(p_{X(T)},p_{\hat{X}_{D,P}(T)})=d\left(p_X,\frac{1}{n}\sum_{i=1}^{n}p_{\hat{X}_{D,P}(i)}\right)\le P,
\end{aligned}$$

it follows that

$$I(X(T);f^{(n)}_U(X^n),U,T)\ge R(\Theta).$$

This completes the proof. □

A.3 Successive Refinement

We now study the case where the rate is not fixed in advance. Bits are sent in two stages as opposed to all at once, with the hope that the reconstructions produced at both stages perform near-optimally in both perception and distortion compared to what can be achieved by one-stage communication at both the lower rate and the higher rate. Two-stage procedures arise frequently under practical constraints, and previous works have considered this only under distortion losses. We address the extension of universal representations to this setting within the successive refinement [9] framework.

Definition 5 (Two-stage Coding)**.** Given two sets of (D,P) pairs Θ_1 and Θ_2, we say rate pair (R_1,R_2) is (operationally) achievable if there exists random variable U, encoding functions

$$f_U:\mathcal{X}\to\mathcal{C}_U,\quad f_{U,f_U(X)}:\mathcal{X}\to\mathcal{C}_{U,f_U(X)},$$

and decoding functions

$$g_{U,D_1,P_1} : \mathcal{C}_U \to \hat{\mathcal{X}}, \quad g_{U,f_U(X),D_2,P_2} : \mathcal{C}_{U,f_U(X)} \to \hat{\mathcal{X}}$$

for each $(D_1, P_1) \in \Theta_1$ and $(D_2, P_2) \in \Theta_2$, such that

$$\begin{aligned} \mathbb{E}[\ell(f_U(X))] \le R_1, \quad \mathbb{E}[\ell(f_{U,f_U(X)}(X))] \le R_2, \\ \mathbb{E}[\Delta(X, \hat{X}_{1,D_1,P_1})] \le D_1, \quad \mathbb{E}[\Delta(X, \hat{X}_{2,D_2,P_2})] \le D_2, \\ d(p_X, p_{\hat{X}_{1,D_1,P_1}}) \le P_1, \quad d(p_X, p_{\hat{X}_{2,D_2,P_2}}) \le P_2, \end{aligned}$$

where $\hat{X}_{1,D_1,P_1} = g_{U,D_1,P_1}(f_U(X))$ and $\hat{X}_{2,D_2,P_2} = g_{U,f_U(X),D_2,P_2}(f_{U,f_U(X)}(X))$. The closure of the set of such (R_1, R_2) is denoted as $\mathcal{R}^*(\Theta_1, \Theta_2)$.

Here, f_U acts with each g_{U,D_1,P_1} forming a low rate encoder-decoder pair to meet each constraint $(D_1, P_1) \in \Theta_1$. Thereafter, $f_{U,f_U(X)}$ encodes additional information about the source which is combined with the low rate encoding to produce a high rate reconstruction through $g_{U,f_U(X),D_2,P_2}$ meeting each constraint $(D_2, P_2) \in \Theta_2$.

Definition 6 (Inner and outer bounds)**.** Define

$$\begin{aligned} \underline{\mathcal{R}}(\Theta_1, \Theta_2) = \bigcup_{p_{Z_1,Z_2|X}} \{(R_1, R_2) \in \mathbb{R}_+^2 : R_1 \ge I(X; Z_1) + \log(I(X; Z_1) + 1) + 5, \\ R_1 + R_2 \ge I(X; Z_1, Z_2) + \log(I(X; Z_1) + 1) + \log(I(X; Z_2|Z_1) + 1) + 10\}, \\ \overline{\mathcal{R}}(\Theta_1, \Theta_2) = \bigcup_{p_{Z_1,Z_2|X}} \{(R_1, R_2) \in \mathbb{R}_+^2 : R_1 \ge I(X; Z_1), R_1 + R_2 \ge I(X; Z_1, Z_2)\} \end{aligned}$$

with the unions taken over $p_{Z_1,Z_2|X}$ such that for any $(D_1, P_1) \in \Theta_1$ and $(D_2, P_2) \in \Theta_2$, there exists

$$p_{\hat{X}_{1,D_1,P_1}|Z_1} \quad \text{and} \quad p_{\hat{X}_{2,D_2,P_2}|Z_2}$$

satisfying

$$\begin{aligned} \mathbb{E}[\Delta(X, \hat{X}_{1,D_1,P_1})] \le D_1, \quad \mathbb{E}[\Delta(X, \hat{X}_{2,D_2,P_2})] \le D_2, \\ d(p_X, p_{\hat{X}_{1,D_1,P_1}}) \le P_1, \quad d(p_X, p_{\hat{X}_{2,D_2,P_2}}) \le P_2. \end{aligned}$$

We now characterize the operational definition in terms of these information rate regions.

Theorem 6. *cl*$(\underline{\mathcal{R}}(\Theta_1, \Theta_2)) \subseteq \mathcal{R}^*(\Theta_1, \Theta_2) \subseteq$ *cl*$(\overline{\mathcal{R}}(\Theta_1, \Theta_2))$.

Proof of Theorem 6. (a) Let Z_1 and Z_2 be jointly distributed with X such that for any $(D_1, P_1) \in \Theta_1$ and $(D_2, P_2) \in \Theta_2$, there exist $p_{\hat{X}_{1,D_1,P_1}|Z_1}$ and $p_{\hat{X}_{2,D_2,P_2}|Z_2}$ satisfying $\mathbb{E}[\Delta(X, \hat{X}_{1,D_1,P_1})] \le D_1$, $\mathbb{E}[\Delta(X, \hat{X}_{2,D_2,P_2})] \le D_2$, $d(p_X, p_{\hat{X}_{1,D_1,P_1}}) \le P_1$, and $d(p_X, p_{\hat{X}_{2,D_2,P_2}}) \le P_2$. It follows by the strong functional representation lemma that there exist a random variable V_1, independent of X, and a deterministic function ϕ_1 such that $Z_1 = \phi_1(X, V_1)$ and $H(\phi_1(X, V_1)|V_1) \le I(X; Z_1) + \log(I(X; Z_1) + 1) + 4$; moreover, there exist a random variable V_2, independent of (X, V_1), and a deterministic function ϕ_2 such that $Z_2 = \phi_2(X, Z_1, V_2)$ and $H(\phi_2(X, Z_1, V_2)|Z_1, V_2) \le I(X; Z_2|Z_1) + \log(I(X; Z_2|Z_1) + 1) + 4$. So with (V_1, V_2) available at both the encoder and the decoder, we can use a class of prefix-free binary codes indexed by V_1 with the expected codeword length no greater than $I(X; Z_1) + \log(I(X; Z_1) + 1) + 5$ to lossless represent Z_1 and then use a class of prefix-free binary codes indexed by (Z_1, V_2) with the expected codeword length no greater than $I(X; Z_2|Z_1) + \log(I(X; Z_2|Z_1) + 1) + 5$ to lossless represent Z_2. Note that in the first stage we can send the codeword used to represent Z_1 and with a certain probability the codeword used to represent Z_2.

Now it suffices for the decoder to simulate $p_{\hat{X}_{1,D_1,P_1}|Z_1}$ and $p_{\hat{X}_{2,D_2,P_2}|Z_2}$. Specifically, it follows by the functional representation lemma that there exist random variables

$$V_{1,D_1,P_1} \quad \text{and} \quad V_{2,D_2,P_2},$$

independent of (X, V_1, V_2), and deterministic functions

$$\psi_{1,D_1,P_1} \quad \text{and} \quad \psi_{2,D_2,P_2}$$

such that

$$\hat{X}_{1,D_1,P_1} = \psi_{1,D_1,P_1}(Z_1, V_{1,D_1,P_1}),$$
$$\hat{X}_{2,D_2,P_2} = \psi_{2,D_2,P_2}(Z_2, V_{2,D_2,P_2}).$$

This proves the desired result.

(b) For any random variable U, encoding functions $f_U : \mathcal{X} \to \mathcal{C}_U$, $f_{U,f_U(X)} : \mathcal{X} \to \mathcal{C}_{U,f_U(X)}$ and decoding functions $g_{U,D_1,P_1} : \mathcal{C}_U \to \hat{\mathcal{X}}$, $(D_1, P_1) \in \Theta_1$, $g_{U,f_U(X),D_2,P_2} : \mathcal{C}_{U,f_U(X)} \to \hat{\mathcal{X}}$, $(D_2, P_2) \in \Theta_2$, satisfying $\mathbb{E}[\Delta(X, \hat{X}_{1,D_1,P_1})] \le D_1$, $\mathbb{E}[\Delta(X, \hat{X}_{2,D_2,P_2})] \le D_2$, $d(p_X, p_{\hat{X}_{1,D_1,P_1}}) \le P_1$, and $d(p_X, p_{\hat{X}_{2,D_2,P_2}}) \le P_2$, we have

$$\begin{aligned}
\mathbb{E}[\ell(f_U(X))] &\ge H(f_U(X)|U) \\
&= I(X; f_U(X)|U) \\
&\ge I(X; f_U(X), U) \\
&= I(X; Z_1)
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}[\ell(f_U(X))] + \mathbb{E}[\ell(f_{U,f_U(X)}(X))] &\ge H(f_U(X)|U) + H(f_{U,f_U(X)}(X)|U, f_U(X)) \\
&\ge I(X; f_U(X)|U) + I(X; f_{U,f_U(X)}(X)|U, f_U(X)) \\
&= I(X; f_U(X), f_{U,f_U(X)}(X)|U) \\
&= I(X; f_U(X), f_{U,f_U(X)}(X), U) \\
&= I(X; Z_1, Z_2),
\end{aligned}$$

where we define $Z_1 = (f_U(X), U)$ and $Z_2 = (f_U(X), f_{U,f_U(X)}(X), U)$. So $(R_1, R_2) \in \overline{\mathcal{R}}(\Theta_1, \Theta_2)$ for any (R_1, R_2) with $R_1 \ge \mathbb{E}[\ell(f_U(X))]$ and $R_2 \ge \mathbb{E}[\ell(f_{U,f_U(X)}(X))]$. This completes the proof. □

Definition 7 (Asymptotic rate region)**.** Given two sets of (D, P) pairs Θ_1 and Θ_2, we say rate pair (R_1, R_2) is asymptotically achievable if there exists a sequence of random variables $U^{(n)}$, encoding functions

$$f^{(n)}_{U^{(n)}} : \mathcal{X}^n \to \mathcal{C}^{(n)}_{U^{(n)}}, \quad f^{(n)}_{U^{(n)}, f^{(n)}_U(X^n)} : \mathcal{X}^n \to \mathcal{C}^{(n)}_{U^{(n)}, f^{(n)}_{U^{(n)}}(X^n)}$$

and decoding functions

$$g^{(n)}_{U^{(n)},D_1,P_1} : \mathcal{C}^{(n)}_{U^{(n)}} \to \hat{\mathcal{X}}^n, \quad g^{(n)}_{U^{(n)}, f^{(n)}_{U^{(n)}}(X^n), D_2, P_2} : \mathcal{C}^{(n)}_{U^{(n)}, f^{(n)}_{U^{(n)}}(X^n)} \to \hat{\mathcal{X}}^n$$

$(D_1, P_1) \in \Theta_1$, $(D_2, P_2) \in \Theta_2$, satisfying

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\Delta(X(i), \hat{X}_{1,D_1,P_1}(i))] \le D_1, \quad \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\Delta(X(i), \hat{X}_{2,D_2,P_2}(i))] \le D_2, \tag{33}$$

$$d\left(p_X, \frac{1}{n}\sum_{i=1}^{n} p_{\hat{X}_{1,D_1,P_1}(i)}\right) \le P_1, \quad d\left(p_X, \frac{1}{n}\sum_{i=1}^{n} p_{\hat{X}_{2,D_2,P_2}(i)}\right) \le P_2 \tag{34}$$

such that

$$\limsup_{n\to\infty} \frac{1}{n}\mathbb{E}[\ell(f^{(n)}_{U^{(n)}}(X^n))] \le R_1, \quad \limsup_{n\to\infty} \frac{1}{n}\mathbb{E}[\ell(f^{(n)}_{U^{(n)}, f^{(n)}(X^n)}(X^n))] \le R_2,$$

where

$$\hat{X}^n_{1,D_1,P_1} = g^{(n)}_{U^{(n)},D_1,P_1}(f^{(n)}_{U^{(n)}}(X^n))$$

and

$$\hat{X}^n_{2,D_2,P_2} = g^{(n)}_{U^{(n)}, f^{(n)}_{U^{(n)}}(X^n), D_2, P_2}(f^{(n)}_{U^{(n)}, f^{(n)}(X^n)}(X^n)).$$

The set of such (R_1, R_2) is denoted as $\mathcal{R}^{(\infty)}(\Theta_1, \Theta_2)$.

Theorem 7. $\mathcal{R}^{(\infty)}(\Theta_1,\Theta_2) = cl(\overline{\mathcal{R}}(\Theta_1,\Theta_2))$.

Remark 2. Remark 1 is applicable here as well.

Proof of Theorem 7. Let Z_1 and Z_2 be jointly distributed with X such that for any $(D_1,P_1)\in\Theta_1$ and $(D_2,P_2)\in\Theta_2$, there exist $p_{\hat{X}_{1,D_1,P_1}|Z_1}$ and $p_{\hat{X}_{2,D_2,P_2}|Z_2}$ satisfying $\mathbb{E}[\Delta(X,\hat{X}_{1,D_1,P_1})]\le D_1$, $\mathbb{E}[\Delta(X,\hat{X}_{2,D_2,P_2})]\le D_2$, $d(p_X,p_{\hat{X}_{1,D_1,P_1}})\le P_1$, and $d(p_X,p_{\hat{X}_{2,D_2,P_2}})\le P_2$. Construct

$$p_{Z_1^nZ_2^n|X^n}\triangleq\prod_{i=1}^n p_{Z_1(i)Z_2(i)|X(i)}$$

with $p_{Z_1(i)Z_2(i)|X(i)}=p_{Z_1Z_2|X}$, $i=1,\cdots,n$. It follows by the strong functional representation lemma that there exist a random variable $V_1^{(n)}$, independent of X^n, and a deterministic function ϕ_1 such that $Z_1^n=\phi_1(X^n,V_1^{(n)})$ and $H(\phi_1(X^n,V_1^{(n)})|V_1^{(n)})\le I(X^n;Z_1^n)+\log(I(X^n;Z_1^n)+1)+4$; moreover, there exist a random variable $V_2^{(n)}$, independent of $(X^n,V_1^{(n)})$, and a deterministic function ϕ_2 such that $Z_2^n=\phi_2(X^n,Z_1^n,V_2^{(n)})$ and $H(\phi_2(X^n,Z_1^n,V_2^{(n)})|Z_1^n,V_2^{(n)})\le I(X^n;Z_2^n|Z_1^n)+\log(I(X^n;Z_2^n|Z_1^n)+1)+4$. So with $(V_1^{(n)},V_2^{(n)})$ available at both the encoder and the decoder, we can use a class of prefix-free binary codes indexed by $V_1^{(n)}$ with the expected codeword length no greater than $I(X^n;Z_1^n)+\log(I(X^n;Z_1^n)+1)+5$ to lossless represent Z_1^n and then use a class of prefix-free binary codes indexed by $(Z_1^n,V_2^{(n)})$ with the expected codeword length no greater than $I(X^n;Z_2^n|Z_1^n)+\log(I(X^n;Z_2^n|Z_1^n)+1)+5$ to lossless represent Z_2^n.

Note that in the first stage we can send the codeword used to represent Z_1^n and with a certain probability the codeword used to represent Z_2^n. Moreover, by the functional representation lemma, there exist random variables V_{1,D_1,P_1} and V_{2,D_2,P_2}, independent of $(X^n;V_1^{(n)},V_2^{(n)})$, and deterministic functions ψ_{1,D_1,P_1} and ψ_{2,D_2,P_2} such that $p_{\psi_{1,D_1,P_1}(Z_1(i),V_{1,D_1,P_1})|Z_1(i)}=p_{\hat{X}_{1,D_1,P_1}|Z_1}$ and $p_{\psi_{2,D_2,P_2}(Z_2(i),V_{2,D_2,P_2})|Z_2(i)}=p_{\hat{X}_{2,D_2,P_2}|Z_2}$. Define $\hat{X}_{1,D_1,P_1}(i)=\psi_{1,D_1,P_1}(Z_1(i),V_{1,D_1,P_1})$ and $\hat{X}_{2,D_2,P_2}(i)=\psi_{2,D_2,P_2}(Z_2(i),V_{2,D_2,P_2})$, $i=1,\cdots,n$. It is easy to verify that

$$\frac{1}{n}\sum_{i=1}^n\mathbb{E}[\Delta(X(i),\hat{X}_{k,D_k,P_k}(i))]=\mathbb{E}[\Delta(X,\hat{X}_{k,D_k,P_k})]\le D_k,\quad k=1,2,$$
$$d\left(p_X,\frac{1}{n}\sum_{i=1}^n p_{\hat{X}_{k,D_k,P_k}(i)}\right)=d(p_X,p_{\hat{X}_{k,D_k,P_k}})\le P_k,\quad k=1,2.$$

Furthermore,

$$\begin{aligned}&\frac{1}{n}I(X^n;Z_1^n)+\frac{1}{n}\log(I(X^n;Z_1^n)+1)+\frac{5}{n}\\&=I(X;Z_1)+\frac{1}{n}\log(nI(X;Z_1)+1)+\frac{5}{n}\\&\overset{n\to\infty}{\to}I(X;Z_1)\end{aligned}$$

and

$$\begin{aligned}&\frac{1}{n}I(X^n;Z_1^n)+\frac{1}{n}\log(I(X^n;Z_1^n)+1)+\frac{5}{n}+\frac{1}{n}I(X^n;Z_2^n|Z_1^n)+\frac{1}{n}\log(I(X^n;Z_2^n|Z_1^n)+1)+\frac{5}{n}\\&=\frac{1}{n}I(X^n;Z_1^n,Z_2^n)+\frac{1}{n}\log(I(X^n;Z_1^n)+1)+\frac{1}{n}\log(I(X^n;Z_2^n|Z_1^n)+1)+\frac{10}{n}\\&=I(X;Z_1,Z_2)+\frac{1}{n}\log(nI(X;Z_1)+1)+\frac{1}{n}\log(nI(X;Z_2|Z_1)+1)+\frac{10}{n}\\&\overset{n\to\infty}{\to}I(X;Z_1,Z_2).\end{aligned}$$

This proves that cl$(\overline{\mathcal{R}}(\Theta_1,\Theta_2))\subseteq R^{(\infty)}(\Theta_1,\Theta_2)$.

For any random variable $U^{(n)}$, encoding functions $f^{(n)}_{U^{(n)}} : \mathcal{X}^n \to \mathcal{C}^{(n)}_{U^{(n)}}$, $f^{(n)}_{U^{(n)},f^{(n)}_{U^{(n)}}(X^n)} : \mathcal{X}^n \to \mathcal{C}^{(n)}_{U^{(n)},f^{(n)}_{U^{(n)}}(X^n)}$ and decoding functions $g^{(n)}_{U^{(n)},D_1,P_1} : \mathcal{C}^{(n)}_{U^{(n)}} \to \hat{\mathcal{X}}^n$, $(D_1,P_1) \in \Theta_1$, $g^{(n)}_{U^{(n)},f^{(n)}_U(X^n),D_2,P_2} : \mathcal{C}_{U^{(n)},f^{(n)}_U(X^n)} \to \hat{\mathcal{X}}^n$, $(D_2,P_2) \in \Theta_2$, satisfying (33) and (34), we have

$$\begin{aligned}
\frac{1}{n}\mathbb{E}[\ell(f^{(n)}_{U^{(n)}}(X^n))] &\geq \frac{1}{n}H(f^{(n)}_{U^{(n)}}(X^n)|U^{(n)}) \\
&= \frac{1}{n}I(X^n; f^{(n)}_{U^{(n)}}(X^n)|U^{(n)}) \\
&= \frac{1}{n}I(X^n; f^{(n)}_{U^{(n)}}(X^n), U^{(n)}) \\
&= \frac{1}{n}\sum_{i=1}^{n} I(X(i); f^{(n)}_{U^{(n)}}(X^n), U^{(n)}|X^{i-1}) \\
&= \frac{1}{n}\sum_{i=1}^{n} I(X(i); f^{(n)}_{U^{(n)}}(X^n), U^{(n)}, X^{i-1}) \\
&\geq \frac{1}{n}\sum_{i=1}^{n} I(X(i); f^{(n)}_{U^{(n)}}(X^n), U^{(n)}) \\
&= I(X(T); f^{(n)}_{U^{(n)}}(X^n), U^{(n)}|T) \\
&= I(X(T); f^{(n)}_{U^{(n)}}(X^n), U^{(n)}, T) \\
&= I(X(T); Z_1)
\end{aligned}$$

and

$$\begin{aligned}
&\frac{1}{n}\mathbb{E}[\ell(f^{(n)}_{U^{(n)}}(X^n))] + \frac{1}{n}\mathbb{E}[\ell(f^{(n)}_{U^{(n)},f^{(n)}_{U^{(n)}}(X^n)}(X^n))] \\
&\geq \frac{1}{n}H(f^{(n)}_{U^{(n)}}(X^n)|U^{(n)}) + \frac{1}{n}H(f^{(n)}_{U^{(n)},f^{(n)}_U(X^n)}(X^n)|U^{(n)}, f^{(n)}_{U^{(n)}}(X^n)) \\
&\geq \frac{1}{n}I(X^n; f^{(n)}_{U^{(n)}}(X^n)|U^{(n)}) + \frac{1}{n}I(X^n; f^{(n)}_{U,f_U(X^n)}(X^n)|U^{(n)}, f^{(n)}_{U^{(n)}}(X^n)) \\
&= \frac{1}{n}I(X^n; f^{(n)}_{U^{(n)}}(X^n), f^{(n)}_{U^{(n)},f^{(n)}_{U^{(n)}}(X^n)}(X^n)|U^{(n)}) \\
&= \frac{1}{n}I(X^n; f^{(n)}_{U^{(n)}}(X^n), f^{(n)}_{U^{(n)},f^{(n)}_{U^{(n)}}(X^n)}(X^n), U^{(n)}) \\
&= \frac{1}{n}\sum_{i=1}^{n} I(X(i); f^{(n)}_{U^{(n)}}(X^n), f^{(n)}_{U^{(n)},f^{(n)}_{U^{(n)}}(X^n)}(X^n), U^{(n)}|X^{i-1}) \\
&= \frac{1}{n}\sum_{i=1}^{n} I(X(i); f^{(n)}_{U^{(n)}}(X^n), f^{(n)}_{U^{(n)},f^{(n)}_{U^{(n)}}(X^n)}(X^n), U^{(n)}, X^{i-1}) \\
&\geq \frac{1}{n}\sum_{i=1}^{n} I(X(i); f^{(n)}_{U^{(n)}}(X^n), f^{(n)}_{U^{(n)},f^{(n)}_{U^{(n)}}(X^n)}(X^n), U^{(n)}) \\
&= I(X(T); f^{(n)}_{U^{(n)}}(X^n), f^{(n)}_{U^{(n)},f^{(n)}_{U^{(n)}}(X^n)}(X^n), U^{(n)}|T) \\
&= I(X(T); f^{(n)}_{U^{(n)}}(X^n), f^{(n)}_{U^{(n)},f^{(n)}_{U^{(n)}}(X^n)}(X^n), U^{(n)}, T) \\
&= I(X(T); Z_1, Z_2),
\end{aligned}$$

where T is uniformly distributed over $\{1, \cdots, n\}$ and is independent of $(X^n, U^{(n)})$, and we define $Z_1 = (f_{U^{(n)}}^{(n)}(X^n), U^{(n)}, T)$ and $Z_2 = (f_{U^{(n)}}^{(n)}(X^n), f_{U^{(n)}, f_{U^{(n)}}^{(n)}(X^n)}^{(n)}(X^n), U^{(n)}, T)$. Since

$$\begin{aligned}
&p_{X(T)} = p_X, \\
&\mathbb{E}[\Delta(X(T), \hat{X}_{k,D_k,P_k}(T))] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\Delta(X(i), \hat{X}_{k,D_k,P_k}(i))] \le D_k, \quad k = 1, 2, \\
&d(p_{X(T)}, p_{\hat{X}_{k,D_k,P_k}(T)}) = d\left(p_X, \frac{1}{n}\sum_{i=1}^{n} p_{\hat{X}_{k,D_k,P_k}(i)}\right) \le P_k, \quad k = 1, 2,
\end{aligned}$$

and $\hat{X}_{k,D_k,P_k}(T)$ is a function of Z_k, $k = 1, 2$, we must have $(R_1, R_2) \in \overline{\mathcal{R}}(\Theta_1, \Theta_2)$ for any (R_1, R_2) with $R_1 \ge \frac{1}{n}\mathbb{E}[\ell(f_U^{(n)}(X^n))]$ and $R_2 \ge \frac{1}{n}\mathbb{E}[\ell(f_{U, f_U^{(n)}(X^n)}^{(n)}(X^n))]$. This completes the proof. □

Definition 8. We say that Θ_1 can be successively refined to Θ_2 if $(R(\Theta_1), R(\Theta_2) - R(\Theta_1)) \in \mathrm{cl}(\overline{\mathcal{R}}(\Theta_1, \Theta_2))$

Remark 3. To show the asymptotic feasibility of successive refinement from Θ_1 to Θ_2, it suffices to find $p_{Z_1,Z_2|X}$ such that

$$I(X; Z_1) = R(\Theta_1), \quad I(X; Z_1, Z_2) = R(\Theta_2),$$

and for any $(D_1, P_1) \in \Theta_1$ and $(D_2, P_2) \in \Theta_2$, there exists

$$p_{\hat{X}_{1,D_1,P_1}|Z_1} \quad \text{and} \quad p_{\hat{X}_{2,D_2,P_2}|Z_2}$$

satisfying

$$\begin{aligned}
\mathbb{E}[\Delta(X; \hat{X}_{1,D_1,P_1})] \le D_1, \quad \mathbb{E}[\Delta(X; \hat{X}_{2,D_2,P_2})] \le D_2, \\
d(p_X, p_{\hat{X}_{1,D_1,P_1}}) \le P_1, \quad d(p_X, p_{\hat{X}_{2,D_2,P_2}}) \le P_2.
\end{aligned}$$

In the Gaussian case, it is easy to show that successive refinement from Θ_1 to Θ_2 is always asymptotically feasible for $R(\Theta_2) \ge R(\Theta_1)$.

Theorem 8. *Let $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ be a scalar Gaussian source and assume MSE and $W_2^2(\cdot,\cdot)$ losses. Let Θ_1 and Θ_2 be arbitrary non-empty sets of (D, P) pairs with $R(\Theta_1) \le R(\Theta_2)$. Then $(R(\Theta_1), R(\Theta_2) - R(\Theta_1)) \in \mathcal{R}^{(\infty)}(\Theta_1, \Theta_2)$, i.e., successive refinement from Θ_1 to Θ_2 is feasible.*

Proof of Theorem 8. Let $Z_2 = Z_1 + N_1$ and $X = Z_2 + N_2$, where

$$\begin{aligned}
Z_1 &\sim \mathcal{N}(\mu_X, \sigma_X^2(1 - 2^{-2R(\Theta_1)})), \\
N_1 &\sim \mathcal{N}(0, \sigma_X^2(2^{-2R(\Theta_1)} - 2^{-2R(\Theta_2)})), \\
N_2 &\sim \mathcal{N}(0, \sigma_X^2 2^{-2R(\Theta_2)})
\end{aligned}$$

are mutually independent. It is easy to verify that $I(X; Z_1) = R(\Theta_1)$ and $I(X; Z_1, Z_2) = I(X; Z_2) = R(\Theta_2)$. In view of Theorem 3, we have $\Theta_i \subseteq \Omega(p_{Z_i|X}) = \Omega(R(\Theta_i))$, $i = 1, 2$. So successive refinement from Θ_1 to Θ_2 is indeed asymptotically feasible. □

Theorem 9 (Approximate refinability under the iRDPF)**.** *Assume MSE loss and any perception measure $d(\cdot,\cdot)$. Let m be the dimension of X and*

$$\delta_R(\sigma_N^2) = R(\Theta_1) - R\left(\frac{\sigma_X^2\sigma_N^2}{\sigma_X^2 + \sigma_N^2}, \infty\right) + \frac{m}{2}\log\frac{(D_1^* + \sigma_N^2)(D_2^* + \sigma_N^2)}{\sigma_N^4},$$

where

$$\begin{aligned}
D_1^* &= \inf\{D_1' : (D_1', P_1') \in \Theta_1 \text{ for some } P_1'\}, \\
D_2^* &= \inf\{D_2' : (D_2', P_2') \in \Theta_1 \text{ for some } P_2'\}.
\end{aligned}$$

Then for any non-empty Θ_1 and Θ_2,

$$(R(\Theta_1), R(\Theta_2) - R(\Theta_1) + \inf_{\sigma_N^2 > 0} \delta_R(\sigma_N^2)) \in \mathcal{R}^{(\infty)}(\Theta_1, \Theta_2).$$

Remark 4. We have

$$\delta_R(\frac{\sigma_X^2 D_1^*}{\sigma_X^2 - D_1^*}) = R(\Theta_1) - R(D_1^*, \infty) + \frac{m}{2} \log \frac{(\sigma_X^2 (D_1^* + D_2^*) - D_1^* D_2^*)(2\sigma_X^2 - D_1^*)}{\sigma_X^4 D_1^*}.$$

In particular, $\delta_R(\frac{\sigma_X^2 D_1^*}{\sigma_X^2 - D_1^*}) \le m$ when $R(\Theta_1) = R(D_1^*, \infty)$ and $D_2^* \le D_1^*$. In the scalar case, this shows that the penalty for refinement (as opposed to sending all bits at once) is not more than 1 bit.

Proof of Theorem 9. This proof is an adaptation of the result from Lastras and Berger [18].

For any $\epsilon > 0$, we can find Z_k with $I(X; Z_k) \le R(\Theta_k) + \epsilon$ such that for any $(D_k, P_k) \in \Theta_k$, there exists $p_{\hat{X}_{k,D_k,P_k}|Z_k}$ satisfying $\mathbb{E}[\|X - \hat{X}_{k,D_k,P_k}\|^2] \le D_k$ and $d(p_X, p_{\hat{X}_{k,D_k,P_k}}) \le P_k$, $k = 1, 2$. We define X, Z_1, and Z_2 in the same probability space such that $Z_1 \leftrightarrow X \leftrightarrow Z_2$ form a Markov chain. It suffices to show that $I(X; Z_1, Z_2) \le R(\Theta_2) + \delta_R(\sigma_N^2) + 2\epsilon$ for any $\sigma_N^2 > 0$.

Let $N \sim \mathcal{N}(0, \frac{\sigma_N^2}{m} I_m)$ be an m-dimensional (multivariate) Gaussian random variable independent of (X, Z_1, Z_2). We have

$$\begin{aligned} I(X; Z_1, Z_2) - I(X; Z_2) &= I(X; Z_1 | Z_2) \\ &\le I(X; Z_1, X + N | Z_2) \\ &= I(X; X + N | Z_2) + I(X; Z_1 | Z_2, X + N). \end{aligned} \tag{35}$$

Note that

$$\begin{aligned} I(X; X + N | Z_2) &= I(X - \mathbb{E}[X|Z_2]; X - \mathbb{E}[X|Z_2] + N | Z_2) \\ &\le I(Z_2, X - \mathbb{E}[X|Z_2]; X - \mathbb{E}[X|Z_2] + N) \\ &= I(X - \mathbb{E}[X|Z_2]; X - \mathbb{E}[X|Z_2] + N) \\ &\le \frac{m}{2} \log \frac{D_2^* + \sigma_N^2}{\sigma_N^2} \end{aligned} \tag{36}$$

and

$$\begin{aligned} & I(X; Z_1 | Z_2, X + N) \\ & \le I(X, Z_2; Z_1 | X + N) \\ & = I(X; Z_1 | X + N) \\ & = I(X; Z_1, X + N) - I(X; X + N) \\ & = I(X; X + N | Z_1) + I(X; Z_1) - I(X; X + N) \\ & \le I(X; X + N | Z_1) + R(\Theta_1) + \epsilon - I(X; X + N) \\ & \overset{(\beta)}{\le} I(X; X + N | Z_1) + R(\Theta_1) + \epsilon - R(\frac{\sigma_X^2 \sigma_N^2}{\sigma_X^2 + \sigma_N^2}, \infty) \\ & = I(X - \mathbb{E}[X|Z_1]; X - \mathbb{E}[X|Z_1] + N | Z_1) + R(\Theta_1) - R(\frac{\sigma_X^2 \sigma_N^2}{\sigma_X^2 + \sigma_N^2}, \infty) + \epsilon \\ & \le I(Z_1, X - \mathbb{E}[X|Z_1]; X - \mathbb{E}[X|Z_1] + N) + R(\Theta_1) - R(\frac{\sigma_X^2 \sigma_N^2}{\sigma_X^2 + \sigma_N^2}, \infty) + \epsilon \\ & = I(X - \mathbb{E}[X|Z_1]; X - \mathbb{E}[X|Z_1] + N) + R(\Theta_1) - R(\frac{\sigma_X^2 \sigma_N^2}{\sigma_X^2 + \sigma_N^2}, \infty) + \epsilon \\ & \le \frac{m}{2} \log \frac{D_1^* + \sigma_N^2}{\sigma_N^2} + R(\Theta_1) - R(\frac{\sigma_X^2 \sigma_N^2}{\sigma_X^2 + \sigma_N^2}, \infty) + \epsilon, \end{aligned} \tag{37}$$

where (β) is because $\mathbb{E}[\|X - \frac{\sigma_X^2}{\sigma_X^2 + \sigma_N^2}(X + N) - \frac{\sigma_N^2}{\sigma_X^2 + \sigma_N^2} \mu_X\|^2] = \frac{\sigma_X^2 \sigma_N^2}{\sigma_X^2 + \sigma_N^2}$ and consequently $I(X; X + N) \ge R(\frac{\sigma_X^2 \sigma_N^2}{\sigma_X^2 + \sigma_N^2}, \infty)$. Substituting (36) and (37) into (35) gives

$$I(X; Z_1, Z_2) - I(X; Z_2) \le \delta_R(\sigma_N^2) + \epsilon,$$

which further implies

$$I(X; Z_1, Z_2) \le R(\Theta_2) + \delta_R(\sigma_N^2) + 2\epsilon.$$

This completes the proof. □

B Experiments

Training lasted 30 epochs for MNIST and 80 epochs for SVHN, and alternates between training the encoder and decoder with the critic fixed and training the critic with the encoder and decode fixed. The learning rate was decayed by a factor of 5 after 20 epochs for MNIST, and after 25 epochs for SVHN. All models were trained with the Adam optimizer. The batch size used was 64. All training was performed on a Tesla V100 GPU. Training a single model takes about 10 minutes and 30 minutes for MNIST and SVHN, respectively. We used the standard train/test splits.

B.1 Comparison of Quantizers

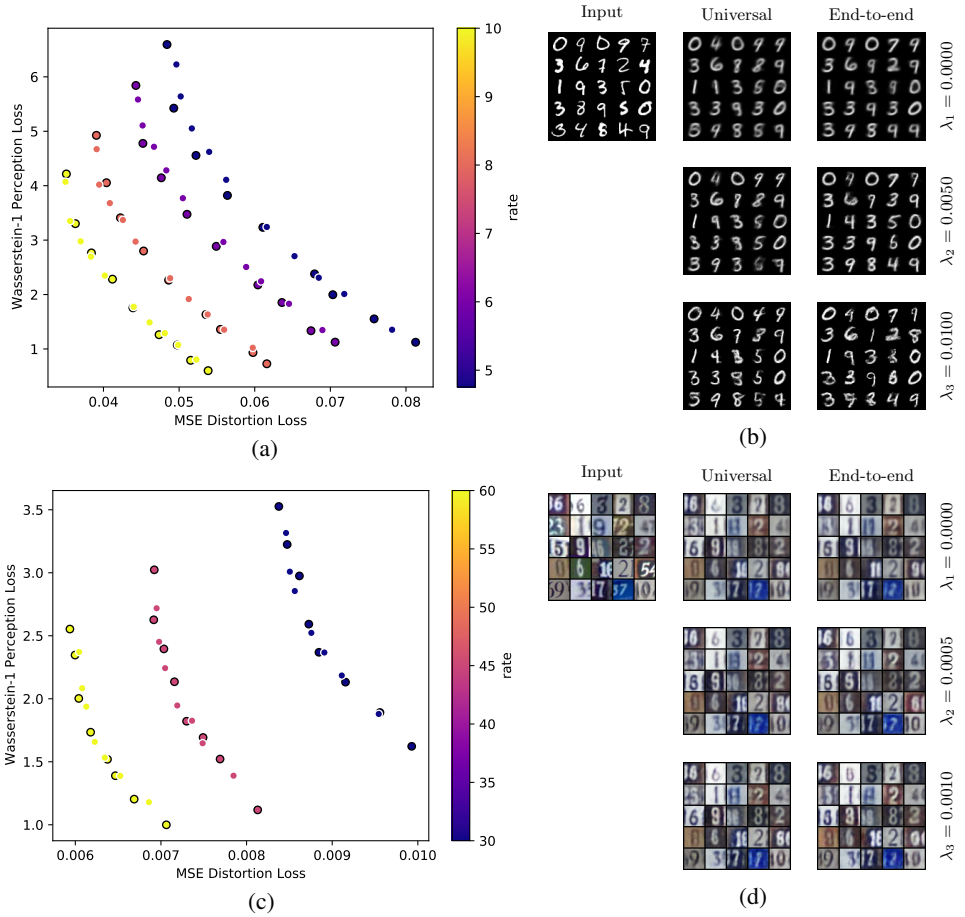

Figure 5: (a) (c) Rate-distortion-perception tradeoffs for NQ. (b) (d) The visual quality of both the end-to-end and universal models are on average comparable for each λ_i (MNIST: $R = 6$, SVHN:$R = 60$.)

Let $\mathcal{C}$ be the set of quantization centers, each containing L levels distributed uniformly between $[-1, +1]$ along each dimension d. Let x be the input and $f(x)$ the output of the encoder before quantization. We compare the performance of deterministic quantization (DQ), universal quantization (UQ), and noisy quantization (NQ). All quantizers use a soft gradient estimator (equation (3) of Mentzer et al. [21]) during backpropogation.

Deterministic quantization (DQ). The sender computes

$$z = \arg\min_{c \in \mathcal{C}} \|f(x) - c\|$$

and sends z to the receiver. The receiver decodes the image by passing z through the decoder. This is the most straightforward method of quantization but lacks the stochasticity required to train an effective generative model.

Quantization with noise added (NQ). The sender computes

$$z = \arg\min_{c \in \mathcal{C}} \|f(x) - c\|$$

and sends z to the receiver. The receiver samples $u \sim U[-1/(L-1), +1/(L-1)]^d$ and decodes the image by passing $z + u$ through the decoder. Note that there is no information loss as the noise range is almost surely below the quantization interval. This scheme was used by Blau and Michaeli [5].

Universal quantization (UQ) [30, 42]. We assume the sender and receiver have access to $u \sim U[-1/(L-1), +1/(L-1)]^d$. The sender computes

$$z = \arg\min_{c \in \mathcal{C}} \|f(x) + u - c\|$$

and sends z to the receiver. The receiver decodes the image by passing $z - u$ through the decoder. This quantization scheme produces stochastic input for the decoder while reducing the quantization error incurred by NQ. This is also known as a subtractive dither [12, 29] in literature.

We demonstrate in Figure 5 that the NQ scheme is still able to produce universal representations within the operational tradeoff it achieves. The results of the comparison when optimizing only for MSE loss are given in Table 1. Both DQ and UQ perform better than NQ. Although DQ performs slightly better, UQ is still highly effective.

Table 1: Comparison of MSE distortion losses using deterministic quantization (DQ), universal quantization (UQ), and noisy quantization (NQ) when optimizing an end-to-end model only for distortion loss ($\lambda = 0$) on MNIST.

R	MSE (DQ)	MSE (UQ)	MSE (NQ)
4.75	0.0442	0.0459	0.0484
6	0.0412	0.0426	0.0443
8	0.0358	0.0362	0.0391
10	0.0315	0.0324	0.0351

B.2 Error Intervals

We provide error intervals across 5 trials for a subset of the universality experiments given in Figure 4 on MNIST here. Each trial consists of training a new end-to-end model ($\lambda = 0.015$, $R = 4.75$), then using the resultant encoder to train universal models across all tradeoff points. The results are very consistent across each trial.

Table 2: MSE distorion losses across 5 trials.

λ	0.0000	0.0025	0.0040	0.0050	0.0060	0.0080	0.0090	0.0100	0.0110	0.0130
max	0.0470	0.0477	0.0493	0.0507	0.0531	0.0583	0.0614	0.0648	0.0681	0.0729
min	0.0466	0.0474	0.0490	0.0504	0.0526	0.0577	0.0606	0.0640	0.0665	0.0715
average	0.0468	0.0475	0.0491	0.0506	0.0528	0.0579	0.0609	0.0643	0.0670	0.0720

Table 3: Wasserstein-1 perception losses across 5 trials.

λ	0.0000	0.0025	0.0040	0.0050	0.0060	0.0080	0.0090	0.0100	0.0110	0.0130
max	5.8014	5.1464	4.6482	4.2692	3.7672	2.9855	2.5596	2.1221	1.9033	1.3256
min	5.5787	5.0612	4.5815	4.1812	3.7576	2.8928	2.5105	2.0020	1.6842	1.1958
average	5.7123	5.1087	4.6177	4.2211	3.7636	2.9446	2.5259	2.0778	1.8296	1.2718

B.3 Refinement Experiments

So far, we have enforced the decoders in universal models to use only the representations produced by the universal encoder, producing a tradeoff curve along perception and distortion at fixed rate. We now consider the scenario where the rate is varied by designing *refinement* models which generalize the universal models in the previous section by taking in extra bits through a (trainable) refining encoder in addition to the bits produced by the initial encoder.

Like the universal models, training the refinement models is broken into two analogous stages. The objective and procedure of the first stage is identical to that of the universal models and produces a universal encoder f to be used across multiple models with frozen weights, and a low-rate decoder g. In the second stage, the refinement model introduces a new high-rate decoder g_1^+ building upon representations from both the universal encoder f and a secondary refining encoder f_1^+. The refining encoder and decoder are both trained along with a critic h_1^+, while the universal encoder is held fixed. We use the alternating training procedure as with the universal models. Bits are sent in two stages so that either low rate or high rate reconstructions

$$\hat{X}_1^{(1)} = g(f(X)), \tag{38}$$

$$\hat{X}_1^{(2)} = g_1^+(f(X), f_1^+(X)) \tag{39}$$

are possible. One may take the view that f_1^+ will embed auxiliary details about the input to supplement the information extracted by f. Since f is held fixed while g_1^+ is being trained, we expect that there should be a performance gap between the refinement model and an end-to-end model with full flexibility in training an encoder. In Figure 6, we find that the gap is not sizeable in practice, with the visual quality of the refinement models similar to the end-to-end models of the same rate.

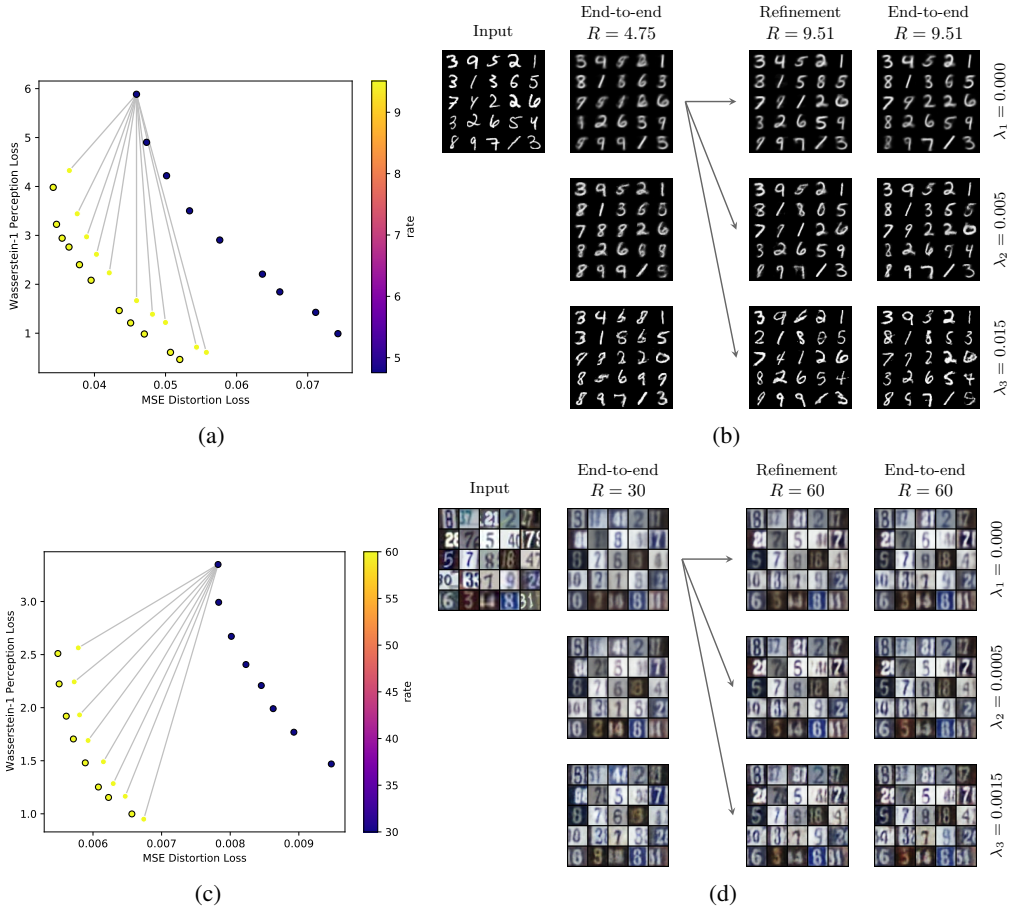

Figure 6: (a) (c) Rate refinability of MNIST and SVHN. Points with black outline are losses for the end-to-end models. Points without outline are the losses for the refinement models, which were trained with a encoder optimized for only distortion loss ($\lambda = 0$). For fair comparison, the parameter count of an end-to-end encoder at high rate is approximately equal to the sum of the parameter counts for the universal encoder and refining encoder in the refinement model. Refinement from $\lambda = 0$ performs closest to end-to-end models of the same rate, but any $\lambda > 0$ can be refined. (b) (d) Outputs of selected models. Visual reconstruction of refinement models is similar to that of high-rate end-to-end models across all tradeoffs.

B.4 Architecture

The architectures used for the experiments are given as follows. Here each row represents a group of layers. d denotes the latent dimension and L the number of quantization levels per dimension, with $R = d \log L$. The widths of the layers may be varied for some experiments (e.g. to facilitate fair comparison in parameter count between the refinement models and end-to-end models). The quantizer performs hard nearest-neighbour quantization on the forward pass and uses a soft relaxation given by Equation (3) in [21] during the backward pass. The bin centers for quantization are spaced evenly in $[-1, 1]$ for each dimension. The type of compression systems are denoted by E for end-to-end, U for (perception-distortion) universal and R for refinement.

B.4.1 MNIST

The universality experiments build off of the encoders produced by the end-to-end experiments of the same rate with $\lambda = 0.015$. The refinement experiment in row 2 of the right table builds off the universal encoder produced by the end-to-end model of row 1 with $\lambda = 0, 0.015$. For fair comparison, the parameter count of an end-to-end encoder at $R = 9.51$ is approximately equal to the sum of the parameter counts for the universal encoder and refining encoder in the refinement model at $R = 9.51$.

Table 4: Network and quantizer settings for MNIST. Left table: models shown in Figure 4(a). Right table: models shown in Figure 6(a).

System	R	d	L
E+U	4.75	3	3
E+U	6	3	4
E+U	8	4	4
E+U	10	5	4

System	R	d	L
E	4.75	3	3
R	9.51	$3+3$	3
E	9.51	6	3

Table 5: The tradeoff coefficients used across all rates in each experiment for MNIST.

System	Tradeoff coefficients
E (Figure 4(a))	$\lambda = 0$, 0.0033, 0.005, 0.0066, 0.008, 0.01, 0.011, 0.013, 0.015
U (Figure 4(a))	$\lambda_i = 0$, 0.0025, 0.004, 0.005, 0.006, 0.008, 0.009, 0.01, 0.011, 0.013
E (Figure 6(a))	$\lambda = 0$, 0.0033, 0.005, 0.0066, 0.008, 0.01, 0.011, 0.013, 0.015
R (Figure 6(a))	$\lambda_i = 0$, 0.0025, 0.004, 0.005, 0.006, 0.008, 0.009, 0.01, 0.013, 0.015

Table 6: Model architectures for MNIST. l-ReLU denotes Leaky ReLU. Refer to code for parameter settings.

Encoder

Input
Flatten
Linear, BatchNorm2D, l-ReLU
Linear, BatchNorm2D, l-ReLU
Linear, BatchNorm2D, l-ReLU
Linear, BatchNorm2D, l-ReLU
Linear, BatchNorm2D, Tanh
Quantizer

Decoder

Input
Linear, BatchNorm1D, l-ReLU
Linear, BatchNorm1D, l-ReLU
Unflatten
ConvT2D, BatchNorm2D, l-ReLU
ConvT2D, BatchNorm2D, l-ReLU
ConvT2D, BatchNorm2D, Sigmoid

Critic

Input
Conv2D, l-ReLU
Conv2D, l-ReLU
Conv2D, l-ReLU
Linear

Table 7: Hyperparameters used for training MNIST models across all rates, including for universal/refining encoders. α is the learning rate, (β_1, β_2) are the parameters for Adam, and λ_{GP} is the gradient penalty coefficient.

	α	β_1	β_2	λ_{GP}
Encoder	10^{-2}	0.5	0.9	-
Decoder	10^{-2}	0.5	0.9	-
Critic	2×10^{-4}	0.5	0.9	10

B.4.2 SVHN

The experiments are similar to MNIST, with the main difference being in the encoder architecture. The universality experiments build off of the encoders produced by the end-to-end experiments of the same rate with $\lambda = 0.002$. The refinement experiment in row 2 of the right table builds off the universal encoder produced by the end-to-end model of row 1 with $\lambda = 0, 0.002$. For fair comparison, the parameter count of an end-to-end encoder at $R = 60$ is approximately equal to the sum of the parameter counts for the universal encoder and refining encoder in the refinement model at $R = 30$.

Table 8: Network and quantizer settings for SVHN. Left table: models shown in Figure 4(c). Right table: models shown in Figure 6(c).

System	R	d	L
E+U	30	10	8
E+U	45	15	8
E+U	60	20	8

System	R	d	L
E	30	10	8
R	60	$10 + 10$	8
E	60	20	8

Table 9: The tradeoff coefficients used across all rates in each experiment for SVHN.

System	Tradeoff coefficients
E (Figure 4(c))	$\lambda = 0$, 0.00025, 0.0005, 0.00075, 0.001, 0.00125, 0.0015, 0.002
U (Figure 4(c))	$\lambda_i = 0$, 0.0003, 0.0005, 0.0008, 0.001, 0.0012, 0.0017
E (Figure 6(c))	$\lambda = 0$, 0.00025, 0.0005, 0.00075, 0.001, 0.00125, 0.0015, 0.002
R (Figure 6(c))	$\lambda_i = 0$, 0.00025, 0.0005, 0.00075, 0.001, 0.00125, 0.0015, 0.002

Table 10: Model architectures for SVHN. Refer to code for parameter settings.

Encoder
Input
Conv2D, l-ReLU
Conv2D, l-ReLU
Conv2D, l-ReLU
Flatten
Linear, Tanh
Quantizer

Decoder
Input
Linear, BatchNorm1D, l-ReLU
Linear, BatchNorm1D, l-ReLU
Unflatten
ConvT2D, BatchNorm2D, l-ReLU
ConvT2D, BatchNorm2D, l-ReLU
ConvT2D, BatchNorm2D, l-ReLU
ConvT2D, BatchNorm2D, Sigmoid

Critic
Input
Conv2D, l-ReLU
Conv2D, l-ReLU
Conv2D, l-ReLU
Linear

Table 11: Hyperparameters used for training. α is the learning rate, (β_1, β_2) are the parameters for Adam, and λ_{GP} is the gradient penalty coefficient.

	α	β_1	β_2	λ_{GP}
Encoder	10^{-4}	0.5	0.999	-
Decoder	10^{-4}	0.5	0.999	-
Critic	10^{-4}	0.5	0.999	10