Rethinking Lossy Compression: The Rate-Distortion-Perception Tradeoff Supplementary Material

Yochai Blau [1] **Tomer Michaeli** [1]

In this supplemental, we first provide proofs for Theorems 1 and 2. We then prove the relation between the rate of an encoder-decoder pair and the rate-distortion-perception function in (4) for a memoryless stationary source. Next, we derive the rate-distortion-perception function $R(D,P)$ of a Bernoulli random variable (see Sec. 3.1), which appears in (6). In the following section, we specify all training and architecture details for the experiments in Sec. 4. Finally, we include details on the choice of the perceptual loss in the experiment of Sec. 4.2.

A. Proof of Theorem 1

The proof of this theorem follows closely that of its rate-distortion analogue (Cover & Thomas (2012), 2nd ed., p. 316).

Monotonicity The value $R(P,D)$ is the minimal mutual information $I(X,\hat{X})$ over a constraint set whose size increases with D and P. This implies that the function $R(D,P)$ is non-increasing in D and P.

Convexity Here, we assume that **A1** holds. That is, the divergence $d(p,q)$ in (4) is convex in its second argument, so that for any $\lambda \in [0,1]$,

$$d(p, \lambda q_1 + (1-\lambda)q_2) \leq \lambda d(p,q_1) + (1-\lambda) d(p,q_2). \tag{S1}$$

To prove the convexity of $R(D,P)$, we will show that

$$\lambda R(D_1,P_1) + (1-\lambda) R(D_2,P_2) \geq R(\lambda D_1 + (1-\lambda)D_2, \lambda P_1 + (1-\lambda)P_2), \tag{S2}$$

for all $\lambda \in [0,1]$. First, by definition, the left hand side of (S2) can be written as

$$\lambda I(X,\hat{X}_1) + (1-\lambda) I(X,\hat{X}_2), \tag{S3}$$

where $\hat{X}_1$ and $\hat{X}_2$ are defined by

$$p_{\hat{X}_1|X} = \arg\min_{p_{\hat{X}|X}} \; I(X,\hat{X}) \quad \text{s.t.} \quad \mathbb{E}[\Delta(X,\hat{X})] \leq D_1,\; d(p_X, p_{\hat{X}}) \leq P_1, \tag{S4}$$

$$p_{\hat{X}_2|X} = \arg\min_{p_{\hat{X}|X}} \; I(X,\hat{X}) \quad \text{s.t.} \quad \mathbb{E}[\Delta(X,\hat{X})] \leq D_2,\; d(p_X, p_{\hat{X}}) \leq P_2. \tag{S5}$$

[1]Technion–Israel Institute of Technology, Haifa, Israel. Correspondence to: Yochai Blau <yochai@campus.technion.ac.il>, Tomer Michaeli <tomer.m@ee.technion.ac.il>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

Since $I(X, \hat{X})$ is convex in $p_{\hat{X}|X}$ for a fixed p_X (Cover & Thomas (2012), 2nd ed., p. 33),

$$\lambda I(X, \hat{X}_1) + (1-\lambda) I(X, \hat{X}_2) \geq I(X, \hat{X}_\lambda), \tag{S6}$$

where $\hat{X}_\lambda$ is defined by

$$p_{\hat{X}_\lambda|X} = \lambda p_{\hat{X}_1|X} + (1-\lambda)\, p_{\hat{X}_2|X}. \tag{S7}$$

Denoting $D_\lambda = \mathbb{E}[\Delta(X, \hat{X}_\lambda)]$ and $P_\lambda = d(p_X, p_{\hat{X}_\lambda})$, we have that

$$I(X, \hat{X}_\lambda) \geq \min_{p_{\hat{X}|X}} \left\{ I(X, \hat{X}) \,:\, \mathbb{E}[\Delta(X, \hat{X})] \leq D_\lambda,\, d(p_X, p_{\hat{X}_\lambda}) \leq P_\lambda \right\} = R(D_\lambda, P_\lambda), \tag{S8}$$

because $\hat{X}_\lambda$ is in the constraint set. The divergence $d(p, q)$ is assumed to be convex in the second argument, thus

$$\begin{aligned} P_\lambda &= d(p_X, p_{\hat{X}_\lambda}) \\ &\leq \lambda d(p_X, p_{X_1}) + (1-\lambda) d(p_X, p_{X_2}) \\ &\leq \lambda P_1 + (1-\lambda) P_2. \end{aligned} \tag{S9}$$

Similarly,

$$\begin{aligned} D_\lambda &= \mathbb{E}\left[\Delta(X, \hat{X}_\lambda)\right] \\ &\overset{(a)}{=} \mathbb{E}\left[\mathbb{E}\left[\Delta(X, \hat{X}_\lambda)|X\right]\right] \\ &\overset{(b)}{=} \mathbb{E}\left[\lambda \mathbb{E}\left[\Delta(X, \hat{X}_1)|X\right] + (1-\lambda)\mathbb{E}\left[\Delta(X, \hat{X}_2)|X\right]\right] \\ &\overset{(c)}{=} \lambda \mathbb{E}[\Delta(X, \hat{X}_1)] + (1-\lambda)\, \mathbb{E}[\Delta(X, \hat{X}_2)] \\ &\leq \lambda D_1 + (1-\lambda) D_2, \end{aligned} \tag{S10}$$

where (a) and (c) are according to the law of total expectation, and (b) is by (S7). Therefore, since $R(D, P)$ is non-increasing in D and P, we have from (S9) and (S10) that

$$R(D_\lambda, P_\lambda) \geq R(\lambda D_1 + (1-\lambda) D_2, \lambda P_1 + (1-\lambda) P_2). \tag{S11}$$

Combining (S3), (S6), (S8) and (S11) proves (S2), thus proving that $R(D, P)$ is convex.

Dependence on the perceptual quality Here, we assume that **A2** holds. In particular, this implies that the function $k(z) = \mathbb{E}_{X \sim p_X}[\Delta(X, z)]$ does not attain its minimum over the entire support of p_X. To prove that $R(\cdot, 0) \neq R(\cdot, \infty)$, let us assume to the contrary that $R(\cdot, 0) = R(\cdot, \infty)$. This implies that for any distortion level, minimizing the rate without a constraint on perception ($P = \infty$), leads to perfect perceptual quality, $p_{\hat{X}} = p_X$, just like with a perfect perception constraint $P = 0$. Let us examine the solution specifically at the distortion level D^* defined by

$$D^* = \min_{p_{\hat{X}|X}} \; \mathbb{E}_{(X, \hat{X}) \sim p_{X, \hat{X}}}[\Delta(X, \hat{X})] \quad \text{s.t.} \quad I(X, \hat{X}) = 0. \tag{S12}$$

Notice that since $I(X,\hat{X})=0$ in this case, X and $\hat{X}$ are independent, so that $p_{\hat{X}|X}=p_{\hat{X}}$. Therefore

$$\begin{aligned} D^* &= \min_{p_{\hat{X}}} \ \mathbb{E}_{(X,\hat{X})\sim p_X p_{\hat{X}}}[\Delta(X,\hat{X})] \\ &= \min_{p_{\hat{X}}} \ \mathbb{E}_{\hat{X}\sim p_{\hat{X}}}[\mathbb{E}_{X\sim p_X}[\Delta(X,\hat{X})]] \\ &= \min_{p_{\hat{X}}} \ \mathbb{E}_{\hat{X}\sim p_{\hat{X}}}[k(\hat{X})]. \end{aligned} \tag{S13}$$

Clearly, the $p_{\hat{X}}$ which minimizes (S13) cannot assign positive probability outside the set where $k(z)$ attains its minimal value. Namely, the support of $p_{\hat{X}}$ must be contained in the set S defined by

$$S = \{z \in \arg\min_{\tilde{z}} \ k(\tilde{z})\}. \tag{S14}$$

But since our encoder-decoder pair achieves perfect perceptual quality, i.e. $p_{\hat{X}}=p_X$, this implies that support$\{p_X\}\subset S$, contradicting Assumption **A2**.

B. Proof of Theorem 2

Assume the MSE distortion, and consider any (R,D) pair on Shannon's classic rate-distortion function (corresponding to $R(D,\infty)$ on the rate-distortion-perception function), and the encoder-decoder mapping $p_{\hat{X}|X}$ which achieves this (R,D) pair. We will prove the theorem by explicitly constructing a modified encoder-decoder, which achieves perfect perceptual quality and has only twice the distortion. To do this, we concatenate a post-processing mapping $p_{\tilde{X}|\hat{X}}$ to produce a new decoded output $\tilde{X}$ by drawing from the posterior distribution $p_{X|\hat{X}}$. That is,

$$p_{\tilde{X}|\hat{X}}(\tilde{x}|\hat{x}) = p_{X|\hat{X}}(\tilde{x}|\hat{x}) = \frac{p_{\hat{X}|X}(\hat{x}|\tilde{x})p_X(\tilde{x})}{p_{\hat{X}}(\hat{x})}. \tag{S15}$$

The distribution $p_{\tilde{X}}$ of this new output $\tilde{X}$ is identical to the distribution of the source signal p_X, as

$$p_{\tilde{X}}(z) = \int p_{\tilde{X}|\hat{X}}(z|\hat{x})p_{\hat{X}}(\hat{x})d\hat{x} = \int p_{X|\hat{X}}(z|\hat{x})p_{\hat{X}}(\hat{x})d\hat{x} = p_X(z). \tag{S16}$$

Therefore, $d(p_X,p_{\tilde{X}})=0$, showing that it achieves perfect perceptual quality. The MSE distortion of the modified encoder-decoder is given by

$$\begin{aligned} \tilde{D} &= \mathbb{E}[\|X-\tilde{X}\|^2] \\ &= \mathbb{E}[\|X\|^2] - 2\mathbb{E}[X^T\tilde{X}] + \mathbb{E}[\|\tilde{X}\|^2] \\ &\overset{(a)}{=} \mathbb{E}[\|X\|^2] - 2\mathbb{E}[\mathbb{E}[X^T\tilde{X}|\hat{X}]] + \mathbb{E}[\mathbb{E}[\|\tilde{X}\|^2|\hat{X}]] \\ &\overset{(b)}{=} \mathbb{E}[\|X\|^2] - 2\mathbb{E}[\|\mathbb{E}[X|\hat{X}]\|^2] + \mathbb{E}[\mathbb{E}[\|X\|^2|\hat{X}]] \\ &= \mathbb{E}[\|X\|^2] - 2\mathbb{E}[\|\mathbb{E}[X|\hat{X}]\|^2]) + \mathbb{E}[\|X\|^2|] \\ &= 2(\mathbb{E}[\|X\|^2] - \mathbb{E}[\|\mathbb{E}[X|\hat{X}]\|^2]) \\ &= 2\,\mathbb{E}[\|X-\mathbb{E}[X|\hat{X}]\|^2] \\ &\overset{(c)}{=} 2\,\mathbb{E}[\|X-\hat{X}\|^2] = 2D, \end{aligned} \tag{S17}$$

where in (a) we used the law of total expectation, in (b) we used the fact that X and $\tilde{X}$ are independent given $\hat{X}$ and both have distribution p_X, and in (c) we used that fact that $\hat{X} = \mathbb{E}[X|\hat{X}]$ as otherwise it would not have lied on the rate-distortion curve (replacing $\hat{X}$ by $\mathbb{E}[X|\hat{X}]$ would lead to a lower MSE without increasing the rate).

The mutual information between the source X and the modified decoded signal $\tilde{X}$ satisfies

$$I(X;\tilde{X}) \leq I(X;\hat{X}), \tag{S18}$$

due to the data processing inequality for the Markov chain $X \to \hat{X} \to \tilde{X}$.

Putting it together, we get

$$\begin{aligned} R(D,\infty) &= \min_{p_{\hat{X}|X}} \{I(X,\hat{X}) \,:\, \mathbb{E}[\Delta(X,\hat{X})] \leq D\} \\ &\overset{(f)}{\geq} I(X,\tilde{X}) \\ &\overset{(g)}{\geq} \min_{p_{\hat{X}|X}} \{I(X,\hat{X}) \,:\, \mathbb{E}[\Delta(X,\hat{X})] \leq \tilde{D}, d(p_X, p_{\hat{X}}) \leq 0\} \\ &= R(\tilde{D},0) \\ &\overset{(h)}{\geq} R(2D,0), \end{aligned} \tag{S19}$$

where (f) is due to (S18), (g) is since $p_{\tilde{X}|X}$ is in the constraint set, and (h) is justified by (S17) and the fact that $R(D,P)$ is non-increasing in D (see Theorem 1). This proves that $R(D,0) \leq R(\frac{1}{2}D,\infty)$.

C. Perception aware lossy compression of a memoryless stationary source

We now prove that when compressing a memoryless stationary source with average distortion D and average perception index P, the rate is lower bounded by $R(D,P)$. This proof follows closely that of its rate-distortion analogue (Cover & Thomas (2012), 2nd ed., p. 316).

Assume a memoryless stationary source. Given a source sequence X^n comprising i.i.d. variables $X_1,\ldots,X_n$ with distribution p_X, the encoder f_n constructs an encoded representation with rate R as $f_n : \mathcal{X}^n \to \{1,2,\ldots,2^{nR}\}$. The decoder g_n outputs an estimate $\hat{X}^n$ of X^n as $g_n : \{1,2,\ldots,2^{nR}\} \to \hat{\mathcal{X}}^n$. We are interested in the the average distortion of the reconstructions, $\frac{1}{n}\sum_{i=1}^n \Delta(X_i,\hat{X}_i)$, and in their average perceptual quality, $\frac{1}{n}\sum_{i=1}^n d(p_{X_i}, p_{\hat{X}_i})$. Assume that

$$\frac{1}{n}\sum_{i=1}^n \Delta(X_i,\hat{X}_i) \leq D, \quad \frac{1}{n}\sum_{i=1}^n d(p_{X_i}, p_{\hat{X}_i}) \leq P. \tag{S20}$$

Then

$$\begin{aligned} nR &\overset{(a)}{\geq} H(f_n(X^n)) \\ &\overset{(b)}{\geq} H(f_n(X^n)) - H(f_n(X^n)|X^n) \\ &= I(X^n; f_n(X^n)) \\ &\overset{(c)}{\geq} I(X^n,\hat{X}^n) \\ &= H(X^n) - H(X^n|\hat{X}^n) \end{aligned}$$

$$
\begin{aligned}
&\overset{(d)}{=} \sum_{i=1}^{n} H(X_i) - H(X^n|\hat{X}^n) \\
&\overset{(e)}{=} \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}^n, X_{i-1}, \dots, X_1) \\
&\overset{(f)}{\geq} \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i|\hat{X}_i) \\
&= \sum_{i=1}^{n} I(X_i, \hat{X}_i) \\
&\overset{(g)}{\geq} \sum_{i=1}^{n} R\left(\mathbb{E}[\Delta(X_i, \hat{X}_i)], d(p_{X_i}, p_{\hat{X}_i})\right) \\
&= n\left(\frac{1}{n}\sum_{i=1}^{n} R\left(\mathbb{E}[\Delta(X_i, \hat{X}_i)], d(p_{X_i}, p_{\hat{X}_i})\right)\right) \\
&\overset{(h)}{\geq} n\, R\left(\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\Delta(X_i, \hat{X}_i)], \frac{1}{n}\sum_{i=1}^{n} d(p_{X_i}, p_{\hat{X}_i})\right) \\
&\overset{(i)}{\geq} n\, R(D, P),
\end{aligned}
\tag{S21}
$$

where (a) is since the size of the range of f_n is 2^{nR}, (b) is since $H(f_n(X^n)|X^n) > 0$, (c) is from the data-processing inequality, (d) is since X_i are independent, (e) is from the chain rule of entropy, (f) is since conditioning reduces entropy, (g) is from the definition of $R(D, P)$ in (4), (h) is from the convexity of $R(D, P)$ (see Theorem 1) and Jensen's inequality, and (i) is from (S20) and the fact that $R(D, P)$ is non-increasing in D, P (see Theorem 1). This proves that the rate of *any* encoder-decoder pair having average distortion $\frac{1}{n}\sum_{i=1}^{n} \Delta(X_i, \hat{X}_i) = D$ and average perceptual quality $\frac{1}{n}\sum_{i=1}^{n} d(p_{X_i}, p_{\hat{X}_i}) = P$, is lower-bounded by $R(D, P)$, the rate-distortion-perception function evaluated at D, P.

To prove that the rate-distortion-perception function describes the optimal rate at distortion level D and perceptual quality P, we would also have to prove that $R(D, P)$ is achievable, which we leave for future work. Yet, the proof that $R(D, P)$ lower-bounds the rate is sufficient for concluding that a tradeoff between rate, distortion and perception necessarily exists. Specifically, in Theorem 1 we prove that (subject to assumptions) the rate-distortion curve elevates when constraining for perceptual quality, i.e. $R(\cdot, 0) > R(\cdot, \infty)$. Now, Shannon's rate-distortion curve $R(\cdot, \infty)$ is known to be achievable (Cover & Thomas (2012), 2nd ed., p. 318) and thus describes the optimal rate R_S when *not* constraining the perceptual quality. As shown above, $R(\cdot, 0)$ lower-bounds the rate R_P when constraining for perfect perceptual quality. Combining these, we get that $R_P > R_S$, indicating that constraining for perceptual quality necessarily leads to an increase in rate (for constant distortion level), thus illustrating the rate-distortion-perception tradeoff.

D. Derivation of the rate-distortion-perception function $R(D, P)$ of a Bernoulli source

Assume that $X \sim \text{Bern}(p)$ with $p \leq \frac{1}{2}$. We seek a conditional distribution $p_{\hat{X}|X}$, which we parameterize by a, b as

$$P(\hat{X} = 0|X = 0) = a, \tag{S22}$$

$$P(\hat{X} = 0|X = 1) = b, \tag{S23}$$

that solves the rate-distortion-perception problem

$$R(D,P) = \min_{a,b} I(X,\hat{X}) \quad \text{s.t.} \quad \mathbb{E}[\Delta(X,\hat{X})] \leq D,\ d(p_X, p_{\hat{X}}) \leq P. \tag{S24}$$

Here we concentrate on the case where $\Delta(\cdot,\cdot)$ is the Hamming distance, and $d(\cdot,\cdot)$ is the total-variation (TV) divergence. The mutual information term $I(X,\hat{X})$ is given by

$$\begin{aligned} I(X,\hat{X}) &= \sum_{x,\hat{x}\in\{0,1\}} P(X=x,\hat{X}=\hat{x}) \log\left(\frac{P(X=x,\hat{X}=\hat{x})}{P(X=x)P(\hat{X}=\hat{x})}\right) \\ &= -a(1-p)\log\left((1-p)+\frac{b}{a}p\right) - (1-a)(1-p)\log\left((1-p)+\frac{1-b}{1-a}p\right) \\ &\quad - bp\log\left(\frac{a}{b}(1-p)+p\right) - (1-b)p\log\left(\frac{1-a}{1-b}(1-p)+p\right), \end{aligned} \tag{S25}$$

the Hamming distance term is given by

$$d_{\mathrm{H}}(X,\hat{X}) = P(X=0,\hat{X}=1) + P(X=1,\hat{X}=0) = (1-a)(1-p) + bp, \tag{S26}$$

and the TV divergence term is given by

$$\begin{aligned} d_{\mathrm{TV}}(p_X, p_{\hat{X}}) &= \tfrac{1}{2}\sum_{z\in 0,1} |p_X(z) - p_{\hat{X}}(z)| \\ &= \tfrac{1}{2}(|P(X=0) - P(\hat{X}=0)| + |P(X=1) - P(\hat{X}=1)|) \\ &= |(1-a)(1-p) - bp|. \end{aligned} \tag{S27}$$

Solution for $P=\infty$ (Shannon's rate-distortion problem) The function $R(D,\infty)$ for Shannon's classic rate-distortion problem is given by (see (Cover & Thomas, 2012), 2nd ed., p. 308)

$$R(D,\infty) = \begin{cases} H_b(p) - H_b(D) & 0 \leq D \leq p, \\ 0 & D > p, \end{cases} \tag{S28}$$

where H_b denotes the binary entropy $H_b(z) = -z\log(z) - (1-z)\log(1-z)$. This optimal solution is obtained by setting the parameters a, b to

$$a_S(D) = \begin{cases} \frac{(1-D)(1-p-D)}{(1-p)(1-2D)} & D \leq p, \\ 1 & D > p, \end{cases} \qquad b_S(D) = \begin{cases} \frac{D(1-p-D)}{p(1-2D)} & D \leq p, \\ 1 & D > p. \end{cases} \tag{S29}$$

Solution for finite P and $I(X,\hat{X})>0$ We now move on to incorporate the additional perception constraint $d(p_X, p_{\hat{X}}) \leq P$. First, notice that the distortion constraint $\mathbb{E}[\Delta(X,\hat{X})] \leq D$ is always active when $I(X,\hat{X}) > 0$, since $I(X,\hat{X}) = 0$ is

achievable for any P when D is not an active constraint[1]. From (S26), the fact that $d_{\mathrm{H}}(X,\hat{X}) = D$ implies that

$$b = \frac{D - (1-a)(1-p)}{p}. \tag{S30}$$

Substituting (S30) into (S27), we get that

$$d_{\mathrm{TV}}(p_X, p_{\hat{X}}) = |2(1-a)(1-p) - D|. \tag{S31}$$

Therefore, the constraint $d_{\mathrm{TV}}(p_X, p_{\hat{X}}) \leq P$ is satisfied when

$$-P \leq 2(1-a)(1-p) - D \leq P \quad \Rightarrow \quad 1 - \frac{D+P}{2(1-p)} \leq a \leq 1 - \frac{D-P}{2(1-p)}. \tag{S32}$$

Below, we show that the lower constraint of (S32) is never active (see **J1**). The upper constraint is obviously active only when $a_S(D)$ of (S29) does not satisfy the upper bound in (S32), which happens when

$$1 - \frac{D-P}{2(1-p)} < \frac{(1-D)(1-p-D)}{(1-p)(1-2D)} \quad \Rightarrow \quad D > \frac{P}{1+2P-2p} \triangleq D_1. \tag{S33}$$

Therefore, when $D \leq D_1$ the solution is independent of P and is given by (S28). When $D > D_1$, the constraint $d_{\mathrm{TV}}(p_X, p_{\hat{X}}) \leq P$ is active, the upper constraint of (S32) is active, and thus

$$a = 1 - \frac{D-P}{2(1-p)}, \tag{S34}$$

and by substituting into (S30) we also get

$$b = \frac{D+P}{2p}. \tag{S35}$$

Note that in (S33) we assumed $D \leq \frac{1}{2}$, below we will justify that this is always the case in this region (see **J2**).

Now, substituting a, b from (S34), (S35) back into (S25) we get

$$\begin{aligned} I(X,\hat{X}) =& (1-p-\frac{D-P}{2}) \log\left(\frac{1-p-\frac{D-P}{2}}{(1-p)(1-p+P)}\right) + (\frac{D-P}{2}) \log\left(\frac{\frac{D-P}{2}}{(1-p)(p-P)}\right) \\ &+ (\frac{D+P}{2}) \log\left(\frac{\frac{D+P}{2}}{p(1-p+P)}\right) + (p-\frac{D+P}{2}) \log\left(\frac{p-\frac{D+P}{2}}{p(p-P)}\right) \\ =& (q-\alpha)\log\left(\frac{q-\alpha}{q(q+P)}\right) + \alpha \log\left(\frac{\alpha}{q(p-P)}\right) \\ &+ \beta \log\left(\frac{\beta}{p(q+P)}\right) + (p-\beta)\log\left(\frac{p-\beta}{p(p-P)}\right) \end{aligned} \tag{S36}$$

where $q = 1-p$, $\alpha = \frac{D-P}{2}$ and $\beta = \frac{D+P}{2}$. This can be further simplified to obtain

$$I(X,\hat{X}) = 2H_b(p) + H_b(p-P) - H_t(\alpha, p) - H_t(\beta, q), \tag{S37}$$

[1]We can always set $p_{\hat{X}|X}(\hat{X} = \hat{x}|X = x) = p_X(\hat{x})$ (i.e. a random draw from p_X disregarding the given input x), which satisfies $d_{\mathrm{TV}}(X,\hat{X}) = 0$ and leads to $I(X,\hat{X}) = 0$ since X and $\hat{X}$ are independent in this case. Only a constraint on the distortion can prevent this solution from being viable.

where $H_t(p_1, p_2)$ is the entropy of a ternary random variable (taking values in a three element alphabet) with probabilities $p_1, p_2, 1-p_1-p_2$.

Solution for finite P and $I(X,\hat{X})=0$ The function $R(D,P)$ is non-increasing in D (see Theorem 1), and will reach $R(D,P)=0$ for $a=b$ since in this case $\hat{X}$ and X are independent. From (S34) and (S35), this happens when

$$1-\frac{D-P}{2(1-p)}=\frac{D+P}{2p} \quad \Rightarrow \quad D=2p(1-p)+(2p-1)P=2pq+(p-q)P \triangleq D_2, \tag{S38}$$

where for $D=D_2$ we get

$$a=b=(1-p)+P. \tag{S39}$$

From this point onward, the solution is fixed, as mutual information $I(X,\hat{X})$ is non-negative and we cannot further decrease the objective of (S24).

Overall solution Putting all the pieces together, the overall solution for $P<p$ is

$$R(D,P)=\begin{cases} H_b(p)-H_b(D) & D\le D_1 \\ 2H_b(p)+H_b(p-P)-H_t(\frac{D-P}{2},p)-H_t(\frac{D+P}{2},q) & D_1<D\le D_2 \\ 0 & D_2<D \end{cases} \tag{S40}$$

where D_1 and D_2 are defined in (S33) and (S38), respectively. For $P\ge p$, the solution is independent of P and is given by the solution to Shannon's classic rate-distortion curve for a Bernoulli source in (S28) (see justification in **J3** below).

Additional justifications

J1 The solution $a_S(D)$ in (S29) does not satisfy this lower constraint of (S32) when

$$1-\frac{D+P}{2(1-p)}>\frac{(1-D)(1-p-D)}{(1-p)(1-2D)}. \tag{S41}$$

When $P<\frac{1-2p}{2}$ this happens for $D<\frac{P}{2p+2P-1}<0$, which never occurs as $D\in[0,1]$. When $P\ge\frac{1-2p}{2}$ this happens for $D>\frac{P}{2p+2P-1}$. However, since $\frac{P}{2p+2P-1}>D_1=\frac{P}{1+2P-2p}$ for all $p\le\frac{1}{2}$ (which is our assumption), the upper constraint of (S32) will always become active before the lower constraint.

J2 Taking the derivative of $D_2=2p(1-p)+(2p-1)P$ with respect to p we obtain

$$\frac{\partial D_2}{\partial p}=2-4p+2P \tag{S42}$$

which is non-negative since $p\le\frac{1}{2}$. Thus, D_2 is increasing in p for all $P>0$, and its largest value in the range $p\in[0,\frac{1}{2}]$, which is $D_2=\frac{1}{2}$, is obtained at $p=\frac{1}{2}$. Thus, in the region where $D_1<D\le D_2$, it is ensured that $D\le\frac{1}{2}$.

J3 Taking the derivative of $D_1=\frac{P}{1+2P-2p}$ with respect to P we obtain

$$\frac{\partial D_1}{\partial P}=\frac{1-2p}{(1+2P-2p)^2} \tag{S43}$$

Table S1. Encoder, decoder, and discriminator architectures. FC is a fully-connected layer, Conv/ConvT is a convolutional/transposed-convolutional layer with "st" denoting stride, BN is a batch-norm layer, and l-ReLU is a leaky-ReLU activation.

Encoder	
Size	Layer
$28 \times 28 \times 1$	Input
784	Flatten
512	FC, BN, l-ReLU
256	FC, BN, l-ReLU
128	FC, BN, l-ReLU
128	FC, BN, l-ReLU
dim	FC, BN, Tanh
dim	Quantize

Decoder	
Size	Layer
dim	Input
128	FC, BN, l-ReLU
512	FC, BN, l-ReLU
$4 \times 4 \times 32$	Unflatten
$11 \times 11 \times 64$	ConvT (st=2), BN, l-ReLU
$25 \times 25 \times 128$	ConvT (st=2), BN, l-ReLU
$28 \times 28 \times 1$	ConvT (st=1), Sigmoid

Discriminator	
Size	Layer
$28 \times 28 \times 1$	Input
$14 \times 14 \times 64$	Conv (st=2), l-ReLU
$7 \times 7 \times 128$	Conv (st=2), l-ReLU
$4 \times 4 \times 256$	Conv (st=2), l-ReLU
4096	Flatten
1	FC

Table S2. Encoder output dimension dim, quantization levels L, and tradeoff coefficients λ used for training the encoder-decoder pairs in the experiments of of Sec. 4

dim	L	λ
2	2	$0, 2, 2.5, 3, 3.5, 4, 4.3, 4.6, 5, 5.5, 6, 8, 10, 15, 20$
3	2	$0, 2, 2.5, 3, 3.5, 4, 4.3, 4.6, 5, 6, 8, 9, 10$
4	2	$0, 2, 2.5, 3, 3.5, 4, 4.3, 4.6, 4.8, 4.9, 5, 6, 8, 10$
4	3	$0, 2, 2.5, 3, 3.5, 4, 4.3, 4.6, 5, 6, 8, 10$
4	4	$0, 2, 2.5, 3, 3.5, 4, 4.3, 4.6, 5, 6, 8, 10$
4	6	$0, 2, 2.5, 3, 3.5, 4, 5, 10$
4	8	$0, 2, 2.5, 3, 4, 5, 6, 10$
4	11	$0, 2, 2.5, 3, 4, 5, 6, 10$
4	16	$0, 2, 2.5, 3, 4, 5, 6, 10$

which is non-negative for $p \leq \frac{1}{2}$, thus D_1 is non-decreasing in P. It is easy to see from (S38) that D_2 in non-increasing in P (for $p \leq \frac{1}{2}$). Thus, $D_1(P) = D_2(P)$ for a single P, which is $P = p$. For any $P \geq p$, there are no D satisfying $D_1 < D \leq D_2$.

E. Architecture and training parameters for the experiments in Sec. 4

The architecture of the encoder, decoder and discriminator nets used for compressing (and decompressing) the MNIST images in Sec. 4 is detailed in Table S1. The optimization objective is given in (10), where $\Delta(x, \hat{x})$ is the squared-error distortion in Sec. 4.1, and a combination of the squared-error and the "perceptual loss" of Johnson et al. (2016) in Sec. 4.2. The encoder output dimension dim, the number of quantization levels L, and values of the tradeoff coefficient λ in (10) used for training the 98 encoder-decoder pairs appear in Table S2. The distortion term in (10) was also multiplied by a constant factor of 10^{-3} for the MSE term (in Sec. 4.1 and Sec. 4.2) and factor of 5×10^{-5} for the perceptual loss (in Sec. 4.2). For each dim and L, an encodoer-decoder with $\lambda = 0$ (only distortion, no adversarial loss) was trained for 25 epochs. The other encoder-decoder pairs with $\lambda > 0$ continued training from this point for another 25 epochs. The ADAM optimizer was used with $\beta_1 = 0.5, \beta_2 = 0.9$. Batch size was 64. Initial learning rates were $10^{-2}/2 \times 10^{-4}$ for the encoder-decoder/discriminator updates in Sec. 4.1, and $5 \times 10^{-3}/2 \times 10^{-4}$ for the encoder-decoder/discriminator updates in Sec. 4.2. These learning rates decreased by $\frac{1}{5}$ after 20 epochs. The convolutional/transposed-convolutional layers filter size (in the decoder and discriminator) was always 5, except for the last convolutional layer in the decoder where the filter size was 4. No padding was used in the decoder, and a padding of 2 was used in each convolutional layer of the discriminator.

The quantization layer (last encoder layer) follows Mentzer et al. (2018). Here, the bin centers $\mathcal{C} = \{c_1, \ldots, c_L\}$ are fixed and evenly spaced in the interval $[-1, 1]$. Denoting by z_i the output of the encoder unit i before quantization (after the Tanh activation), the encoder output $\hat{z}_i$ in the forward pass is given by nearest-neighbor assignment, i.e. $\hat{z}_i = \arg\min_{c_j} \|z_i - c_j\|$.

Table S3. Architecture of the pre-trained MNIST digit classification net used with the perceptual loss of Johnson et al. (2016) in the experiment in Sec. 4.2. FC is a fully-connected layer, Conv is a convolutional layer with k denoting the kernel size, MP is a max-pooling layer with w denoting the window size over which the maximum is taken.

Size	Layer
$28 \times 28 \times 1$	Input
$12 \times 12 \times 10$	Conv (k=5), MP (w=2), l-ReLU
$4 \times 4 \times 20$	Conv (k=5), Dropout, MP (w=2), l-ReLU
320	Flatten
50	FC, ReLU, Dropout
10	FC, Softmax

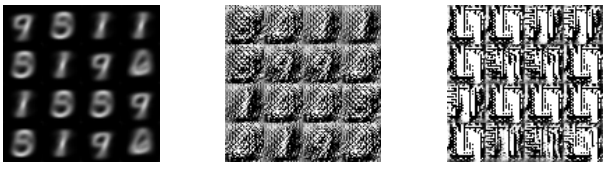
To compute the gradients in the backward pass, we use a differential "soft" assignment

$$\tilde{z}_i = \sum_{j=1}^{L} \frac{\exp(-\sigma \|z_i - c_j\|_1)}{\sum_{l=1}^{L} \exp(-\sigma \|z_i - c_l\|_1)} c_j, \tag{S44}$$

where we use $\sigma = 2/L$. Uniformly distributed noise $\mathcal{U}(-\frac{a}{2}, \frac{a}{2})$ is added to the encoder output before it is passed on to the decoder, with $a = 2/(L-1)$.

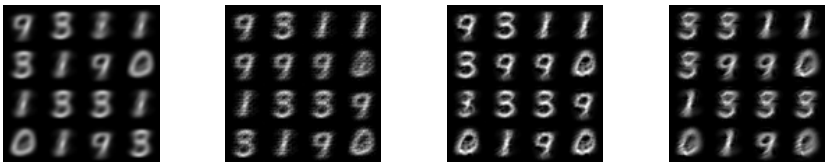
F. The perceptual loss in the experiment in Sec. 4.2

In the experiment of Sec. 4.2, the distortion term of the optimization objective (10) is taken as a combination of the squared-error and the perceptual loss of Johnson et al. (2016). We use this combination since minimizing the perceptual loss alone does not lead to pleasing results (Fig. S1), and is commonly used in combination with an additional distortion term, e.g. ℓ_2/ℓ_1/contextual loss (Ledig et al., 2017; Mechrez et al., 2018a;b; Liu et al., 2018; Wang et al., 2018; Shama et al., 2018; Shoshan et al., 2018). As shown in (11), this perceptual loss is in essence the squared-error in *the deep-feature space* of a pre-trained convolutional net. The standard pre-trained net used with the perceptual loss is the VGG net (Simonyan & Zisserman, 2014), which is trained on natural images from the ImageNet dataset, and is not appropriate for assessing the similarity between MNIST digit images. We therefore pre-train a simple net for classifying MNIST digit images, which achieves over 99% accuracy. The architecture of this pre-trained net is presented in Table S3. The perceptual loss in our experiment is taken as the MSE on the outputs of the second convolutional layer, as this leads to the best perceptual quality (see Fig. S2). We trained with stochastic gradient descent for 30 epochs with a batch size of 30. The learning rate was initialized to 10^{-2} and decreased by $\frac{1}{5}$ after 20 epochs.



(a) (b) (c)

Figure S1. **Minimizing the perceptual loss alone (without an additional MSE term).** The perceptual loss is evaluated on the outputs of: (a) the first convolutional layer, (b) the second convolutional layer, and (c) the first fully-connected layer.



(a) (b) (c) (d)

Figure S2. **Visual comparison of assessing the perceptual loss on the outputs of different layers.** (a) Minimizing the MSE alone. (b)-(d) Minimizing a combination of the MSE and percpetual loss, where the perceptual loss is evaluated on the outputs of: (b) the first convolutional layer, (c) the second convolutional layer, and (d) the first fully-connected layer. The weights of each term (MSE, perceptual) in the loss were optimized for visual quality.

References

Cover, T. M. and Thomas, J. A. *Elements of information theory*. John Wiley & Sons, 2012.

Johnson, J., Alahi, A., and Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A. P., Tejani, A., Totz, J., Wang, Z., and Shi, W. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Liu, G., Reda, F. A., Shih, K. J., Wang, T.-C., Tao, A., and Catanzaro, B. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

Mechrez, R., Talmi, I., Shama, F., and Zelnik-Manor, L. Maintaining natural image statistics with the contextual loss. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2018a.

Mechrez, R., Talmi, I., and Zelnik-Manor, L. The contextual loss for image transformation with non-aligned data. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018b.

Mentzer, F., Agustsson, E., Tschannen, M., Timofte, R., and Van Gool, L. Conditional probability models for deep image compression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Shama, F., Mechrez, R., Shoshan, A., and Zelnik-Manor, L. Adversarial feedback loop. *arXiv preprint arXiv:1811.08126*, 2018.

Shoshan, A., Mechrez, R., and Zelnik-Manor, L. Dynamic-Net: Tuning the objective without re-training. *arXiv preprint arXiv:1811.08760*, 2018.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Loy, C. C., Qiao, Y., and Tang, X. ESRGAN: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision Workshops (ECCVW)*, 2018.