Supplementary Material for “The Information Sieve”

A Detail of the optimization of $TC(X;Y)$

We need to optimize the following objective.

$$\max_{p(y|x)} \sum_i I(X_i;Y) - I(X;Y)$$

If we take the derivative of this expression (along with the constraint that $p(y|x)$ should be normalized) and set it equal to zero, the following simple fixed point equation emerges.

$$p(y|x) = \frac{p(y)}{Z(x)} \prod_{i=1}^{n} \frac{p(x_i|y)}{p(x_i)}$$

Surprisingly, optimizing this objective over possible functions has a fixed point solution with a simple form. This leads to an iterative solution procedure that actually corresponds to a special case of the one considered in (Ver Steeg & Galstyan, 2015). There it is shown that each iterative update of the fixed-point equation increases the objective and that we are therefore guaranteed to converge to a local optimum of the objective. In short, we consider the empirical distribution over observed samples. For each sample, we start with a random probabilistic label. Then we use these labels to estimate the marginals, $p(x_i|y)$, then we use the fixed point to re-estimate $p(y|x)$, and so on until convergence.

Also, note that we can estimate the value of the objective in a simple way. The normalization term, $Z(x)$ is computed for each sample by just summing over the two values of $Y = y$, since Y is binary. The expected logarithm of Z, or the free energy is an estimate of the objective (Ver Steeg & Galstyan, 2015).

Algorithmic details The code implementing this optimization is included as a module in the sieve code (Ver Steeg). The algorithm is described in Alg. 1. Note we use δ as the discrete delta function. The complexity is $O(k \times N \times n)$, where n is the number of variables, N is the number of samples, and k is the cardinality of the latent factor, Y. Because the solution only depends on estimation of marginals between X_i and Y, the number of samples needed for accurate estimation is small (Ver Steeg & Galstyan, 2015).

Labeling test data The fixed point equation above essentially gives us a simple representation of the labeling function in terms of some parameters which, in this case, just correspond to the marginal probability distributions. We simply input values of x from a test set into that equation, and then round y to the most likely value to generate labels.

Algorithm 1 Optimizing $TC(X;Y)$

```
Input: Data matrix, $x_i^l$,
$i = 1, \ldots, n$ variables, $l = 1, \ldots, N$ samples.
Specify $k$: Cardinality of $Y = 1, \ldots, k$
repeat
  Randomly initialize $p(Y = y|X = x^l)$
  $p(Y = y) = 1/n \sum_l p(y|x^l)$
  for $i = 1$ to $n$ do
    $p(X_i = x_i|Y = y) = 1/N \sum_l p(Y = y|X = x^l)\delta_{x_i, x_i^l}/p(Y = y)$
  end for
  $p(Y = y|X = x^l) = \frac{p(Y=y)}{Z(x)} \prod_{i=1}^n \frac{p(X_i=x_i^l|Y=y)}{p(X_i=x_i^l)}$
until Convergence
```

Missing data Note that missing data is handled quite gracefully in this scenario. Imagine that some subset of the X_i's are observed. Denote the subset of indices for which we have observed data on a given sample with G and the subset of random variables as x_G. If we solved the optimization problem for this subset only, we would get a form for the solution like this:

$$p(y|x_G) = \frac{p(y)}{Z(x)} \prod_{i \in G} \frac{p(x_i|y)}{p(x_i)}.$$

In other words, we simply omit the contribution from unobserved variables in the product.

B Proof of Theorem 2.1

We begin by adopting a general definition for “representations” and recalling a useful theorem concerning them.

Definition The random variables $Y \equiv Y_1, \ldots, Y_m$ constitute a *representation* of X if the joint distribution factorizes, $p(x,y) = \prod_{j=1}^m p(y_j|x)p(x), \forall x \in \mathcal{X}, \forall j \in \{1, \ldots, m\}, \forall y_j \in \mathcal{Y}_j$. A representation is completely defined by the domains of the variables and the conditional probability tables, $p(y_j|x)$.

Theorem B.1. Basic Decomposition of Information *(Ver Steeg & Galstyan, 2015)*

If Y is a representation of X and we define,

$$TC_L(X;Y) \equiv \sum_{i=1}^n I(Y:X_i) - \sum_{j=1}^m I(Y_j:X), \tag{7}$$

then the following bound and decomposition holds.

$$TC(X) \geq TC(X;Y) = TC(Y) + TC_L(X;Y) \tag{8}$$

Theorem. Incremental Decomposition of Information

Let Y be some (deterministic) function of $X_1, \ldots, X_n$ and for each $i = 1, \ldots, n$, $\bar{X}_i$ is a probabilistic function of

X_i, Y. *Then the following upper and lower bounds on $TC(X)$ hold.*

$$\begin{gathered} -\sum_{i=1}^{n} I(\bar{X}_i;Y) \le \\ TC(X) - \big(TC(\bar{X}) + TC(X;Y)\big) \le \\ \sum_{i=1}^{n} H(X_i|\bar{X}_i, Y) \end{gathered} \tag{9}$$

Proof. We refer to Fig. 1(a) for the structure of the graphical model. We set $\bar{X} \equiv \bar{X}_1, \ldots, \bar{X}_n, Y$ and we will write $\bar{X}_{1:n}$ to pick out all terms except Y. Note that because Y is a deterministic function of X, we can view $\bar{X}_i$ as a probabilistic function of X_i, Y or of X (as required by Thm. B.1). Applying Thm. B.1, we have

$$TC(X;\bar{X}) = TC(\bar{X}) + TC_L(X;\bar{X}).$$

On the LHS, note that $TC(X;\bar{X}) = TC(X) - TC(X|\bar{X})$, so we can re-arrange to get

$$\begin{aligned} &TC(X) - (TC(\bar{X}) + TC(X;Y)) \\ &= TC(X|\bar{X}) + TC_L(X;\bar{X}) - TC(X;Y). \end{aligned} \tag{10}$$

The LHS is the quantity we are trying to bound, so we focus on expanding the RHS and bounding it.

First we expand $TC_L(X;\bar{X}) = \sum_{i=1}^{n} I(X_i;\bar{X}) - \sum_{i=1}^{n} I(\bar{X}_i;X) - I(Y;X)$. Using the chain rule for mutual information we expand the first term.

$$\begin{aligned} TC_L(X;\bar{X}) = &\sum_{i=1}^{n} I(X_i;Y) \\ &+ \sum_{i=1}^{n} I(X_i;\bar{X}_{1:n}|Y) \\ &- \sum_{i=1}^{n} I(\bar{X}_i;X) - I(Y;X). \end{aligned}$$

Rearranging, we take out a term equal to $TC(X;Y)$.

$$\begin{aligned} TC_L(X;\bar{X}) =&TC(X;Y) + \\ &\sum_{i=1}^{n} I(X_i;\bar{X}_{1:n}|Y) - \sum_{i=1}^{n} I(\bar{X}_i;X). \end{aligned}$$

We use the chain rule again to write $I(X_i;\bar{X}_{1:n}|Y) = I(X_i;\bar{X}_i|Y) + I(X_i;\bar{X}_{\bar{i}}|Y\bar{X}_i)$, where $\bar{X}_{\bar{i}} \equiv \bar{X}_1, \ldots, \bar{X}_n$ with $\bar{X}_i$ (and Y) excluded.

$$\begin{aligned} TC_L(X;\bar{X}) =&TC(X;Y) + \sum_{i=1}^{n} (I(X_i;\bar{X}_i|Y) \\ &+ I(X_i;\bar{X}_{\bar{i}}|Y\bar{X}_i) - I(\bar{X}_i;X)). \end{aligned}$$

The conditional mutual information, $I(A;B|C) = I(A;BC) - I(A;C)$. We expand the first instance of CMI in the previous expression.

$$\begin{aligned} TC_L(X;\bar{X}) =&TC(X;Y) + \sum_{i=1}^{n} (I(\bar{X}_i;X_i,Y) \\ &- I(\bar{X}_i;Y) + I(X_i;\bar{X}_{\bar{i}}|Y\bar{X}_i) \\ &- I(\bar{X}_i;X)). \end{aligned}$$

Since $Y = f(X)$, the first and fourth terms cancel. Finally, this leaves us with

$$\begin{aligned} TC_L(X;\bar{X}) =&TC(X;Y) - \sum_{i=1}^{n} I(\bar{X}_i;Y) \\ &+ \sum_{i=1}^{n} I(X_i;\bar{X}_{\bar{i}}|Y\bar{X}_i). \end{aligned}$$

Now we can replace all of this back in to Eq. 10, noting that the $TC(X;Y)$ terms cancel.

$$\begin{aligned} &TC(X) - (TC(\bar{X}) + TC(X;Y)) \\ &= TC(X|\bar{X}) - \sum_{i=1}^{n} I(\bar{X}_i;Y) + \sum_{i=1}^{n} I(X_i;\bar{X}_{\bar{i}}|Y\bar{X}_i). \end{aligned} \tag{11}$$

First, note that total correlation, conditional total correlation, mutual information, conditional mutual information, and entropy (for discrete variables) are non-negative. Therefore we trivially have the lower bound, $LHS \ge -\sum_{i=1}^{n} I(\bar{X}_i;Y)$. All that remains is to find the upper bound. We drop the negative mutual information, expand the definition of TC in the first line, then drop the negative of an entropy in the second line.

$$\begin{aligned} LHS &\le \sum_{i=1}^{n} H(X_i|\bar{X}) - H(X|\bar{X}) + \sum_{i=1}^{n} I(X_i;\bar{X}_{\bar{i}}|Y\bar{X}_i) \\ &\le \sum_{i=1}^{n} \big(H(X_i|\bar{X}) + I(X_i;\bar{X}_{\bar{i}}|Y\bar{X}_i)\big) \\ &= \sum_{i=1}^{n} H(X_i|\bar{X}_i, Y) \end{aligned}$$

The equality in the last line can be seen by just expanding all the definitions of conditional entropies and conditional mutual information. These provide the upper and lower bounds for the theorem. □

C An algorithm for perfect reconstruction of remainder information

We will use the notation of Fig. 1(a) to construct remainder information for one variable in one layer of the sieve. The goal is to construct the remainder information, $\bar{X}_i$, as

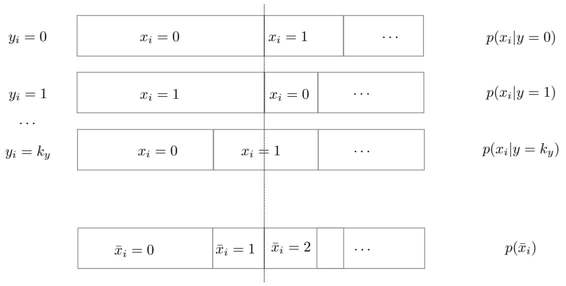

Figure B.1. An illustration of how the remainder information, $\bar{x}_i$, is constructed from statistics about $p(x_i, y)$.

a probabilistic function of X_i, Y so that we satisfy the conditions of Lemma 2.2,

$$(i) \qquad I(\bar{X}_i; Y) = 0 \qquad (ii) \qquad H(X_i|\bar{X}_i, Y) = 0.$$

We need to write down a probabilistic function $p(\bar{x}_i|x_i, y)$ so that, for the observed statistics, $p(x_i, y)$, these conditions are satisfied. There are many ways to accomplish this, and we sketch out one solution here. The actual code we use to generate remainder information for results in this paper are available (Ver Steeg).

We start with the picture in Fig. B.1 that visualizes the conditional probabilities $p(x_i|y)$. Note that the order of the x_i for each value of y can be arbitrary for this scheme to succeed. For concreteness, we sort the values of x_i for each y in order of descending likelihood. Next, we construct the marginal distribution, $p(\bar{x}_i)$. Every time we see a split in one of the histograms of $p(x_i|y)$, we introduce a corresponding split for $p(\bar{x}_i)$. Now, to construct $p(\bar{x}_i|x_i, y)$, for each $\bar{x}_i = q$, for each $y = j$, we find the unique value of $x_i = k(j, q)$ that is directly above the histogram for $p(\bar{x}_i = q)$. Then we set $p(\bar{x}_i = q|x_i, y) = p(\bar{x}_i = q)/p(x_i = k(j, q)|y = j)$. Now, marginalizing over x_i, $p(\bar{x}_i|y) = p(\bar{x}_i)$, ensuring that $I(\bar{X}_i; Y) = 0$. Visually, it can be seen that $H(X_i|\bar{X}_i, Y) = 0$ by picking a value of $\bar{x}_i$ and y and noting that it picks out a unique value of x_i in Fig. B.1.

Note that the function to construct $\bar{x}_i$ is probabilistic. Therefore, when we construct the remainder information at the next layer of the sieve, we have to draw $\bar{x}_i$ stochastically from this distribution. In the example in Sec. 3 the functions for the remainder information happened to be deterministic. In general, though, probabilistic functions inject some noise to ensure that correlations with Y are forgotten at the next level of the sieve. In Sec. 6 we point out that this scheme is detrimental for lossless compression and we point out an alternative.

Controlling the cardinality of $\bar{x}_i$ It is easy to imagine scenarios in Fig. B.1 where the cardinality of $\bar{x}_i$ becomes very large. What we would like is to be able to approximately satisfy conditions (i) and (ii) while keeping the cardinality of the variables, $\bar{X}_i$, small (so that we can accurately estimate probabilities from samples of data). To guide intuition, consider two extreme cases. First, imagine setting $\bar{x}_i = 0$, regardless of x_i, y. This satisfies condition (i) but maximally violates (ii). The other extreme is to set $\bar{x}_i = x_i$. In that case, (ii) is satisfied, but $I(\bar{X}_i; Y) = I(X_i; Y)$. This is only problematic if X_i is related to Y to begin with. If it is, and we set $\bar{X}_i = X_i$, then the same dependence can be extracted at the next layer as well (since we pass X_i to the next layer unchanged).

In practice we would like to find the best solution with a cardinality of fixed size. Note that this can be cast as an optimization problem where $p(\bar{x}_i = |x_i, y)$ represent $\bar{k} \times k_x \times k_y$ variables to optimize over if those are the respective cardinalities of the variables. Then we can minimize a nonlinear objective like $\mathcal{O} = H(X_i|\bar{X}_i, Y) + I(\bar{X}_i; Y)$ over these variables. While off-the-shelf solvers will certainly return local optima for this problem, the optimization is quite slow, especially if we let k's get big.

For the results in this paper, instead of directly solving the optimization problem above to get a representation with cardinality of fixed size, we first construct a perfect solution without limiting the cardinality. Then we modify that solution to let either (i) or (ii) grow somewhat while reducing the cardinality of $\bar{x}_i$ to some target. To keep $I(\bar{X}_i; Y) = 0$ while reducing the cardinality of $\bar{x}_i$, we just pick the $\bar{x}_i$ with the smallest probability and merge it with another value for $\bar{x}_i$. On the other hand, to reduce the cardinality while keeping $H(X_i|\bar{X}_i, Y) = 0$, we again start by finding the $\bar{x}_i = k$ with the lowest probability. Then we take the probability mass for $p(\bar{x}_i = k|x_i, y)$ for each x_i and y and add it to the $p(\bar{x}_i \neq k|x_i, y)$ that already has the highest likelihood for that x_i, y combination. Note that $I(\bar{X}_i; Y)$ will no longer be zero after doing so. For both of these schemes (keeping (i) fixed or keeping (ii) fixed) we reduce cardinality until we achieve some target. For the results in this paper we alway picked $k_{\bar{x}_i} = k_{x_i} + 1$ as the target and we always used the strategy where (ii) was satisfied and we let (i) be violated. In cases where perfect remainder information is impractical due to issues of finite data, we have to define "good remainder information" based on how well it preserves the bounds in Thm. 2.1. The best way to do this may depend on the application, as we saw in Sec. 6.

D More MNIST results

Fig. B.2 shows the same type of results as Fig. 7 but using test data that was never seen in training. Note that no labels were used in any training.

There are several plausible to generate new, never before

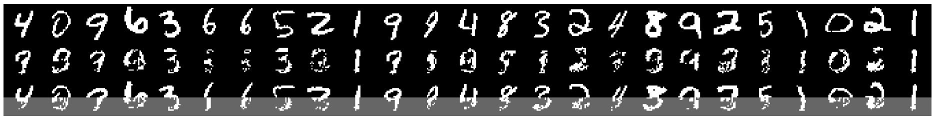

Figure B.2. The same results as Fig. 7 but using samples from a test set instead of the training set.

seen images using the sieve. Here we chose to draw the variables at the last layer of the sieve randomly and independently according to each of their marginal distributions over the training data. Then we inverted the sieve to recover hallucinated images. Some example results are shown in Fig. D.1.

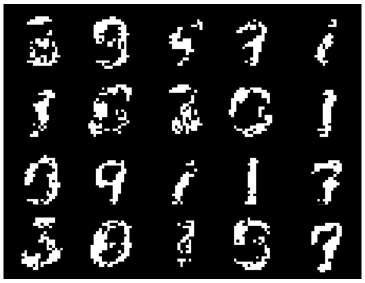

Figure D.1. An attempt to generate new images using the sieve.