Supplementary Information for "Decomposing information into copying versus transformation"

Artemy Kolchinsky and Bernat Corominas-Murtra

A. D_x^{copy} SATISFIES THE FOUR AXIOMS

D_x^{copy} satisfies Axiom 1 by non-negativity of KL.

It satisfies Axiom 2 when $p_{Y|x}(x) > p_Y(x)$ because $\mathsf{d}(p_{Y|x}(x), p_Y(x)) \leq D_{\mathsf{KL}}(p_{Y|x}\|p_Y)$ by the data processing inequality for KL divergence [1, Lemma 3.11]. Otherwise, when $p_{Y|x}(x) \leq p_Y(x)$, D_x^{copy} vanishes and thus satisfies Axiom 2 trivially.

It satisfies Axiom 3 when $p_{Y|x}(x) \leq p_Y(x)$ because in that case $D_x^{\mathsf{copy}}(p_{Y|x}\|p_Y) = 0 \leq D_x^{\mathsf{copy}}(p_{Y|x}(x)\|p_Y)$. If $p_{Y|x}(x) \leq p_Y(x)$, then note that the derivative of $\mathsf{d}(a,b)$ with respect to a is $\frac{d}{da}\mathsf{d}(a,b) = \log\frac{a}{b} - \log\frac{1-a}{1-b}$, which is strictly positive when $a > b$.

Finally, we show that D_x^{copy} satisfies Axiom 4. For any prior distribution p_Y, define the following posterior distribution $p^{\alpha}_{Y|x}(y)$:

$$p^{\alpha}_{Y|x}(y) = \begin{cases} \alpha & \text{if } y = x \\ \frac{1-\alpha}{1-p_Y(x)} p_Y(y) & \text{if } y \neq x \end{cases}, \tag{A1}$$

where α is a parameter that can vary from 0 to 1. It is easy to verify that for all $\alpha \in [p_Y(x), 1]$,

$$D_{\mathsf{KL}}(p^{\alpha}_{Y|x}\|p_Y) = \mathsf{d}(\alpha, p_Y(x)) = D_x^{\mathsf{copy}}(p^{\alpha}_{Y|x}\|p_Y), \tag{A2}$$

and that $D_x^{\mathsf{copy}}(p^{\alpha}_{Y|x}\|p_Y)$ ranges in a continuous manner from 0 (for $\alpha = p_Y(x)$) to $-\log p_Y(x)$ (for $\alpha = 1$).

B. PROOF OF THEOREM 1

Before proceeding, we first prove two useful lemmas.

Lemma B1. *Given Axiom 3, $F(p_{Y|x}, p_Y, x) = F(q_{Y|x}, p_Y, x)$ if $p_{Y|x}(x) = q_{Y|x}(x)$.*

Proof. Follows from applying Axiom 3 in both directions. □

Lemma B2. *Given Axioms 1 to 3, if $p_{Y|x}(x) \leq p_Y(x)$, then $F(p_{Y|x}, p_Y, x) = 0$.*

Proof. If $p_{Y|x}(x) \leq p_Y(x)$, then $F(p_{Y|x}, p_Y, x) \leq F(p_Y, p_Y, x)$ by Axiom 3. By Axiom 2, $F(p_Y, p_Y, x) \leq D_{\mathsf{KL}}(p_Y\|p_Y) = 0$. Combining gives $F(p_{Y|x}, p_Y, x) \leq 0$, while $F(p_{Y|x}, p_Y, x) \geq 0$ by Axiom 1. □

We then show that D_x^{copy} is the largest possible measure that satisfies Axioms 1 to 3.

Proposition B1. *Any F which satisfies Axioms 1 to 3 must obey $F(p_{Y|x}, p_Y, x) \leq D_x^{\mathsf{copy}}(p_{Y|x}\|p_Y)$.*

Proof. Given Lemma B2, without loss of generality we restrict our attention to the case where $p_{Y|x}(x) > p_Y(x)$. Define the posterior $p^{\alpha}_{Y|x}$ as in Eq. (A1), while taking $\alpha = p_{Y|x}(x)$. Then, by Lemma B1,

$$F(p_{Y|x}, p_Y, x) = F(p^{\alpha}_{Y|x}, p_Y, x).$$

At the same time,

$$\begin{aligned} F(p^{\alpha}_{Y|x}, p_Y, x) &\leq D_{\mathsf{KL}}(p^{\alpha}_{Y|x}\|p_Y) \\ &= \mathsf{d}(p_{Y|x}(x)\|p_Y(x)) = D_x^{\mathsf{copy}}(p_{Y|x}\|p_Y), \end{aligned}$$

where the first inequality follows from Axiom 2, and the second equality from Eq. (A2). □

We are now ready to prove the main result from Section II B.

Proof of Theorem 1. Consider some $p_{Y|x}, p_Y, x$, and assume $p_{Y|x}(x) > p_Y(x)$ (without loss of generality by Lemma B2). By Axiom 4, there must exist a posterior $q_{Y|x}$ such that $q_{Y|x}(x) = p_{Y|x}(x)$ and

$$F(q_{Y|x}, p_Y, x) = D_{\mathsf{KL}}(q_{Y|x}\|p_Y). \tag{B3}$$

Note that by the data processing inequality for KL divergence, $D_{\mathsf{KL}}(q_{Y|x}\|p_Y) \geq D_x^{\mathsf{copy}}(q_{Y|x}\|p_Y)$.

Then, by Lemma B1, $F(p_{Y|x}, p_Y, x) = F(q_{Y|x}, p_Y, x)$ since $p_{Y|x}(x) = q_{Y|x}(x)$. Similarly, it can be verified that $D_x^{\mathsf{copy}}(q_{Y|x}\|p_Y) = D_x^{\mathsf{copy}}(p_{Y|x}\|p_Y)$. Combining the above results shows that $F(p_{Y|x}, p_Y, x) \geq D_x^{\mathsf{copy}}(p_{Y|x}\|p_Y)$. the theorem follows by combining with Proposition B1.

□

C. AXIOMATIC DERIVATION AND SOLUTION OF EQ. 15

1. Axiomatic derivation

We first demonstrate that the generalized copy information defined in Eq. (15), $G_x^{\mathsf{copy}}(p_{Y|x}\|p_Y)$, is the unique measure that satisfies Axioms 1 and 2 and our modified Axioms 3* and 4*. Our derivation has the same structure as the one in Section B, and we proceed more quickly.

First, we verify that G_x^{copy} satisfies the four axioms. It satisfies Axiom 1 by non-negativity of KL. It satisfies Axiom 2 because $p_{Y|x}$ falls within the feasibility set of Eq. (15), therefore the minimum $G_x^{\mathsf{copy}}(p_{Y|x}\|p_Y)$ has to be less than or equal to $D_{\mathsf{KL}}(p_{Y|x}\|p_Y)$. It satisfies Axiom 3* because $\mathbb{E}_{p_{Y|x}}[\ell(x,Y)] \geq \mathbb{E}_{q_{Y|x}}[\ell(x,Y)]$ means that the feasibility set of Eq. (15) for $q_{Y|x}$ is a subset of the feasibility set for $p_{Y|x}$, so the minimum $G_x^{\mathsf{copy}}(q_{Y|x}\|p_Y)$ has to be greater than or equal to the minimum $\hat{F}(p_{Y|x}, p_Y, x)$. To show that it satisfies Axiom 4*, note that the distribution w_Y which optimizes Eq. (15) will achieve $\mathbb{E}_{w_Y}[\ell(x,Y)] = \mathbb{E}_{p_{Y|x}}[\ell(x,Y)]$

whenever $\mathbb{E}_{p_{Y|x}}[\ell(x,Y)] \leq \mathbb{E}_{p_Y}[\ell(x,Y)]$ [2, pp.299-300]. Note also that $\mathbb{E}_{p_{Y|x}}[\ell(x,Y)]$ can vary from $\min_y \ell(x,y)$ (for $p_{Y|x}(y|x) = \delta(y, \arg\min_{y'} \ell(x,y'))$ to $\mathbb{E}_{p_Y}[\ell(x,Y)]$ (for $p_{Y|x} = p_Y$).

We now demonstrate that G_x^{copy} is the unique measure that satisfies the four axioms. We begin by showing that $F(p_{Y|x}, p_Y, x) \leq G_x^{\mathsf{copy}}(p_{Y|x}\|p_Y)$ for any F. Given a choice of $p_{Y|x}$, p_Y, and x, let w_Y be the solution to Eq. (15), so

$$G_x^{\mathsf{copy}}(p_{Y|x}\|p_Y) = D_{\mathsf{KL}}(w_Y\|p_Y). \tag{C4}$$

Given the definition of G_x^{copy}, $\mathbb{E}_{w_Y}[\ell(x,Y)] \leq \mathbb{E}_{p_{Y|x}}[\ell(x,Y)]$. Then, by Axiom 3*, Axiom 2, and Eq. (C4),

$$\begin{aligned} F(p_{Y|x}, p_Y, x) &\leq F(w_Y, p_Y, x) \\ &\leq D_{\mathsf{KL}}(w_Y\|p_Y) = G_x^{\mathsf{copy}}(p_{Y|x}\|p_Y). \end{aligned}$$

We finish by showing that $F(p_{Y|x}, p_Y, x) \geq G_x^{\mathsf{copy}}(p_{Y|x}\|p_Y)$ for any F. First consider the case $\mathbb{E}_{p_{Y|x}}[\ell(x,Y)] \geq \mathbb{E}_{p_Y}[\ell(x,Y)]$. Then, $G_x^{\mathsf{copy}}(p_{Y|x}\|p_Y) = 0$ by construction, and therefore $F(p_{Y|x}, p_Y, x) \geq G_x^{\mathsf{copy}}(p_{Y|x}\|p_Y)$ by Axiom 1.

When $\mathbb{E}_{p_{Y|x}}[\ell(x,Y)] < \mathbb{E}_{p_Y}[\ell(x,Y)]$, by Axiom 4* there must exist a posterior $q_{Y|x}$ such that $\mathbb{E}_{q_{Y|x}}[\ell(x,Y)] = \mathbb{E}_{p_{Y|x}}[\ell(x,Y)]$ and

$$F(q_{Y|x}, p_Y, x) = D_{\mathsf{KL}}(q_{Y|x}\|p_Y). \tag{C5}$$

Then, by definition of G_x^{copy},

$$D_{\mathsf{KL}}(q_{Y|x}\|p_Y) \geq G_x^{\mathsf{copy}}(p_{Y|x}\|p_Y). \tag{C6}$$

Finally, by Axiom 3*,

$$F(p_{Y|x}, p_Y, x) \geq F(q_{Y|x}, p_Y, x) \tag{C7}$$

Combining Eqs. (C5) to (C7) shows that $F(p_{Y|x}, p_Y, x) \geq G_x^{\mathsf{copy}}(p_{Y|x}\|p_Y)$.

Thus, G_x^{copy} is the unique measure that satisfies Axioms 1 and 2 and our generalized Axioms 3* and 4*.

2. D_x^{copy} as the solution to Eq. 15 for the 0-1 loss function

Consider the optimization problem:

$$\min_{r_Y \in \Delta:\, r_Y(x) \geq p_{Y|x}(x)} D_{\mathsf{KL}}(r_Y\|p_Y). \tag{C8}$$

When $p_Y(x) \geq p_{Y|x}(x)$, then the solution $r_Y = p_Y$ satisfies the constraint and achieves $D_{\mathsf{KL}}(p_Y\|p_Y) = 0$, the minimum possible. When $p_Y(x) < p_{Y|x}(x)$, we use the chain rule for KL divergence [3] to write

$$\begin{aligned} D_{\mathsf{KL}}(r_Y\|p_Y) = \mathsf{d}(r_Y(x), p_Y(x)) + \\ (1 - r_Y(x)) D_{\mathsf{KL}}(r_Y(Y|Y \neq x)\|p_Y(Y|Y \neq x)). \end{aligned}$$

The second term is minimized by setting $r_Y(y) \propto p_Y(y)$ for $y \neq x$, so that $r_Y(y|Y \neq x) = p_Y(y|Y \neq x)$ and $D_{\mathsf{KL}}(r_Y(Y|Y \neq x)\|p_Y(Y|Y \neq x)) = 0$. Thus, in the case that $p_Y(x) < p_{Y|x}(x)$, we have reduced the optimization problem of Eq. (C8) to the equivalent problem

$$\min_{a \in \mathbb{R}:\, a \geq p_{Y|x}(x)} \mathsf{d}(a, p_Y(x)). \tag{C9}$$

Note that the derivative $\mathsf{d}(a,b)$ with respect to a is $\frac{d}{da}\mathsf{d}(a,b) = \log\frac{a}{b} - \log\frac{1-a}{1-b}$, which is strictly positive when $a > b$. Given the assumption that $p_{Y|x}(x) > p_Y(x)$, Eq. (C9) is minimized by $a = p_{Y|x}(x)$. Thus, $\mathsf{d}(p_{Y|x}(x), p_Y(x))$ is the solution to Eq. (C8) when $p_Y(x) < p_{Y|x}(x)$.

Combining these two results shows that $D_x^{\mathsf{copy}}(p_{Y|x}\|p_Y)$, as defined in Eq. (8), is the solution to Eq. (C8).

3. Vector-valued loss functions

One can also generalize the approach described in Section III to vector-valued loss functions, $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^n$, where we use $\mathcal{X}$ and $\mathcal{Y}$ to indicate the sets of outcomes of X and Y, respectively (recall that these can be different, in the context of our generalized copy and transformation information measures). As we'll see below, one application of vector-valued loss functions is to define measures of copy and transformation information that are additive when independent channels are concatenated.

We first discuss which axioms might be expected to hold for generalized copy information measures with vector-valued loss functions. Axiom 1 and Axiom 2 do not make reference to the loss function, and remain unmodified. Then, Axiom 3* is still a meaningful requirement, as long as the inequality $\mathbb{E}_{p_{Y|x}}[\ell(x,Y)] \geq \mathbb{E}_{q_{Y|x}}[\ell(x,Y)]$ is taken in an element-wise fashion. Axiom 4* should be dropped for vector-valued functions, for reasons explained below.

Using the derivation found in Section C 1, it can be shown that the largest measure which satisfies Axiom 1, Axiom 2, and Axiom 3* for a vector-valued loss function is given by

$$\begin{aligned} G_x^{\mathsf{copy}}(p_{Y|x}\|p_Y) := \min_{r_Y}\ & D_{\mathsf{KL}}(r_Y\|p_Y) \\ \text{s.t.}\quad & \mathbb{E}_{r_Y}[\ell_i(x,Y)] \leq \mathbb{E}_{p_{Y|x}}[\ell_i(x,Y)] \text{ for } i = 1..n, \end{aligned} \tag{C10}$$

where ℓ_i indicates the i^{th} component of the loss function ℓ. Eq. (C10) is a minimum cross-entropy problem with n different constraints. The general solution to this problem will have the following form [2]:

$$w(y) = \frac{1}{Z(\lambda_1, \ldots, \lambda_n)} p_Y(y) e^{-\sum_i \lambda_i \ell_i(x,y)}, \tag{C11}$$

where $\lambda_i \geq 0$ is the Lagrange multiplier for constraint i and $Z(\lambda_1, \ldots, \lambda_n)$ is a normalization constant. The Lagrange multipliers can be found by using standard convex optimization techniques. Note that all $\lambda_i = 0$ if $\mathbb{E}_{p_{Y|x}}[\ell_i(x,Y)] \geq \mathbb{E}_{p_Y}[\ell_i(x,Y)]$ for all i, in which case $w_Y = p_Y$. Even if $\mathbb{E}_{p_{Y|x}}[\ell(x,Y)] < \mathbb{E}_{p_Y}[\ell(x,Y)]$, however, it may be impossible to make all of the constraints simultaneously tight up to equality. In other words, it will not always be the case that

$\mathbb{E}_{w_Y}[\ell_i(x,Y)] = \mathbb{E}_{p_{Y|x}}[\ell_i(x,Y)]$ for all $i = 1..n$, and some (but not all) of the multipliers λ_i will be equal to 0. For this reason, Axiom 4* is not generally achievable for copy information defined with vector-valued loss functions, and we drop it from our requirements. This means G_x^{copy}, as defined in Eq. (C10), is not the unique measure which satisfies the remaining three axioms (Axiom 1, Axiom 2, and Axiom 3*). For example, they are also satisfied by the trivial measure $F(p_{Y|x}, p_Y, x) = 0$ for all $p_{Y|x}$, p_Y, and x.

Vector-valued loss functions can be used to derive an additive measure of copy information. Imagine that source and destination messages consists of sequences of n symbols. If the source symbols are chosen independently, $s(x) = \prod_{i=1}^n s_i(x_i)$, and transmitted across n independent channels, $p(y|x) = \prod_{i=1}^n p_i(y_i|x_i)$, then one can verify that the destination marginal distribution will also have a product form,

$$p(y) = \prod_{i=1}^n p_i(y_i). \tag{C12}$$

In that case, one may desire a measure of copy information that is additive across the n transmission (see also discussion in Section II D). This can be achieved by choosing an n-dimensional loss function, $\ell(x,y) = \langle \ell(x_1,y_1), \ell(x_2,y_2), \ldots, \ell(x_n,y_n)\rangle$. It can be seen from Eq. (C12) and Eq. (C11) that the optimal distribution will have a product form, $w(y) = \prod_{i=1}^n w_i(y_i)$. By Eq. (C10), it can also be checked that the resulting copy information will have an additive form,

$$G_x^{\mathsf{copy}}(p_{Y|x}\|p_Y) = \sum_{i=1}^n G_x^{\mathsf{copy}}(p_{Y_i|x_i}\|p_{Y_i}), \tag{C13}$$

where $G_x^{\mathsf{copy}}(p_{Y_i|x_i}\|p_{Y_i})$ is the generalized copy information defined for dimension i and loss function $\ell_i(x_i,y_i)$. Note that in this case $D_{\mathsf{KL}}(p_{Y|x}\|p_Y) = \sum_i D_{\mathsf{KL}}(p_{Y_i|x_i}\|p_Y)$. Therefore, by Eqs. (C13) and (17), the generalized transformation information G_x^{trans} will also be additive.

D. PROOF OF PROP. 1

Before proving Proposition 1, we prove several intermediate results. We start by proving some useful properties of the roots of the quadratic polynomial $ax^2 - (a+s)x + sc$. In particular, we consider the two roots

$$f_\pm(a,s,c) = \frac{a+s \pm \sqrt{(a+s)^2 - 4asc}}{2a} \tag{D14}$$

where $a \in \mathbb{R}\setminus\{0\}$, $s \in (0,1]$, $c \in (0,1]$.

Lemma D1. *$f_+(a,s,c) < 0$ when $a < 0$ and $f_+(a,s,c) \geq 1$ when $a > 0$.*

Proof. When $a < 0$, $f_+(a,s,c) \leq f_-(a,s,c)$. Vieta's formula states that

$$f_-(a,s,c) f_+(a,s,c) = \frac{sc}{a} < 0. \tag{D15}$$

This implies $f_+(a,s,c) < 0$. When $a > 0$, we lower bound the determinant,

$$(a+s)^2 - 4asc \geq a^2 + 2as + s^2 - 4as = (a-s)^2. \tag{D16}$$

This implies

$$f_+(a,s,c) \geq \frac{a+s+|a-s|}{2a} = \begin{cases} 1 & \text{if } a \geq s \\ \frac{s}{a} > 1 & \text{if } s > a > 0 \end{cases}$$

□

Lemma D2. $\lim_{a\to 0} f_-(a,s,c) = c$.

Proof. By L'Hôpital's rule,

$$\begin{aligned}\lim_{a\to 0} f_-(a,s,c) &= \lim_{a\to 0} \frac{\frac{d}{da}\left(a+s-\sqrt{(a+s)^2-4asc}\right)}{\frac{d}{da}(2a)} \\ &= \frac{1}{2} - \lim_{a\to 0} \frac{2(a+s)-4sc}{2\cdot 2\sqrt{(a+s)^2-4asc}} \\ &= \frac{1}{2} - \frac{s-2sc}{2s} = c.\end{aligned}$$

□

Lemma D3. *$f_-(a,s,c)$ is continuous and monotonically decreasing in a. It is strictly monotonically decreasing in a when $f_-(a,s,c) < 1$.*

Proof. First consider the the case when $c = 1$,

$$f_-(a,s,c) = \frac{a+s-|a-s|}{2a} = \begin{cases} \frac{s}{a} & \text{if } a \geq s \\ 1 & \text{otherwise} \end{cases}$$

which is continuous and monotonically decreasing in a, and strictly so when $f_-(a,s,c) < 1$ (so $a > s$).

When $c < 1$, define the square root of the determinant

$$\eta := \sqrt{(a+s)^2 - 4asc} \overset{(a)}{>} |a-s| \geq 0.$$

Inequality (a) is strict because Eq. (D16) is strict when $c < 1$. Then, consider the derivative,

$$\begin{aligned}&\tfrac{\partial}{\partial a} f_-(a,s,c) \\ &= \frac{1}{4a^2}\left[\left(1 - \frac{1}{2}\frac{2a+2s-4sc}{\eta}\right)2a - 2(a+s-\eta)\right] \\ &= \frac{1}{2a^2}\left[-\frac{a^2+sa-2sca}{\eta} - s + \eta\right] \\ &\propto -a^2 - sa + 2sca - s\eta + \eta^2 \qquad \text{(D17)} \\ &= s\left[a - 2ac + s - \eta\right] \qquad \text{(D18)} \\ &\propto \frac{a-2ac+s}{\eta} - 1 \qquad \text{(D19)} \\ &\leq \frac{|a-2ac+s|}{\eta} - 1\end{aligned}$$

$$= \sqrt{\frac{(a-2ac+s)^2}{\eta^2}} - 1$$
$$= \sqrt{1 - 4a^2c\frac{1-c}{\eta^2}} - 1 < 0,$$

where in Eq. (D17) we multiplied by the (positive) term $2a^2\eta$, in Eq. (D18) we plugged in the definition of η and simplified, and in Eq. (D19) we divided by the (strictly positive) term ηs. The inequality in the last line uses the fact that $4a^2c\frac{1-c}{\eta^2} > 0$ given that $a \neq 0$ and $0 < c < 1$, and that $\sqrt{1-x} < 1$ for $x > 0$. □

We now prove the following.

Theorem D1. *Let $c(x) \in [0,1]$ indicate a set of values for all $x \in \mathcal{A}$. Then, for any source distribution s_X with full support, there is a channel $p_{Y|X}$ that satisfies*

$$p(y|x) = \begin{cases} c(x) & \text{if } x = y \\ \frac{1-c(x)}{1-p_Y(x)} p_Y(y) & \text{otherwise,} \end{cases} \tag{D20}$$

where p_Y is the marginal $p_Y(y) = \sum_x s(x)p(y|x)$. The channel $p_{Y|X}$ is unique if $c(x) > 0$ for all x. Moreover, $I_p(Y:X) = I_p^{\text{copy}}(X \to Y)$ if and only if $\sum_x c(x) \geq 1$.

Proof. We will show that there exists a marginal p_Y that satisfies the consistency conditions of Eq. (D20).

We first eliminate a few edge cases. The solution is trivial for $|\mathcal{A}| = 1$, so we assume that $|\mathcal{A}| \geq 1$. If $c(x) = 0$ for all x, then for any two states $x, x' \in \mathcal{A}$, the following is a solution: $p_Y(x) = s(x')/(s(x)+s(x'))$, $p_Y(x') = s(x)/(s(x)+s(x'))$, $p_Y(x'') = 0$ for all $x'' \in \mathcal{A} \setminus \{x, x'\}$. If $c(x) = 0$ for some but not all x, then the problem can be solved for the reduced outcome space $\mathcal{S} = \{x \in \mathcal{A} : c(x) > 0\}$, using the procedure below. It can then be extended to all outcomes by keeping $p_Y(x)$ fixed for $x \in \mathcal{S}$ and setting $p_Y(x) = 0$ for all $x \in \mathcal{A} \setminus \mathcal{S}$. Therefore, without loss of generality, below we assume $c(x) > 0$ for all x.

We first plug Eq. (D20) into $p_Y(y) = \sum_x s(x)p(y|x)$,

$$p_Y(x) = s(x)c(x) + p_Y(x) \sum_{x':x' \neq x} s(x') \frac{1-c(x')}{1-p_Y(x')}. \tag{D21}$$

Define $a := 1 - \sum_{x'} s(x') \frac{1-c(x')}{1-p_Y(x')}$ and rearrange Eq. (D21) to give

$$0 = s(x)c(x) + p_Y(x)\left(-a - s(x)\frac{1-c(x)}{1-p_Y(x)}\right).$$

Multiplying both sides by $1 - p_Y(x)$ and simplifying gives

$$\begin{aligned} 0 &= s(x)c(x) - s(x)c(x)p_Y(x) - ap_Y(x) + ap_Y(x)^2 - \\ &\quad [p_Y(x)s(x) - p_Y(x)s(x)c(x)] \\ &= ap_Y(x)^2 - (a + s(x))p_Y(x) + s(x)c(x). \end{aligned} \tag{D22}$$

Dividing by $s(x)$, then summing over x and rearranging gives

$$a\left[\sum_x \frac{p_Y(x) - p_Y(x)^2}{s(x)}\right] = \left[\sum_x c(x)\right] - 1\,. \tag{D23}$$

Note that the sum inside the brackets on the left hand side is strictly positive. Thus,we have

$$a \geq 0 \text{ iff } \sum_x c(x) \geq 1 \quad ; \quad a < 0 \text{ iff } \sum_x c(x) < 1 \tag{D24}$$

Note also that $a = 0$ if $\sum_x c(x) = 1$, in which case $p_Y(x) = c(x)$ is the unique solution to Eq. (D22) for all x. Below, we disregard this simple special case, and assume that $\sum_x c(x) \neq 1$ and $a \neq 0$.

We now solve Eq. (D22) for $p_Y(x)$. First, note that $p_Y(x) = \sum_{x'} s(x')p(x|x') \geq c(x)s(x) > 0$ for all x, since we assume that $s(x) > 0$ and $c(x) > 0$ for all x. Given that $|\mathcal{A}| > 1$, this also means that $p_Y(x) < 1$ for all x (if this were not the case, then it would be that $p_Y(x) = 0$ for all except one x). We then solve the quadratic equation,

$$p_Y^a(x) = \frac{a + s(x) - \sqrt{(a+s(x))^2 - 4as(x)c(x)}}{2a}, \tag{D25}$$

where we include the superscript a in p_Y^a to make the dependence on a explicit. We chose the negative solution of the quadratic equation because, by Lemma D1, it is the only one compatible with the requirement that $0 < p_Y^a(x) < 1$.

We wish to find the value of a satisfies $\sum_x p_Y^a(x) = 1$, which is defined implicitly via

$$1 = \sum_x \frac{a + s(x) - \sqrt{(a+s(x))^2 - 4as(x)c(x)}}{2a} \tag{D26}$$

Note that each $p_Y^a(x)$ is continuous and strictly monotonically decreasing in a (Lemma D3), and therefore so is the right hand side of Eq. (D26). Moreover, a must lie between -1 and 1. To see why, evaluate the right hand side of Eq. (D26) for $a = -1$,

$$\begin{aligned} &\sum_x \frac{1 - s(x) + \sqrt{(1+s(x))^2 + 4s(x)c(x)}}{2} \\ &\geq \sum_x \frac{1 - s(x) + (1 + s(x))}{2} = n \geq 1 \end{aligned}$$

Then, evaluate it for $a = 1$,

$$\begin{aligned} &\sum_x \frac{1 + s(x) - \sqrt{(1+s(x))^2 - 4s(x)c(x)}}{2} \\ &\leq \sum_x \frac{1 + s(x) - \sqrt{(1+s(x))^2 - 4s(x)}}{2} \\ &= \sum_x \frac{1 + s(x) - (1 - s(x))}{2} = \sum_x \frac{2s(x)}{2} = 1 \end{aligned}$$

Thus, there is a unique $a \in [-1, 1]$ that satisfies Eq. (D26), resulting in a unique p_Y^a and corresponding $p_{Y|X}$ in Eq. (D20).

Now, by definition of I^{copy}, $I_p(Y:X) = I_p^{\mathsf{copy}}(X \to Y)$ if $c(x) \geq p_Y(x)$ for all x. By Lemma D2 and Lemma D3, the right hand side of Eq. (D25) is greater than $c(x)$ if and only if $a \geq 0$. By Eq. (D24), $a \geq 0$ if and only if $\sum_x c(x) \geq 1$. □

In practice, the value a^* in the proof of Theorem D1 can be found by a numerical root finding algorithm, or by trying values from -1 to 1 in small intervals and selecting the first value that makes the LHS of Eq. (D26) less than or equal to 1. The marginal p_Y and channel $p_{Y|X}$ can then be computed in closed form using Eqs. (D20) and (D25).

We are now ready to prove Proposition 1.

Proposition 1. *For any source distribution s_X with $H(X) < \infty$, there exist channels p for all levels of mutual information $I_p(Y:X) \in [0, H(X)]$ such that $I_p^{\mathsf{copy}}(X \to Y) = I_p(Y:X)$.*

Proof. Consider the proof of Theorem D1. Note that for each $x \in \mathcal{A}$ and any $\gamma \in [0,1]$, Eq. (D22) is satisfied by taking $p_Y(x) = s(x)$ and $c(x) = \gamma + s(x) - \gamma s(x)$.

Let $p^{\gamma}_{Y|X}$ represent the channel corresponding to each γ, as defined in Eq. (D20). It is easy to check that $I^{\mathsf{copy}}_{p^\gamma}(X \to Y) = I_{p^\gamma}(Y:X)$, with $I^{\mathsf{copy}}_{p^\gamma}(X \to Y) = 0$ for $\gamma = 0$ and $I^{\mathsf{copy}}_{p^\gamma}(X \to Y) = H(s_X)$ for $\gamma = 1$. Note that $c(x)$ is increases monotonically in γ for all x, from $c(x) = s(x)$ for $\gamma = 0$ to $c(x) = 1$ for $\gamma = 1$. This means that for all γ,

$$\begin{aligned} I^{\mathsf{copy}}_{p^\gamma}(X \to Y) &= \sum_x s(x) \mathsf{d}(c(x), s(x)) \\ &\leq -\sum_x s(x) \ln s(x) = H(s_X) < \infty. \end{aligned}$$

Thus, the sums that define $I^{\mathsf{copy}}_{p^\gamma}(X \to Y)$ for each γ converge uniformly. Therefore, $I^{\mathsf{copy}}_{p^\gamma}(X \to Y)$ is continuous in γ. The proposition follows from the intermediate value theorem. □

E. THE BINARY SYMMETRIC CHANNEL

The BSC is a channel over a two-state space ($\mathcal{A} = \{0,1\}$) parameterized by a "probability of error" $\epsilon \in [0,1]$. The BSC can be represented in matrix form as

$$p^{\epsilon}_{Y|X} = \begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix}.$$

When $\epsilon = 0$, the BSC is a noiseless channel which copies the source without error. In this extreme case, MI is large, and we expect it to consist entirely of copy information. On the other hand, when $\epsilon = 1$, the BSC is a noiseless "inverted" channel, where messages are perfectly switched between the source and the destination. In this case, MI is again large, but we now expect it to consist entirely of transformation information. Finally, $\epsilon = 1/2$ defines a completely noisy channel, for which mutual information (and thus copy and transformation information) must be 0.

For simplicity, we assume a uniform source distribution, $s_X(0) = s_X(1) = 1/2$, which by symmetry implies a marginal probability $p_Y(0) = p_Y(1) = 1/2$ at the destination for any ϵ. For the BSC with this source distribution, Eq. (8) states that for both $x = 0$ and $x = 1$, $D_x^{\mathsf{copy}}(p^{\epsilon}_{Y|x} \| p_Y) = I_{p^\epsilon}(Y : X{=}x)$ and $D_x^{\mathsf{trans}}(p^{\epsilon}_{Y|x} \| p_Y) = 0$ when $\epsilon \leq 1/2$, and $D_x^{\mathsf{copy}}(p^{\epsilon}_{Y|x} \| p_Y) = 0$ and $D_x^{\mathsf{trans}}(p^{\epsilon}_{Y|x} \| p_Y) = I_{p^\epsilon}(Y : X{=}x)$ otherwise. Using the definition of the (total) copy and transformation components of total MI, Eqs. (10) and (11), it then follows that

$$I^{\mathsf{copy}}_{p^\epsilon}(X \to Y) = \begin{cases} I_{p^\epsilon}(Y:X) & \text{if } \epsilon \leq 1/2 \\ 0 & \text{otherwise} \end{cases}$$

$$I^{\mathsf{trans}}_{p^\epsilon}(X \to Y) = \begin{cases} I_{p^\epsilon}(Y:X) & \text{if } \epsilon \geq 1/2 \\ 0 & \text{otherwise} \end{cases}.$$

This confirms intuitions about the BSC discussed in the beginning of this section. The behavior of MI, $I^{\mathsf{copy}}(X \to Y)$ and $I^{\mathsf{trans}}(X \to Y)$ for the BSC with a uniform source distribution is shown visually in Fig. 2 of the main text.

[1] Imre Csiszar and János Körner. *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011.

[2] Reuven Y. Rubinstein and Dirk P. Kroese. *Simulation and the Monte Carlo method*, volume 10. John Wiley & Sons, 2016.

[3] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.