A Rate-Distortion Approach to Caching

Roy Timo, Shirin Saeedi Bidokhti, Michèle Wigger and Bernhard C. Geiger

Abstract

In this paper we consider a lossy single-user caching problem with correlated sources. We first describe the fundamental interplay between the source correlations, the capacity of the user's cache, the user's reconstruction distortion requirements, and the final delivery-phase (compression) rate. We then illustrate this interplay using a multivariate Gaussian source example and a binary symmetric source example. To fully explore the effect of the user's distortion requirements, we formulate the caching problem using f-separable distortion functions recently introduce by Shkel and Verdú. The class of f-separable distortion functions includes separable distortion functions as a special case, and our analysis covers both the expected- and excess-distortion settings in detail. We also determine what "common information" should be placed in the cache, and what information should be transmitted during the delivery phase. To this end, two new common-information measures are introduced for caching, and their relationship to the common-information measures of Wyner, Gács and Körner is discussed in detail.

I. Introduction

THIS paper takes a rate-distortion (RD) approach to understanding the information-theoretic laws governing cache-aided communications systems. To help fix ideas, let us start by outlining some of the applications that motivated this paper.

A. Motivation

1) Streaming media: Consider the problem of streaming media to millions of users. A common problem is that the users will most likely request and stream media during periods of high congestion. For example, most users would prefer to watch a movie during the evening, rather than during the early hours of the morning. Downloading bandwidth hungry media files during such periods leads to further congestion, high latency, and poor user experience.

To help overcome this problem, content providers often cache useful information about the media library in small storage systems at the network edge (with fast user connections) during periods of low congestion. Naturally these small caches cannot host the entire media library, so the provider must carefully cache information that will be useful to the users' future requests.

2) Distributed databases: Now imagine a large database that is distributed over a vast global disk-storage network. Such a database might contain measurements taken by weather or traffic sensors spread across several countries; the time-series prices of companies' stock at different exchanges; the shopping history of customers; or the mobility patterns of mobile devices in cellular networks.

Now suppose that a user queries the database and requests an approximate copy of one file (or, perhaps, a function of several files). Since the database is large and distributed, we can expect that it will need to make several network calls to load relevant data in memory before it can communicate the file to the user. Such network calls are performance bottlenecks, potentially leading to high latency and network traffic costs.

Modern database systems handle such problems by smartly caching the most common queries in fast memory. For example, a user is more likely to request the weather forecast of its hometown rather than of a remote location, hence we can simply cache this forecast in memory close to this user. Obviously, however, we cannot always know in advance what data will be requested, so we should carefully cache information that is useful to many different requests.

B. Focus and modelling assumptions

Our study will focus on the lossy single-user system illustrated in Figure 1. This basic caching problem consists of two distinct phases: A *caching phase* where information about the library is transported (e.g., during a period of low congestion) to a *cache* near the user; and a *delivery phase* where the particular source/file requested by the user is compressed, transported to the user (e.g., during a period of peak-congestion), and reconstructed in a lossy manner subject to some distortion constraint. The main purpose of this paper is to answer the following questions:

1) What are the fundamental tradeoffs between the cache capacity, delivery-phase (compression) rate, and the user's reconstruction distortion requirements?

R. Timo is with Ericsson Research, Stockholm, roy.timo@ericsson.com. S. Saeedi Bidokhti is with the Department of Electrical Engineering, Stanford University, saeedi@stanford.edu. M. Wigger is with the Communications and Electronics Department, Telecom ParisTech, michele.wigger@telecom-paristech.fr. B. C. Geiger is with the Institute for Communications Engineering, Technical University of Munich, bernhard.geiger@tum.de.

Some of the material in this paper was completed by R. Timo at the Technical University of Munich and presented at the International Zurich Seminar on Communications (IZS), March, 2016.

S. Saeedi Bidokhti was supported by the Swiss National Science Foundation Fellowship no. 158487. Bernhard C. Geiger was supported by the Erwin Schrödinger Fellowship J 3765 of the Austrian Science Fund.

2) What “common information” should be placed in the cache, and what information should be transmitted during the delivery phase?

We have chosen this basic single-user setup because it focuses on the interplay between the user's distortion constraints and various notions of common information between the sources. Indeed, we are particularly interested in understanding how probabilistic dependencies between sources affect this interplay and common information. We will see, for example, that the particular choice of distortion function greatly influences what common information should be placed in the cache.

We assume throughout the paper that the library, which we denote by $\boldsymbol{X}^n$, consists of $(L \geq 1)$ different sources:

$$\boldsymbol{X}^n = (X_1^n, X_2^n, \ldots, X_L^n).$$

The ℓ-th source X_ℓ^n, where $\ell \in \mathcal{L} := \{1, 2, \ldots, L\}$, consists of n symbols chosen from a discrete and finite alphabet[1] $\mathcal{X}_\ell$:

$$X_\ell^n = (X_{\ell,1}, X_{\ell,2}, \ldots, X_{\ell,n}).$$

We accordingly assume that the cache can reliably store up to nC bits, and we say that it has a capacity of C bits per source symbol. The number of source symbols n (also called the blocklength) will be allowed to grow without bound so as to enable an information-theoretic analysis. Thus, we are interested in libraries consisting of a fixed number of large sources/files.

We further assume that the library $\boldsymbol{X}^n$ is randomly generated by a discrete memoryless source (DMS); that is, $\boldsymbol{X}^n$ is a sequence of n independent and identically distributed (iid) tuples $\boldsymbol{X} = (X_1, \ldots, X_L)$ defined by an arbitrary joint pmf $p_{\boldsymbol{X}}(\boldsymbol{x})$ defined on $\boldsymbol{\mathcal{X}} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_L$. This assumption is quite common in the multi-terminal RD theory literature, as it admits rigorous proofs and gives some insight to more complicated models. Although it is somewhat restrictive, some important transformations (e.g. Burrows-Wheeler) are known to emit almost memoryless processes [3, 4].

So as to fully explore the influence of the user's distortion constraints in the caching problem, we will study both *separable distortion functions* and the more general f-*separable distortion functions* under both *expected*- and *excess-distortion constraints*. Specifically, for each $\ell \in \mathcal{L}$ let $\hat{\mathcal{X}}_\ell$ denote the user's reconstruction alphabet for the ℓ-th source, and let $\mathsf{d}_\ell : \hat{\mathcal{X}}_\ell \times \mathcal{X}_\ell \to [0, \infty)$ be an arbitrary *symbol distortion function*[2]. For example, d_ℓ can be the *Hamming distortion function* where $\hat{\mathcal{X}}_\ell = \mathcal{X}_\ell$ and

$$\mathsf{d}_\ell(\hat{x}_\ell, x_\ell) = \begin{cases} 1 & \text{if } \hat{x}_\ell \neq x_\ell \\ 0 & \text{if } \hat{x}_\ell = x_\ell. \end{cases}$$

The n-symbol *separable distortion* between a source sequence $x_\ell^n \in \mathcal{X}_\ell^n$ and reconstruction sequence $\hat{x}_\ell^n \in \hat{\mathcal{X}}_\ell^n$ is then

$$\bar{\mathsf{d}}_\ell(\hat{x}_\ell^n, x_\ell^n) := \frac{1}{n} \sum_{i=1}^{n} \mathsf{d}_\ell(\hat{x}_{\ell,i}, x_{\ell,i}). \tag{1}$$

Separable distortion functions (1) are widely used in the multi-terminal RD theory literature primarily because they yield single-letter (i.e., computable) solutions to optimal RD trade-off problems. Unfortunately, distortion functions used in practice are often not separable. The more general class of f-separable distortion functions, recently proposed by Shkel and Verdú [1], provides more flexibility in this regard. Let $\mathsf{f}_\ell : [0, \infty) \to [0, \infty)$ be any continuous and strictly increasing function. The n-symbol f-*separable distortion* between $x_\ell^n \in \mathcal{X}_\ell^n$ and $\hat{x}_\ell^n \in \hat{\mathcal{X}}_\ell^n$ is then

$$\overline{\mathsf{fd}}_\ell(\hat{x}^n, x^n) := \mathsf{f}_\ell^{-1}\left(\frac{1}{n} \sum_{i=1}^{n} \mathsf{f}_\ell\big(\mathsf{d}_\ell(\hat{x}_i, x_i)\big)\right). \tag{2}$$

The basic idea here is to design f_ℓ to assign appropriate (possibly non-linear) frequency costs to different quantization error events. If f_ℓ is the identity mapping, then $\overline{\mathsf{fd}}_\ell$ reduces to the usual separable distortion function $\bar{\mathsf{d}}_\ell$ corresponding to d_ℓ. Several interesting connections between f-separable distortions and Rényi entropy, compression with linear costs, and sub-additive distortion functions are discussed in [1]. Perhaps the most appealing motivation for using f-separable distortions, however, is the axiomatic argument provided by the following proposition (for a more detailed discussion, see [1]).

Proposition 1 (Kolmogorov [2]): Let $\{a_1, \ldots, a_n\}$ be any set of n real numbers and $\bar{\mathsf{M}}_n : \mathbb{R}^n \to \mathbb{R}$ satisfy the following four *axioms of mean*: (1) $\bar{\mathsf{M}}_n(a_1, \ldots, a_n)$ is a continuous and strictly increasing function of each argument a_i. (2) $\bar{\mathsf{M}}_n(a_1, \ldots, a_n)$ is a symmetric function of its arguments. (3) $\bar{\mathsf{M}}_n(a, \ldots, a) = a$. (4) For any integer $m \leq n$, $\bar{\mathsf{M}}_n(a_1, \ldots, a_m, \ldots, a_n) = \bar{\mathsf{M}}_n(a, \ldots, a, a_{m+1}, \ldots, a_n)$, where $a = \bar{\mathsf{M}}_m(a_1, \ldots, a_m)$. Then $\bar{\mathsf{M}}_n$ must take the form [2, p. 144]

$$\bar{\mathsf{M}}_n(a_1, \ldots, a_n) = \mathsf{f}^{-1}\left(\frac{1}{n} \sum_{i=1}^{n} \mathsf{f}(a_i)\right)$$

for some continuous and strictly increasing f.

[1] We will drop this assumption for two Gaussian source examples.

[2] To simplify the presentation, we assume throughout that each d_ℓ satisfies the following two conditions: (1) For each source symbol $x_\ell \in \mathcal{X}_\ell$ there exists a reconstruction symbol $\hat{x}_\ell \in \hat{\mathcal{X}}_\ell$ such that $\mathsf{d}_\ell(\hat{x}_\ell, x_\ell) = 0$; and (2) there exists a finite $D_{\max} > 0$ such that $\mathsf{d}_\ell(\hat{x}_\ell, x_\ell) \leq D_{\max}$ for all $x_\ell \in \mathcal{X}_\ell$ and $\hat{x}_\ell \in \hat{\mathcal{X}}_\ell$.

Thus, if we have any n-symbol distortion function that computes some mean of per-symbol distortions (satisfying the above axioms), then it must be an f-separable distortion function.

Remark 1: Although f-separable distortion functions are more general than separable distortions, we will not state and prove our main results directly using f-separable distortion functions. Instead, we will first consider separable distortion functions and then generalize to f-separable distortions. The reason for this approach is that the f-separable distortion proofs will need to bootstrap results for separable distortions.

C. Related literature and main contributions

Cache-aided communication systems have been of interest in the recent information-theoretic literature, e.g., [5]–[7, 29]–[39]. The works [5]–[7] consider correlated sources, and, among these, the work that is most related to our setup is by Wang, Lim and Gastpar [5]. A key difference to [5], however, is the source request model: Wang *et al.* assumed that at each time instant i the user(s) randomly select a symbol from the tuple $(X_{1,i}, X_{2,i}, \ldots, X_{L,i})$ in an iid manner. They then leveraged connections to the Gray-Wyner network to establish some interesting trade-offs between the optimal compression rate and cache capacity under a lossless[3] reconstruction constraint. In contrast to [5], we will require that the user requests one source in its entirety, we do not place prior probabilities on the user's selection, and we allow for lossy reconstructions. We thus consider a lossy worst-demand (i.e., compound source) scenario, while [5] considered a lossless ergodic iid-demand scenario.

Hassanzadeh, Erkip, Llorca and Tulino [6] studied cache-aided communications systems for transmitting independent memoryless Gaussian sources under mean-squared error distortion constraints. Their caching schemes exploited successive-refinement techniques to minimize the mean-squared error of the users' reconstructions, and they presented a useful "reverse filling-type solution" to the minimum distortion problem. Yang and Gündüz [7] consider the same cache-aided Gaussian problem, but instead focussed on the minimum delivery-phase rate for a given distortion requirement. They presented a numerical method to determine the minimum delivery rate, and proposed two efficient caching algorithms.

In the light of this, the main contributions of our work are:

- In Section II, we show that the single-user caching problem, assuming that the user's reconstructions are subject to expected (separable) distortion constraints, is related to the lossy Gray-Wyner network. We then present a coding theorem that characterizes the interplay between the delivery rate, cache capacity and reconstruction distortion with a single-letter optimization problem.
- In Section III, we evaluate (or, bound) the above optimization problem for three different examples: 1) A multivariate Gaussian source with respect to separable squared error distortions, 2) a bivariate Gaussian source with respect to separable squared error distortions, and 3) a doubly-symmetric binary source with respect to Hamming distortions.
- The three examples outlined above all use some idea of "common information" to specify the best information to place in the cache. In Section IV we elaborate on this idea, and provide two new common-information measures for caching. The new measures both have operational meaning for caching and can be computed via single-letter expressions. We then describe how the new measures relate to (and differ from) the well-known common information measures of Wyner, Gács and Körner that often appear in studies related to the Gray-Wyner network.
- The above results are all derived w.r.t. expected (separable) distortions. In Section V, we study excess (separable) distortions, and our main result is a new strong converse. The new converse does not automatically follow from the strong converse of the standard RD problem, and, instead, uses a perturbed source idea that is motivated by the work of Watanabe [25]. Based on this new converse we study f-separable distortion functions in Section VI.

D. Basic Informational RD functions

The following functions will be used throughout the paper. The *informational RD function* of the ℓ-th source X_ℓ w.r.t. the symbol distortion function $\mathsf{d}_\ell : \hat{\mathcal{X}}_\ell \times \mathcal{X}_\ell \to [0, \infty)$ is

$$\mathsf{R}_{X_\ell}(D_\ell) := \min_{p_{\hat{X}_\ell|X_\ell}:\ \mathbb{E}[\mathsf{d}_\ell(\hat{X}_\ell, X_\ell)] \leq D_\ell} I(X_\ell; \hat{X}_\ell),$$

where the minimization is over all test channels $p_{\hat{X}_\ell|X_\ell}$ from $\mathcal{X}_\ell$ to $\hat{\mathcal{X}}_\ell$ satisfying the indicated distortion constraint. The *informational joint RD function of* $\boldsymbol{X} = (X_1, \ldots, X_L)$ w.r.t. the symbol distortion functions $\mathbf{d} = (\mathsf{d}_1, \ldots, \mathsf{d}_L)$ is [8]

$$\mathsf{R}_{\boldsymbol{X}}(\boldsymbol{D}) := \min_{p_{\hat{\boldsymbol{X}}|\boldsymbol{X}}:\ \mathbb{E}[\mathsf{d}_\ell(\hat{X}_\ell, X_\ell)] \leq D_\ell,\ \forall \ell \in \mathcal{L}} I(\boldsymbol{X}; \hat{\boldsymbol{X}}),$$

where the minimization is over all test joint channels $p_{\hat{\boldsymbol{X}}|\boldsymbol{X}}$ from $\boldsymbol{\mathcal{X}}$ to $\hat{\boldsymbol{\mathcal{X}}}$ satisfying all L indicated distortion constraints. Finally, the *informational conditional RD function* [8] of X_ℓ with side information U is

$$\mathsf{R}_{X_\ell|U}(D_\ell) := \min_{p_{\hat{X}_\ell|X_\ell U}:\ \mathbb{E}[\mathsf{d}_\ell(\hat{X}_\ell, X_\ell)] \leq D_\ell} I(X_\ell; \hat{X}_\ell|U),$$

[3]Specifically, Wang *et al.* required that a function of the source is reliably reconstructed (otherwise known as a deterministic distortion function).

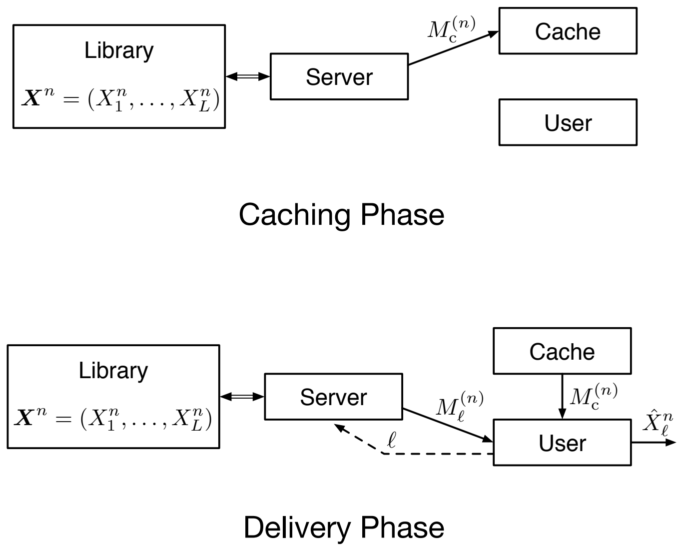

Fig. 1. A cache-aided communications system with a single user.

where the minimization is over all test channels $p_{\hat{X}_\ell|X_\ell U}$ from $\mathcal{X}_\ell \times \mathcal{U}$ to $\hat{\mathcal{X}}_\ell$ satisfying the indicated distortion constraint.

Remark 2: The above minima exist by the continuity of Shannon's information measures, the assumption of bounded single-symbol distortion functions d, and the fact that each (conditional) mutual information is minimized over a compact set.

II. Caching w.r.t. expected (separable) distortions

A. Problem setup

A joint *rate-distortion-cache (RDC) code* for a given blocklength n is a collection of $(2L+1)$ mappings:

(i) A *cache-phase encoder* at the server $\phi_{\mathrm{c}}^{(n)} : \boldsymbol{\mathcal{X}}^n \to \mathcal{M}_{\mathrm{c}}^{(n)}$. Here $\mathcal{M}_{\mathrm{c}}^{(n)}$ is a finite (index) set with an appropriate cardinality for the cache capacity.

(ii) A *delivery-phase encoder* at the server $\phi_\ell^{(n)} : \boldsymbol{\mathcal{X}}^n \to \mathcal{M}^{(n)}$ for each user request $\ell \in \mathcal{L}$. Here $\mathcal{M}^{(n)}$ is a finite (index) set with an appropriate cardinality for the delivery phase.

(iii) A *delivery-phase decoder* at the user $\varphi_\ell^{(n)} : \mathcal{M}^{(n)} \times \mathcal{M}_{\mathrm{c}}^{(n)} \to \hat{\mathcal{X}}_\ell^n$ for each possible user request $\ell \in \mathcal{L}$.

We call the above collection of encoders and decoders an $(n, |\mathcal{M}^{(n)}|, |\mathcal{M}_{\mathrm{c}}^{(n)}|)$*-code*.

During the *caching phase*, the server places the message $M_{\mathrm{c}}^{(n)} = \phi_{\mathrm{c}}^{(n)}(\boldsymbol{X}^n)$ in the cache. Later, during the *delivery phase*, the user picks $\ell \in \mathcal{L}$ arbitrarily and requests the corresponding source X_ℓ^n from the server. The server responds to the user's request with the message $M_\ell^{(n)} = \phi_\ell^{(n)}(\boldsymbol{X}^n)$, and the user attempts to reconstruct X_ℓ^n by computing $\hat{X}_\ell^n = \varphi_\ell^{(n)}(M_\ell^{(n)}, M_{\mathrm{c}}^{(n)})$. This encoding and decoding process is illustrated in Figure 1.

Suppose that we would like the caching system to operate with a delivery-phase rate R, cache capacity C, and reconstruction distortions $\boldsymbol{D} = (D_1, \ldots, D_L)$, where D_ℓ is the desired expected distortion of the ℓ-th source X_ℓ^n.

Definition 1: We say that the rate-distortion-cache tuple $(R, \boldsymbol{D}, C)$ is *achievable w.r.t. expected (separable) distortions* if there exists a sequence of $(n, |\mathcal{M}^{(n)}|, |\mathcal{M}_{\mathrm{c}}^{(n)}|)$-codes such that

$$\limsup_{n\to\infty} \frac{1}{n} \log |\mathcal{M}_{\mathrm{c}}^{(n)}| \leq C, \tag{3a}$$

$$\limsup_{n\to\infty} \frac{1}{n} \log |\mathcal{M}^{(n)}| \leq R, \text{ and} \tag{3b}$$

$$\limsup_{n\to\infty} \mathbb{E}\Big[\bar{\mathsf{d}}_\ell(\hat{X}_\ell^n, X_\ell^n)\Big] \leq D_\ell, \quad \forall\, \ell \in \mathcal{L}. \tag{3c}$$

The *RDC function w.r.t. expected (separable) distortions* $\mathsf{R}^\dagger(\boldsymbol{D}, C)$ is the infimum of all rates $R \geq 0$ such that $(R, \boldsymbol{D}, C)$ is achievable.

The next lemma summarizes some basic properties of $\mathsf{R}^\dagger(\boldsymbol{D}, C)$ that will be useful later. We omit the proof.

Lemma 2:

(i) $\mathsf{R}^\dagger(\boldsymbol{D}, C)$ is convex, non-increasing and continuous in $(\boldsymbol{D}, C) \in [0, \infty)^{L+1}$.

(ii) If the cache capacity is larger than the informational joint RD function $C > \mathsf{R}_{\boldsymbol{X}}(\boldsymbol{D})$, then $\mathsf{R}^\dagger(\boldsymbol{D}, C) = 0$.

(iii) If the cache has zero capacity $C = 0$, then $\mathsf{R}^\dagger(\boldsymbol{D}, 0) = \max_{\ell \in \mathcal{L}} \mathsf{R}_{X_\ell}(D_\ell)$.

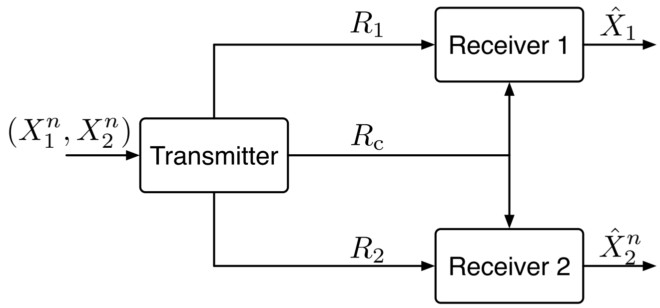

Fig. 2. The Gray-Wyner network.

B. A single-letter expression for $\mathsf{R}^\dagger(\boldsymbol{D}, C)$

A computable single-letter expression for $\mathsf{R}^\dagger(\boldsymbol{D}, C)$ can easily be obtained by leveraging known results for the Gray-Wyner network shown in Figure 2. The Gray-Wyner network is a multi-user RD problem with a single transmitter and two receivers. The transmitter is connected to the receivers via a *single* common link with rate R_c and two private links of rates R_1 and R_2 respectively. Receiver ℓ is required to reconstruct the ℓ-th source X_ℓ^n to within an expected (separable) distortion D_ℓ. The set of all achievable RD tuples $(R_\mathrm{c}, R_1, R_2, D_1, D_2)$ was established by Gray and Wyner in [12]. It is straightforward to extend this result to the case of $(L \geq 2)$-receivers (with one common rate R_c and L private rates $\boldsymbol{R} = (R_1, \ldots, R_L)$): The set of all achievable RD tuples $(R_\mathrm{c}, \boldsymbol{R}, \boldsymbol{D})$ for the L-receiver Gray-Wyner network is given by

$$\mathcal{R}_{\mathrm{GW}}(\boldsymbol{D}) := \bigcup_{p_{U|\boldsymbol{X}}} \left\{ (R_\mathrm{c}, \boldsymbol{R}) : \begin{array}{lll} R_\mathrm{c} & \geq & I(\boldsymbol{X}; U) \\ R_\ell & \geq & \mathsf{R}_{X_\ell|U}(D_\ell) \quad \forall\, \ell \in \mathcal{L}. \end{array} \right\},$$

where the union is over all test channels $p_{U|\mathbf{X}}$ from $\boldsymbol{\mathcal{X}}$ to $\mathcal{U}$ with $|\mathcal{U}| \leq |\boldsymbol{\mathcal{X}}| + 2L$. The next lemma shows that our RDC function $\mathsf{R}^\dagger(\boldsymbol{D}, C)$ can be expressed as a minimization over the achievable rate region $\mathcal{R}_{\mathrm{GW}}(\boldsymbol{D})$.

Lemma 3: $\mathsf{R}^\dagger(\boldsymbol{D}, C) = \mathsf{R}(\boldsymbol{D}, C)$, where

$$\mathsf{R}(\boldsymbol{D}, C) = \min_{U:\ I(\boldsymbol{X};U) \leq C}\ \max_{\ell \in \mathcal{L}}\ \mathsf{R}_{X_\ell|U}(D_\ell) \tag{4}$$

and the minimization is over all test channels $p_{U|\mathbf{X}}$ from $\boldsymbol{\mathcal{X}}$ to $\mathcal{U}$ with $|\mathcal{U}| \leq |\boldsymbol{\mathcal{X}}| + 2L$.

We call $\mathsf{R}(\boldsymbol{D}, C)$ the *informational RDC function*. This function will play a central role in this paper.

Proof of Lemma 3: We need only show that

$$\mathsf{R}^\dagger(\boldsymbol{D}, C) = \min_{(C, \boldsymbol{R}) \in \mathcal{R}_{\mathrm{GW}}(\boldsymbol{D})}\ \max_{\ell \in \mathcal{L}}\ R_\ell. \tag{5}$$

If $(C, \boldsymbol{R}) \in \mathcal{R}_{\mathrm{GW}}(\boldsymbol{D})$, then we can use the corresponding Gray-Wyner encoder and decoders to achieve a delivery phase-rate of $\max_\ell R_\ell$ in the caching problem; thus, $\mathsf{R}^\dagger(\boldsymbol{D}, C)$ cannot be larger than the R.H.S. of (5). Now suppose $\mathsf{R}^\dagger(\boldsymbol{D}, C)$ is strictly smaller than the R.H.S. of (5): There would then exist an encoder and decoders in the Gray-Wyner problem that can operate outside of the rate region $\mathcal{R}_{\mathrm{GW}}(\boldsymbol{D})$. ■

III. Examples of $\mathsf{R}(\boldsymbol{D}, C)$

We now evaluate/bound the informational RDC function $\mathsf{R}(\boldsymbol{D}, C)$ for some common sources and symbol distortion functions.

A. Identical and independent sources

Suppose that $\boldsymbol{X} = (X_1, \ldots, X_L)$ consists of L mutually independent instances of a random variable X on $\mathcal{X}$. If the symbol distortion functions are identical $\mathsf{d}_1 = \cdots = \mathsf{d}_L = \mathsf{d}$ and the distortion constraints are symmetric $\boldsymbol{D} = (D, \ldots, D)$, then informational RDC function is given by

$$\mathsf{R}(\boldsymbol{D}, C) = \left[\mathsf{R}_X(D) - \frac{C}{L} \right]^+,$$

where $[a] := \max\{a, 0\}$. The optimal caching strategy for this case is simple: Take an optimal RD code for each $(X, \bar{\mathsf{d}}_\ell, D)$; compress each X_ℓ^n to the RD limit $\mathsf{R}_X(D)$; cache C/L of the compressed bits output by each RD code; and transmit the remaining bits during the delivery phase.

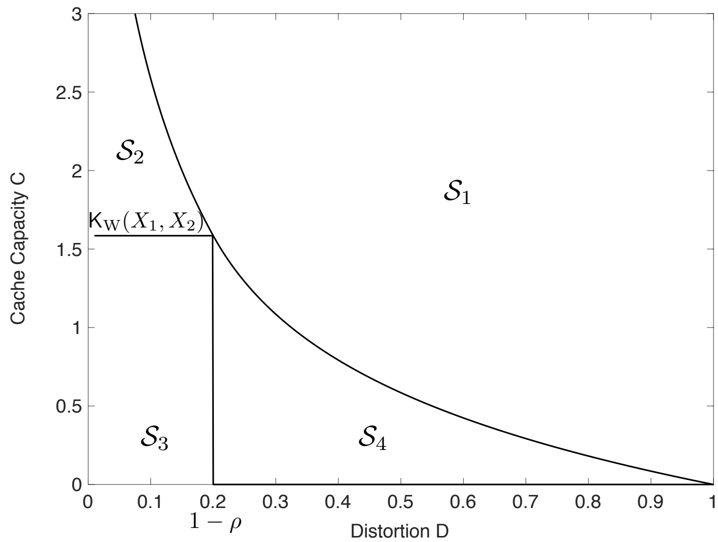

Fig. 3. Illustration of the distortion-cache regions $\mathcal{S}_1$, $\mathcal{S}_2$, $\mathcal{S}_3$ and $\mathcal{S}_4$ used in Proposition 6 with $\rho = 0.8$.

B. Multivariate Gaussian sources with squared error distortion functions

The discussion so far has been restricted to sources defined on finite alphabets. However, it can be shown that the above ideas extend to multivariate Gaussian sources with squared-error distortions, e.g. [14]. Let $\boldsymbol{X} = (X_1, \ldots, X_L) \in \mathbb{R}^L$ be a zero mean multivariate Gaussian with covariance matrix $\boldsymbol{K_X}$ and $\mathsf{d}_\ell(\hat{x}_\ell, x_\ell) = (\hat{x}_\ell - x_\ell)^2$ for all $\ell \in \mathcal{L}$. Let $\mathsf{R}_{\mathrm{G}}^{\dagger}(\boldsymbol{D}, C)$ denote the corresponding operational RDC function w.r.t. the expected (separable) distortion constraints

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{X}_{\ell,i} - X_{\ell,i})^2\right] \leq D_i, \quad \forall\ \ell \in \mathcal{L}.$$

Now let

$$\mathsf{R}_{\mathrm{G}}(\boldsymbol{D}, C) = \inf_{(U,\hat{\boldsymbol{X}})} \ \max_{\ell \in \mathcal{L}} I(X_\ell; \hat{X}_\ell | U), \tag{6}$$

where the infimum is taken over all tuples $(U, \hat{\boldsymbol{X}})$ jointly distributed with $\boldsymbol{X}$ such that

$$I(\boldsymbol{X}; U) \leq C \tag{7a}$$

and

$$\mathbb{E}\big[(X_\ell - \hat{X}_\ell)^2\big] \leq D_\ell, \quad \forall\ \ell \in \mathcal{L}. \tag{7b}$$

The next lemma is the Gaussian counterpart of Lemma 3. Its proof is omitted.

Lemma 4: $\mathsf{R}_{\mathrm{G}}^{\dagger}(\boldsymbol{D}, C) = \mathsf{R}_{\mathrm{G}}(\boldsymbol{D}, C)$.

The next lemma gives a lower bound on $\mathsf{R}_{\mathrm{G}}(\boldsymbol{D}, C)$ for symmetric distortions. For each subset $\mathcal{S} \subseteq \mathcal{L}$, let $X_{\mathcal{S}} = (X_\ell;\ \ell \in \mathcal{S})$ denote the tuple of random variables with indices in $\mathcal{S}$, and let $\boldsymbol{K}_{X_{\mathcal{S}}}$ denote the covariance matrix of $X_{\mathcal{S}}$.

Proposition 5: If $\boldsymbol{D} = (D, \ldots, D)$, then

$$\mathsf{R}_{\mathrm{G}}(\boldsymbol{D}, C) \geq \max_{\mathcal{S} \subseteq \mathcal{L}} \left[\frac{1}{2|\mathcal{S}|} \log \frac{\det \boldsymbol{K}_{X_{\mathcal{S}}}}{D^{|\mathcal{S}|}} - \frac{C}{|\mathcal{S}|}\right].$$

Proof: Proposition 5 is proved in Appendix A. ■

C. Bivariate Gaussian Sources

Fix $\rho \in (0, 1)$ and consider a zero mean bivariate Gaussian source $\boldsymbol{X} = (X_1, X_2)$ with the covariance matrix

$$\boldsymbol{K}_{X_1 X_2} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}. \tag{8}$$

We wish to evaluate the Gaussian RDC function in (6) with symmetric distortions $D_1 = D_2 = D$. To do this, we will consider distortion-cache pairs (D, C) separately for each one of the regions $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3$ and $\mathcal{S}_4$ defined shortly. There are two key quantities defining these regions: The Gaussian joint RD function $\mathsf{R}_{\mathrm{G},X_1X_2}$ and the Wyner common information between X_1 and X_2 (Wyner's common information will be discussed in detail in the next section). For symmetric[4] distortions $D_1 = D_2 = D$, the joint RD function $\mathsf{R}_{\mathrm{G},X_1X_2}$ is given by [17, Thm. III.1] and [18]:

(i) If $0 < D \leq 1-\rho$, then
$$\mathsf{R}_{\mathrm{G},X_1X_2}(D,D) = \frac{1}{2}\log\frac{1-\rho^2}{D^2}.$$

(ii) If $1-\rho \leq D \leq 1$, then
$$\mathsf{R}_{\mathrm{G},X_1X_2}(D,D) = \frac{1}{2}\log\frac{1+\rho}{2D-(1-\rho)}.$$

(iii) If $D > 1$, then
$$\mathsf{R}_{\mathrm{G},X_1X_2}(D,D) = 0.$$

The Wyner common information of the Gaussian pair X_1 and X_2 is given by [9, 16]
$$\mathsf{K}_{\mathrm{W}}(X_1,X_2) = \frac{1}{2}\log\frac{1+\rho}{1-\rho}. \tag{9}$$

Consider the following four regions $\mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3, \mathcal{S}_4$:
$$\mathcal{S}_1 := \Big\{(D,C) : C \geq \mathsf{R}_{\mathrm{G},X_1X_2}(D,D)\Big\},$$
$$\mathcal{S}_2 := \Big\{(D,C) : \mathsf{K}_{\mathrm{W}}(X_1,X_2) \leq C \leq \mathsf{R}_{\mathrm{G},X_1X_2}(D,D)\Big\},$$
$$\mathcal{S}_3 := \Bigg\{(D,C) : D \leq 1-\rho,\ C \leq \mathsf{K}_{\mathrm{W}}(X_1,X_2)\Bigg\},$$
and
$$\mathcal{S}_4 := \Bigg\{(D,C) : 1-\rho \leq D \leq 1,\ C \leq \mathsf{R}_{\mathrm{G},X_1X_2}(D,D)\Bigg\}.$$

These four regions are illustrated in Figure 3.

Proposition 6: For the zero mean bivariate Gaussian source (X_1,X_2) with the covariance matrix $\boldsymbol{K}_{X_1X_2}$ in (8) and squared error distortion constraints, we have
$$\mathsf{R}_{\mathrm{G}}((D,D),C) = \begin{cases} 0, & (C,D) \in \mathcal{S}_1, \\ \dfrac{1}{4}\log\dfrac{1-\rho^2}{D^2} - \dfrac{C}{2}, & (C,D) \in \mathcal{S}_2, \end{cases}$$
and
$$\mathsf{R}_{\mathrm{G}}((D,D),C) \leq \frac{1}{2}\log\frac{1-\frac{1}{2}(1+\rho)(1-2^{-2C})}{D}, \qquad (C,D) \in \mathcal{S}_3 \cup \mathcal{S}_4.$$

Proof: Proposition 6 is proved in Appendix B. ■

Figure 4 illustrates an example of Proposition 6.

D. Doubly Symmetric Binary Source

We now evaluate the RDC function for a *doubly symmetric binary source* (DSBS) under Hamming distortion functions. Fix $0 \leq \rho \leq 1/2$ and let (X_1,X_2) be defined by $\mathcal{X}_1 = \mathcal{X}_2 = \hat{\mathcal{X}}_1 = \hat{\mathcal{X}}_2 = \{0,1\}$ and
$$p_{\boldsymbol{X}}(x_1,x_2) = \frac{1}{2}(1-\rho)\mathbb{1}\{x_1 = x_2\} + \frac{1}{2}\rho\mathbb{1}\{x_1 \neq x_2\}.$$

The Wyner common information of the pair (X_1,X_2) is given by [15]
$$\mathsf{K}_{\mathrm{W}}(X_1,X_2) = 1 + h(\rho) - 2h(\rho^*)$$
where
$$\rho^* = \frac{1}{2} - \frac{1}{2}\sqrt{1-2\rho}.$$

[4]Here we only recall the joint RD function of (X_1,X_2) for the case of symmetric distortions, $D_1 = D_2 = D$. A treatment of the RD function for arbitrary distortion pairs can be found in [17] and the references therein.

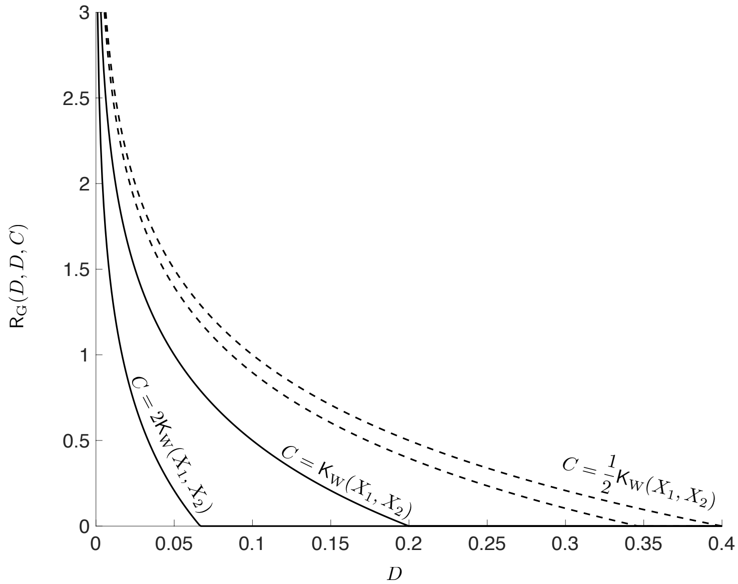

Fig. 4. Illustration of the informational RDC functions in Proposition 6 for a zero mean bivariate Gaussian source (X_1, X_2) with the covariance matrix $\boldsymbol{K}_{X_1 X_2}$ in (8) ($\rho = 0.8$), and symmetric distortion constraints $D_1 = D_2 = D$. The RDC function $\mathsf{R}_{\mathrm{G}}(D, D, C)$ is plotted as a function of the distortion D for three different cache capacities $C = 2\mathsf{K}_{\mathrm{W}}(X_1, X_2)$, $\mathsf{K}_{\mathrm{W}}(X_1, X_2)$ and $(1/2)\mathsf{K}_{\mathrm{W}}(X_1, X_2))$, where $\mathsf{K}_{\mathrm{W}}(X_1, X_2)$ denotes the Wyner common information (9). For $C = (1/2)\mathsf{K}_{\mathrm{W}}(X_1, X_2)$, Proposition 5 and Proposition 6 only give lower and upper bounds, and these are shown with dashed lines.

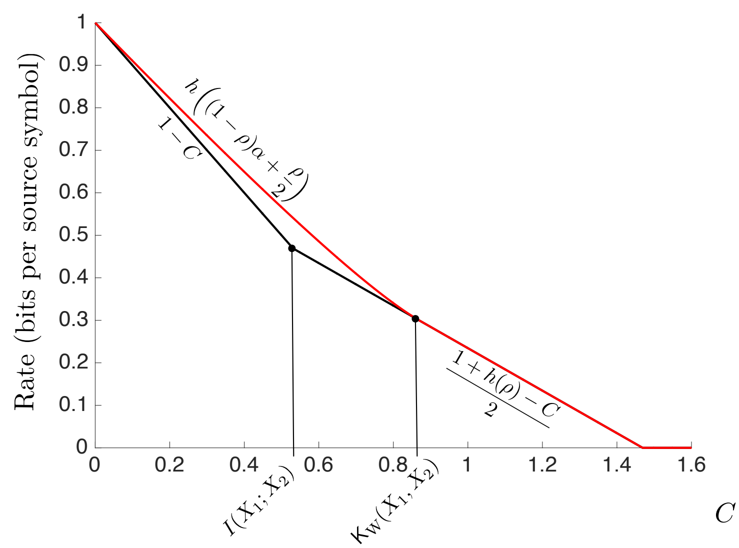

Fig. 5. Illustration of the upper (achievable) and lower (converse) bounds in Proposition 7 for the DSBS RDC function $\mathsf{R}(\mathbf{0}, C)$ with $\rho = 0.1$.

Here, the binary entropy function is denoted and defined by $h(\rho) := -\rho \log_2 \rho - (1-\rho)\log_2(1-\rho)$ for $\rho \in (0, 1)$ and $h(0) = h(1) := 0$. The next proposition can be proved in a similar way to the DSBS examples in [12, Sec. 1.5], [19, Sec. III.C] and [5, Ex. 1], so we omit the proof.

Proposition 7:

(i) If $\mathsf{K}_{\mathrm{W}}(X_1, X_2) \le C \le 1 + h(\rho)$, then $\mathsf{R}(\mathbf{0}, C) = (1 + h(\rho) - C)/2$.

(ii) If $0 < C < 1 + h(\rho)$, then $\mathsf{R}(\mathbf{0}, C) > [1 - C]^+$.

(iii) If $0 < C \le \mathsf{K}_{\mathrm{W}}(X_1, X_2)$, then

$$\frac{1 + h(\rho) - C}{2} \le \mathsf{R}(\mathbf{0}, C) \le h\left((1-\rho)\alpha + \frac{\rho}{2}\right),$$

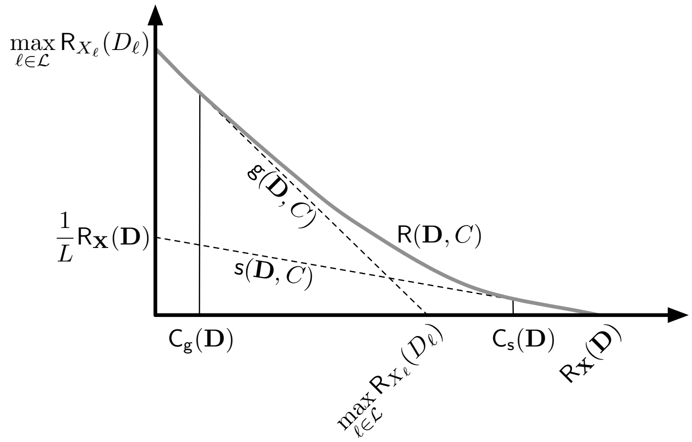

Fig. 6. An illustration of some typical characteristics of the RDC function $\mathsf{R}(\boldsymbol{D}, C)$ for a fixed distortion tuple $\boldsymbol{D} = (D_1, \ldots, D_L)$. The function describes the optimal (minimum) delivery-phase rate (vertical axis) for a given cache capacity (horizontal axis). The bounds (10) and (18) are plotted with dashed lines.

where

$$\alpha := h^{-1}\left(\frac{1-\rho-C}{1-\rho}\right).$$

The above bounds are illustrated in Figure 5.

Remark 3: It is worth noting that, in this special case, the informational RDC function $\mathsf{R}(\boldsymbol{D}, C)$ particularizes to the same expression as in [5, Ex. 1] (see also [12, Sec. 1.5]). This equivalence is a consequence of the DSBS's symmetry and does not hold when the source and/or the distortion constraints are asymmetric.

IV. Common-information measures for caching

In this section we give two new operational definitions of common information for the caching problem. The first definition relates to a "genie-aided" caching system where the encoder knows in advance which source the user will select. The second system relates to a "super-user" caching system.

A. Genie-aided caching

Imagine that, before the caching phase, a genie tells the server which $\ell \in \mathcal{L}$ the user will choose in the future. The optimal caching strategy for this hypothetical *genie-aided* system is obvious: We should compress the ℓ-th source X_ℓ^n using an optimal RD code for that source, cache nC bits of the code's output, and then send the remaining bits during the delivery phase. The RDC function of the genie-aided problem is clearly

$$\mathsf{g}(\boldsymbol{D}, C) = \left[\max_{\ell\in\mathcal{L}}\ \mathsf{R}_{X_\ell}(D_\ell) - C\right]^+.$$

In the main caching problem at hand, however, the server does not know in advance which $\ell \in \mathcal{L}$ the user will select, and this uncertainty may cost additional rate in either the caching or delivery phases. Consequently,

$$\mathsf{R}(\boldsymbol{D}, C) \geq \mathsf{g}(\boldsymbol{D}, C). \tag{10}$$

We have equality in (10) whenever $C = 0$, so it is natural to define the *critical cache capacity*[5]

$$\mathsf{C}_{\mathsf{g}}(\boldsymbol{D}) := \max\Big\{C \geq 0 : \mathsf{R}(\boldsymbol{D}, C) = \mathsf{g}(\boldsymbol{D}, C)\Big\}. \tag{11}$$

We can view $\mathsf{C}_{\mathsf{g}}(\boldsymbol{D})$ as a type of common information for caching: It is the maximum information that can be extracted from every source and placed in the cache without needing redundant information to be transmitted during the delivery phase (w.r.t. the hypothetical genie-aided system). Figure 6 illustrates some typical characteristics of $\mathsf{R}(\boldsymbol{D}, C)$ and $\mathsf{g}(\boldsymbol{D}, C)$.

We now give a single-letter expression for $\mathsf{C}_{\mathsf{g}}(\boldsymbol{D})$. Let $\mathcal{L}^*(\boldsymbol{D}) := \{\ell^* \in \mathcal{L} : \mathsf{R}_{X_{\ell^*}}(D_{\ell^*}) = \max_{\ell\in\mathcal{L}} \mathsf{R}_{X_\ell}(D_\ell)\}$. Define

$$\mathsf{C}_{\mathsf{g}}^*(\boldsymbol{D}) := \max_{U} I(\boldsymbol{X}; U), \tag{12}$$

[5]The maximum indicated in (11) exists because $\mathsf{R}(\boldsymbol{D}, C)$ is convex and $\mathsf{g}(\boldsymbol{D}, C)$ is linear for C in the interval $[0, \mathsf{R}_{\boldsymbol{X}}(\boldsymbol{D})]$.

where the maximization is over the set of all auxiliary random variables U jointly distributed with $\boldsymbol{X}$ such that for all $\ell^* \in \mathcal{L}^*$

$$I(\boldsymbol{X};U) = \mathsf{R}_{X_{\ell^*}}(D_{\ell^*}) - \mathsf{R}_{X_{\ell^*}|U}(D_{\ell^*}) \tag{13a}$$

and

$$\mathsf{R}_{X_{\ell^*}|U}(D_{\ell^*}) = \max_{\ell \in \mathcal{L}} \mathsf{R}_{X_\ell|U}(D_\ell). \tag{13b}$$

Theorem 8: $\mathsf{C}_{\mathrm{g}}(\boldsymbol{D}) = \mathsf{C}^*_{\mathrm{g}}(\boldsymbol{D})$.

Corollary 8.1: For almost lossless Hamming distortions we have $\mathsf{C}_{\mathrm{g}}(\mathbf{0}) = \mathsf{C}^*_{\mathrm{g}}(\mathbf{0}) = \max_U I(\boldsymbol{X};U)$, where the maximization is taken over the set of all U satisfying $U \leftrightarrow X_{\ell^*} \leftrightarrow X_{\mathcal{L}\setminus\ell^*}$ and $H(X_{\ell^*}|U) = \max_{\ell\in\mathcal{L}} H(X_\ell|U)$ for all $\ell^* \in \mathcal{L}^*$.

Proof: Theorem 8 and Corollary 8.1 are proved in Appendices C-A and C-B respectively. ■

B. Gács-Körner common information and the Gray-Wyner network

The Gray-Wyner network in Figure 2 has often been used to provide operational meaning for Gács-Körner common information. Since this network is closely related to our caching problem, it is useful to relate these ideas to $\mathsf{C}_{\mathrm{g}}(\boldsymbol{D})$. To this end, consider the L-receiver Gray-Wyner network with one common rate R_{c} and L private rates $\boldsymbol{R} = (R_1, \ldots, R_L)$ and almost lossless (separable) Hamming distortions. It is not too hard to show that the achievable rate region of this network is equal to the set of all $(L+1)$-rate tuples $(R_{\mathrm{c}}, \boldsymbol{R})$ for which there exists an auxiliary random variable U such that $R_{\mathrm{c}} \geq I(\boldsymbol{X};U)$ and $R_\ell \geq H(X_\ell|U)$ for all $\ell \in \mathcal{L}$.

For any receiver $\ell \in \mathcal{L}$, the smallest sum rate $R_{\mathrm{c}} + R_\ell$ that can be achieved is clearly $H(X_\ell)$. Let us call this smallest sum rate the *cut-set rate* for receiver ℓ. Now consider the maximum common rate R_{c} for which there exists private rates $\boldsymbol{R}$ such that $(R_{\mathrm{c}}, \boldsymbol{R})$ simultaneously meets all L cut-set rates. It is not difficult to show that this maximum common rate is given by

$$\mathsf{K}_{\mathrm{GK}}(\boldsymbol{X}) = \max_{U \leftrightarrow X_\ell \leftrightarrow X_{\mathcal{L}\setminus\ell},\ \forall \ell\in\mathcal{L}} I(\boldsymbol{X};U). \tag{14}$$

For the special case of $(L=2)$-variables, it is well-known that $\mathsf{K}_{\mathrm{GK}}(X_1, X_2)$ simplifies to the *Gács-Körner common information*

$$\mathsf{K}_{\mathrm{GK}}(X_1, X_2) = \max_{H(U|X_1)=0 \text{ and } H(U|X_2)=0} H(U). \tag{15}$$

The next lemma extends (15) to $(L \geq 2)$-variables. To the best of our knowledge, this result has not been shown before.

Proposition 9:

$$\mathsf{K}_{\mathrm{GK}}(\boldsymbol{X}) = \max_{U:\ H(U|X_\ell)=0,\ \forall \ell\in\mathcal{L}} H(U)$$

Proof: Proposition 9 is proved in Appendix D. ■

Thus, the L-variable Gács-Körner common information $\mathsf{K}_{\mathrm{GK}}(\boldsymbol{X})$ can be viewed as the maximum common information that can be extracted from every variable in $\boldsymbol{X}$ and transmitted over the common link, without needing redundant information to be transmitted over the private links.

Viswanatham, Akyol and Rose [9] generalized the above idea (for two receivers) from lossless to lossy reconstructions, and, in doing so, proposed a new lossy version of (14). The next definition is the natural generalization of this lossy common information applied to L variables[6].

Definition 2: We define the *lossy Gács-Körner common information* of $\boldsymbol{X}$ w.r.t. the symbol distortion functions $\mathbf{d}$ by[7]

$$\mathsf{K}_{\mathrm{GK}}(\boldsymbol{X};\boldsymbol{D}) := \max_{(U,\hat{\boldsymbol{X}})} I(\boldsymbol{X};U), \tag{16}$$

where the maximum is taken over all tuples $(U, \hat{\boldsymbol{X}})$ on $\mathcal{U} \times \hat{\boldsymbol{\mathcal{X}}}$ jointly distributed with $\boldsymbol{X}$ and satisfying

(i) $\forall\ \ell \in \mathcal{L}:\ U \leftrightarrow X_\ell \leftrightarrow X_{\mathcal{L}\setminus\ell}$

(ii) $\forall\ \ell \in \mathcal{L}:\ U \leftrightarrow \hat{X}_\ell \leftrightarrow X_\ell$

(iii) $\forall\ \ell \in \mathcal{L}:\ \mathbb{E}[\mathsf{d}_\ell(\hat{X}_\ell, X_\ell)] \leq D_\ell$

(iv) $\forall\ \ell \in \mathcal{L}:\ I(X_\ell;\hat{X}_\ell) = \mathsf{R}_{X_\ell}(D_\ell)$.

The next theorem relates the critical cache capacity to lossy Gács-Körner common information.

Theorem 10: $\mathsf{C}^*_{\mathrm{g}}(\boldsymbol{D}) \geq \mathsf{K}_{\mathrm{GK}}(\boldsymbol{X};\boldsymbol{D})$ with equality whenever $\mathsf{R}_{X_1}(D_1) = \cdots = \mathsf{R}_{X_L}(D_L)$.

Proof: Theorem 10 is proved in Appendix E. ■

[6]Setting $L=2$ gives the original definition in [9].

[7]The indicated maximum in Definition 2 exists because the set of all tuples $(U, \hat{\boldsymbol{X}})$ satisfying (i)–(iv) can be viewed as a compact subset of the corresponding probability simplex.

C. Super-user caching

Now imagine that a *superuser* is connected to the server by L independent rate R noiseless links, and suppose that the superuser requests every source. The optimal caching strategy for this superuser problem is again clear: Take an optimal code for the joint RD function of $\boldsymbol{X}$, cache C bits of the code's output, and distribute the remaining bits equally over the L links in the delivery phase. The RDC function of this superuser problem is

$$\mathsf{s}(\boldsymbol{D}, C) = \left[\frac{\mathsf{R}_{\boldsymbol{X}}(\boldsymbol{D}) - C}{L}\right]^{+}. \tag{17}$$

Since the average of L non-negative numbers cannot be larger than the maximum, we have

$$\mathsf{R}(\boldsymbol{D}, C) \geq \mathsf{s}(\boldsymbol{D}, C). \tag{18}$$

Clearly the superuser bound (18) is achievable by the caching system at $C = \mathsf{R}_{\boldsymbol{X}}(\boldsymbol{D})$. It is natural to consider the smallest cache capacity for which there is *no rate loss* with respect to the optimal superuser system[8]

$$\mathsf{C}_{\mathsf{s}}(\boldsymbol{D}) := \min\left\{C \geq 0 : \mathsf{R}(\boldsymbol{D}, C) = \mathsf{s}(\boldsymbol{D}, C)\right\}. \tag{19}$$

We now give a single-letter expression for $\mathsf{C}_{\mathsf{s}}(\boldsymbol{D})$. For a given $\boldsymbol{D}$, let

$$\mathsf{C}^*_{\mathsf{s}}(\boldsymbol{D}) := \min_{(U,\hat{\boldsymbol{X}})} I(\boldsymbol{X}; U)$$

where the minimum is taken over all tuples $(U, \hat{\boldsymbol{X}})$ on $\mathcal{U} \times \hat{\boldsymbol{\mathcal{X}}}$ such that the following five properties hold

(i) $\boldsymbol{X} \leftrightarrow \hat{\boldsymbol{X}} \leftrightarrow U$
(ii) $I(X_1; \hat{X}_1|U) = \cdots = I(X_L; \hat{X}_L|U)$
(iii) $\forall\ \ell \in \mathcal{L}:\ \hat{X}_\ell \leftrightarrow U \leftrightarrow \hat{X}_{\mathcal{L}\setminus\ell}$
(iv) $\forall\ \ell \in \mathcal{L}:\ \mathbb{E}[\mathsf{d}_\ell(\hat{X}_\ell, X_\ell)] \leq D_\ell$
(v) $I(\boldsymbol{X}; \hat{\boldsymbol{X}}) = \mathsf{R}_{\boldsymbol{X}}(\boldsymbol{D})$.

Theorem 11: $\mathsf{C}_{\mathsf{s}}(\boldsymbol{D}) = \mathsf{C}^*_{\mathsf{s}}(\boldsymbol{D})$.

Proof: Theorem 11 is proved in Appendix F. ■

D. Wyner common information and the Gray-Wyner Network

The Gray-Wyner network in Figure 2 with almost lossless (separable) Hamming distortions is also often used to provide an operation meaning for Wyner's common information [15]

$$\mathsf{K}_{\mathrm{W}}(X_1, X_2) := \min_{U: X_1 \leftrightarrow U \leftrightarrow X_2} I(X_1, X_2; U). \tag{20}$$

Specifically, $\mathsf{K}_{\mathrm{W}}(X_1, X_2)$ is equal to the minimum common rate R_{c} for which it is possible to achieve the so called Pangloss plane $R_{\mathrm{c}} + R_1 + R_2 = H(X_1, X_2)$. The natural extension of Wyner's common information to L variables $\boldsymbol{X}$ is

$$\mathsf{K}_{\mathrm{W}}(\boldsymbol{X}) := \min_{U:\ X_\ell\ \leftrightarrow\ U\ \leftrightarrow\ X_{\mathcal{L}\setminus\ell},\ \forall \ell\in\mathcal{L}} I(\boldsymbol{X}; U).$$

Viswanatha, Akyol and Rose's [9] generalized the above idea from lossless to reconstructions, and, in doing so, proposed the following *lossy Wyner common information*.

Definition 3: For a given distortion tuple $\boldsymbol{D}$ and single-symbol distortion functions **d**, the *lossy Wyner common information* of $\boldsymbol{X}$ is given by

$$\mathsf{K}_{\mathrm{W}}(\boldsymbol{X}; \boldsymbol{D}) := \min_{(U,\hat{\boldsymbol{X}})} I(\boldsymbol{X}; U)$$

where the minimum is taken over all tuples $(U, \hat{\boldsymbol{X}})$ on $\mathcal{U} \times \hat{\boldsymbol{\mathcal{X}}}$ such that the following four properties hold

(i) $\boldsymbol{X} \leftrightarrow \hat{\boldsymbol{X}} \leftrightarrow U$
(ii) $\forall\ \ell \in \mathcal{L}:\ \hat{X}_\ell \leftrightarrow U \leftrightarrow \hat{X}_{\mathcal{L}\setminus\ell}$
(iii) $\forall\ \ell \in \mathcal{L}:\ \mathbb{E}[\mathsf{d}_\ell(\hat{X}_\ell, X_\ell)] \leq D_\ell$
(iv) $I(\boldsymbol{X}; \hat{\boldsymbol{X}}) = \mathsf{R}_{\boldsymbol{X}}(\boldsymbol{D})$.

The next proposition and corollary relate Wyner common information measures to the caching problem, and they trivially follow from the above definitions.

Proposition 12: $\mathsf{C}^*_{\mathsf{s}}(\boldsymbol{D}) \geq \mathsf{K}_{\mathrm{W}}(\boldsymbol{X}; \boldsymbol{D})$.

[8] The minimum in (19) exists because $\mathsf{R}_{\boldsymbol{X}}(\boldsymbol{D})$ is convex and $\mathsf{s}(\boldsymbol{D}, C)$ is linear for C in the interval for $[0, \mathsf{R}_{\boldsymbol{X}}(\boldsymbol{D})]$. Figure 6 depicts the superuser bound and the critical cache capacity $\mathsf{C}_{\mathsf{s}}(\boldsymbol{D})$.

Corollary 12.1: $\mathsf{C}_\mathsf{s}^*(\mathbf{0}) \geq \mathsf{K}_\mathrm{W}(\boldsymbol{X})$, with equality whenever the caching problem is symmetric in the sense that $\mathsf{K}_\mathrm{W}(\boldsymbol{X}) = I(\boldsymbol{X}; U^*)$ for some U^* satisfying $H(X_1|U^*) = \cdots = H(X_L|U^*)$ and $X_\ell \leftrightarrow U^* \leftrightarrow U^*_{\mathcal{L}\setminus\ell}$ for all $\ell \in \mathcal{L}$.

Remark 4: The lossy Wyner common information $\mathsf{K}_\mathrm{W}(\boldsymbol{X}; \boldsymbol{D})$ as well as Wyner's original common information $\mathsf{K}_\mathrm{W}(\boldsymbol{X})$ are both defined for discrete and continuous random vectors $\boldsymbol{X}$. In the latter case, the lossy Wyner common information is only defined when the RD function in (iv) is finite, $\mathsf{R}_{\boldsymbol{X}}(\boldsymbol{D}) < \infty$. It is also is worth noting that, in general, the lossy Wyner common information $\mathsf{K}_\mathrm{W}(\boldsymbol{X}; \boldsymbol{D})$ is neither convex/concave nor monotonic in $\boldsymbol{D}$. Moreover, it is generally the case that $\mathsf{K}_\mathrm{W}(\boldsymbol{X}; \boldsymbol{D})$ can be larger/smaller than the Wyner common information $\mathsf{K}_\mathrm{W}(\boldsymbol{X})$. A nice treatment of this issue for $L = 2$ variables is given by Viswanatha *et al.* in [9, Sec. III.B].

V. Caching w.r.t. excess (separable) distortions

In this section we reconsider the caching problem formulation from Section II with the expected distortion constraints replaced by an excess distortion constraints. We will show that under this more restrictive criteria, a strong converse holds.

Definition 4: We say that a rate-distortion-cache tuple $(R, \boldsymbol{D}, C)$ is **d**-*achievable w.r.t. excess (separable) distortions* if there exists a sequence of $(n, |\mathcal{M}^{(n)}|, |\mathcal{M}_\mathsf{c}^{(n)}|)$-codes such that (3a) and (3b) hold and

$$\lim_{n\to\infty} \mathbb{P}\left[\bigcup_{\ell\in\mathcal{L}} \left\{\bar{\mathsf{d}}_\ell(\hat{X}_\ell^n, X_\ell^n) \geq D_\ell\right\}\right] = 0. \tag{21}$$

The *RDC function w.r.t. excess (separable) distortions* $\mathsf{R}^\ddagger(\boldsymbol{D}, C)$ is the infimum of all rates $R \geq 0$ such that the $(R, \boldsymbol{D}, C)$ is **d**-achievable.

It is not too hard to show that the RDC functions of the excess and expected distortion problems coincide (assuming that the symbol distortion functions **d** are bounded). We omit the proof.

Lemma 13: $\mathsf{R}^\dagger(\boldsymbol{D}, C) = \mathsf{R}^\ddagger(\boldsymbol{D}, C) = \mathsf{R}(\boldsymbol{D}, C)$.

Lemma 13 provides us only with the following *weak converse*: If the delivery-phase rate R is strictly smaller than the informational RDC function $\mathsf{R}(\boldsymbol{D}, C)$, then the excess-distortion probability of any sequence of $(n, |\mathcal{M}_\mathsf{c}^{(n)}|, |\mathcal{M}^{(n)}|)$ codes satisfying (3a) and (3b) will be bounded away from zero; that is,

$$\limsup_{n\to\infty} \mathbb{P}\left[\bigcup_{\ell\in\mathcal{L}} \left\{\bar{\mathsf{d}}_\ell(\hat{X}_\ell^n, X_\ell^n) \geq D_\ell\right\}\right] > 0.$$

The next theorem strengthens this weak converse to a strong converse.

Theorem 14: Fix any cache capacity C and distortion tuple $\boldsymbol{D}$ such that $\mathsf{R}(\boldsymbol{D}, C) > 0$. Any sequence of $(n, |\mathcal{M}_\mathsf{c}^{(n)}|, |\mathcal{M}^{(n)}|)$-codes satisfying

$$\limsup_{n\to\infty} \frac{1}{n} \log |\mathcal{M}^{(n)}| < \mathsf{R}(\boldsymbol{D}, C) \tag{22}$$

and

$$\limsup_{n\to\infty} \frac{1}{n} \log |\mathcal{M}_\mathsf{c}^{(n)}| \leq C \tag{23}$$

must also satisfy

$$\limsup_{n\to\infty} \mathbb{P}\left[\bigcup_{\ell\in\mathcal{L}} \left\{\bar{\mathsf{d}}_\ell(\hat{X}_\ell^n, X_\ell^n) \geq D_\ell\right\}\right] = 1. \tag{24}$$

Proof: Theorem 14 is proved in Appendix G. ■

Remark 5: The strong converse in Theorem 14 applies to the probability of the *union* of excess-distortion events in (24). One might wonder if a similar strong converse can be proved for the *maximum probability of excess distortion* scenario in which union probability in (24) is replaced by $\max_{\ell\in\mathcal{L}} \mathbb{P}[\bar{\mathsf{d}}_\ell(\hat{X}_\ell^n, X_\ell^n) \geq D_\ell]$. If the cache capacity is smaller than the critical cache capacity $C \leq \mathsf{C}_\mathsf{g}(\boldsymbol{D})$, then one can easily show a new converse in which (24) is replaced by

$$\limsup_{n\to\infty} \max_{\ell\in\mathcal{L}} \mathbb{P}\left[\bar{\mathsf{d}}_\ell(\hat{X}_\ell^n, X_\ell^n) \geq D_\ell\right] = 1. \tag{25}$$

This result essentially just employs the strong converse for the standard point-to-point RD problem with separable distortion functions. For larger values of C it is unclear, at least to us, whether (25) still holds.

VI. Caching w.r.t f-separable distortion functions

We now consider the caching problem w.r.t. f-separable distortion functions and both expected and excess distortions. The corresponding RDC functions are defined in exactly the same way as in Definitions 1 and 4, except that the f-separable distortion function $\overline{\mathsf{fd}}_\ell$ replaces the separable distortion function $\bar{\mathsf{d}}_\ell$. We denote the corresponding RDC function under expected and excess distortions by $\mathsf{R}^{\dagger}_{\mathbf{f}}(\boldsymbol{D}, C)$ and $\mathsf{R}^{\ddagger}_{\mathbf{f}}(\boldsymbol{D}, C)$ respectively.

For each request $\ell \in \mathcal{L}$ let $\mathsf{d}^*_\ell : \hat{\mathcal{X}}_\ell \times \mathcal{X}_\ell \to [0, \infty)$ be the single-symbol distortion function obtained by setting

$$\mathsf{d}^*_\ell(\hat{x}_\ell, x_\ell) = \mathsf{f}_\ell\big(\mathsf{d}_\ell(\hat{x}_\ell, x_\ell)\big). \tag{26}$$

Now let $\mathsf{R}_{\mathbf{d}^*}(\mathbf{f}(\boldsymbol{D}), C)$ denote the informational RDC function in (4) evaluated w.r.t. the single-symbol distortion functions $\mathbf{d}^* = (\mathsf{d}^*_1, \ldots, \mathsf{d}^*_L)$ and distortion tuple $\mathbf{f}(\boldsymbol{D}) = \big(\mathsf{f}_1(D_1), \ldots, \mathsf{f}_L(D_L)\big)$. Modifying the strong converse for the usual point-to-point RD problem (see, for example, Kieffer [13]), and using ideas in [1], it is not too difficult to obtain the following proposition. We omit the proof.

Proposition 15: For f-separable distortion functions and all cache capacities $C \leq \mathsf{C}_{\mathsf{g}}(\boldsymbol{D})$, we have

$$\mathsf{R}^{\dagger}_{\mathbf{f}}(\boldsymbol{D}, C) = \mathsf{R}^{\ddagger}_{\mathbf{f}}(\boldsymbol{D}, C) = \mathsf{R}_{\mathbf{d}^*}(\mathbf{f}(\boldsymbol{D}), C).$$

Proposition 15 is quite intuitive, and a natural question is whether or not it extends to cache capacities larger than $\mathsf{C}_{\mathsf{g}}(\boldsymbol{D})$. The next result considers such cases, but it requires a slightly more restricted version of the *expected distortions* operational model. Specifically, let us consider the following definition:

Definition 5: We say that a rate-distortion-cache tuple $(R,\, \boldsymbol{D}, C)$ is *achievable w.r.t. the expected max-distortion criterium* if there exists a sequence of $(n, \mathcal{M}^{(n)}_{\mathsf{c}}, \mathcal{M}^{(n)})$-codes such that (3a) and (3b) hold and

$$\limsup_{n\to\infty} \mathbb{E}\Big[\max_{\ell\in\mathcal{L}} \big(\overline{\mathsf{fd}}_\ell(\hat{X}^n_\ell, X^n_\ell) - D_\ell\big)\Big] \leq 0. \tag{27}$$

The *RDC function w.r.t. expected max-distortions criterion* is

$$\tilde{\mathsf{R}}^{\dagger}_{\mathbf{f},\text{max-exc}}(\boldsymbol{D}, C) := \inf\Big\{R \geq 0 : (R, \boldsymbol{D}, C) \text{ is achievable w.r.t. expected max-distortions}\Big\}$$

Theorem 16: $\tilde{\mathsf{R}}^{\dagger}_{\mathbf{f},\text{max-exc}}(\boldsymbol{D}, C) = \mathsf{R}^{\ddagger}_{\mathbf{f}}(\boldsymbol{D}, C) = \mathsf{R}_{\mathbf{d}^*}(\mathbf{f}(\boldsymbol{D}), C)$.

Proof: Theorem 16 is proved in Appendix I. ■

VII. Conclusion

We studied cache-aided systems with correlated source files and characterized the tradeoff between delivery rate, cache memory, and reconstruction distortion. This trade-off is formalized in terms of an auxiliary random variable, and, therefore, its computation is non-trivial. Moreover, it does not provide an explicit answer to what type of "common information" among the sources should be cached. We investigated two new notions of common information and their operational meaning for the caching problem and showed that it is optimal to cache these common informations in some regimes. Our approach is motivated by the operational meaning of Wyner's common information and Gács-Körner common information on the Gray-Wyner network. Under some very special symmetry conditions, our new definitions coincide with the previous ones. In general, however, the definitions are different.

We also extended our results to excess-distortion criteria and f-separable distortion measures introduced in [1]. A key component of this extension is a new strong converse for a union (over all sources) excess separable distortions criteria. The new strong converse is needed because, in general, it is possible to non-trivially trade distortions between the sources by modifying what information is placed in the cache.

Our approach can also be generalized to cache-aided multi-user settings (see, e.g., [40, Section IX], [41]). In general, however, finding exact tradeoffs is challenging and it is interesting to seek approximate solutions.

Future interesting directions on the problem include addressing practical requirements such as latency, security/privacy, and complexity of code design.

Appendix A
Proof of Proposition 5

Fix $\boldsymbol{D} = (D, D, \ldots, D)$ for some $D \geq 0$, and consider any tuple $(U, \hat{\boldsymbol{X}})$ satsifying (7). Fix $\mathcal{S} \subseteq \mathcal{L}$ and let $S := |\mathcal{S}|$. Then

$$\begin{aligned}
&\max_{\ell\in\mathcal{L}} I(X_\ell; \hat{X}_\ell|U) \\
&\quad \geq \max_{\ell\in\mathcal{S}} \Big[h(X_\ell|U) - h(X_\ell|\hat{X}_\ell)\Big] \\
&\quad \overset{\text{a}}{\geq} \frac{1}{S} h(X_{\mathcal{S}}|U) - \frac{1}{2}\log(2\pi e D)
\end{aligned}$$

$$\stackrel{\text{b}}{\geq} \frac{1}{S}\left(\frac{1}{2}\log\left((2\pi e)^S \det \boldsymbol{K}_{X_{\mathcal{S}}}\right) - C\right) - \frac{1}{2}\log(2\pi e D)$$
$$= \frac{1}{2S}\log\frac{\det \boldsymbol{K}_{X_{\mathcal{S}}}}{D^S} - \frac{C}{S}.$$

Step (a) follows because

$$\begin{aligned} h(X_\ell|\hat{X}_\ell) &\stackrel{\text{a.1}}{=} h(X_\ell - \hat{X}_\ell|\hat{X}_\ell) \\ &\stackrel{\text{a.2}}{\leq} h\big(\mathcal{N}(0, \mathbb{E}(\hat{X}_\ell - X_\ell)^2)\big) \\ &\stackrel{\text{a.3}}{\leq} h\big(\mathcal{N}(0, D)\big) \\ &\stackrel{\text{a.4}}{\leq} \frac{1}{2}\log(2\pi e D), \end{aligned}$$

where (a.1) follows by the *translation property of differential entropy* [27, Thm. 10.18]; (a.2) uses the fact that the normal distribution maximizes differential entropy for a given second moment [27, Thm. 10.43], and (a.3) invokes the distortion constraint in (7). Moreover, for the first term, we have

$$\max_{\ell\in\mathcal{S}} h(X_\ell|U) \stackrel{\text{a.5}}{\geq} \frac{1}{S}\sum_{\ell\in\mathcal{S}} h(X_\ell|U) \stackrel{\text{a.6}}{\geq} \frac{1}{S}\, h(X_{\mathcal{S}}|U),$$

where (a.5) follows because the maximum cannot be smaller than the average, and (a.6) follows by the *independence bound for differential entropy* [27, Thm. 10.34]

Step (b) follows from the cache capacity constraint in (7)

$$\begin{aligned} C &\geq I(\boldsymbol{X};U) \geq I(X_{\mathcal{S}};U) \\ &= h(X_{\mathcal{S}}) - h(X_{\mathcal{S}}|U) \\ &= \frac{1}{2}\log\left((2\pi e)^S \det \boldsymbol{K}_{X_{\mathcal{S}}}\right) - h(X_{\mathcal{S}}|U) \end{aligned}$$

■

APPENDIX B
PROOF OF PROPOSITION 6

A. *Case 1:* $(D, C) \in \mathcal{S}_1$

If $(D, C) \in \mathcal{S}_1$, then it trivially follows from the definition of $\mathsf{R}_{\mathrm{G},X_1X_2}(D, D)$ that $\mathsf{R}_{\mathrm{G}}(D, D, C) = 0$.

B. *Case 2:* $(D, C) \in \mathcal{S}_2$

Since $\mathsf{R}_{\mathrm{G},X_1X_2}(D, D)$ is strictly decreasing in D, it follows that for a given $C \leq \mathsf{R}_{\mathrm{G},X_1X_2}(D, D)$ the distortion D must satisfy

$$0 < D \leq 2^{-C}\sqrt{1-\rho^2}.$$

Define

$$\alpha = 1 - \rho - 2^{-C}\sqrt{1-\rho^2} \tag{28}$$

and note that $0 \leq \alpha < 1 - \rho$ for all finite

$$C > \frac{1}{2}\log\frac{1+\rho}{1-\rho}.$$

Now let $W, N_1, N_2, \tilde{N}_1, \tilde{N}_2, Z_1$ and Z_2 be mutually independent standard Gaussians $\mathcal{N}(0, 1)$, and notice that our bivariate Gaussian source (X_1, X_2) can be written as

$$X_i = \sqrt{\rho}\ W + \sqrt{\alpha}\ N_i + \sqrt{1-\rho-D-\alpha}\ \tilde{N}_i + \sqrt{D}\ Z_i, \quad i = 1, 2$$

Choose $U = (U_1, U_2)$, where

$$U_i = \sqrt{\rho}\ W + \sqrt{\alpha}\ N_i, \quad i = 1, 2.$$

Define the reconstructions $\hat{X}_1$ and $\hat{X}_2$ to be

$$\hat{X}_i := U_i + \sqrt{1-\rho-\alpha-D}\ \tilde{N}_i, \quad i = 1, 2.$$

We notice that

$$X_1 \leftrightarrow \hat{X}_1 \leftrightarrow U_1 \leftrightarrow U \leftrightarrow U_2 \leftrightarrow \hat{X}_2 \leftrightarrow X_2 \tag{29}$$

forms a Markov chain. Additionally,

$$
\begin{aligned}
&I(X_1, X_2; U) \\
&= h(X_1, X_2) - h(X_1, X_2|U) \\
&\stackrel{\text{a}}{=} h(X_1, X_2) - h(X_1|U) - h(X_2|U) \\
&\stackrel{\text{b}}{=} h(X_1, X_2) - h(X_1|U_1) - h(X_2|U_2) \\
&= h(X_1, X_2) - 2h(X_1|U_1) \\
&\stackrel{\text{c}}{=} h(X_1, X_2) - 2h(X_1 - U_1|U_1) \\
&= \frac{1}{2}\log\left((2\pi\mathsf{e})^2(1-\rho^2)\right) - \log\left(2\pi\mathsf{e}(1-\rho-\alpha)\right) \\
&= \frac{1}{2}\log\frac{1-\rho^2}{(1-\rho-\alpha)^2} \\
&\stackrel{\text{d}}{=} C,
\end{aligned}
$$

where (a) and (b) follow from (29), (c) follows by symmetry, and (d) substitutes (28). Similarly,

$$
\begin{aligned}
I(X_1; \hat{X}_1|U) &= h(X_1|U) - h(X_1|\hat{X}_1, U) \\
&\stackrel{\text{a}}{=} h(X_1|U_1) - h(X_1|\hat{X}_1) \\
&= h(X_1 - U_1|U_1) - h(X_1 - \hat{X}_1|\hat{X}_1) \\
&= \frac{1}{2}\log\Big(2\pi\mathsf{e}(1-\rho-\alpha)\Big) - \frac{1}{2}\log\Big(2\pi\mathsf{e}D\Big) \\
&= \frac{1}{2}\log\left(\frac{1-\rho-\alpha}{D}\right) \\
&= \frac{1}{4}\log\left(\frac{1-\rho^2}{D^2}\right) - \frac{C}{2},
\end{aligned}
$$

where (a) uses the Markov chain (29) and (b) substitutes (28). Finally, we notice that the above achievable rate is equal to the superuser lower bound from Proposition 5.

C. Case 3: $(D, C) \in \mathcal{S}_3$

Let

$$\alpha = \frac{1}{2}(1+\rho)(1-2^{-2C}),$$

and note that $0 \le \alpha \le \rho$. Now let $W, \tilde{W}, Z_1, Z_2, N_1$ and N_2 be mutually independent standard Gaussians $\mathcal{N}(0,1)$. Choose

$$U = \sqrt{\alpha}\; W + \sqrt{\rho-\alpha}\; \tilde{W}$$

and

$$\hat{X}_i = \sqrt{\rho}\; W + \sqrt{1-\rho-D}\; Z_i, \quad i = 1, 2.$$

We may now write our bivariate Gaussian source (X_1, X_2) as

$$X_i = \hat{X}_i + \sqrt{D}\; N_i, \quad i = 1, 2.$$

The pair (X_1, U) and the pair (X_2, U) are both zero mean bivariate Gaussians with identical covariance matrices

$$\boldsymbol{K}_{X_1,U} = \boldsymbol{K}_{X_2,U} = \begin{bmatrix} 1 & \sqrt{\alpha\rho} \\ \sqrt{\alpha\rho} & \rho \end{bmatrix}.$$

Similarly, (X_1, X_2, U) is a zero mean multivariate normal with the covariance matrix

$$\boldsymbol{K}_{X_1X_2U} = \begin{bmatrix} 1 & \rho & \sqrt{\alpha\rho} \\ \rho & 1 & \sqrt{\alpha\rho} \\ \sqrt{\alpha\rho} & \sqrt{\alpha\rho} & \rho \end{bmatrix}.$$

Thus,

$$
\begin{aligned}
I(X_1, X_2; U) &= h(X_1, X_2) + h(U) - h(X_1, X_2, U) \\
&= \frac{1}{2}\log\left((2\pi e)^2 \det \boldsymbol{K}_{X_1X_2}\right) + \frac{1}{2}\log\left(2\pi e\rho\right)
\end{aligned}
$$

$$
\begin{aligned}
&\quad - \frac{1}{2}\log\left((2\pi e)^3 \det \boldsymbol{K}_{X_1X_2U}\right)\\
&= \frac{1}{2}\log\frac{1+\rho}{1+\rho-2\alpha}\\
&= C,
\end{aligned}
$$

and

$$
\begin{aligned}
I(X_1;\hat{X}_1|U) &= h(X_1|U) - h(X_1|\hat{X}_1, U)\\
&= h(X_1,U) - h(U) - h(X_1|\hat{X}_1)\\
&= \frac{1}{2}\log\left((2\pi e)^2 \det \boldsymbol{K}_{X_1U}\right) - \frac{1}{2}\log(2\pi e\rho)\\
&\qquad - \frac{1}{2}\log(2\pi e D)\\
&= \frac{1}{2}\log\frac{1-\alpha}{D}.
\end{aligned}
$$

D. Case 4: $(D,C) \in \mathcal{S}_4$

Suppose that $(D,C) \in \mathcal{S}_4$. Since (D,C) lies below the Gaussian joint RD function $\mathsf{R}_{\mathrm{G},X_1X_2}(D,D)$, it follows that for any given distortion $D \in [1-\rho, 1]$ the cache capacity C must lie within

$$
0 \le C \le \frac{1}{2}\log\frac{1+\rho}{2D-1+\rho}.
$$

Define

$$
\alpha = \frac{1}{2}(1+\rho)(1-2^{-2C})
$$

and

$$
\beta = 1-\alpha-D,
$$

where we notice that

$$
0 \le \alpha, \beta \le 1-D \quad \text{and} \quad \alpha+\beta = 1-D \le \rho.
$$

In this case, we may write

$$
X_i = \sqrt{\alpha}\ A + \sqrt{\beta}\ B + \sqrt{\rho-(\alpha+\beta)}\ W + \sqrt{1-\rho}\ N_i, \quad i = 1,2,
$$

where A, B, W, N_1 and N_2 are mutually independent standard Gaussians $\mathcal{N}(0,1)$. Now let

$$
U = \sqrt{\alpha}\ A,
$$

and

$$
\hat{X}_1 = \hat{X}_2 = \hat{X} := U + \sqrt{\beta}\ B.
$$

Here (X_1, X_2, U) is a zero mean multivariate Gaussian with covariance matrix

$$
\boldsymbol{K}_{X_1,X_2,U} = \begin{bmatrix} 1 & \rho & \alpha \\ \rho & 1 & \alpha \\ \alpha & \alpha & \alpha. \end{bmatrix}
$$

Then,

$$
\begin{aligned}
I(X_1,X_2;U) &= h(X_1,X_2) - h(X_1,X_2,U) - h(U)\\
&= \frac{1}{2}\log\left((2\pi \mathrm{e})^2 \det \boldsymbol{K}_{X_1,X_2}\right)\\
&\qquad - \frac{1}{2}\log\left((2\pi \mathrm{e})^3 \det \boldsymbol{K}_{X_1,X_2,U}\right)\\
&\qquad\qquad + \frac{1}{2}\log\left(2\pi \mathrm{e}\rho\right)\\
&= \frac{1}{2}\log\frac{1+\rho}{1+\rho-2\alpha}\\
&= C.
\end{aligned}
$$

Moreover,

$$
I(X_1;\hat{X}|U) = h(X_1|U) - h(X_1|U,\hat{X})
$$

$$
\begin{aligned}
&= h(X_1|U) - h(X_1|\hat{X}) \\
&= \frac{1}{2}\log\big(2\pi\mathsf{e}(1-\alpha)\big) - \frac{1}{2}\log\big(2\pi\mathsf{e}(1-\alpha-\beta)\big) \\
&= \frac{1}{2}\log\frac{1-\alpha}{D}.
\end{aligned}
$$

■

APPENDIX C
PROOF OF THEOREM 8

A. *Proof of Theorem 8*

Choose the cache capacity to be $C = \mathsf{C}_\mathsf{g}(\boldsymbol{D})$ and assume that $\mathsf{R}(\boldsymbol{D}, \mathsf{C}_\mathsf{g}(\boldsymbol{D})) > 0$. By the definition of $\mathsf{C}_\mathsf{g}(\boldsymbol{D})$:

$$
\mathsf{R}(\boldsymbol{D}, C) = \max_{\ell\in\mathcal{L}} \mathsf{R}_{X_\ell}(D_\ell) - C. \tag{30}
$$

Let U be an optimal auxiliary random variable for the informational RDC function $\mathsf{R}(\boldsymbol{D}, C)$, i.e., U is so that

$$
\mathsf{R}(\boldsymbol{D}, C) = \max_{\ell\in\mathcal{L}} \mathsf{R}_{X_\ell|U}(D_\ell) \tag{31}
$$

and

$$
I(\boldsymbol{X}; U) \le C. \tag{32}
$$

Let $\ell^* \in \mathcal{L}^*$, i.e., ℓ^* attains the maximum in (30). We have the following:

$$
\begin{aligned}
\mathsf{R}(\boldsymbol{D}, C) &\overset{\text{a}}{=} \max_{\ell\in\mathcal{L}} \mathsf{R}_{X_\ell|U}(D_\ell) \\
&\ge \mathsf{R}_{X_{\ell^*}|U}(D_{\ell^*}) \\
&= \min_{q_{\hat{X}_{\ell^*}|X_{\ell^*},U}\,:\ \mathbb{E}[d(\hat{X}_{\ell^*}, X_{\ell^*})]\le D_\ell} I(X_{\ell^*}; \hat{X}_{\ell^*}|U) \\
&\overset{\text{b}}{\ge} \min_{q_{\hat{X}_{\ell^*}|X_{\ell^*},U}\,:\ \mathbb{E}[d(\hat{X}_{\ell^*}, X_{\ell^*}]\le D_\ell} I(X_{\ell^*}; U, \hat{X}_{\ell^*}) - I(\boldsymbol{X}; U) \\
&\overset{\text{c}}{\ge} \mathsf{R}_{X_{\ell^*}}(D_{\ell^*}) - I(\boldsymbol{X}; U) \\
&\overset{\text{d}}{\ge} \mathsf{R}_{X_{\ell^*}}(D_{\ell^*}) - C \\
&\overset{\text{e}}{=} \mathsf{R}(\boldsymbol{D}, C),
\end{aligned}
$$

where (a) is identical to (31); (b) follows by adding the negative term $I(X_{\ell^*}; U) - I(\boldsymbol{X}; U)$; (c) holds because $I(X_{\ell^*}; U, \hat{X}_{\ell^*}) \ge I(X_{\ell^*}; \hat{X}_{\ell^*})$; (d) holds by (32); and (e) holds by (30) and because $\ell^* \in \mathcal{L}^*$.

The above inequalities must all hold with equality and so the chosen U must satisfy $I(\boldsymbol{X}; U) = C = \mathsf{C}_\mathsf{g}(\boldsymbol{D})$, (13a) and (13b). Therefore,

$$
\mathsf{C}_\mathsf{g}(\boldsymbol{D}) \le \mathsf{C}^*_\mathsf{g}(\boldsymbol{D}). \tag{33}
$$

Choose now the cache capacity $C = \mathsf{C}^*_\mathsf{g}(\boldsymbol{D})$, and let U be an optimal auxiliary random variable for $\mathsf{C}^*_\mathsf{g}(\boldsymbol{D})$. That means, U satisfies (13a) and (13b) and

$$
I(\boldsymbol{X}; U) = \mathsf{C}^*_\mathsf{g}(\boldsymbol{D}) = C. \tag{34}
$$

The following holds for all $\ell^* \in \mathcal{L}^*$:

$$
\begin{aligned}
\mathsf{R}(\boldsymbol{D}, C) &\overset{\text{a}}{\le} \max_{\ell\in\mathcal{L}} \mathsf{R}_{X_\ell|U}(D_\ell) \\
&\overset{\text{b}}{=} \mathsf{R}_{X_{\ell^*}|U}(D_{\ell^*}) \\
&\overset{\text{c}}{=} \mathsf{R}_{X_{\ell^*}}(D_{\ell^*}) - I(\boldsymbol{X}; U) \\
&\overset{\text{d}}{=} \mathsf{R}_{X_{\ell^*}}(D_{\ell^*}) - C,
\end{aligned}
$$

where (a) follows because U need not be optimal for $\mathsf{R}(\boldsymbol{D}, C)$, (b) follows from (13b), (c) follows from (13a), and (d) from (34).

Therefore, at the cache capacity $C = I(\boldsymbol{X}; U) = \mathsf{C}_{\mathsf{g}}^*(\boldsymbol{D})$ we have $\mathsf{R}(D, C) = \mathsf{R}_{X_{\ell^*}}(D_{\ell^*}) - C$ and consequently

$$\mathsf{C}_{\mathsf{g}}(\boldsymbol{D}) \geq \mathsf{C}_{\mathsf{g}}^*(\boldsymbol{D}). \tag{35}$$

The theorem follows from (33) and (35). ■

B. Proof of Corollary 8.1

The conditional RD function particularises to the conditional entropy function: $\mathsf{R}_{X_\ell|U}(0) = H(X_\ell|U)$. Similarly, the constraint (13a) particularises to

$$I(\boldsymbol{X}; U) = H(X_{\ell^*}) - H(X_{\ell^*}|U) = I(X_{\ell^*}; U),$$

which is equivalent to $U \leftrightarrow X_{\ell^*} \leftrightarrow X_{\mathcal{L}\setminus\ell^*}$. ■

Appendix D
Proof of Proposition 9

We have

$$\max_{U:\ H(U|X_\ell)=0,\ \forall \ell\in\mathcal{L}} H(U) \leq \max_{U:\ U\ \leftrightarrow\ X_\ell\ \leftrightarrow\ X_{\mathcal{L}\setminus\ell},\ \forall\ell\in\mathcal{L}} I(\boldsymbol{X}; U)$$

since any U satisfying $H(U|X_\ell) = 0$ for all $\ell \in \mathcal{L}$ must also satisfy $U \leftrightarrow X_\ell \leftrightarrow X_{\mathcal{L}\setminus\ell}$ for all $\ell \in \mathcal{L}$. The reverse inequality follows by the next lemma, which is a multivariate extension of [24, Lem. A.1]. ■

Lemma 17: If U is jointly distributed with $\boldsymbol{X}$ such that $U \leftrightarrow X_\ell \leftrightarrow X_{\mathcal{L}\setminus\ell}$ for all $\ell \in \mathcal{L}$, then there exists U' jointly distributed with $(U, \boldsymbol{X})$ such that $U \leftrightarrow U' \leftrightarrow \boldsymbol{X}$ and $H(U'|X_\ell) = 0$ for all $\ell \in \mathcal{L}$.

Proof: Let $p_{U|\boldsymbol{X}}$ denote the conditional distribution of U given $\boldsymbol{X}$, and suppose that

$$U \leftrightarrow X_\ell \leftrightarrow X_{\mathcal{L}\setminus\ell}, \quad \forall\ \ell \in \mathcal{L}. \tag{36}$$

We first generate an L-partite graph

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}),$$

with vertices

$$\mathcal{V} = \bigcup_{\ell\in\mathcal{L}} \mathcal{X}_\ell.$$

The edge set $\mathcal{E}$ contains an edge

$$\{x, x'\}, \quad x \in \mathcal{X}_i,\ x' \in \mathcal{X}_j, \quad i, j \in \mathcal{L} \text{ with } i \neq j,$$

if and only if there exists an $\tilde{\boldsymbol{x}} \in \boldsymbol{\mathcal{X}}$ with $\tilde{x}_i = x$ and $\tilde{x}_j = x'$ and $p_{\boldsymbol{X}}(\tilde{\boldsymbol{x}}) > 0$.

Let $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_{N_{\text{cc}}}$ denote the connected components of $\mathcal{G}$, and let $c(x)$ denote the index of the connected component that contains vertex x.

Let us now construct a new auxiliary random variable U' on $\{1, \ldots, N_{\text{cc}}\}$ that is jointly distributed with $\boldsymbol{X}$ by setting

$$U' = c(X_1).$$

Now, for any $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ with $p_{\boldsymbol{X}}(\boldsymbol{x}) > 0$, the corresponding set of vertices $\{x_1, \ldots, x_L\}$ forms a clique and, therefore, is a subgraph of some connected component. Therefore,

$$U' = c(X_\ell) \quad \text{a.s.},\ \forall\ \ell \in \{2, \ldots, L\}.$$

This, of course, implies $H(U'|X_\ell) = 0$ for all ℓ.

To complete the proof, we need only to show that U can be generated by some conditional distribution $q_{U|U'} : \{1, \ldots, N_{\text{cc}}\} \to \mathcal{U}$. We first notice that the Markov chain (36) is equivalent to the following condition: For all $\boldsymbol{x} \in \boldsymbol{\mathcal{X}}$ with $p_{\boldsymbol{X}}(\boldsymbol{x}) > 0$, we have

$$p_{U|\boldsymbol{X}}(u|\boldsymbol{x}) = p_{U|X_1}(u|x_1) = \cdots = p_{U|X_L}(u|x_L), \quad \forall\ u \in \mathcal{U}.$$

Now consider any connected component $\mathcal{C}_i$ and any $u \in \mathcal{U}$. By the above method of constructing $\mathcal{G}$, we may conclude that

$$p_{U|X_\ell}(u|x_\ell) = \text{constant}, \quad \forall\ \ell \in \mathcal{L} \text{ and } x_\ell \in \mathcal{C}_i \cap \mathcal{X}_\ell.$$

That is, $p_{U|X_\ell}(u|x_\ell)$ depends only on the connected component $c(x_\ell)$ and the particular $u \in \mathcal{U}$, and we can write the above constant as $q_{c(x_\ell)}(u)$. Choose $p_{U|U'}(u|u') := q_{u'}(u)$ to complete the proof. ■

Appendix E
Proof of Theorem 10

Let $\ell \in \mathcal{L}$. For any $(\boldsymbol{X}, U) \sim p_{\boldsymbol{X}}\ p_{U|\boldsymbol{X}}$ on $\boldsymbol{\mathcal{X}} \times \mathcal{U}$, the following inequalities hold:

$$\begin{aligned}
\mathsf{R}_{\boldsymbol{X}_\ell|U}(D_\ell) &= \min_{q_{\hat{X}'_\ell|U,X_\ell}:\ \mathbb{E}[\mathsf{d}_\ell(\hat{X}'_\ell, X_\ell)] \le D_\ell} I(X_\ell; \hat{X}'_\ell|U) \\
&\ge \min_{q_{\hat{X}'_\ell|X_\ell}:\ \mathbb{E}[\mathsf{d}_\ell(\hat{X}'_\ell, X_\ell)] \le D_\ell} I(X_\ell; \hat{X}'_\ell) - I(\boldsymbol{X}; U) \\
&= \mathsf{R}_{X_\ell}(D_\ell) - I(\boldsymbol{X}; U).
\end{aligned} \tag{37}$$

Now suppose that we have $(U, \hat{\boldsymbol{X}}) \sim p_{\hat{\boldsymbol{X}},U|\boldsymbol{X}}$ on $\mathcal{U} \times \hat{\boldsymbol{\mathcal{X}}}$ satisfying conditions (i), (ii), (iii), and (iv) in Definition 2. Then,

$$\begin{aligned}
\mathsf{R}_{X_\ell|U}(D_\ell) &\overset{\text{a}}{\le} I(X_\ell; \hat{X}_\ell|U) \\
&\overset{\text{b}}{=} I(X_\ell; \hat{X}_\ell) - I(\boldsymbol{X}; U) \\
&\overset{\text{c}}{=} \mathsf{R}_{X_\ell}(D_\ell) - I(\boldsymbol{X}; U),
\end{aligned} \tag{38}$$

where (a) follows from property (iii) of Definition 2; (b) follows by properties (i) and (ii) of Definition 2; and (c) follows from property (iv) of Definition 2.

Inequalities (37) and (38) combine to

$$\mathsf{R}_{X_\ell|U}(D_\ell) = \mathsf{R}_{X_\ell}(D_\ell) - I(\boldsymbol{X}; U), \quad \forall \ell \in \mathcal{L}. \tag{39}$$

Thus, the pair $(U, \hat{\boldsymbol{X}})$ satisfies (13a). Moreover, since the mutual information $I(\boldsymbol{X}; U)$ does not depend on $\ell \in \mathcal{L}$, the conditional rate-distortion function $\mathsf{R}_{X_\ell|U}(D_\ell)$ is largest for the same indices ℓ as the standard rate-distortion function $\mathsf{R}_{X_\ell}(D_\ell)$. Since $\mathsf{R}_{X_\ell}(D_\ell)$ is maximum for indices $\ell^* \in \mathcal{L}^*$, this proves that the pair $(U, \hat{\boldsymbol{X}})$ also satisfies (13b). To conclude: If $(U, \hat{\boldsymbol{X}})$ satisfies (i), (ii), (iii), and (iv) in Definition 2, then U is a valid tuple for $\mathsf{C}^*_{\mathsf{g}}(\boldsymbol{D})$ and $\mathsf{K}_{\mathsf{GK}}(\boldsymbol{D}) \le \mathsf{C}^*_{\mathsf{g}}(\boldsymbol{D})$.

Now suppose that

$$\mathsf{R}_{X_1}(D_1) = \mathsf{R}_{X_2}(D_2) = \cdots = \mathsf{R}_{X_L}(D_L)$$

and, therefore, $\mathcal{L}^* = \mathcal{L}$. Let $U \sim p_{U|\boldsymbol{X}}$ on $\mathcal{U}$ be any auxiliary random variable satisfying (13a) for every $\ell \in \mathcal{L}$. (Condition (13b) automatically follows because $\mathcal{L}^* = \mathcal{L}$.) For each $\ell \in \mathcal{L}$, let $p_{\hat{X}_\ell|UX_\ell}$ be any test channel that is optimal for the informational conditional RD function

$$\mathsf{R}_{X_\ell|U}(D_\ell) = \min_{q_{\hat{X}'_\ell|UX_\ell}:\ \mathbb{E}[\mathsf{d}_\ell(\hat{X}'_\ell, X_\ell)] \le D_\ell} I(X_\ell; \hat{X}'_\ell|U).$$

Now consider the tuple

$$(\boldsymbol{X}, U, \hat{\boldsymbol{X}}) \sim p_{\boldsymbol{X}}\ p_{U|\boldsymbol{X}} \prod_{\ell \in \mathcal{L}} p_{\hat{X}_\ell|UX_\ell}.$$

For all $\ell \in \mathcal{L}$ we have

$$\begin{aligned}
\mathsf{R}_{X_\ell|U}(D_\ell) &\overset{\text{a}}{=} \mathsf{R}_{X_\ell}(D_\ell) - I(\boldsymbol{X}; U) \\
&\overset{\text{b}}{\le} I(X_\ell; \hat{X}_\ell) - I(\boldsymbol{X}; U) \\
&\le I(X_\ell; \hat{X}_\ell|U) \\
&\overset{\text{c}}{=} \mathsf{R}_{X_\ell|U}(D_\ell),
\end{aligned}$$

where (a) follow because U was originally chosen to satisfy (13a); (b) follows because $(\boldsymbol{X}, U, \hat{\boldsymbol{X}})$ need not be optimal for the informational RD functions $\mathsf{R}_{X_\ell}(D_\ell)$; and (c) follows because $p_{\hat{X}_\ell|UX_\ell}$ achieves $\mathsf{R}_{X_\ell|U}(D_\ell)$. The above inequalities must be equalities and, therefore, $(\boldsymbol{X}, U, \hat{\boldsymbol{X}})$ satisfies the following four conditions:

- $\forall\ \ell \in \mathcal{L}:\ \ U \leftrightarrow X_\ell \leftrightarrow X_{\mathcal{L}\setminus\ell}$
- $\forall\ \ell \in \mathcal{L}:\ \ U \leftrightarrow \hat{X}_\ell \leftrightarrow X_\ell$
- $\forall\ \ell \in \mathcal{L}:\ \ I(X_\ell; \hat{X}_\ell) = \mathsf{R}_{X_\ell}(D_\ell)$
- $\forall\ \ell \in \mathcal{L}:\ \ \mathbb{E}[\mathsf{d}_\ell(\hat{X}_\ell, X_\ell)] \le D_\ell$.

To conclude: Given any $(\boldsymbol{X}, U) \sim p_{\boldsymbol{X}}\ p_{U|\boldsymbol{X}}$ satisfying (13a) for all $\ell \in \mathcal{L}$ we can always find a test channel $p_{\hat{\boldsymbol{X}}|U\boldsymbol{X}}$ such that $(\boldsymbol{X}, U, \hat{\boldsymbol{X}}) \sim p_{\boldsymbol{X}}\ p_{U|\boldsymbol{X}}\ p_{\hat{\boldsymbol{X}}|U\boldsymbol{X}}$ satisfies the conditions of Definition 2. ■

APPENDIX F
PROOF OF THEOREM 11

Choose the cache capacity $C = \mathsf{C}_\mathsf{s}(\boldsymbol{D})$. Let U be an optimal auxiliary random variable for the informational RDC function; that is,

$$\mathsf{R}(\boldsymbol{D}, C) = \max_{\ell \in \mathcal{L}} \mathsf{R}_{X_\ell|U}(D_\ell).$$

Now, for each $\ell \in \mathcal{L}$, let $p_{\hat{X}_\ell|U}$ be an optimal test channel for the informational conditional RD function $\mathsf{R}_{X_\ell|U}(D_\ell)$. Define

$$(\boldsymbol{X}, U, \hat{\boldsymbol{X}}) \sim p_{\boldsymbol{X}}\ p_{U|\boldsymbol{X}}\ \prod_{\ell \in \mathcal{L}} p_{\hat{X}_\ell|U X_\ell},$$

and note that

$$\hat{X}_\ell \leftrightarrow (U, X_\ell) \leftrightarrow (X_{\mathcal{L}\setminus\ell}, \hat{X}_{\mathcal{L}\setminus\ell}), \quad \forall\, \ell \in L. \tag{40}$$

Then,

$$\begin{aligned}
\mathsf{R}(\boldsymbol{D}, C) &= \max_{\ell \in \mathcal{L}} I(X_\ell; \hat{X}_\ell | U) \\
&\geq \frac{1}{L} \sum_{\ell=1}^{L} I(X_\ell; \hat{X}_\ell | U) \\
&\overset{\text{a}}{\geq} \frac{1}{L} \sum_{\ell=1}^{L} I(\boldsymbol{X}; \hat{X}_\ell | U, \hat{X}_1^{\ell-1}) \\
&= \frac{1}{L} I(\boldsymbol{X}; \hat{\boldsymbol{X}} | U) \\
&\overset{\text{b}}{\geq} \frac{1}{L} \Big(I(\boldsymbol{X}; \hat{\boldsymbol{X}}) - \mathsf{C}_\mathsf{s}(\boldsymbol{D}) \Big) \\
&\overset{\text{c}}{\geq} \frac{1}{L} \Big(\mathsf{R}_{\boldsymbol{X}}(\boldsymbol{D}) - \mathsf{C}_\mathsf{s}(\boldsymbol{D}) \Big) \\
&\overset{\text{d}}{=} \mathsf{R}(\boldsymbol{D}, C),
\end{aligned}$$

where (a) follows from (40); (b) follows because $I(\boldsymbol{X}; U) \leq \mathsf{C}_\mathsf{s}(\boldsymbol{D})$; (c) follows because $\mathbb{E}[\mathsf{d}_\ell(X_\ell, \hat{X}_\ell)] \leq D_\ell$; and (d) follows from the definition of $\mathsf{C}_\mathsf{s}(\boldsymbol{D})$.

The above inequalities are equalities and consequently $I(X_1; \hat{X}_1 | U) = \cdots = I(X_L; \hat{X}_L | U)$, $\hat{X}_\ell \leftrightarrow U \leftrightarrow \hat{X}_{\ell-1}$ (and therefore $\hat{X}_\ell \leftrightarrow U \leftrightarrow \hat{X}_{\mathcal{L}\setminus\ell}$ since the chain rule expansion order is arbitrary), $\boldsymbol{X} \leftrightarrow \hat{\boldsymbol{X}} \leftrightarrow U$ and $C = I(\boldsymbol{X}; U)$. We can thus conclude that the tuple $(\boldsymbol{X}, U, \hat{\boldsymbol{X}})$ satisfies conditions (i)–(v) in the definition of $\mathsf{C}_\mathsf{s}^*(\boldsymbol{D})$ and $\mathsf{C}_\mathsf{s}^*(\boldsymbol{D}) \leq \mathsf{C}_\mathsf{s}(\boldsymbol{D})$.

Now suppose that $(\boldsymbol{X}, U, \hat{\boldsymbol{X}})$ satisfies conditions (i)–(v) in the definition of $\mathsf{C}_\mathsf{s}^*(\boldsymbol{D})$ and $I(\boldsymbol{X}; U) = \mathsf{C}_\mathsf{s}^*(\boldsymbol{D})$. Then

$$\begin{aligned}
\mathsf{R}(\boldsymbol{D}, C) &\leq \max_{\ell \in \mathcal{L}} I(X_\ell; \hat{X}_\ell | U) \\
&\overset{\text{a}}{=} \frac{1}{L} \sum_{\ell=1}^{L} I(X_\ell; \hat{X}_\ell | U) \\
&\overset{\text{b}}{\leq} \frac{1}{L} \sum_{\ell=1}^{L} I(\boldsymbol{X}; \hat{X}_\ell | U, \hat{X}_1^{\ell-1}) \\
&= \frac{1}{L} \Big(I(\boldsymbol{X}; \hat{\boldsymbol{X}}, U) - I(\boldsymbol{X}; U) \Big) \\
&\overset{\text{c}}{\leq} \frac{1}{L} \Big(\mathsf{R}_{\boldsymbol{X}}(\boldsymbol{D}) - \mathsf{C}_\mathsf{s}^*(\boldsymbol{D}) \Big),
\end{aligned}$$

where (a) follows because from condition (ii); (b) follows from condition (iii); (c) follows from conditions (i) and (v). Thus, we can achieve the superuser bound at $C = \mathsf{C}_\mathsf{s}^*(\boldsymbol{D})$ and $\mathsf{C}_\mathsf{s}^*(\boldsymbol{D}) \leq \mathsf{C}_\mathsf{s}(\boldsymbol{D})$. ■

APPENDIX G
PROOF OF THEOREM 14

We need the following lemma.

Lemma 18: Take any sequence of $(n, |\mathcal{M}_\mathsf{c}^{(n)}|, |\mathcal{M}^{(n)}|)$-codes and any positive real sequence $\{\alpha_n\} \downarrow 0$. If for every sufficiently large blocklength n we have

$$\mathbb{P}\left[\bigcap_{\ell \in \mathcal{L}} \Big\{ \bar{\mathsf{d}}_\ell(\hat{X}_\ell^n, X_\ell^n) < D_\ell \Big\} \right] \geq 2^{-n\alpha_n},$$

then there exists real sequence $\{\zeta_n\} \to 0$ such that

$$\frac{1}{n}\log|\mathcal{M}^{(n)}| \geq \mathsf{R}\left(\boldsymbol{D}+\zeta_n, \frac{1}{n}\log|\mathcal{M}_{\mathrm{c}}^{(n)}|+\zeta_n\right)-\zeta_n.$$

Proof: Lemma 18 is proved in Appendix H. ■

Now consider Theorem 14 and any sequence of $(n, \mathcal{M}_{\mathrm{c}}^{(n)}, \mathcal{M}^{(n)})$-codes satisfying (22) and (23). Pick a positive real sequence $\{\alpha_n\} \downarrow 0$ satisfying

$$\lim_{n\to\infty} 2^{-n\alpha_n} = 0.$$

Suppose that there exists a large blocklength n^* so that for all $n > n^*$:

$$\mathbb{P}\left[\bigcap_{\ell\in\mathcal{L}}\left\{\bar{\mathsf{d}}_\ell(\hat{X}_\ell^n, X_\ell^n) < D_\ell\right\}\right] \geq 2^{-n\alpha_n}. \tag{41}$$

Pick $\gamma > 0$ arbitrarily. By assumptions (22) and (23), and by Lemma 18, we can pick n^* sufficiently large so that $\forall n \geq n^*$ the following chain of inequalities holds:

$$\begin{aligned}\mathsf{R}(\boldsymbol{D}, C)+\gamma &\overset{\text{a}}{>} \frac{1}{n}\log|\mathcal{M}^{(n)}| \\ &\overset{\text{b}}{\geq} \mathsf{R}\left(\boldsymbol{D}+\gamma, \frac{1}{n}\log|\mathcal{M}_{\mathrm{c}}^{(n)}|+\gamma\right)-\gamma \\ &\overset{\text{c}}{\geq} \mathsf{R}\left(\boldsymbol{D}+\gamma, C+2\gamma\right)-\gamma,\end{aligned} \tag{42}$$

where step (a) follows by assumption (22); step (b) follows from Lemma 18; and step (c) follows by assumption (23) and the fact that the informational RDC function is non-increasing in the cache capacity.

Since the RDC function $\mathsf{R}(\boldsymbol{D}, C)$ is a continuous function of $\boldsymbol{D} \in [0,\infty)^L$ and $C \in [0,\infty)$ and by choosing γ sufficiently close to 0, for any desired $\epsilon > 0$ we can obtain from (42) that

$$\begin{aligned}&\mathsf{R}(\boldsymbol{D}, C)-\frac{1}{n}\log|\mathcal{M}^{(n)}| \\ &\quad\leq \mathsf{R}(\boldsymbol{D}, C)-\mathsf{R}\left(\boldsymbol{D}+\gamma, C+2\gamma\right)+\gamma \\ &\quad< \epsilon.\end{aligned} \tag{43}$$

This contradicts assumption (22). We therefore conclude that assumption (41) was wrong and holds with a strict inequality in the reverse direction for some $n \geq n^*$ and consequently

$$\limsup_{n\to\infty}\mathbb{P}\left[\bigcup_{\ell\in\mathcal{L}}\left\{\bar{\mathsf{d}}_\ell(\hat{X}_\ell^n, X_\ell^n) \geq D_\ell\right\}\right] = 1. \tag{44}$$

■

Appendix H
Proof of Lemma 18

A. Proof setup and outline

Assume that we have a sequence of $(n, \mathcal{M}_{\mathrm{c}}^{(n)}, \mathcal{M}^{(n)})$-codes for the RDC problem. For each blocklength n and RDC code $(\phi_{\mathrm{c}}^{(n)}, \phi_\ell^{(n)}, \varphi_\ell^{(n)})$, let

$$\mathcal{G}^{(n)} := \left\{\boldsymbol{x}^n \in \boldsymbol{\mathcal{X}}^n : \bar{\mathsf{d}}_\ell\Big(\varphi_\ell^{(n)}\big(\mathsf{f}(\boldsymbol{x}^n), \phi_{\mathrm{c}}^{(n)}(\boldsymbol{x}^n)\big), x_\ell^n\Big) < D_\ell, \forall\, \ell \in \mathcal{L}\right\}$$

denote the set of all "good" sequences that the code will reconstruct with acceptable distortions. Let $\{\alpha_n\} \downarrow 0$ be a sequence of positive real numbers, and suppose that the above mentioned sequence of RDC codes satisfies

$$\mathbb{P}\big[\boldsymbol{X}^n \in \mathcal{G}^{(n)}\big] \geq 2^{-n\alpha_n} \tag{45}$$

for every blocklength n. For example, we are free to choose $\{\alpha_n\}$ such that $\{2^{-n\alpha_n}\} \to 0$ or $\{2^{-n\alpha_n}\} \to 1$.

The basic idea of the following proof is to show that (45) implies that the delivery-phase rate of the sequence of RDC codes satisfies

$$\frac{1}{n}\log|\mathcal{M}^{(n)}| \geq \mathsf{R}\left(\boldsymbol{D}+\zeta_n, \frac{1}{n}\log|\mathcal{M}_{\mathrm{c}}^{(n)}|+\zeta_n\right)-\zeta_n \tag{46}$$

for some sequence $\{\zeta_n\} \to 0$. The key idea in proving this inequality will be to use the RDC code on a hypothetical "perturbed" source that is constructed from the good set $\mathcal{G}^{(n)}$ and the DMS of pmf $p_{\boldsymbol{X}}$.

B. Construction of the perturbed source

The following construction is similar to that used by Watanabe [25] and Gu and Effros [26]. Let us call the DMS

$$\boldsymbol{X}^n \sim p_{\boldsymbol{X}}^n(\boldsymbol{x}^n) = \prod_{i=1}^{n} p_{\boldsymbol{X}}(\boldsymbol{x}_i), \quad \boldsymbol{x}^n \in \boldsymbol{\mathcal{X}}^n$$

the *real source*. The *perturbed source*

$$\boldsymbol{Y}^n \sim q_{\boldsymbol{Y}^n}(\boldsymbol{y}^n) = \mathbb{P}[\boldsymbol{Y}^n = \boldsymbol{y}^n] \quad \boldsymbol{y}^n \in \boldsymbol{\mathcal{X}}^n$$

is defined as follows: If $\boldsymbol{y} \in \mathcal{G}_n$, then

$$q_{\boldsymbol{Y}^n}(\boldsymbol{y}^n) = \frac{2^{n(\alpha_n+\frac{1}{\sqrt{n}})} p_{\boldsymbol{X}}^n(\boldsymbol{y}^n)}{2^{n(\alpha_n+\frac{1}{\sqrt{n}})}\mathbb{P}[\boldsymbol{X}^n \in \mathcal{G}_n] + \mathbb{P}[\boldsymbol{X}^n \notin \mathcal{G}_n]}. \tag{47a}$$

Otherwise if $\boldsymbol{y}^n \notin \mathcal{G}_n$, then

$$q_{\boldsymbol{Y}^n}(\boldsymbol{y}^n) = \frac{p_{\boldsymbol{X}}^n(\boldsymbol{y}^n)}{2^{n(\alpha_n+\frac{1}{\sqrt{n}})}\mathbb{P}[\boldsymbol{X}^n \in \mathcal{G}_n] + \mathbb{P}[\boldsymbol{X}^n \notin \mathcal{G}_n]} \tag{47b}$$

It is worth noting that $q_{\boldsymbol{Y}^n}$ need not be a product distribution on $\boldsymbol{\mathcal{X}}^n$. It is, however, not too difficult to see that $q_{\boldsymbol{Y}^n}$ is "close" to the product distribution $p_{\boldsymbol{X}}^n$ of the real DMS in the following sense. For every sequence $\boldsymbol{y}^n \in \boldsymbol{\mathcal{X}}^n$:

$$2^{-n(\alpha_n+\frac{1}{\sqrt{n}})}\ p_{\boldsymbol{X}}^n(\boldsymbol{y}^n) \le q_{\boldsymbol{Y}^n}(\boldsymbol{y}^n) \le 2^{n(\alpha_n+\frac{1}{\sqrt{n}})}\ p_{\boldsymbol{X}}^n(\boldsymbol{y}^n). \tag{48}$$

C. Caching the perturbed source — distortion bounds

We now take the $(n, \mathcal{M}_{\mathrm{c}}^{(n)}, \mathcal{M}^{(n)})$-code $(\phi_{\mathrm{c}}^{(n)}, \phi_\ell^{(n)}, \varphi_\ell^{(n)})$ from the above mentioned sequence, and use it to cache the perturbed source $\boldsymbol{Y}^n \sim q_{\boldsymbol{Y}^n}$. For each $\ell \in \mathcal{L}$, let

$$\hat{Y}_\ell^n = \varphi_\ell^{(n)}\big(\phi_{\mathrm{c}}^{(n)}(\boldsymbol{Y}^n), \phi_\ell^{(n)}(\boldsymbol{Y}^n)\big)$$

denote the corresponding output at the decoder. A lower bound on the probability of the decoding success for this RDC code on $\boldsymbol{Y}^n$ can be obtained as follows:

$$\begin{aligned}
&\mathbb{P}\big[\boldsymbol{Y}^n \in \mathcal{G}^{(n)}\big] \\
&= \sum_{\boldsymbol{y}^n \in \mathcal{G}^{(n)}} q_{\boldsymbol{Y}^n}(\boldsymbol{y}^n) \\
&\overset{\text{a}}{=} \sum_{\boldsymbol{y}^n \in \mathcal{G}^{(n)}} \frac{2^{n(\alpha_n+\frac{1}{\sqrt{n}})} p_{\boldsymbol{X}}(\boldsymbol{y}^n)}{2^{n(\alpha_n+\frac{1}{\sqrt{n}})}\ \mathbb{P}[\boldsymbol{X}^n \in \mathcal{G}^{(n)}] + 1 - \mathbb{P}[\boldsymbol{X}^n \in \mathcal{G}^{(n)}]} \\
&= \frac{2^{n(\alpha_n+\frac{1}{\sqrt{n}})}\mathbb{P}[\boldsymbol{X}^n \in \mathcal{G}^{(n)}]}{2^{n(\alpha_n+\frac{1}{\sqrt{n}})}\ \mathbb{P}[\boldsymbol{X}^n \in \mathcal{G}^{(n)}] + 1 - \mathbb{P}[\boldsymbol{X}^n \in \mathcal{G}^{(n)}]} \\
&= \frac{2^{n(\alpha_n+\frac{1}{\sqrt{n}})}}{2^{n(\alpha_n+\frac{1}{\sqrt{n}})} + \frac{1}{\mathbb{P}[\boldsymbol{X}^n \in \mathcal{G}^{(n)}]} - 1} \\
&\overset{\text{b}}{\ge} \frac{2^{n(\alpha_n+\frac{1}{\sqrt{n}})}}{2^{n(\alpha_n+\frac{1}{\sqrt{n}})} + 2^{n\alpha_n} - 1} \\
&= \frac{2^{\sqrt{n}}}{2^{\sqrt{n}} + 1 - 2^{-n\alpha_n}} \\
&\ge \frac{2^{\sqrt{n}}}{2^{\sqrt{n}} + 1},
\end{aligned}$$

where (a) substitutes the definition of $q_{\boldsymbol{Y}^n}(\boldsymbol{y}^n)$ from (47) and (b) invokes the assumption (45). Therefore,

$$\lim_{n\to\infty} \mathbb{P}\big[\boldsymbol{Y}^n \in \mathcal{G}^{(n)}\big] = 1.$$

The expected distortion performance of the RDC code on $\boldsymbol{Y}^n \sim q_{\boldsymbol{Y}^n}$ can be upper bounded by

$$\begin{aligned}
\mathbb{E}\Big[\bar{\mathsf{d}}_\ell\big(\hat{Y}_\ell^n, Y_\ell^n\big)\Big] &= \mathbb{E}\Big[\bar{\mathsf{d}}_\ell\big(\hat{Y}_\ell^n, Y_\ell^n\big)\Big|\boldsymbol{Y}^n \in \mathcal{G}^{(n)}\Big]\mathbb{P}\Big[\boldsymbol{Y}^n \in \mathcal{G}^{(n)}\Big] \\
&\quad + \mathbb{E}\Big[\bar{\mathsf{d}}_\ell\big(\hat{Y}_\ell^n, Y_\ell^n\big)\Big|\boldsymbol{Y}^n \notin \mathcal{G}^{(n)}\Big]\mathbb{P}\Big[\boldsymbol{Y}^n \notin \mathcal{G}^{(n)}\Big]
\end{aligned}$$

$$\leq D_\ell + D_{\max}\left(1 - \frac{2^{\sqrt{n}}}{2^{\sqrt{n}}+1}\right). \tag{49}$$

Therefore,

$$\limsup_{n\to\infty} \mathbb{E}\big[\bar{\mathsf{d}}_\ell(\hat{Y}_\ell^n, Y_\ell^n)\big] \leq D_\ell, \quad \forall\ \ell \in \mathcal{L}.$$

D. Caching the perturbed source — A lower bound on the caching rate

We now give a single-letter lower bound on the caching rate for the perturbed source. Let $M_{\mathrm{c}}^{(n)} = \phi_{\mathrm{c}}^{(n)}(\boldsymbol{Y}^n)$ in $\mathcal{M}_{\mathrm{c}}^{(n)}$ denote the corresponding cache message. We have

$$\begin{aligned}
\frac{1}{n}\log|\mathcal{M}_{\mathrm{c}}^{(n)}| &\geq \frac{1}{n}H(M_{\mathrm{c}}^{(n)}) \geq \frac{1}{n}I(\boldsymbol{Y}^n; M_{\mathrm{c}}^{(n)})\\
&= \frac{1}{n}\sum_{i=1}^n I(\boldsymbol{Y}_i; M_{\mathrm{c}}^{(n)}|\boldsymbol{Y}_1^{i-1})\\
&\overset{\mathrm{a}}{=} \frac{1}{n}\sum_{i=1}^n I(\boldsymbol{Y}_i; M_{\mathrm{c}}^{(n)}, \boldsymbol{Y}_1^{i-1}) - I(\boldsymbol{Y}_i; \boldsymbol{Y}_1^{i-1})\\
&\overset{\mathrm{b}}{=} \frac{1}{n}\sum_{i=1}^n I(\boldsymbol{Y}_i; U_i) - \frac{1}{n}\sum_{i=1}^n H(\boldsymbol{Y}_i) + \frac{1}{n}\sum_{i=1}^n H(\boldsymbol{Y}_i|\boldsymbol{Y}_1^{i-1})\\
&= \frac{1}{n}\sum_{i=1}^n I(\boldsymbol{Y}_i; U_i) - \frac{1}{n}\sum_{i=1}^n H(\boldsymbol{Y}_i) + \frac{1}{n}H(\boldsymbol{Y}^n),
\end{aligned} \tag{50}$$

where in (a) we note that $q_{\boldsymbol{Y}^n}$ need not be a product measure and (b) substitutes

$$U_i = (M_{\mathrm{c}}^{(n)}, \boldsymbol{Y}_1^{i-1}) \quad \text{on} \quad \mathcal{U}_i = \mathcal{M}_{\mathrm{c}}^{(n)} \times \boldsymbol{\mathcal{X}}^{i-1}.$$

E. Caching the perturbed source — A lower bound on the delivery rate

Now consider an arbitrary request $\ell \in \mathcal{L}$, and let $M_\ell^{(n)} = \phi_\ell^{(n)}(\boldsymbol{Y}^n)$ in $\mathcal{M}^{(n)}$ denote the corresponding delivery phase message. The delivery-phase rate can be lower bound as follows:

$$\begin{aligned}
\frac{1}{n}\log|\mathcal{M}^{(n)}| &\geq \frac{1}{n}H(M_\ell^{(n)}|M_{\mathrm{c}}^{(n)})\\
&\geq \frac{1}{n}I(\boldsymbol{Y}^n; M_\ell^{(n)}|M_{\mathrm{c}}^{(n)})\\
&\overset{\mathrm{a}}{\geq} \frac{1}{n}I(\boldsymbol{Y}^n; \hat{Y}_\ell^n|M_{\mathrm{c}}^{(n)})\\
&= \frac{1}{n}\sum_{i=1}^n I(\boldsymbol{Y}_i; \hat{Y}_\ell^n|M_{\mathrm{c}}^{(n)}, \boldsymbol{Y}_1^{i-1})\\
&\geq \frac{1}{n}\sum_{i=1}^n I(Y_{\ell,i}; \hat{Y}_{\ell,i}|U_i),
\end{aligned} \tag{51}$$

where (a) follows because $\hat{Y}_\ell^n \leftrightarrow (M_\ell^{(n)}, M_{\mathrm{c}}^{(n)}) \leftrightarrow \boldsymbol{Y}^n$ forms a Markov chain; and (b) substitutes U_i.

F. Caching the perturbed source — timesharing and cardinality reduction

Consider the tuple of random variables $(\boldsymbol{Y}^n, U^n, \hat{\boldsymbol{Y}}^n)$ constructed in the above sections. Let $J \in \{1, 2, \ldots, n\}$ be a uniform random variable that is independent of $(\boldsymbol{Y}^n, U^n, \hat{\boldsymbol{Y}}^n)$, and let

$$\bar{\mathcal{U}}^{(n)} = \left(\bigcup_{i=1}^n \mathcal{U}_i\right) \times \{1, 2, \ldots, n\}.$$

Let $\big(\bar{\boldsymbol{Y}}, \bar{U}, \hat{\bar{\boldsymbol{Y}}}\big) \in \boldsymbol{\mathcal{X}} \times \bar{\mathcal{U}} \times \hat{\boldsymbol{\mathcal{X}}}$, denote the random tuples generated by setting

$$\bar{\boldsymbol{Y}} = \boldsymbol{Y}_J, \quad \bar{U} = (U_J, J) \text{ and } \hat{\bar{\boldsymbol{Y}}} = \hat{\boldsymbol{Y}}_J.$$

With this choice, it then follows from (50) that

$$\frac{1}{n}\log|\mathcal{M}_{\mathrm{c}}^{(n)}| \geq I(\boldsymbol{Y}_J; U_J|J) - H(\boldsymbol{Y}_J) + \frac{1}{n}H(\boldsymbol{Y}^n)$$

$$= I(\bar{\boldsymbol{Y}};\bar{U}) - H(\bar{\boldsymbol{Y}}) + \frac{1}{n}H(\boldsymbol{Y}^n) \tag{52}$$

and from (51) that

$$\begin{aligned}\frac{1}{n}\log|\mathcal{M}^{(n)}| &\geq I(Y_{\ell,J};\hat{Y}_{\ell,J}|U_J,J) \\ &= I(\bar{Y}_\ell;\hat{\bar{Y}}_\ell|\bar{U}).\end{aligned} \tag{53}$$

Finally, from (49) the expected distortion for satisfies

$$\begin{aligned}\mathbb{E}\big[\mathsf{d}_\ell(\hat{\bar{Y}}_\ell,\bar{Y}_\ell)\big] &= \mathbb{E}\Big[\,\bar{\mathsf{d}}(\hat{Y}_\ell^n,Y_\ell^n)\Big] \\ &\leq D_\ell + D_{\max}\left(1-\frac{2^{\sqrt{n}}}{2^{\sqrt{n}}+1}\right).\end{aligned} \tag{54}$$

Let $q_{\bar{\boldsymbol{Y}}\bar{U}\hat{\bar{\boldsymbol{Y}}}}$ denote the joint distribution of the variables $(\bar{\boldsymbol{Y}},\bar{U},\hat{\bar{\boldsymbol{Y}}})$. The cardinality of $\bar{\mathcal{U}}^{(n)}$ grows without bound in n, and the next lemma uses the convex cover method [22, Appendix C] to bound this cardinality by a finite number.

Lemma 19: There exists a random tuple $(\bar{\bar{\boldsymbol{Y}}},\bar{\bar{U}},\hat{\bar{\bar{\boldsymbol{Y}}}}) \sim q_{\bar{\bar{\boldsymbol{Y}}}\bar{\bar{U}}\hat{\bar{\bar{\boldsymbol{Y}}}}}$ defined on $\boldsymbol{\mathcal{X}} \times \bar{\bar{\mathcal{U}}} \times \hat{\boldsymbol{\mathcal{X}}}$ for which the following is true:

- $|\bar{\bar{\mathcal{U}}}| \leq |\boldsymbol{\mathcal{X}}| + 2L$,
- $q_{\bar{\bar{\boldsymbol{Y}}}} = q_{\bar{\boldsymbol{Y}}}$,
- $I(\bar{\bar{\boldsymbol{Y}}};\bar{\bar{U}}) = I(\bar{\boldsymbol{Y}};\bar{U})$,
- $I(\bar{\bar{Y}}_\ell;\hat{\bar{\bar{Y}}}_\ell|\bar{\bar{U}}) = I(\bar{Y}_\ell;\hat{\bar{Y}}_\ell|\bar{U})$ for all $\ell \in \mathcal{L}$, and
- $\mathbb{E}[\mathsf{d}_\ell(\hat{\bar{\bar{Y}}}_\ell,\bar{\bar{Y}}_\ell)] = \mathbb{E}[\mathsf{d}_\ell(\hat{\bar{Y}}_\ell,\bar{Y}_\ell)]$ for all $\ell \in \mathcal{L}$.

Combining Lemma 19 with (52), (53) and (54) yields the following: There exists some tuple

$$(\bar{\bar{\boldsymbol{Y}}},\bar{\bar{U}},\hat{\bar{\bar{\boldsymbol{Y}}}}) \sim q_{\bar{\bar{\boldsymbol{Y}}}\bar{\bar{U}}\hat{\bar{\bar{\boldsymbol{Y}}}}} \quad \text{on} \quad \boldsymbol{\mathcal{X}} \times \bar{\bar{\mathcal{U}}} \times \hat{\boldsymbol{\mathcal{X}}}$$

such that cache rate is lower bounded by

$$\frac{1}{n}\log|\mathcal{M}_{\mathrm{c}}^{(n)}| \geq I(\bar{\bar{\boldsymbol{Y}}};\bar{\bar{U}}) - H(\bar{\bar{\boldsymbol{Y}}}) + \frac{1}{n}H(\boldsymbol{Y}^n); \tag{55}$$

the expected distortion is upper bounded by

$$\mathbb{E}\Big[\mathsf{d}_\ell(\hat{\bar{\bar{Y}}}_\ell,\bar{\bar{Y}}_\ell)\Big] \leq D_\ell + D_{\max}\left(1-\frac{2^{\sqrt{n}}}{2^{\sqrt{n}}+1}\right); \tag{56}$$

and the delivery phase rate is lower bounded by

$$\begin{aligned}\frac{1}{n}\log|\mathcal{M}^{(n)}| &\geq I(\bar{\bar{Y}}_\ell;\hat{\bar{\bar{Y}}}_\ell|\bar{\bar{U}}) \\ &\geq \mathsf{R}_{\bar{\bar{Y}}_\ell|\bar{\bar{U}}}\left(D_\ell + D_{\max}\left(1-\frac{2^{\sqrt{n}}}{2^{\sqrt{n}}+1}\right)\right),\end{aligned} \tag{57}$$

where the second inequality follows from the definition of the conditional RD function.

G. Convergence of $H(\bar{\bar{\boldsymbol{Y}}})$ to $H(\boldsymbol{X})$

Fix $\gamma > 0$ arbitrarily small. The set of *γ-letter typical sequences* [23] with respect to the DMS $p_{\boldsymbol{X}}^n$ will be useful in the following arguments. This set is given by

$$\mathcal{A}_\gamma^{(n)}(p_{\boldsymbol{X}}^n) = \left\{\boldsymbol{x}^n \in \boldsymbol{\mathcal{X}}^n : \left|\frac{1}{n}\mathsf{N}(\boldsymbol{a}|\boldsymbol{x}^n) - p_{\boldsymbol{X}}(\boldsymbol{a})\right| \leq \gamma\; p_{\boldsymbol{X}}(\boldsymbol{a}),\ \forall\ \boldsymbol{a} \in \boldsymbol{\mathcal{X}}\right\}.$$

Lemma 20: The probability that the real DMS $\boldsymbol{X}^n \sim p_{\boldsymbol{X}}$ does not emit an γ-letter typical sequence satisfies [23, Thm. 1.1]

$$\mathbb{P}\Big[\boldsymbol{X}^n \notin \mathcal{A}_\gamma^{(n)}(p_{\boldsymbol{X}})\Big] \leq 2|\boldsymbol{\mathcal{X}}|2^{-n\gamma^2\mu(p_{\boldsymbol{X}})},$$

where $\mu(p_{\boldsymbol{X}})$ is the smallest value of $p_{\boldsymbol{X}}$ on its support set $\mathrm{supp}(p_{\boldsymbol{X}})$.

Let us now return to the perturbed source $\boldsymbol{Y}^n \sim q_{\boldsymbol{Y}^n}$. For each $\boldsymbol{a} \in \boldsymbol{\mathcal{X}}$ we have

$$q_{\bar{\bar{\boldsymbol{Y}}}}(\boldsymbol{a})$$

$$
\begin{aligned}
&\stackrel{\text{a}}{=} q_{\bar{\boldsymbol{Y}}}(\boldsymbol{a}) \\
&\stackrel{\text{b}}{=} \sum_{\boldsymbol{y}^n \in \boldsymbol{\mathcal{X}}^n} q_{\boldsymbol{Y}^n}(\boldsymbol{y}^n) \mathbb{P}\big[\bar{\boldsymbol{Y}} = \boldsymbol{a} \big| \boldsymbol{Y}^n = \boldsymbol{y}^n\big] \\
&\stackrel{\text{c}}{=} \sum_{\boldsymbol{y}^n \in \boldsymbol{\mathcal{X}}^n} q_{\boldsymbol{Y}^n}(\boldsymbol{y}^n) \frac{\mathsf{N}(\boldsymbol{a}|\boldsymbol{y}^n)}{n} \\
&= \sum_{\boldsymbol{y}^n \in \mathcal{A}_\gamma^{(n)}} q_{\boldsymbol{Y}^n}(\boldsymbol{y}^n) \frac{N(\boldsymbol{a}|\boldsymbol{y}^n)}{n} + \sum_{\boldsymbol{y}^n \notin \mathcal{A}_\gamma^{(n)}} q_{\boldsymbol{Y}^n}(\boldsymbol{y}^n) \frac{N(\boldsymbol{a}|\boldsymbol{y}^n)}{n} \\
&\stackrel{\text{d}}{\leq} p_{\boldsymbol{X}}(\boldsymbol{a})(1+\gamma)\mathbb{P}\big[\boldsymbol{X}^n \in \mathcal{A}_\gamma^{(n)}\big] + \mathbb{P}\big[\boldsymbol{X}^n \notin \mathcal{A}_\gamma^{(n)}\big] \\
&\stackrel{\text{e}}{\leq} p_{\boldsymbol{X}}(\boldsymbol{a})(1+\gamma) + 2|\boldsymbol{\mathcal{X}}|2^{-n\gamma^2 \mu(p_{\boldsymbol{X}})}
\end{aligned} \tag{58}
$$

where (a) applies Lemma 19; (b) and (c) use the fact that $\bar{\boldsymbol{Y}}$ is generated by uniformly at random selecting symbols from $\boldsymbol{Y}^n$ (the timesharing argument above); (d) uses the definition of γ-letter typical sequences; and (e) invokes Lemma 20. Using similar arguments, we obtain

$$
q_{\bar{\bar{\boldsymbol{Y}}}}(\boldsymbol{a}) \geq p_{\boldsymbol{X}}(\boldsymbol{a})(1-\gamma)\big(1 - 2^{-n\gamma^2\mu(p_{\boldsymbol{X}})}\big). \tag{59}
$$

From (58) and (59), we have

$$
(1-\gamma)p_{\boldsymbol{X}}(\boldsymbol{a}) \leq \liminf_{n\to\infty} q_{\bar{\bar{\boldsymbol{Y}}}}(\boldsymbol{a}) \leq \limsup_{n\to\infty} q_{\bar{\bar{\boldsymbol{Y}}}}(\boldsymbol{a}) \leq (1+\gamma)p_{\boldsymbol{X}}(\boldsymbol{a}). \tag{60}
$$

Since (60) holds for every $\gamma > 0$, and the sequence $\{q_{\bar{\bar{\boldsymbol{Y}}}}\}$ does not dependent on γ, we have

$$
\lim_{n\to\infty} q_{\bar{\bar{\boldsymbol{Y}}}}(\boldsymbol{a}) = p_{\boldsymbol{X}}(\boldsymbol{a}), \quad \forall\ \boldsymbol{a} \in \boldsymbol{\mathcal{X}}. \tag{61}
$$

Therefore, by the continuity of entropy [27, Chap. 2.3] we have

$$
\lim_{n\to\infty} H(\bar{\bar{\boldsymbol{Y}}}) = H(\boldsymbol{X}). \tag{62}
$$

H. Convergence of $(1/n)H(\boldsymbol{Y}^n)$ *to* $H(\boldsymbol{X})$

It follows from (48) that for all $\boldsymbol{a}^n \in \boldsymbol{\mathcal{X}}^n$ we have

$$
\begin{aligned}
-\alpha_n - \frac{1}{\sqrt{n}} \leq \frac{1}{n}\log p_{\boldsymbol{X}}^n(\boldsymbol{a}^n) - \frac{1}{n}\log q_{\boldsymbol{Y}^n}(\boldsymbol{a}^n) \qquad (63) \\
\leq \alpha_n + \frac{1}{\sqrt{n}}. \qquad (64)
\end{aligned}
$$

Moreover, for every $\boldsymbol{a}^n \in \mathcal{A}_\gamma^{(n)}(p_{\boldsymbol{X}})$ we have

$$
\begin{aligned}
\frac{1}{n}\log\frac{1}{p_{\boldsymbol{X}}^n(\boldsymbol{a}^n)} &\stackrel{\text{a}}{=} \frac{1}{n}\log\left(\prod_{i=1}^n \frac{1}{p_{\boldsymbol{X}}(\boldsymbol{a}_i)}\right) \\
&= \frac{1}{n}\sum_{i=1}^n \log\frac{1}{p_{\boldsymbol{X}}(\boldsymbol{a}_i)} \\
&= \frac{1}{n}\sum_{\boldsymbol{a}' \in \boldsymbol{\mathcal{X}}} \mathsf{N}(\boldsymbol{a}'|\boldsymbol{a}^n)\log\frac{1}{p_{\boldsymbol{X}}(\boldsymbol{a}')} \\
&\stackrel{\text{b}}{\leq} \big(1+\gamma\big)\sum_{\boldsymbol{a}' \in \boldsymbol{\mathcal{X}}} p_{\boldsymbol{X}}(\boldsymbol{a}')\log\frac{1}{p_{\boldsymbol{X}}(\boldsymbol{a}')} \\
&= \big(1+\gamma\big)H(\boldsymbol{X}),
\end{aligned} \tag{65}
$$

where (a) follows because $p_{\boldsymbol{X}}^n$ is a product measure and (b) follows because $\boldsymbol{a}^n \in \mathcal{A}_\gamma^{(n)}(p_{\boldsymbol{X}})$. Similarly, we have

$$
\frac{1}{n}\log\frac{1}{p_{\boldsymbol{X}}^n(\boldsymbol{a}^n)} \geq (1-\gamma)H(\boldsymbol{X}) \tag{66}
$$

for all $\boldsymbol{a}^n \in \boldsymbol{\mathcal{X}}^n$.

Now consider the joint entropy $H(\boldsymbol{Y}^n)$. With a few manipulations, we obtain the upper bound in (67). Here step (a) uses (63). Step (b) uses the upper bound in (65) on the first logarithmic term, and

$$
\frac{1}{n}\log\frac{1}{p_{\boldsymbol{X}}^n(\boldsymbol{a}^n)} = \frac{1}{n}\sum_{i=1}^n \log\frac{1}{p_{\boldsymbol{X}}(\boldsymbol{a}_i)}
$$

$$
\begin{aligned}
\frac{1}{n}H(\boldsymbol{Y}^n) &= \frac{1}{n}\sum_{\boldsymbol{a}^n \in \mathrm{supp}(q_{\boldsymbol{Y}^n})} q_{\boldsymbol{Y}^n}(\boldsymbol{a}^n)\log\frac{1}{q_{\boldsymbol{Y}}(\boldsymbol{a}^n)} \\
&\overset{\text{a}}{\leq} \sum_{\boldsymbol{a}^n \in \mathrm{supp}(q_{\boldsymbol{Y}^n})} q_{\boldsymbol{Y}^n}(\boldsymbol{a}^n)\left(\frac{1}{n}\log\frac{1}{p^n_{\boldsymbol{X}}(\boldsymbol{a}^n)} + \alpha_n + \frac{1}{\sqrt{n}}\right) \\
&= \sum_{\boldsymbol{a}^n \in \mathcal{A}^{(n)}_\gamma(p_{\boldsymbol{X}}) \,\cap\, \mathrm{supp}(q_{\boldsymbol{Y}^n})} q_{\boldsymbol{Y}^n}(\boldsymbol{a}^n)\left(\frac{1}{n}\log\frac{1}{p^n_{\boldsymbol{X}}(\boldsymbol{a}^n)} + \alpha_n + \frac{1}{\sqrt{n}}\right) \\
&\qquad + \sum_{\boldsymbol{a}^n \notin \mathcal{A}^{(n)}_\gamma(p_{\boldsymbol{X}}) \,\cap\, \mathrm{supp}(q_{\boldsymbol{Y}^n})} q_{\boldsymbol{Y}^n}(\boldsymbol{a}^n)\left(\frac{1}{n}\log\frac{1}{p^n_{\boldsymbol{X}}(\boldsymbol{a}^n)} + \alpha_n + \frac{1}{\sqrt{n}}\right) \\
&\overset{\text{b}}{\leq} \sum_{\boldsymbol{a}^n \in \mathcal{A}^{(n)}_\gamma(p_{\boldsymbol{X}}) \,\cap\, \mathrm{supp}(q_{\boldsymbol{Y}^n})} q_{\boldsymbol{Y}^n}(\boldsymbol{a}^n)\left((1+\gamma)H(\boldsymbol{X}) + \alpha_n + \frac{1}{\sqrt{n}}\right) \\
&\qquad + \sum_{\boldsymbol{a}^n \notin \mathcal{A}^{(n)}_\gamma(p_{\boldsymbol{X}}) \,\cap\, \mathrm{supp}(q_{\boldsymbol{Y}^n})} q_{\boldsymbol{Y}^n}(\boldsymbol{a}^n)\left(\log\frac{1}{\mu(p_{\boldsymbol{X}})} + \alpha_n + \frac{1}{\sqrt{n}}\right) \\
&\overset{\text{c}}{\leq} (1+\gamma)H(\boldsymbol{X}) + \alpha_n + \frac{1}{\sqrt{n}} + 2|\boldsymbol{X}|2^{-n\gamma\mu(p_{\boldsymbol{X}})}\left(\log\frac{1}{\mu(p_{\boldsymbol{X}})} + \alpha_n + \frac{1}{\sqrt{n}}\right)
\end{aligned}
\tag{67}
$$

$$
\begin{aligned}
&\leq \frac{1}{n}\sum_{i=1}^{n}\log\frac{1}{\mu(p_{\boldsymbol{X}})} \\
&= \log\frac{1}{\mu(p_{\boldsymbol{X}})}
\end{aligned}
$$

on the second term[9]. Finally, step (c) applies Lemma 20. Using similar arguments, we also have

$$
\begin{aligned}
&\frac{1}{n}H(\boldsymbol{Y}^n) \\
&= \frac{1}{n}\sum_{\boldsymbol{a}^n \in \mathrm{supp}(q_{\boldsymbol{Y}^n})} q^n_{\boldsymbol{Y}}(\boldsymbol{a}^n)\log\frac{1}{q_{\boldsymbol{Y}}(\boldsymbol{a}^n)} \\
&\overset{\text{a}}{\geq} \sum_{\boldsymbol{a}^n \in \mathcal{A}^{(n)}_\gamma(p_{\boldsymbol{X}})} q^n_{\boldsymbol{Y}}(\boldsymbol{a}^n)\left(\frac{1}{n}\log\frac{1}{p^n_{\boldsymbol{X}}(\boldsymbol{a}^n)} - \alpha_n - \frac{1}{\sqrt{n}}\right) \\
&\overset{\text{b}}{\geq} \sum_{\boldsymbol{a}^n \in \mathcal{A}^{(n)}_\gamma(p_{\boldsymbol{X}})} q^n_{\boldsymbol{Y}}(\boldsymbol{a}^n)\left((1-\gamma)H(\boldsymbol{X}) - \alpha_n - \frac{1}{\sqrt{n}}\right) \\
&\overset{\text{c}}{\geq} \left((1-\gamma)H(\boldsymbol{X}) - \alpha_n - \frac{1}{\sqrt{n}}\right)\left(1 - 2|\boldsymbol{X}|2^{-n\gamma\mu(p_{\boldsymbol{X}})}\right).
\end{aligned}
\tag{68}
$$

Step (a) follows from (63); step (b) follows from (66); and step (c) applies Lemma 20. From (67) and (68) we have for every fixed $\gamma > 0$

$$(1-\gamma)H(\boldsymbol{X}) \leq \liminf_{n\to\infty}\frac{1}{n}H(\boldsymbol{Y}^n) \leq \limsup_{n\to\infty}\frac{1}{n}H(\boldsymbol{Y}^n) \leq (1+\gamma)H(\boldsymbol{X}),$$

which, in turn, implies

$$\lim_{n\to\infty}\frac{1}{n}H(\boldsymbol{Y}^n) = H(\boldsymbol{X}). \tag{69}$$

I. Completing the Proof

The above arguments show that there exists a sequence of random variables[10]

$$\left\{(\bar{\bar{\boldsymbol{Y}}}_n, \bar{\bar{U}}_n) \sim q_{\bar{\bar{\boldsymbol{Y}}}_n}(\cdot)\; q_{\bar{\bar{U}}_n|\bar{\bar{\boldsymbol{Y}}}_n}(\cdot|\cdot)\right\},$$

[9] If $p^n_{\boldsymbol{X}}(\boldsymbol{a}^n) = 0$, then by definition $q_{\boldsymbol{Y}^n}(\boldsymbol{a}^n) = 0$ and $\boldsymbol{a}^n \notin \mathrm{supp}(q_{\boldsymbol{Y}^n})$.

[10] Here, for clarity, we have added the subscript n on the random variables to identify the corresponding blocklength n.

with each $(\bar{\bar{\boldsymbol{Y}}}_n, \bar{\bar{U}}_n)$ defined on $\boldsymbol{\mathcal{X}} \times \mathcal{U}$, such that

$$\lim_{n\to\infty} q_{\bar{\bar{\boldsymbol{Y}}}_n}(\boldsymbol{a}) = p_{\boldsymbol{X}}(\boldsymbol{a}), \quad \forall\ \boldsymbol{a} \in \boldsymbol{\mathcal{X}}$$

and

$$\begin{aligned}
&\frac{1}{n}\log|\mathcal{M}_{\mathrm{c}}^{(n)}| \geq I(\bar{\bar{\boldsymbol{Y}}};\bar{\bar{U}}) - \epsilon_{1,n} \\
&\frac{1}{n}\log|\mathcal{M}_{\mathrm{c}}^{(n)}| \geq \mathsf{R}_{\bar{\bar{Y}}_{\ell,n}|\bar{\bar{U}}_n}(D_\ell + \epsilon_{2,n}), \quad \forall\ \ell \in \mathcal{L},
\end{aligned}$$

where

$$\epsilon_{1,n} = \left|\frac{1}{n}H(\boldsymbol{Y}^n) - H(\bar{\bar{\boldsymbol{Y}}})\right| \tag{70}$$

$$\epsilon_{2,n} = D_{\max}\left(1 - \frac{2^{\sqrt{n}}}{2^{\sqrt{n}} - 1}\right). \tag{71}$$

Let $(\boldsymbol{X}, \bar{\bar{U}}_n) \sim p_{\boldsymbol{X}}(\cdot)\ q_{\bar{\bar{U}}_n|\bar{\bar{\boldsymbol{Y}}}_n}(\cdot|\cdot)$, and define

$$\epsilon_{3,n} = \left|\mathsf{R}_{\bar{\bar{\boldsymbol{Y}}}_n|\bar{\bar{U}}_n}(D_\ell + \epsilon_{2,n}) - \mathsf{R}_{\boldsymbol{X}|\bar{\bar{U}}_n}(D_\ell + \epsilon_{2,n})\right|.$$

Finally, choose $\zeta_n = \max\{\epsilon_{1,n}, \epsilon_{2,n}, \epsilon_{2,n}\}$ so that the lemma follows from (61), (62) and (69) and the continuity of the informational conditional RD function. ■

APPENDIX I
PROOF OF THEOREM 16

The proof of Theorem 16 will bootstrap the achievability part of Lemma 13 and the strong converse in Theorem 14. Take the single-symbol distortion functions $\mathbf{d}^*$ from (26), and consider $\mathsf{R}^{\dagger}_{\mathbf{d}^*}(\boldsymbol{D}, C)$ and $\mathsf{R}^{\ddagger}_{\mathbf{d}^*}(\boldsymbol{D}, C)$ — the respective operational RDC functions in the expected and excess distortion settings w.r.t. the separable distortion functions

$$\bar{\mathbf{d}}^* = (\,\bar{\mathsf{d}}_1^*, \ldots,\, \bar{\mathsf{d}}_L^*),$$

where

$$\bar{\mathsf{d}}_\ell^*(\hat{x}_\ell^n, x_\ell^n) = \frac{1}{n}\sum_{i=1}^{n}\mathsf{d}_\ell^*(\hat{x}_{\ell,i}, x_{\ell,i}) = \frac{1}{n}\sum_{i=1}^{n}\mathsf{f}_\ell\big(\mathsf{d}_\ell(\hat{x}_{\ell,i}, x_{\ell,i})\big).$$

Lemma 21:

$$\mathsf{R}^{\dagger}_{\mathbf{d}^*}(\boldsymbol{D}, C) = \mathsf{R}^{\ddagger}_{\mathbf{d}^*}(\boldsymbol{D}, C) = \mathsf{R}_{\mathbf{d}^*}(\boldsymbol{D}, C).$$

Proof: Apply Lemma 13 with $\bar{\mathbf{d}}^*$. ■

Lemma 22:

$$\mathsf{R}^{\ddagger}_{\mathbf{f}}(\boldsymbol{D}, C) = \mathsf{R}^{\ddagger}_{\mathbf{d}^*}\big(\mathbf{f}(\boldsymbol{D}), C\big).$$

Proof: For every $(n, \mathcal{M}_{\mathrm{c}}^{(n)}, \mathcal{M}^{(n)})$-code we have

$$\begin{aligned}
&\mathbb{P}\left[\bigcup_{\ell\in\mathcal{L}}\left\{\overline{\mathsf{fd}}_\ell(\hat{X}_\ell^n, X_\ell^n) \geq D_\ell\right\}\right] \\
&\overset{\text{a}}{=} \mathbb{P}\left[\bigcup_{\ell\in\mathcal{L}}\left\{\mathsf{f}_\ell^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathsf{f}_\ell\Big(\mathsf{d}_\ell(\hat{X}_\ell^n, X_\ell^n)\Big)\right) \geq D_\ell\right\}\right] \\
&= \mathbb{P}\left[\bigcup_{\ell\in\mathcal{L}}\left\{\frac{1}{n}\sum_{i=1}^{n}\mathsf{d}_\ell^*(\hat{X}_\ell^n, X_\ell^n) \geq \mathsf{f}_\ell\big(D_\ell\big)\right\}\right] \\
&\overset{\text{b}}{=} \mathbb{P}\left[\bigcup_{\ell\in\mathcal{L}}\left\{\bar{\mathsf{d}}_\ell^*(\hat{X}_\ell^n, X_\ell^n) \geq \mathsf{f}_\ell(D_\ell)\right\}\right].
\end{aligned}$$

The left hand side of (a) corresponds to the excess-distortion event for $\mathsf{R}^{\ddagger}_{\mathbf{f}}(\boldsymbol{D}, C)$, and the right hand side of (b) corresponds to the excess-distortion event for $\mathsf{R}^{\ddagger}_{\mathbf{d}^*}(\mathbf{f}(\boldsymbol{D}), C)$. Therefore, a sequence of $(n, \mathcal{M}_{\mathrm{c}}^{(n)}, \mathcal{M}^{(n)})$-codes can achieve vanishing error probabilities w.r.t. the f-separable distortion functions $\bar{\mathbf{d}}_{\mathbf{f}}$ if and only if it achieves vanishing error probabilities w.r.t. the separable distortion functions $\bar{\mathbf{d}}^*$. ■

Lemma 23:

$$\tilde{\mathsf{R}}^{\dagger}_{\mathbf{f},\text{max-exc}}(\boldsymbol{D}, C) \leq \mathsf{R}^{\ddagger}_{\mathbf{f}}(\boldsymbol{D}, C).$$

Proof: Recall Definition 4 and fix the distortion tuple $\boldsymbol{D}$ and cache capacity C. If $R > \mathsf{R}^{\ddagger}_{\mathbf{f}}(\boldsymbol{D}, C)$ then there exists a sequence of $(n, \mathcal{M}_{\mathsf{c}}^{(n)}, \mathcal{M}^{(n)})$-codes satisfying (3a), (3b) and (21). For this sequence of codes, let

$$\mathcal{G}_n = \bigcap_{\ell \in \mathcal{L}} \Big\{ \bar{\mathsf{d}}_{\mathsf{f}_\ell}(\hat{X}_\ell^n, X_\ell^n) < D_\ell \Big\},$$

and let $\mathcal{G}_n^c$ denote the complement of $\mathcal{G}_n$. Then

$$\begin{aligned}
&\mathbb{E}\left[\max_{\ell \in \mathcal{L}} \big(\overline{\mathsf{fd}}_\ell(\hat{X}_\ell^n, X_\ell^n) - D_\ell\big)\right] \\
&= \mathbb{E}\Big[\max_{\ell \in \mathcal{L}} \big(\overline{\mathsf{fd}}_\ell(\hat{X}_\ell^n, X_\ell^n) - D_\ell\big)\Big|\mathcal{G}_n\Big] \mathbb{P}\Big[\mathcal{G}_n\Big] \\
&\quad + \mathbb{E}\Big[\max_{\ell \in \mathcal{L}} \big(\overline{\mathsf{fd}}_\ell(\hat{X}_\ell^n, X_\ell^n) - D_\ell\big)\Big|\mathcal{G}_n^c\Big] \mathbb{P}\Big[\mathcal{G}_n^c\Big] \\
&\le D_{\max}\ \mathbb{P}[\mathcal{G}_n^c].
\end{aligned} \tag{72}$$

Since $D_{\max}$ is finite and $\mathbb{P}[\mathcal{G}_n^c] \to 0$ by (21), we have

$$\limsup_{n\to\infty} \mathbb{E}\left[\max_{\ell \in \mathcal{L}} \Big(\overline{\mathsf{fd}}_\ell(\hat{X}_\ell^n, X_\ell^n) - D_\ell\Big)\right] \le 0$$

and $R \ge \tilde{\mathsf{R}}^{\dagger}_{\mathbf{f},\text{max-exc}}(\boldsymbol{D}, C)$ by Definition 5. ∎

Lemma 24:

$$\tilde{\mathsf{R}}^{\dagger}_{\mathbf{f},\text{max-exc}}(\boldsymbol{D}, C) \ge \mathsf{R}_{\mathbf{d}^*}(\mathbf{f}(\boldsymbol{D}), C).$$

Proof: If $\mathsf{R}_{\mathbf{d}^*}(\boldsymbol{D}^*, C) = 0$, then the lemma immediately follows because we always have $\tilde{\mathsf{R}}^{\dagger}_{\mathbf{f},\text{max-exc}}(\boldsymbol{D}, C) \ge 0$. We henceforth restrict attention to the nontrivial case $\mathsf{R}_{\mathbf{d}^*}(\mathbf{f}(\boldsymbol{D}), C) > 0$.

Suppose, to the contrary of Lemma 24, that $\tilde{\mathsf{R}}^{\dagger}_{\mathbf{f},\text{max-exc}}(\boldsymbol{D}, C)$ is strictly smaller than $\mathsf{R}_{\mathbf{d}^*}(\mathbf{f}(\boldsymbol{D}), C)$ and, therefore, there exists some $\gamma > 0$ such that

$$\tilde{\mathsf{R}}^{\dagger}_{\mathbf{f},\text{max-exc}}(\boldsymbol{D}, C) \le \mathsf{R}_{\mathbf{d}^*}(\mathbf{f}(\boldsymbol{D}), C) - \gamma. \tag{73}$$

By the continuity and monotonicity of $\mathsf{R}_{\mathbf{d}^*}(\mathbf{f}(\boldsymbol{D}), C)$ and each f_ℓ, there exists some distortion tuple $\boldsymbol{D}'$ such that

$$\mathsf{R}_{\mathbf{d}^*}(\mathbf{f}(\boldsymbol{D}'), C) = \mathsf{R}_{\mathbf{d}^*}(\mathbf{f}(\boldsymbol{D}), C) - \frac{\gamma}{2} \tag{74}$$

where $D'_\ell > D_\ell$ for all $\ell \in \mathcal{L}$.

Now recall Definition 5 and the operational meaning of $\tilde{\mathsf{R}}^{\dagger}_{\mathbf{f},\text{max-exc}}(\boldsymbol{D}, C)$. There exists a sequence of $(n, \mathcal{M}_{\mathsf{c}}^{(n)}, \mathcal{M}^{(n)})$-codes satisfying (3a), (3b) and (27). On combining (3b), (73) and (74), we see that the delivery-phase rates of this sequence of codes satisfy

$$\limsup_{n\to\infty} \frac{1}{n} \log|\mathcal{M}^{(n)}| \le \mathsf{R}_{\mathbf{d}^*}\big(\mathbf{f}(\boldsymbol{D}'), C\big) - \frac{\gamma}{2}. \tag{75}$$

Now consider the excess-distortion performance of the sequence of $(n, \mathcal{M}_{\mathsf{c}}^{(n)}, \mathcal{M}^{(n)})$-codes w.r.t. the separable distortion functions $\mathbf{d}^*$. Let

$$\mathcal{B}_n = \bigcup_{\ell \in \mathcal{L}} \Big\{ \bar{\mathsf{d}}^*_\ell(\hat{X}_\ell^n, X_\ell^n) \ge \mathsf{f}_\ell(D'_\ell) \Big\},$$

and let $\mathcal{B}_n^c$ denote the complement of $\mathcal{B}_n$. Notice that we have

$$\mathcal{B}_n = \bigcup_{\ell \in \mathcal{L}} \Big\{ \overline{\mathsf{fd}}_\ell(\hat{X}_\ell^n, X_\ell^n) \ge D'_\ell \Big\}.$$

Since the asymptotic delivery-phase rate is strictly smaller than the informational RDC function (75), the strong converse in Theorem 14 yields

$$\limsup_{n\to\infty} \mathbb{P}\big[\mathcal{B}_n\big] = 1.$$

Let

$$\zeta = \min_{\ell \in \mathcal{L}} \big(D'_\ell - D_\ell\big).$$

We now have

$$\begin{aligned}
&\mathbb{E}\Big[\max_{\ell \in \mathcal{L}} \big(\overline{\mathsf{fd}}_\ell(\hat{X}_\ell^n, X_\ell^n) - D_\ell\big)\Big] \\
&= \mathbb{E}\Big[\max_{\ell \in \mathcal{L}} \big(\overline{\mathsf{fd}}_\ell(\hat{X}_\ell^n, X_\ell^n) - D_\ell\big)\Big|\mathcal{B}_n\Big]\ \mathbb{P}\Big[\mathcal{B}_n\Big]
\end{aligned}$$

$$
\begin{aligned}
&+ \mathbb{E}\Big[\max_{\ell \in \mathcal{L}} \big(\overline{\mathsf{fd}}_\ell(\hat{X}_\ell^n, X_\ell^n) - D_\ell \big) \Big| \mathcal{B}_n^c \Big] \ \mathbb{P}\Big[\mathcal{B}_n^c \Big] \\
&\overset{\text{a}}{\geq} \mathbb{E}\Big[\max_{\ell \in \mathcal{L}} \big(\overline{\mathsf{fd}}_\ell(\hat{X}_\ell^n, X_\ell^n) - D_\ell \big) \Big| \mathcal{B}_n \Big] \ \mathbb{P}\Big[\mathcal{B}_n \Big] \\
&- \big(\min_{\ell \in \ell} D_\ell \big) \mathbb{P}\Big[\mathcal{B}_n^c \Big] \\
&\overset{\text{b}}{\geq} \zeta \ \mathbb{P}\Big[\mathcal{B}_n \Big] - \big(\min_{\ell \in \ell} D_\ell \big) \mathbb{P}\Big[\mathcal{B}_n^c \Big],
\end{aligned} \tag{76}
$$

where (a) follows because $\overline{\mathsf{fd}}_\ell(\hat{X}_\ell^n, X_\ell^n)$ is nonnegative; and (b) follows because, conditioned on $\mathcal{B}_n$, there must exist at least one $\ell' \in \mathcal{L}$ such that

$$
\overline{\mathsf{fd}}_{\ell'}(\hat{X}_{\ell'}^n, X_{\ell'}^n) \geq D'_{\ell'} > D_{\ell'}
$$

and thus

$$
\max_{\ell \in \mathcal{L}} \big(\overline{\mathsf{fd}}_\ell(\hat{X}_\ell^n, X_\ell^n) - D_\ell \big) \geq \overline{\mathsf{fd}}_{\ell'}(\hat{X}_{\ell'}^n, X_{\ell'}^n) - D_{\ell'} > \zeta.
$$

Finally, we have

$$
\begin{aligned}
0 &\overset{\text{a}}{=} \limsup_{n \to \infty} \mathbb{E}\Big[\max_{\ell \in \mathcal{L}} \big(\overline{\mathsf{fd}}_\ell(\hat{X}_\ell^n, X_\ell^n) - D_\ell \big) \Big] \\
&\overset{\text{b}}{\geq} \limsup_{n \to \infty} \Big[\zeta \ \mathbb{P}\big[\mathcal{B}_n \big] - (\min_{\ell \in \mathcal{L}} D_\ell) \mathbb{P}\big[\mathcal{B}_n^c \big] \Big] \\
&\overset{\text{c}}{>} 0,
\end{aligned}
$$

where (a) follows from (27), (b) follows from (76), and (c) follows because $\mathbb{P}[\mathcal{B}_n] \to 1$ by the strong converse Theorem 14 and $\zeta > 0$. The above contradiction implies that $\tilde{\mathsf{R}}^\dagger_{\mathbf{f},\text{max-exc}}(\boldsymbol{D}, C)$ *cannot* be strictly smaller than $\mathsf{R}_{\mathbf{d}^*}(\mathbf{f}(\boldsymbol{D}), C)$. ∎

To complete the proof of Theorem I we need only combine the above lemmas:

$$
\begin{aligned}
\tilde{\mathsf{R}}^\dagger_{\mathbf{f},\text{max-exc}}(\boldsymbol{D}, C) &\overset{\text{a}}{\leq} \mathsf{R}^\ddagger_{\mathbf{f}}(\boldsymbol{D}, C) \\
&\overset{\text{b}}{=} \mathsf{R}^\ddagger_{\mathbf{d}^*}\big(\mathbf{f}(\boldsymbol{D}), C\big) \\
&\overset{\text{c}}{=} \mathsf{R}_{\mathbf{d}^*}(\mathbf{f}(\boldsymbol{D}), C) \\
&\overset{\text{d}}{\leq} \tilde{\mathsf{R}}^\dagger_{\mathbf{f},\text{max-exc}}(\boldsymbol{D}, C),
\end{aligned}
$$

where (a) uses Lemma 23, (b) uses Lemma 22, (c) uses Lemma 21, and (d) uses Lemma 24. ∎

References

[1] Y. Shkel and S. Verdú, "A coding theorem for f-separable distortion measures," in *Information Theory and Applications Workshop (ITA)*, San Diego, USA, 2016.

[2] V. Tikhomirov, "On the notion of mean," *Selected works of A. N. Kolmogorov series Mathematics and its Applications*, vol. 25, 1991.

[3] K. Visweswariah, S. R. Kulkarni, and S. Verdú, "Output distribution of the Burrows-Wheeler transform," in proceedings *IEEE International Symposium on Information Theory*, Sorrento, Italy, 2000.

[4] M. Effros, K. Visweswariah, S. R. Kulkarni and S. Verdú, "Universal lossless source coding with the Burrows Wheeler transform," *IEEE Transactions on Information Theory*, vol. 48, no. 5, 2002.

[5] C. Y. Wang, S. H. Lim and M. Gastpar, "Information-theoretic caching: sequential coding for computing," *arXiv*, vol. 1504.00553, 2015.

[6] P. Hassanzadeh, E. Erkip, J. Llorca and A. Tulino, "Distortion-memory tradeoffs in cache-aided wireless video delivery," *arXiv*, 1511.03932, 2015.

[7] Q. Yang and D. Gündüz, "Centralized coded caching for Heterogenous lossy requests," *arXiv*, 1604.08178, 2016.

[8] R. Gray, "Conditional rate-distortion theory," Stanford University Technical Report, October, 1972.

[9] K. B. Viswanatha, E. Akyol and K. Rose, "The lossy common information of correlated sources," *IEEE Transactions on Information Theory*, vol. 60, no. 6, pp. 3238 – 3253, 2014.

[10] P. Gács and J. Körner, "Common information is far less than mutual information," *Problems of Control and Information Theory*, vol. 2, no. 2, p. 149 –162, 1973.

[11] A. Wyner, "The common information of two dependent random variables," *IEEE Transactions on Information Theory*, vol. 21, no. 2, pp 163 – 179, 1975.

[12] R. Gray and A. Wyner, "Source coding for a simple network," *Bell Systems Technical Journal*, vol. 53, no. 9, pp. 1681 – 1721, 1974.

[13] J. C. Kieffer, "Sample converses in source coding theory," *IEEE Transactions on Information Theory*, vol. 37, no. 2, pp. 263 – 268, 1991.

[14] A. Wyner, "The rate-distortion function for source coding with side-information at the decoder-II: general sources," *Information and Control*, vol. 38, pp. 60 – 80, 1978.

[15] A. Wyner, "The common information of two dependent random variables," *IEEE Transactions on Information Theory*, vol. 21, no. 2, pp. 163 – 179, 1975.

[16] G. Xu, W. Liu, and B. Chen, "A lossy source coding interpretation of Wyner's common information" *IEEE Transactions on Information Theory*, vol. 62, no. 2, pp. 754–768, 2016.

[17] A. Lapidoth and S. Tinguely, "Sending a bivariate Gaussian over a Gaussian MAC," *IEEE Transactions on Information Theory*, vol. 56, no. 6, pp. 2714 – 2752, 2010.

[18] J. Xiao, Z. Luo, "Compression of correlated Gaussian sources under individual distortion criteria," in *Allerton Conference on Communications, Control, and Computing*, Monticello, IL, September, 2005.

[19] R. Timo, A. Grant, and G. Kramer, "Lossy broadcasting with complementary side information," *IEEE Transactions on Information Theory*, vol. 59, no. 1, pp. 104 – 131, 2013.

[20] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Transactions on Information Theory*, vol. 22, no. 1, pp. 1 – 10, 1976.

[21] E. Tuncel, "Slepian-Wolf coding over broadcast channels," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1469 – 1482, 2006.

[22] A. El Gamal and Y.-H. Kim, *Network Information Theory*, Cambridge University Press, 2011.

[23] G. Kramer, "Topics in multi-user information theory," *Foundations and Trends in Communications and Information Theory*, vol. 4, no. 45, pp. 265 – 444, 2008.

[24] V. M. Prabhakaran and M. M. Prabhakaran, "Assisted common information with an application to secure two-party sampling," *IEEE Transactions on Information Theory*, vol. 60, no. 6, pp. 3413 – 3434, 2014.

[25] S. Watanabe, "Second-order region for Gray-Wyner network," *arXiv*, 1508.04227, 2015.

[26] W. H. Gu and M. Effros, "A strong converse for a collection of network source coding problems," in proceedings *IEEE International Symposium on Information Theory*, Seoul, South Korea, 2009.

[27] R. W. Yeung, *Information theory and network coding*, Springer, 2008.

[28] C. Y. Wang, S. H. Lim and M. Gastpar, "A New Converse Bound for Coded Caching," *arXiv*, vol. 1601.05690, 2016.

[29] M. A. Maddah-Ali, U. Niesen, "Fundamental limits of caching," in *IEEE Transactions on Information Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.

[30] U. Niesen and M. A. Maddah-Ali, "Coded caching with nonuniform demands," *arXiv*, 1308.0178, 2013.

[31] H. Ghasemi and A. Ramamoorthy, "Improved lower bounds for coded caching," *IEEE Transactions on Information Theory*, vol. 63, no. 7, pp. 4388–4413, July 2017.

[32] A. Amiri, Q. Yang, and D. Gündüz, "Coded caching for a large number of users," in proceedings *IEEE Information Theory Workshop*, Cambridge, UK, July 2016, pp. 171–175.

[33] M. Ji, A. M. Tulino, J. Llorca, and G. Caire, "Order-optimal rate of caching and coded multicasting with random demands," *arXiv*, 1502.03124, Feb. 2015.

[34] S. Saeedi Bidokhti, M. Wigger, and R. Timo, "Noisy broadcast networks with receiver caching," *arXiv*, 1605.02317, May, 2016.

[35] S. Saeedi Bidokhti, M. Wigger, and A. Yener, "Benefits of cache assignment on degraded broadcast channels," *arXiv*, 1702.08044, February, 2017.

[36] A. S. Cacciapuoti, M. Caleffi, M. Ji, J. Llorca, A. M. Tulino,"Speeding up future video distribution via channel-aware caching-aided coded multicast," *IEEE Journal Selected Areas in Communications*, vol. 34, no. 8, pp. 2207–2218, August, 2016.

[37] A. Ghorbel, M. Kobayashi, and S. Yang "Content delivery in erasure broadcast channels with cache and feedback," *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 6407–6422, November, 2016.

[38] J. Zhang and P. Elia, "Fundamental limits of cache-aided wireless BC: interplay of coded-caching and CSIT feedback," *IEEE Transactions on Information Theory*, vol. 63, no. 5, pp. 3142–3160, May, 2017.

[39] J. Hachem, N. Karamchandani, and S. Diggavi, "Content caching and delivery over heterogeneous wireless networks," in proceedings *IEEE International Conference on Computer Communications*, Kowloon, China, March, 2015.

[40] R. Timo, S. Saeedi Bidokhti, and M. Wigger, "A Rate-Distortion Approach to Caching," in proceedings *arXiv*, 1610.07304, October, 2016.

[41] P. Hassanzadeh, A. Tulino, J. Llorca, and E. Erkip, "Rate-Memory Trade-off for the Two-User Broadcast Caching Network with Correlated Sources," *arXiv*, 1705.04616, May, 2017.