State-Adaptive Coded Caching for Symmetric Broadcast Channels

†Shirin Saeedi Bidokhti, ‡Michèle Wigger, †*Aylin Yener, and ‡*Abbas El Gamal

† University of Pennsylvania, saeedi@seas.upenn.edu, †*The Pennsylvania State University, yener@engr.psu.edu
‡ LTCI, Telecom ParisTech, Université Paris-Saclay, 75013 Paris, France, michele.wigger@telecom-paristech.fr
‡*Stanford University, abbas@ee.stanford.edu

***Abstract*—Coded-caching delivery is considered over a symmetric noisy broadcast channel whose state is unknown at the transmitter during the cache placement phase. In particular, the delivery phase is modeled by a state-dependent broadcast channel where the state remains constant over each transmission block and is learned by the transmitter (and the receivers) only at the beginning of each block. A *state-adaptive coded caching* scheme is proposed that improves either on rate or decoding latency over two baseline schemes that are based on standard coded caching.**

I. Introduction

Coded caching [1] has recently emerged as a means to improve content delivery in multiuser networks. The performance gains offered by coded-caching scale with the number of users and go beyond those so-called local gains stemming from the fact that part of the data is locally stored at the receivers. While earlier works studied network models with noiseless channels for delivery [1], caching has more recently been studied in noisy channels, including broadcast channels (BCs) that are most related to this work. In particular, [2]–[5] consider static (and known) degraded BCs and propose joint cache-channel coding schemes that improve rate of communication and attain new global caching gains when the users have unequal channel qualities and the weaker receivers have larger cache memories (or demand less data). Time-varying (fading) BCs and the interplay between feedback, channel state information and spatial multiplexing with caching have also been studied in [6]–[10]. These works apply separate cache-channel coding architectures; hence, the performance of communication in the delivery phase is limited by the weakest users. By contrast, in this work we illustrate the benefits of joint cache-channel coding schemes for state-dependent BCs even when different users have equal size caches and i.i.d. channel statistics.

In this work, we model the delivery phase as a state-dependent BC in which the state sequence is constant over a coherence block and changes from block to block in an i.i.d. manner. This channel model subsumes the standard block fading channel model. The transmitter and receivers learn the realization of the state at the *beginning of each block*. This can be done using pilot symbols and feedback. For clarity of presentation, we consider state-symmetric BCs in which all users have equal size caches and statistically equivalent channels and we assume that the channel is degraded in each state. Since the state realizations vary over blocks, a receiver that is strongest in one block can be weakest in the next block.

We propose a coding scheme for state-dependent BCs termed *state-adaptive coded caching* hereafter. The caching phase of our scheme is performed in an uncoded manner, following the original work of [1]. Our delivery scheme applies (i) opportunistic user scheduling across blocks and (ii) generalized coded caching [5] in each block. Specifically, only the $t+1$ receivers with the best channel conditions are served in each block, t being the coded caching parameter used in the cache placement [1]. The proposed scheme serves each of the chosen receivers with a transmission rate that is proportional to its channel quality; i.e., each chosen receiver k is served at a rate that approaches $I(X;Y_k|S=s_b)$, where $X, Y_k, S=s_b$ denote the input, output, and state variables in block b, respectively. This performance is achieved by a variation of Tuncel coding [11] where for each receiver k, the transmitter only encodes bits that are stored in the cache memories of all the other receivers in the chosen subset. This implies a state-adaptive virtual cache allocation at the receivers that allocates a larger portion of cache memories for decoding at weaker receivers than at stronger receivers. Note that for the state-symmetric BC considered in this paper, the total rate and the total required cache size at each user are the same on average (in the long run) across all users.

The proposed strategy is compared to two baseline schemes that combine standard coded caching with the opportunistic BC codes [14] in a separate cache-channel coding architecture. The first baseline scheme, which we term *blockwise coded caching*, operates on a per-block basis and is limited by the worst channel in each block. A variant of this baseline scheme in which opportunistic user selection policy is replaced by a threshold-based user selection policy is discussed in [9]. Our proposed strategy also operates on a per-block basis but employs a joint cache-channel coding architecture such that the communication to stronger users is not limited by weaker users. It therefore achieves higher rates than blockwise coded caching. The second baseline scheme, which we term *ergodic coded caching*, codes over the entire communication duration, i.e., over many blocks. This results in symmetric channel conditions for all the receivers and eliminates the rate-bottleneck issue of weak receivers in a coherence block. It has, however, the drawback that decoding is performed only at the end of transmission. In state-adaptive coded caching (as well as in the first baseline scheme), decoding can be performed after each block so that a part of the message bits can be recovered earlier. This is particularly beneficial

in video streaming in which one wishes to start watching a movie as soon as some of the bits are recovered[1]. We quantify this notion by a new delay measure termed the *decoding latency factor* that describes the extent to which decoding is performed sequentially. We show a factor of two improvement in decoding latency factor of the new state-adaptive coded caching scheme over the second baseline scheme.

II. Problem Definition

Consider a state-dependent K-receiver broadcast channel (BC) with (finite) input, output, and state alphabets $\mathcal{X}, \mathcal{Y}$, and $\mathcal{S}$. Given the time-i channel input $X_i \in \mathcal{X}$ and state $S_i \in \mathcal{S}$, receiver $k \in \mathcal{K} := \{1, \ldots, K\}$'s time-$i$ output $Y_{k,i} \in \mathcal{Y}$ follows the broadcast channel law

$$p_{Y_k|X,S}(y_{k,i}|x_i, s_i). \tag{1}$$

For simplicity, we consider state-symmetric BCs in which for any permutation on users $\nu\colon \mathcal{K} \to \mathcal{K}$ there exists a permutation on states $\pi_\nu\colon \mathcal{S} \to \mathcal{S}$, so that for all $s \in \mathcal{S}$, $x \in \mathcal{X}$, $y \in \mathcal{Y}$, we have (see [13, Definition 1] for the two-user definition):

$$p_S(s) = p_S(\pi_\nu(s)) \tag{2a}$$

$$p_{Y_k|X,S}(y|x,s) = p_{Y_{\nu(k)}|X,S}(y|x, \pi_\nu(s)), \quad \forall k \in \mathcal{K}. \tag{2b}$$

Moreover, we assume that the BC is stochastically degraded [12] in any state realization $S = s$.

The state sequence $S_1, S_2, \ldots,$ stays constant over a coherence interval of T_s channel uses and then changes in an independent and identically distributed (i.i.d.) manner. I.e.,

$$S_{(b-1)T_s+1} = \ldots = S_{bT_s} = S'_b, \quad b = 1, 2, \ldots$$

where $S'_1, S'_2, \ldots$ is an i.i.d. sequence distributed according to a given distribution $p_{S'}(\cdot)$.

The transmitter has access to a database with D independent messages (files) $W_1, \ldots, W_D$, each consisting of nR i.i.d. random bits. Here, n denotes the blocklength and R the message rate. Each receiver k demands exactly one of the messages, which we denote by W_{d_k}. Receiver $k \in \mathcal{K}$ has access to a local cache memory of $n\mathsf{M}$ bits.

Communication takes place in two phases. The first *cache placement phase* is assumed to take place during a period of low network congestion and is thus assumed error free. In this phase, the transmitter stores information about the messages in each of the K receivers' cache memory. So, in receiver k's cache memory, it stores

$$\mathbb{V}_k := g_k(W_1, \ldots, W_D)$$

for some function $g_k\colon \{1, \ldots, 2^{nR}\}^D \to 2^{n\mathsf{M}}$ that is known to all terminals. The cache content $\mathbb{V}_k$ is known only to the transmitter and receiver k. During the placement phase, it is unknown which messages are demanded by the users; so g_k cannot depend on the demands.

The subsequent *delivery phase* takes place during periods of high network congestion and is modeled by the state-dependent BC in (1). At the beginning of the delivery phase, each receiver demands one of the messages in the library; i.e., receiver k demands message W_{d_k}. At this time, the transmitter and all receivers get informed about all receivers' demands, $\mathbf{d} = (d_1, d_2, \ldots, d_K)$. The transmitter then computes the sequence of channel inputs as

$$X_i := f^{(i)}(\mathbf{d}, W_1, \ldots, W_D, S^i), \quad i \in \{1, \ldots, n\},$$

where $f^{(i)} : \{1, \ldots, D\}^K \times \{1, \ldots, 2^{nR}\}^D \times \mathcal{S}^i \to \mathcal{X}$.

Decoding is performed *online*. In particular, we present coding schemes in which receivers recover a certain number of message bits after each coherence interval T_s. Let

$$B = \frac{n}{T_s} \tag{3}$$

denote the number of coherence blocks encountered when communicating over n channel uses. The online decoding procedure is described as follows. After each coherence block $b \in \{1, \ldots, B\}$, receiver k recovers $m_{k,b}$ new bits of its desired message W_{d_k} using the decoding operation

$$\hat{W}_{k,b} := \varphi_{k,b}(\mathbf{d}, Y_k^{bT_s}, \mathbb{V}_k, S^{bT_s}),$$

where $\varphi_{k,b} : \{1, \ldots, D\}^K \times \mathcal{Y}^{bT_s} \times \mathcal{V}_k \times \mathcal{S}^{bT_s} \to \{1, \ldots, 2^{m_{k,b}}\}$. The final estimate of the receiver for message W_{d_k} is then composed of the concatenation of all the estimates $(\hat{W}_{k,1}, \hat{W}_{k,2}, \ldots)$.

To capture the nature of online decoding, we study the following average expected delay per bit

$$\bar{L}_{\text{bit}} := \max_{\mathbf{d}} \frac{1}{K} \sum_{k=1}^{K} \mathbb{E}\left[\sum_{b=1}^{n} m_{k,b} \cdot bT_s\right] \frac{1}{(R - \mathsf{M}/D)n}, \tag{4}$$

where the worst case over all possible demands is considered. Normalization is by $(R - M/D)n$ because we wish to average only over the number of transmitted bits but not over the bits that are already stored in the cache memory. Expectation is over the random state, channel realizations and messages.

We consider the worst-case error probability over demands:

$$P_{\text{e}}^{(n)} := \max_{\mathbf{d} \in \{1,\ldots,D\}^K} \mathbb{P}\Big[\bigcup_{k=1}^{K} \{\hat{W}_k \neq W_{d_k}\} \Big].$$

We also assume that the coherence time T_s and the number of blocks B tend to infinity, i.e., $T_s, B \to \infty$. Under this assumption, for positive rates $R > 0$, the delay $\bar{L}_{\text{bit}}$ also tends to infinity. We, therefore, further normalize it by the blocklength n, yielding the *decoding latency factor* $\bar{\rho}$*:*

$$\bar{\rho} \triangleq \lim_{n\to\infty} \frac{\bar{L}_{\text{bit}}}{n}. \tag{5}$$

Definition 1 *A triple* (M, R, ρ) *is achievable, if there exists a sequence (in* n*) of caching and delivery encoders and decoders with cache and message rates* M *and* R *such that*

$$\lim_{n\to\infty} P_{\text{e}}^{(n)} = 0 \qquad \text{and} \qquad \bar{\rho} \leq \rho. \tag{6}$$

III. State-Adaptive Coded-Caching

Our proposed scheme has a parameter $t \in \{0, \ldots, K-1\}$, where $t+1$ indicates the number of users that are simultaneously served in the delivery phase. E.g., parameter $t = 0$

[1]The assumption here is that the movie is encoded using multi-description coding and thus the order of the bits is not relevant. Otherwise, it is also possible to prioritize the bits.

corresponds to opportunistic broadcasting, which is known to achieve the maximum sum-rate and symmetric rate [14, Chapter 6] when there are no cache memories.

We start with some definitions. Fix $t \in \{0,\dots,K-1\}$. Let $\mathcal{G}_1^t,\dots,\mathcal{G}_{\binom{K}{t}}^t$ denote all size-t subsets of $\mathcal{K}$, i.e., all sets of t users. Choose a conditional probability law $p_{X|S}$ so that

$$p_{X|S}(x|s) = p_{X|S}(x|\pi_\nu(s)), \qquad \forall x \in \mathcal{X},\ s \in \mathcal{S}, \tag{7}$$

for any set of permutations π_ν introduced in (2).

Define the mapping $G^{t+1}\colon \mathcal{S} \to \mathcal{K}^{t+1}$ such that for all $s \in \mathcal{S}$, $k \in G^{t+1}(s)$, and $j \in (\mathcal{K}\backslash G^{t+1}(s))$, the channel $p_{Y_j|X,S=s}$ is stochastically degraded with respect to $p_{Y_k|X,S=s}$. In our scheme, $G^{t+1}(s)$ is the set of "active" receivers in a block with state $S'=s$. The symmetry condition (2) ensures that $G(S)$ is uniformly distributed over all $t+1$-user sets of $\mathcal{K}$.

Define for $k \in \mathcal{K}$:

$$R = \frac{t+1}{K-t} \cdot I\big(X;Y_k|S,\{k \in G^{t+1}(S)\}\big) - \epsilon. \tag{8}$$

Note by (2) and (7) that the choice of k does not matter. Here, $\{k \in G^{t+1}(S)\}$ denotes the event that index k is an element of $G^{t+1}(s)$. Let the cache size be

$$\mathsf{M} = \frac{\binom{K-1}{t-1}}{\binom{K}{t}} RD = \frac{t}{K} RD. \tag{9}$$

Distribute the nR bits of file W_d into $\binom{K}{t}$ queues $Q_{d,\mathcal{G}_1^t},\dots,Q_{d,\mathcal{G}_{\binom{K}{t}}^t}$, each consisting of $nR\cdot\binom{K}{t}^{-1}$ bits[2]

Placement Phase: For each k and $\ell \in \big\{1,\dots,\binom{K}{t}\big\}$ such that $k \in \mathcal{G}_\ell^t$, store all the bits of queue $Q_{d,\mathcal{G}_\ell^t}$ in receiver k's cache memory. The cache content of user k is thus:

$$\mathbb{V}_k = \Big\{Q_{d,\mathcal{G}_\ell^t}\colon d \in \{1,\dots,D\} \text{ and } \ell \in \Big\{1,\dots,\binom{K}{t}\Big\} \text{ s.t. } k \notin \mathcal{G}_\ell^t\Big\}. \tag{10}$$

Notice that each sub-message is stored at exactly t receivers. Moreover, the placement of the information does not depend on the realization of the channel state. By (9), this placement satisfies the memory constraint of $n\mathsf{M}$ bits.

Delivery Phase: Delivery is block-by-block in our scheme. Consider the coherence block $b \in \{1,\dots,B\}$ and assume that the channel state is realized to be $S'_b = s_b$. At the beginning of each coherence block, the transmitter retrieves the next

$$\mu_{k,G^{t+1}(s_b)} \triangleq T_s \cdot \left(I\big(X;Y_k|S=s_b\big) - \epsilon \cdot \frac{K-t}{t+1}\right) \tag{11}$$

bits from queue $Q_{d_k,G^{t+1}(s_b)\backslash\{k\}}$, for $k \in G^{t+1}(s_b)$. Denote the bits retrieved from queue $Q_{d_k,G^{t+1}(s_b)\backslash\{k\}}$ by $W_{k,b}$. If the queue is empty, let $W_{k,b}$ be the all-zero string.

Use a random codebook

$$\mathcal{C}_b^{T_s} = \Big\{\mathbf{x}_b^{T_s}(w)\colon\ w \in \big\{1,\dots,2^{T_s r(s_b)}\big\}\Big\} \tag{12}$$

of rate

$$r(s_b) \triangleq \max_{k \in G^{t+1}(s_b)} I(X;Y_k|S=s_b) - \epsilon \cdot \frac{K-t}{t+1} \tag{13}$$

[2]We assume $n \geq \binom{K}{t}$ since our interest is in the regime $n \to \infty$.

with entries drawn i.i.d. according to a given law $p_{X|S}(\cdot|s_b)$. The transmitter sends the codeword

$$\mathbf{x}_b^{T_s}\Big(\overline{\bigoplus}_{k \in G^{t+1}(s_b)} W_{k,b}\Big) \tag{14}$$

over the channel. Here $\overline{\bigoplus}$ describes the XOR operation of the submessages after zero-padding to the same length.

Decoding is done sequentially after each block $b = 1,2,\dots$. Consider decoding at receiver $k \in \mathcal{K}$. Suppose $S'_b = s_b$ and $k \in G^{t+1}(s_b)$. Receiver k can retrieve bits from the queues

$$\left\{Q_{d_k,\mathcal{G}_\ell^t}\colon\ \ell \in \left\{1,\dots,\binom{K}{t}\right\} \text{ s.t. } k \in \mathcal{G}_\ell^t\right\} \tag{15}$$

that are stored in its local cache. To recover the missing bits, it uses the retrieved bits to form the XOR-message

$$W_{\text{XOR},b}(k) := \overline{\bigoplus}_{i \in G^{t+1}(s_b)\backslash\{k\}} W_{i,b}. \tag{16}$$

It then extracts a subcodebook $\tilde{\mathcal{C}}_{b,k}\big(W_{\text{XOR},b}(k)\big)$ from $\mathcal{C}_b$ that contains all codewords that are compatible with $W_{\text{XOR},b}(k)$:

$$\tilde{\mathcal{C}}_{b,k}\big(W_{\text{XOR},b}(k)\big) := \Big\{\mathbf{x}_b^{T_s}\big(w \oplus W_{\text{XOR},b}(k)\big)\Big\}.$$

Finally, it collects the outputs in coherence block b, and applies a maximum likelihood decoder based on the extracted subcodebook $\tilde{\mathcal{C}}_{b,k}\big(W_{\text{XOR},b}(k)\big)$ to recover the bits $W_{k,b}$. If $k \notin G^{t+1}(s_b)$, receiver k does not decode anything in this block b.

Performance Analysis: Given that $S'_b = s_b$, the number of bits $m_{k,b}$ recovered by a given receiver $k \in \mathcal{K}$ at the end of coherence block b is

$$m_{k,b} = \begin{cases} 0, & \text{if } k \notin G^{t+1}(s_b) \\ T_s I\big(X;Y_k|S=s_b\big) - \frac{\epsilon T_s (K-t)}{t+1} & \text{if } k \in G^{t+1}(s_b). \end{cases}$$

These bits pertain to queue $Q_{d_k,G^{t+1}(s_b)\backslash\{k\}}$ and are useful information bits unless this queue is empty.

Notice that the symmetry conditions (2) and (7) imply that $\sum_{s\colon k \in G^{t+1}(s)} p_S(s) I(X;Y_k|S=s)$ does not depend on the receiver index k. Moreover, (2) ensures that the set $G^{t+1}(S)$ is uniformly distributed over all $t+1$-user subsets of $\mathcal{K}$. As a consequence, when averaged over the random state realization, for each block b the same expected number of bits is transmitted from each of the queues $\big\{Q_{d_k,\mathcal{G}_\ell^t}\colon k \notin \mathcal{G}_\ell^t\big\}$. By the ergodicity of the process $\{S'_b\}$ and because during the initialization procedure each queue is allocated the same number of bits, when $B \to \infty$, almost all transmitted bits are useful information bits and all queues will be emptied at the end of the transmission as long as the message rate R satisfies:

$$\begin{aligned} R &< \varlimsup_{B\to\infty} \frac{1}{BT_s} \sum_{b=1}^{B} m_{k,b} + \frac{\mathsf{M}}{D} \\ &= \sum_{s\in\mathcal{S}\colon\ k \in G^{t+1}(s)} p_S(s) \left(I\big(X;Y_k|S=s\big) - \frac{\epsilon(K-t)}{t+1}\right) + \frac{\mathsf{M}}{D} \\ &\overset{(a)}{=} I(X;Y_k|S,\{k \in G^{t+1}(S)\})\mathbb{P}[k \in G^{t+1}(S)] \\ &\quad - \frac{\epsilon(K-t)}{t+1}\mathbb{P}[k \in G^{t+1}(S)] + \frac{\mathsf{M}}{D} \end{aligned}$$

$$\stackrel{(b)}{=} \frac{t+1}{K} I(X;Y_k|S,\{k\in G^{t+1}(S)\}) + \frac{\mathsf{M}}{D} - \frac{K-t}{K}\epsilon$$
$$\stackrel{(c)}{=} \frac{t+1}{K-t} I(X;Y_k|S,\{k\in G^{t+1}(S)\}) - \epsilon, \quad (17)$$

where (a) holds because $\mathbb{P}[S=s, k\in G^{t+1}(s)] = \mathbb{P}[S=s]$ and $I(X;Y_k|S=s,\{k\in G^{t+1}(s)\}) = I(X;Y_k|S=s)$ for all s such that $k\in G^{t+1}(s)$; (b) holds because $\mathbb{P}[k\in G^{t+1}(S)] = \frac{t+1}{K}$ and (c) holds by (9).

Notice further that by the choice in (11), the probability of decoding error in each block tends to 0 as $T_s\to\infty$.

For finite $n = T_sB$ the average expected delay per bit is:

$$\bar{L}_{\text{bit}} \stackrel{(a)}{=} \frac{1}{K}\sum_{k=1}^{K}\frac{1}{(R-\mathsf{M}/D)n}\sum_{b=1}^{B}\mathbb{E}_{S_b}[m_{k,b}]\cdot bT_s$$
$$\stackrel{(b)}{=} \frac{1}{BT_s}\cdot T_s^2\cdot\sum_{b=1}^{B} b = \frac{T_s(B+1)}{2}, \quad (18)$$

where (a) follows by the definition in (4); and (b) because $\mathbb{E}_{S_b}[m_{k,b}] = R - \mathsf{M}/D$ and $n = BT_s$. The decoding latency factor of the proposed state-adaptive coded caching is thus:

$$\bar{\rho} = \lim_{T_s,B\to\infty}\frac{(B+1)T_s}{2n} = \lim_{T_s,B\to\infty}\frac{(B+1)T_s}{2BT_s} = \frac{1}{2}. \quad (19)$$

For each $t\in\{0,1,\ldots,K-1\}$, define

$$R_t := \frac{t+1}{K-t}\cdot\max_{p_{X|S}:\,(7)\text{ holds}} I\big(X;Y_1|S,\{1\in G^{t+1}(S)\}\big) \quad (20)$$
$$\mathsf{M}_t := \frac{t}{K}DR_t \quad (21)$$

By time/memory-sharing arguments [1] and by optimizing over $p_{X|S}$, the presented analysis (with $\epsilon\to 0$) establishes the following theorem.

Theorem 1 *State-adaptive coded caching achieves all rate-memory pairs on the upper convex envelope of*

$$\{(R_t,\mathsf{M}_t)\colon t=0,1,\ldots,K-1\} \quad (22)$$

with decoding latency factor $\bar{\rho}=\frac{1}{2}$.

Remark 1 *Including input distributions* $p_{X|S}$ *that satisfy* (7) *but don't maximize* (20) *does not increase the set of achievable rate-memory pairs in Theorem 1, because they are subsumed by the upper convex envelope operation.*

Remark 2 *The maximization in* (20) *can be re-written as:*

$$\max_{p_{X|S}}\frac{1}{K}\sum_{k=1}^{K} I\big(X;Y_k|S,\{k\in G^{t+1}(S)\}\big) \quad (23)$$

where the maximization is over all (also non-symmetric) input distributions. This follows from the symmetry condition (2).

Remark 3 *The proposed scheme only serves the best* $t+1$ *receivers in each block. We could combine transmissions to various sets of* $t+1$ *receivers in a single block by means of superposition or Marton coding. But since for each state realization the BC is assumed degraded, these techniques do not increase the set of achievable rate-memory-latency triples.*

IV. Comparison to Baseline Schemes

Two baseline schemes derived from standard coded caching are described and compared to the proposed state-adaptive coded caching scheme. The results are summarized in Table I.

A. Blockwise Coded Caching

Fix a parameter $t\in\{0,\ldots,K-1\}$. Consider a separate cache-channel coding scheme with placement strategy as in Section III and a delivery strategy that combines standard coded caching [1] of parameter t with an opportunistic BC code that in each coherence block serves only the $t+1$ strongest receivers. Specifically, it sends an XOR-message produced by the coded caching algorithm to these strongest $t+1$ receivers. With this scheme, the performance in each block is limited by the worst of the $t+1$ best receivers. In fact, at the end of coherence block b with state $S'_b = s_b$, the number of bits recovered at receiver $k\in G^{t+1}(s_b)$ is:

$$m_{k,b} = \begin{cases} 0, & \text{if } k\notin G^{t+1}(s_b)\\ T_s\min\limits_{j\in G^{t+1}(s_b)} I\big(X;Y_j|S=s_b\big) & \\ -\frac{\epsilon T_s(K-t)}{t+1} & \text{if } k\in G^{t+1}(s_b).\end{cases}$$

By symmetry, and when $B\to\infty$, for any $k\in\mathcal{K}$ the message rate to receiver k is:

$$R = \sum_{s\in\mathcal{S}:\ 1\in G^{t+1}(s)} p_S(s)\left(\min_{j\in G^{t+1}(s_b)} I(X;Y_j|S=s) - \frac{\epsilon(K-t)}{t+1}\right) + \frac{\mathsf{M}}{D}. \quad (24)$$

The required cache size M and the decoding latency factor are similar as for the state-adaptive coded caching scheme:

$$\mathsf{M} = \frac{t}{K}RD \qquad \text{and} \qquad \bar{\rho} = \frac{1}{2}. \quad (25)$$

Plugging (25) into (24), taking $\epsilon\to 0$, and optimizing over $p_{X|S}$ yields the desired value for the rate R_t in Table I.

B. Ergodic Coded Caching

Fix a parameter $t\in\{0,\ldots,K-1\}$. The scheme combines standard coded caching with an opportunistic BC code that codes over the entire blocklength n. That means, in each block transmission is only to the best $t+1$ receivers, but decoding is performed only at the end of the entire blocklength n. That means, the XOR-message sent to a given a set of $t+1$ receivers is decoded based on *all the blocks* where the opportunistic scheduling chooses to transmit to these $t+1$ receivers. This allows to exploit the ergodic behaviour of the blocks. Ergodic coded caching achieves the same rate-memory pairs as state-adaptive coded caching. The price to pay is the worst case decoding latency factor $\bar{\rho}=1$.

V. Gaussian Fading Channels

Consider a Rayleigh block-fading channel

$$Y_{k,i} = h_{k,i}X_i + Z_{k,i}, \quad (26)$$

Scheme	Expected Rate R_t	Decoding Latency Factor ρ
State-Adaptive Coded Caching	$\frac{t+1}{K-t} \cdot \max_{p_{X\|S}\,:\,(7)\text{ holds}} I(X;Y_1\|S,\{1\in G^{t+1}(S)\})$	1/2
Blockwise Coded Caching	$\sum_{s\in\mathcal{S}:\ 1\in G^{t+1}(s)} p_S(s) \max_{p_{X\|S}\,:\,(7)\text{ holds}} \min_{j\in G^{t+1}(s)} I(X;Y_j\|S=s)$	1/2
Ergodic Coded Caching	$\frac{t+1}{K-t} \cdot \max_{p_{X\|S}\,:\,(7)\text{ holds}} I(X;Y_1\|S,\{1\in G^{t+1}(S)\})$	1

TABLE I: Comparison of rate and decoding latency factor for the different coded-caching adaptations.

with channel coefficients that remain constant over a block,

$$h_{k,i} = h'_{k,b}, \qquad \forall i = (b-1)T_s + 1, \dots, bT_s, \tag{27}$$

and with $\{h'_{k,b}\}$ an i.i.d. complex Gaussian sequence with zero-mean unit-variance symbols. The noise sequence $\{Z_{k,i}\}$ is also i.i.d. complex Gaussian of unit variance. Inputs $X_1, \dots, X_n$ are subject to an expected average block power constraint P.

Let $\mathbf{h}' := (h'_1, \dots, h'_k)$ and fix $t \in \{0, \dots, K-1\}$. Here, $G^{t+1}(\mathbf{h}')$ denotes the set of $t+1$ users with largest channel coefficients in $\mathbf{h}'$. The maximum in Theorem 1 is attained by a zero-mean Gaussian input of state-dependent instantaneous power $P(\mathbf{h}')$, which can be found using the Karush-Kuhn-Tucker conditions on the equivalent maximization problem (23). This proves achievability of the upper convex envelope of all rate-memory pairs

$$R_t = \frac{t+1}{K-t}\mathbb{E}_{\mathbf{h}'}\left[\log\left(1+|h'_1|^2 P(\mathbf{h}')\right)\Big|\{1\in G^{t+1}(\mathbf{h}')\}\right] \tag{28}$$

$$M_t = \frac{t}{K}DR. \tag{29}$$

where $P(\mathbf{h}')$ is the waterfilling solution characterized by:

$$\lambda = \sum_{k\in G^{t+1}(\mathbf{h}')} \frac{1}{x(\mathbf{h}') + \frac{1}{|h'_k|^2}} \tag{30}$$

$$P(\mathbf{h}') = [x(\mathbf{h}')]^+ \tag{31}$$

$$P = \mathbb{E}_{\mathbf{h}'}\left[P(\mathbf{h}')\right]. \tag{32}$$

Figure 1 compares the rates achieved by state-adaptive and blockwise coded caching (CC) under opportunistic and non-opportunistic designs. A non-opportunistic design refers to a variation of the schemes where time-sharing is applied in each block to serve all subsets of $t+1$ users during the same fraction of time. Each marked memory-rate point corresponds to a choice of the parameter t, with the left-most point corresponding to $t=0$ and the right-most point corresponding to $t=K-1$. The curve is obtained by time/memory-sharing between the points. The rate-memory pairs lying to the right of the right-most ($t=K-1$) point are achieved by a scheme that stores a part of each message in every cache memory and applies placement and delivery strategies with parameter $t=K-1$ to the remaining part of the files.

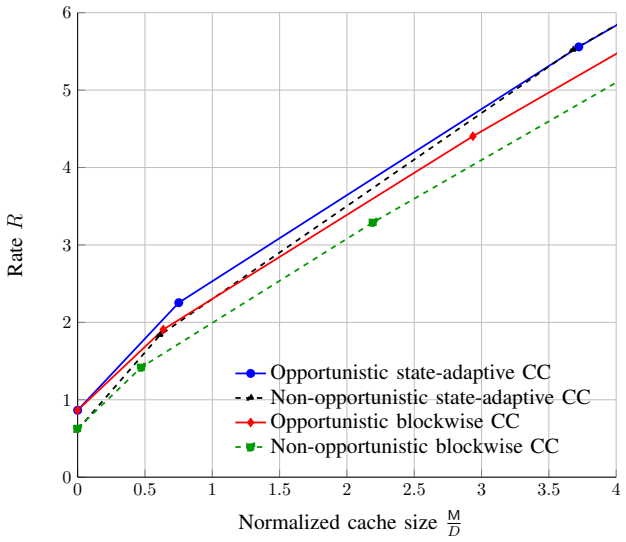

Fig. 1: Comparison of achievable rates on a 3-user example with power constraint $P=4$.

Acknowledgment

M. Wigger was supported by the ERC-project CTO-Com.

References

[1] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.
[2] S. Saeedi Bidokhti, M. Wigger, and R. Timo, "Noisy broadcast networks with receiver caching," *arXiv*, 1605.02317, May 2016.
[3] 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 J. Selected Areas Commun.*, vol. 34, no. 8, Aug 2016.
[4] M. M. Amiri and D. Gündüz, "Cache-aided content delivery over erasure broadcast channels," *arXiv*, 1702.05454, Feb 2017.
[5] S. Saeedi Bidokhti, M. A. Wigger, and A. Yener, "Benefits of cache assignment on degraded broadcast channels," *arXiv* 1702.08044, Feb 2017.
[6] S. Wang, X. Tian and H. Liu, "Exploiting the unexploited of coded caching for wireless content distribution," in *Proc. IEEE Int. Conf. Computing, Networking and Commun.*, Feb 2015.
[7] J. Zhang and P. Elia, "Fundamental limits of cache-aided wireless bc: Interplay of coded-caching and CSIT feedback," *IEEE Trans. Inf. Theory*, vol. 63, no. 5, pp. 3142–3160, May 2017.
[8] A. Ghorbel, M. Kobayashi, and S. Yang, "Content delivery in erasure broadcast channels with cache and feedback," *IEEE Trans. Inf. Theory*, vol. 62, no. 11, pp. 6407–6422, Nov 2016.
[9] K. Ngo, S. Yang, and M. Kobayashi, "Scalable content delivery with coded caching in multi-antenna fading channels," *arXiv*, 1703.06538, Mar 2017.
[10] E. Piovano, H. Joudeh, B. Clerckx, "On coded caching in the overloaded MISO broadcast channel," *arXiv* 1702.01672, Feb 2017.
[11] E. Tuncel, "Slepian-Wolf coding over broadcast channels," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1469–1482, Apr. 2006.
[12] A. El Gamal and Y. H. Kim, Network Information Theory, 2011, Cambridge Univ. Press.
[13] H. Kim, Y. K. Chia, and A. El Gamal, "A note on the broadcast channel with stale state information at the transmitter," *IEEE Trans. Inf. Theory*, vol. 61, no. 7, pp. 3622–3631, July 2015.
[14] D. Tse and P. Viswanath, "Fundamentals of wireless communications," 2004.