Information Embedding on Actions

Behzad Ahmadi, Student Member, IEEE, Himanshu Asnani, Student Member, IEEE, Osvaldo Simeone, Member, IEEE, and Haim H. Permuter, Member, IEEE

Abstract

The problem of optimal actuation for channel and source coding was recently formulated and solved in a number of relevant scenarios. In this class of models, actions are taken at encoders or decoders, either to acquire side information in an efficient way or to control or probe effectively the channel state. In this paper, the problem of embedding information on the actions is studied for both the source and the channel coding set-ups. In both cases, a decoder is present that observes only a function of the actions taken by an encoder or a decoder of an action-dependent point-to-point link. For the source coding model, this decoder wishes to reconstruct a lossy version of the source being transmitted over the point-to-point link, while for the channel coding problem the decoder wishes to retrieve a portion of the message conveyed over the link.

For the problem of source coding with actions taken at the decoder, a single letter characterization of the set of all achievable tuples of rate, distortions at the two decoders and action cost is derived, under the assumption that the mentioned decoder observes a function of the actions non-causally, strictly causally or causally. A special case of the problem in which the actions are taken by the encoder is also solved. A single-letter characterization of the achievable capacity-cost region is then obtained for the channel coding set-up with actions. Examples are provided that shed light into the effect of information embedding on the actions for the action-dependent source and channel coding problems.

B. Ahmadi and O. Simeone are with the Center for Wireless Communications and Signal Processing Research (CWCSPR), ECE Department, New Jersey Institute of Technology (NJIT), Newark, NJ 07102, USA (email: behzad.ahmadi@njit.edu, osvaldo.simeone@njit.edu). H. Asnani is with Information Systems Lab (ISL), Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA (email: asnani@stanford.edu). H. H. Permuter is with the Department of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva 84105, Israel (e-mail: haimp@bgu.ac.il)

The work of O. Simeone was supported in part by the U.S. National Science Foundation under Grant No. 0914899. H. Asnani was supported by Scott A. and Geraldine D. Macomber Stanford Graduate Fellowship. H. H. Permuter was supported in part by the Marie Curie Reintegration fellowship.
Index Terms

Action-dependent source coding, action-dependent channel coding, block Markov decoding, cribbing, forward encoding, information embedding, source-channel separation, side information, side information vending machine.

I. INTRODUCTION

The recent works [1], [2] study the problem of optimal actuation for source and channel coding for resource-constrained systems. Specifically, in [1], an extension of the Wyner-Ziv source coding problem is considered in which the decoder or the encoder can take actions that affect the quality of the side information available at the decoder’s side. When the actions are taken by the decoder, the latter operates in two stages. In the first stage, based on the message received from the encoder, the decoder selects cost-constrained actions $A$ that affect the measurement of the side information $Y$. This effect is modelled by a channel $p_{Y|X,A}(y|x,a)$, where $X$ represents the source available at the encoder. In the second stage, the decoder produces an estimate of source $X$ based on the side information $Y$ as in the standard Wyner-Ziv problem (see, e.g., [3]). A similar formulation also applies when the actions are taken at the encoder’s side. This model can account, as an example, for computer networks in which the acquisition of side information from remote data bases is costly in terms of system resources and thus should be done efficiently. We refer to this class of problems as having actions for side information acquisition.

In [2], a related channel coding problem is studied in which the encoder in a point-to-point channel can take actions to affect the state of a channel. The encoder operates in two stages. In the first stage, based on the message to be conveyed to the decoder, cost-constrained actions $A$ are selected by the encoder that affect the channel state $S$ of the channel $p_{Y|X,S}(y|x,s)$ used for communication to the decoder in the second stage. In the second stage, the channel $p_{Y|X,S}(y|x,s)$ is used in a standard way based on the available information about the state $S$ (which can be non-causal or causal, see, e.g., [3]). We refer to this problem as having actions for channel state control. As shown in [4], this model can be used to account for an encoder that in the first stage probes the channel to acquire state information.
A. Information Embedding on Actions

As discussed above, *optimal actuation for channel and source coding*, as proposed in [1], [2], prescribes the selection of the actions $A$ towards the goal of improving the performance of the resource-constrained communication link between encoder and decoder. This can be done by acquiring side information in an efficient way for source coding problems, and by controlling or probing effectively the channel state for channel coding problems.

This work starts from the observations that the actions $A$ often entail the use of physical resources for communication within the system encompassing the link under study. For instance, acquiring information from a data base requires the receiver to exchange control signals with a server, and probing the congestion state of a network (modelled as a channel) requires transmission of training packets to the closest router. In all these cases, the “recipient” of the actions, e.g., the server or a router in the examples above, may request to obtain partial information about the source or message being communicated on the link. To illustrate this point, the server in the data base application might need to acquire some explicit information about the file being transmitted in the link before granting access to the server. Similarly, the router might need to obtain the header of the packet (message) that the transmitter intends to deliver to the end receiver.

In the scenarios discussed above, the action $A$ thus serves a double purpose: on the one hand, it should be designed to improve the performance of the communication link at hand as in [1], [2], [4], and, on the other, it should provide explicit information about source or message for a separate decoder (the server or router in the examples above). A relevant question thus is: *How much information can be embedded in the actions $A$ without affecting the performance of the link?* Or, to turn the question around, *what is the performance loss for the link as a function of the amount of information that is encoded in the actions $A$?* This work aims at answering these questions for both the source and channel coding scenarios discussed above (see Fig. 1, Fig. 2, Fig. 3 and Sec. I-C).

B. Related Work

The interplay between communication and actuation, or control, is recognized to arise at different levels. As mentioned, the main theme in the papers [1], [2], [4] is “*control for communication*”: in [1], [2], [4], actuation is instrumental in improving the performance of a resource-
Figure 1. Source coding with decoder-side actions for information acquisition and with information embedding on actions. A function of the actions $f(A^n) = (f(A_1), ..., f(A_n))$ is observed in full (“non-causally”) by Decoder 2 before decoding. See Fig. 4 and Fig. 5 for the corresponding models with strictly causal and causal observation of the actions at Decoder 2, respectively.

constrained communication system. Extensions of this research direction include models with additional design constraints [5], [6], with adaptive actions [7], with memory [8], [6] and with multiple terminals [9],[14]. Somewhat related, but distinct, is the line of work including [15]-[16], in which control-theoretic tools are leveraged to design effective communication schemes. An altogether different theme is instead central in work such as [17], [18] that can be referred to as “communication for control”. In fact, in a reversed way, in [17], [18] (and references therein), communication is instrumental in carrying out control tasks such as stabilization of a plant. For instance, [18] shows that an implicit message communicated between two controllers can greatly improve the performance of the control task.

The idea of embedding information in the actions is related to the classical problem of information hiding (see, e.g., [19] and references therein). In information hiding, a message is embedded in a host data under distortion constraints. The message is then retrieved by a decoder that observes the host signal through a noisy channel. Note that the (host) signal onto which the message is embedded is a given process. Instead, in the set-up of information embedding on actions considered here, the (action) signal on which information is embedded is designed to optimize the given communication task.

The set-up at hand is also related to the source coding model of [20], in which an encoder communicates to two decoders and one of the decoders is able to observe the source estimate produced by the other. For its duality with the classical channel coding model studied in [22], the operation of the first decoder was referred to in [20] as cribbing. Although the problem
of interest here (in the source coding part) is significantly different, in that the recipient of the embedded information is a decoder “cribbing” the actions and not the estimates of another decoder, the solutions of the two problems turn out to be related, as it will be discussed.

C. Contributions and Paper Organization

The main contributions of this paper are as follows.

- **Decoder-side actions for side information acquisition:** We first consider the model in Fig. 1 in which the problem of source coding with actions taken at the decoder (Decoder 1) is generalized by including an additional decoder (Decoder 2). Decoder 2 is the recipient of a function of the action sequence and is interested in reconstructing a lossy version of the source measured at the encoder. A single-letter characterization of the set of all achievable tuples of rate, distortions at the two decoders and action cost is derived in Sec. II under the assumption that Decoder 2 observes a function of the actions non-causally (Sec. II-B), strictly causally (Sec. II-C) or causally (Sec. II-D). An example is provided to shed light
into the effect of information embedding on actions in Sec. II-E.

- **Encoder-side actions for side information acquisition:** We then consider the set-up in Fig. 2 in which an additional decoder observing the actions is added to the problem of source coding with actions taken at the encoder. Sec. III derives the achievable rate-distortion-cost region in the special case in which the channel \( p_{Y|X,A}(y|x,a) \) with source and action \((X,A)\) as inputs and side information \( Y \) as output is such that \( Y \) is a deterministic function of \( A \);

- **Actions for channel control and probing:** Finally, we consider the impact of information embedding on actions for channel control by studying the set-up in Fig. 3 which generalizes [2]. Specifically, a decoder (Decoder 1) is added to the model in [2], that observes a function of the actions taken by the encoder and wishes to decode part of the message that is intended for the channel decoder (Decoder 2). A single-letter characterization of the achievable capacity-cost region is obtained in Sec. IV. Finally, the special case of actions for channel probing [4] is elaborated on with an example in Sec. IV-C.

II. **Decoder-Side Actions for Side Information Acquisition**

In this section, we first describe the system model for the set-up illustrated in Fig. 1, Fig. 4 and Fig. 5 of source coding with decoder-side actions. Then, a single letter characterization of the set of all achievable tuples of rate, distortions at the two decoders and action cost is derived under the assumption that Decoder 2 observes the actions fully (non-causally) in Sec. II-B, strictly causally in Sec. II-C and causally in Sec. II-D. An example is provided in Sec. II-E.

A. **System Model**

We present here the problem corresponding to full observation of a function of the actions as per Fig. 1. We refer to this model as having non-causal action observation. The changes necessary to account for causal or strictly causal as illustrated in Fig. 4 and Fig. 5 will be discussed in the appropriate sections later. It is remarked that this definition does not entail any non-causal operation, but only a larger estimation delay for Decoder 2 as compared to the causal cases in Fig. 4 and Fig. 5. The model is defined by the probability mass functions (pmfs) \( p_X(x) \) and \( p_{Y|AX}(y|a,x) \), by the function \( f: \mathcal{A} \to \mathcal{B} \), and by discrete alphabets \( \mathcal{X}, \mathcal{Y}, \mathcal{A}, \mathcal{B}, \hat{X}_1, \hat{X}_2, \mathcal{Y}_1, \mathcal{Y}_2 \).
as follows. The source sequence $X^n$ is such that $X_i \in \mathcal{X}$ for $i \in [1, n]$ is independent and identically distributed (i.i.d.) with pmf $p_X(x)$. The Encoder measures sequence $X^n$ and encodes it in a message $M$ of $nR$ bits, which is delivered to Decoder 1. Decoder 1 receives message $M$ and selects an action sequence $A^n$, where $A^n \in A^n$. The action sequence affects the quality of the measurement $Y^n$ of sequence $X^n$ obtained at the Decoder 1. Specifically, given $A^n = a^n$ and $X^n = x^n$, the sequence $Y^n$ is distributed as $p(y^n|a^n, x^n) = \prod_{i=1}^{n} p_{Y|A,X}(y_i|a_i, x_i)$. The cost of the action sequence is defined by a cost function $\Lambda: A \rightarrow [0, \Lambda_{\text{max}}]$ with $0 \leq \Lambda_{\text{max}} < \infty$, as $\Lambda(a^n) = \sum_{i=1}^{n} \Lambda(a_i)$. The estimated sequence $\hat{X}_1^n \in \hat{X}_1^n$ is then obtained as a function of $M$ and $Y^n$. Decoder 2 observes a function of the action sequence $A^n$, thus obtaining $f(A^n) = (f(A_1), \ldots, f(A_n)) \in B^n$. Based on $f(A^n)$, Decoder 2 obtains an estimate $\hat{X}_2^n \in \hat{X}_2^n$ within given distortion requirements. The estimated sequences $\hat{X}_j^n$ for $j = 1, 2$ must satisfy distortion constraints defined by functions $d_j(x, \hat{x}_j): \mathcal{X} \times \hat{X}_j \rightarrow [0, D_{j,\text{max}}]$ with $0 \leq D_{j,\text{max}} < \infty$ for $j = 1, 2$, respectively. A formal description of the operations at encoder and decoder follows.

**Definition 1.** An $(n, R, D_1, D_2, \Gamma)$ code for the set-up of Fig. 1 consists of a source encoder

$$h^{(e)}: \mathcal{X}^n \rightarrow [1, 2^{nR}],$$

which maps the sequence $X^n$ into a message $M$; an “action” function

$$h^{(a)}: [1, 2^{nR}] \rightarrow A^n,$$

which maps the message $M$ into an action sequence $A^n$; two decoders, namely

$$h_1^{(d)}: [1, 2^{nR}] \times Y^n \rightarrow \hat{X}_1^n,$$

which maps the message $M$ and the measured sequence $Y^n$ into the estimated sequence $\hat{X}_1^n$;

$$h_2^{(d)}: B^n \rightarrow \hat{X}_2^n,$$

which maps the observed sequence $f(A^n)$ into the the estimated sequence $\hat{X}_2^n$, such that the action cost constraint $\Gamma$ and distortion constraints $D_j$ for $j = 1, 2$ are satisfied, i.e.,

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ \Lambda(A_i) \right] \leq \Gamma$$

and

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left[ d_j(X_{ji}, \hat{X}_{ji}) \right] \leq D_j \text{ for } j = 1, 2.$$
Definition 2. Given a distortion-cost tuple \((D_1, D_2, \Gamma)\), a rate \(R\) is said to be achievable if, for any \(\epsilon > 0\), and sufficiently large \(n\), there exists a \((n, R, D_1 + \epsilon, D_2 + \epsilon, \Gamma + \epsilon)\) code.

Definition 3. The rate-distortion-cost function \(R(D_1, D_2, \Gamma)\) is defined as \(R(D_1, D_2, \Gamma) = \inf \{ R : \text{the tuple } (R, D_1, D_2, \Gamma)\text{ is achievable} \}\).

In the rest of this section, for simplicity of notation, we drop the subscripts from the definition of the pmfs, thus identifying a pmf by its argument.

B. Non-Causal Action Observation

In this section, a single-letter characterization of the rate-distortion region is derived for the set-up in Fig. 1 in which Decoder 1 observes the entire sequence \(f^n(A^n)\) prior to decoding.

Proposition 1. The rate-distortion-cost function \(R(D_1, D_2, \Gamma)\) for the source coding problem with decoder-side actions and non-causal observation of the actions at Decoder 2 illustrated in Fig. 2 is given by

\[
R(D_1, D_2, \Gamma) = \min_{p(\hat{x}_2, a, u|x), g(U,Y)} \left[ I(X; \hat{X}_2, A) + I(X; U|\hat{X}_2, A, Y) \right],
\]

where the mutual information is evaluated with respect to the joint pmf

\[
p(x, y, a, \hat{x}_2, u) = p(x)p(\hat{x}_2, a, u|x)p(y|x, a),
\]

for some pmf \(p(\hat{x}_2, a, u|x)\) such that the inequalities

\[
E[d_j(X, \hat{X}_j)] \leq D_j, \text{ for } j = 1, 2, \quad (9a)
\]

\[
E[\Lambda(A)] \leq \Gamma, \quad (9b)
\]

and

\[
I(X; \hat{X}_2, f(A)) \leq H(f(A)) \quad (9c)
\]

are satisfied for \(\hat{X}_1 = g(U,Y)\). Finally, \(U\) is an auxiliary random variable whose alphabet cardinality can be constrained as \(|U| \leq |X| |\hat{X}_2| |A| + 1\) without loss of optimality.

At an intuitive level, in (7), the term \(I(X; \hat{X}_2, A)\) accounts for the rate needed to instruct Decoder 1 about the actions \(A\) to be taken for the acquisition of the side information \(Y\), which are selected on the basis of the source \(X\), and, at the same time, to communicate the
reconstruction $\hat{X}_2$ to Decoder 2. The additional rate $I(X; U|\hat{X}_2, A, Y)$ is instead required to refine the description of the source $X$ provided via $(\hat{X}_2, A)$ using an auxiliary codebook $U$ for Decoder 1. Note that this rate is conditioned on the side information $Y$, thanks to the rate saving obtained through Wyner-Ziv binning. The condition (9c) ensures that, based on the observation of $f(A)$, Decoder 2 is able to reconstruct $\hat{X}_2$. The details of achievability follow as a combination of the techniques proposed in [1] and [24], [20]. Below we briefly outline the main ideas, since the technical details follow from standard arguments. The proof of the converse is provided in Appendix A.

**Sketch of the achievability proof:** We fix a pmf (8) and define a random variable $B = f(A)$. The joint pmf $p(x, y, a, \hat{x}_2, u, b)$ of variables $(X, Y, A, \hat{X}_2, U, B)$ is obtained by multiplying the right-hand side of (8) by the term $1_{\{b = f(a)\}}$. In the scheme at hand, the Encoder first maps sequence $X^n$ into a sequence $\hat{X}_2^n \in \hat{X}_2^n$ using the joint typicality criterion with respect to the joint pmf $p(x, \hat{x}_2)$. This mapping requires a codebook of rate $I(X; \hat{X}_2)$ (see, e.g., [3, pp. 62-63]). Given the sequence $\hat{X}_2^n$, the sequence $X^n$ is further mapped into a sequence $B^n \in B^n$ using the joint typicality criterion with respect to the joint pmf $p(x, b|\hat{x}_2)$ where $B = f(A)$, which requires a codebook of rate $I(X; f(A)|\hat{X}_2)$ for each sequence $\hat{X}_2^n$. For later reference, we refer to every such codebook as a bin in the following. Note that we have one bin for every sequence $\hat{X}_2^n$. For each pair $(\hat{X}_2^n, B^n)$, the sequence $X^n$ is mapped into an action sequence $A^n$ using joint typicality with respect to the joint pmf $p(x, a|\hat{x}_2, b)$, which requires a codebook of rate $I(X; A|\hat{X}_2, f(A))$. Note that, by construction, we have that $B^n = f(A^n)$ for each generated $A^n$. Finally, the source sequence $X^n$ is mapped into a sequence $U^n$ using the joint typicality criterion with respect to the joint pmf $p(x, u|\hat{x}_2, a)$, which requires a codebook of rate $I(X; U|\hat{X}_2, A)$ for each pair $(\hat{X}_2^n, A^n)$.

The indices of codewords $\hat{X}_2^n, B^n$ and $A^n$ are sent to Decoder 1, along with the index for the codeword $U^n$. For the latter, by leveraging the side information $Y^n$ available at Decoder 1, the rate can be reduced to $I(X; U|\hat{X}_2, A, Y)$ by the Wyner-Ziv theorem [3, p. 280]. Decoder 2 estimates the sequence $\hat{X}_2^n$ from the observed sequence $f(A^n)$ as follows: if there is only one bin containing the observed sequence $f(A^n)$, then $\hat{X}_2^n$ equals the sequence corresponding to such bin (recall that each bin corresponds to one sequence $\hat{X}_2^n$). Otherwise, an error is decoded. To obtain a vanishing

---

1The notation $1_{\{S\}}$ is used for the indicator function of the event $S$. 

May 2, 2014 DRAFT
probability of error, the sequence $f^n(A^n)$ should thus not lie within more than one bin with high probability. The probability of the latter event can be upper bounded by $2^n(I(X;\hat{X}_2,f(A)) - H(f(A)))$ since each sequence $B^n$ is generated with probability approximately $2^{-nH(f(A))}$ and there are $2^nI(X;\hat{X}_2,f(A))$ sequences $B^n$ [20]. Therefore, as long as $I(X;\hat{X}_2,f(A)) \leq H(f(A))$, Decoder 1 is able to infer the conveyed bin index with high probability. Finally, Decoder 1 produces the estimate $\hat{X}_1^n$ through a symbol-by-symbol function as $\hat{X}_i = g(U_i, Y_i)$ for $i \in [1,n]$. □

Remark 1. Assume that the action $A_i$ is allowed to be a function, not only of the message $M$ as per (2), but also of the previous values of the side information $Y_{i-1}$, which we refer to as adaptive actions. Then, the rate-distortion-cost function derived in Proposition 2 can generally be improved. This can be seen by considering the case in which $R = 0$. In this case, if the actions were selected as per (2), then the distortion at Decoder 2 would be forced to be maximal, i.e., $D_2 = D_{2,\max}$, since the actions $A$ cannot depend in any way on the source $X$. Instead, by selecting $A$ as a function of the previously observed values of $Y$, Decoder 1 can provide Decoder 2 with information about $X$, thus decreasing the distortion $D_2$. It is noted that the usefulness of adaptive actions in this setting contrasts with the known fact that, in the absence of Decoder 2, adaptive actions do not decrease the rate-distortion function [7].

C. Strictly Causal Action Observation

The system model for the set-up in Fig. 4 is similar to the one described in Sec. II-A with the only difference the decoding function for Decoder 2 a time $i$ is given as

$$h_{2i}^{(d)}: B^{i-1} \rightarrow \hat{X}_2,$$

which maps the strictly causally observed sequence $f(A^{i-1}) = (f(A_1), ..., f(A_{i-1}))$ into the $i$th estimated symbol $\hat{X}_{2i}$.

Proposition 2. The rate-distortion-cost function $R(D_1, D_2, \Gamma)$ for the source coding problem with decoder-side actions and strictly causal observation of the actions at Decoder 1 as illustrated in Fig. 4 is given by

$$R(D_1, D_2, \Gamma) = \min_{p(\hat{X}_2, a, u|x), g(U,Y)} I(X;\hat{X}_2, A) + I(X;U|\hat{X}_2, A, Y),$$

where the mutual information is evaluated with respect to the joint pmf

$$p(x, y, a, \hat{x}_2, u) = p(x)p(\hat{x}_2, a, u|x)p(y|x, a),$$

May 2, 2014 DRAFT
Figure 4. Source coding with decoder-side actions for information acquisition and with information embedding on actions. At time $i$, Decoder 2 has available the samples $f(A^{i-1}) = (f(A_1), ..., f(A_{i-1}))$ in a strictly causal fashion.

For some pmf $p(\hat{x}_2, a, u|x)$ such that the inequalities

\begin{align}
\text{E}[d_j(X, \hat{X}_j)] & \leq D_j, \text{ for } j = 1, 2, \tag{13a} \\
\text{E}[\Lambda(A)] & \leq \Gamma, \tag{13b} \\
\text{and } I(X; \hat{X}_2, f(A)) & \leq H(f(A)|\hat{X}_2) \tag{13c}
\end{align}

are satisfied for $\hat{X}_1 = g(U, Y)$ for some function $g: \mathcal{U} \times \mathcal{Y} \rightarrow \hat{X}_1$. Finally, $U$ is an auxiliary random variable whose alphabet cardinality can be constrained as $|U| \leq |X||\hat{X}_2||A| + 1$ without loss of optimality.

The only difference between the rate-distortion-cost function of Proposition 1 with non-causal action observation with respect to the case with strictly causal action observation of Proposition 2 is the constraint (13c). Recall that the latter is needed to ensure that Decoder 2 is able to recover the reconstruction $\hat{X}_2$. As detailed below, the strict causality of the observation of the action at Decoder 2 calls for a block-based encoding in which the actions carries information about the source sequence as observed in two different blocks, namely the current block for Decoder 1 and the future block for Decoder 2. This additional requirement causes the conditioning on $\hat{X}_2$ in (13c), which generally increases the rate (11) with respect to the counterpart (20) achievable with non-causal action observation. A sketch of the achievability proof is provided below and is based on the techniques proposed in [24], [20] (see also [21]). The proof of the converse is provided in Appendix B.

**Sketch of the achievability proof:** We fix a pmf (12) and define a random variable $B = f(A)$. The joint pmf $p(x, y, a, \hat{x}_2, u, b)$ of variables $(X, Y, A, \hat{X}_2, U, B)$ is obtained by multiplying the
right-hand side of (12) by the term \(1_{\{b=f(a)\}}\). We use the “Forward Encoding” and “Block Markov Decoding” strategy of [24], [20] (see also [21]) and combine it with the coding scheme of [1]. The scheme operates over multiple blocks and we denote by \(X^n(l)\) the portion of the source sequence encoded in block \(l\). The sequence \(f(A^n(l))\) observed during block \(l\) is used in block \(l+1\) by Decoder 2 due to the strict causality constraint. To this end, the action sequence \(f(A^n(l))\) produced in block \(l\) must carry information about the source sequence \(X^n(l+1)\) corresponding to the next block \(l+1\). Note that this is possible since encoder knows the entire sequence \(X^n\). At the same time, sequence \(A^n(l)\) should also perform well as an action sequence to be used by Decoder 1 to estimate sequence \(X^n(l)\) for the current block. This is accomplished as follows; In each block \(l\), \(2^{nI(X;\hat{X}_2)}\) codewords \(\hat{X}_2^n \in \hat{X}_2^n\) are generated according to the pmf \(p(\hat{x}_2)\). Next, \(2^{nI(X;\hat{X}_2)}\) bins are assigned to each codeword \(\hat{X}_2^n\), where each bin contains \(2^{nI(X;A|\hat{X}_2)}\) codewords \(B^n \in B^n\), generated according to pmf \(p(b|\hat{x}_2)\). For each pair \((\hat{X}_2^n, B^n)\), a codebook of \(2^{nI(X;A|\hat{X}_2;f(A))}\) codewords \(A^n \in A^n\) is generated according to the joint pmf \(p(x, a|\hat{x}_2, b)\). Finally, a codebook of \(2^{nI(X;U|\hat{X}_2;A)}\) codewords \(U^n \in U^n\) is generated according to the joint pmf \(p(x, u|\hat{x}_2, a)\). The latter codebook is further binned into a codebook of rate \(I(X; U|\hat{X}_2, A, Y)\) to leverage the side information \(Y^n\) available at Decoder 1 via the Wyner-Ziv theorem [3, p. 280].

For encoding, in each block \(l\), a sequence \(\hat{X}_2^n\) is selected from the \(\hat{X}_2\)—codebook of block \(l\) to be jointly typical with the source sequence \(X^n(l)\) in the current block. Instead, the bin index describes a \(\hat{X}_2^n\) sequence in the \(\hat{X}_2\)—codebook of block \((l + 1)\)th that is jointly typical with the source sequence \(X^n(l+1)\) of the \((l + 1)\)th block. Moreover, given \(\hat{X}_2^n\) and the bin index, a sequence \(A^n\) is chosen such that \((A^n, X^n(l))\) are jointly typical. Similarly, a sequence \(U^n\) is selected for block \(l\) to be jointly typical with the sequence of \(X^n(l)\) of block \(l\).

Thanks to the observation of the actions, at block \(l+1\) Decoder 2 knows the functions \(f(A^n(l))\), and aims to find the bin index in which the corresponding codeword \(B^n\) lies. As shown in [20], this is possible with vanishing probability of error, if \(I(X; \hat{X}_2, f(A)) \leq H(f(A)|\hat{X}_2)\). Note that the conditioning in the right-hand side is due to the fact that the sequences \(B^n\) are generated conditioned on the sequence \(\hat{X}_2^n\) representing a compressed version of the source for the current block \(l\). The latter does not bring any information regarding the desired sequence \(X^n(l + 1)\).□

Remark 2. From the proof of the converse in Appendix B, it follows, similarly to [7], that, adaptive actions (see Remark [1]) do not increase the rate-distortion-cost function derived in May 2, 2014 DRAFT
Figure 5. Source coding with decoder-side actions for information acquisition and with information embedding on actions. At time $i$, Decoder 2 has available the samples $f(A^i) = (f(A_1), ..., f(A_i))$ in a causal fashion.

Proposition 2.

**D. Causal Action Observation**

The system model for the set-up in Fig. 5 is similar to the one described in Sec. II-A with the only difference the decoding function for Decoder 2 is

$$h_{2i}^{(d)} : \mathcal{B}^i \rightarrow \hat{X}_2,$$

which maps the causally observed sequence $f(A^i) = (f(A_1), ..., f(A_i))$ into the $i$th estimated symbol $\hat{X}_{2i}$.

**Proposition 3.** The rate-distortion-cost function $R(D_1, D_2, \Gamma)$ for the source coding problem with decoder-side actions and causal observation of the actions illustrated in Fig. 5 is given by

$$R(D_1, D_2, \Gamma) = \min_{p(v,a,u|x), g_1(U,Y), g_2(V,f(A))} I(X; V, A) + I(X; U|V, A, Y),$$

where the mutual information is evaluated with respect to the joint pmf

$$p(x, y, a, u, v) = p(x)p(v, a, u|x)p(y|x, a),$$

for some pmf $p(v, a, u|x)$ such that the inequalities

$$\text{E}[d_j(X, \hat{X}_j)] \leq D_j, \text{ for } j = 1, 2,$$

$$\text{E}[\Lambda(A)] \leq \Gamma,$$

and $I(X; V, f(A)) \leq H(f(A)|V)$

May 2, 2014
are satisfied for $\tilde{X}_1 = g_1(U, Y)$ and $\tilde{X}_2 = g_2(V, f(A))$ with some functions $g_1: U \times Y \to \tilde{X}_1$ and $g_2: V \times B \to \tilde{X}_2$, respectively. Finally, $U$ and $V$ are auxiliary random variables whose alphabet cardinalities can be constrained as $|U| \leq |X||Y||A| + 1$ and $|V| \leq |X| + 3$, respectively, without loss of optimality.

The difference between the rate-distortion-cost function above with causal and strictly causal action observation is given by the fact that, with causal action observation, Decoder 2 can use the current value of the function $f(A)$ for the estimate of $\tilde{X}_2$. This is captured by the fact that the Encoder provides Decoder 2 with an auxiliary source description $V$, which is then combined with $f(A)$ via a function $\tilde{X}_2 = g_2(V, f(A))$ to obtain $\tilde{X}_2$. The rate (15) and the constraint (17c) are changed accordingly. The proof of the converse is given in Appendix B.

Remark 3. As seen in Appendix B, with adaptive actions, the rate-distortion-cost function derived in Proposition 3 remains unchanged.

E. Binary Example

In this section, an example is provided to illustrate the effect of the communication requirements of the additional decoder (Decoder 2) that observes a function of the actions on the system performance. We assume binary alphabets as $\mathcal{X} = \mathcal{A} = \mathcal{Y} = \{0, 1\}$ and a source distribution $X \sim \text{Bern}(\frac{1}{2})$. The distortion metrics are assumed to be Hamming, i.e., $d_j(x, \hat{x}_j) = 0$ if $x = \hat{x}_j$ and $d_j(x, \hat{x}_j) = 1$ otherwise for $j = 1, 2$. Moreover, as shown in Fig. 6 the side information $Y$ at Decoder 1 is observed through a Z-channel for $A = 0$ or an S-channel for $A = 1$. We assume no cost constraint on the actions taken by Decoder 1 (which can be enforced by choosing $\Lambda(A) = A$ and $\Gamma = 1$), and we set $f(A) = A$. The example extends that of [1, Sec. II-D] to a set-up with the additional Decoder 2. Under the requirement of lossless reconstruction at Decoder 1, i.e., $D_1 = 0$, the rate-distortion-cost function $R(0, D_2, \Gamma = 1)$ with non-causal action observation is obtained from Proposition 1 by setting $U = \tilde{X}_1 = X$, obtaining

$$R(0, D_2, 1) = \min_{p(\hat{x}_2, a|x)} I(X; \hat{X}_2, A) + H(X|\hat{X}_2, A, Y),$$

(18)

where the minimization is done under the constraints $E[d_2(X, \hat{X}_2)] \leq D_2$ and $I(X; \hat{X}_2, A) \leq H(A)$. 

May 2, 2014 DRAFT
The minimization in (18) can be done over the parameters \( p(a = 1, \hat{x}_2 = 0|x = 0) \triangleq \alpha_1 \), \( p(a = 1, \hat{x}_2 = 1|x = 0) \triangleq \alpha_2 \) and \( p(a = 0, \hat{x}_2 = 1|x = 0) \triangleq \alpha_3 \) with \( \sum_{i=1}^{3} \alpha_i \leq 1 \) and \( \alpha_i \geq 0 \) for \( i = 1, 2, 3 \), since by symmetry, we can set \( p(a = 0, \hat{x}_2 = 1|x = 1) = \alpha_1 \), \( p(a = 0, \hat{x}_2 = 0|x = 1) = \alpha_2 \) and \( p(a = 1, \hat{x}_2 = 0|x = 1) = \alpha_3 \) without loss of optimality. Explicit expressions can be easily found and have been optimized numerically.

Fig. 7 depicts the rate-distortion function versus the distortion \( D_2 \) of Decoder 2 for values of \( \delta = 0.2, \delta = 0.5 \) and \( \delta = 0.8 \). It can be seen that if the distortion \( D_2 \) tolerated by Decoder 2 is sufficiently large (e.g., \( D_2 \geq 0.4 \) for \( \delta = 0.5 \)), then the communication requirements of Decoder 2 do not increase the required rate. This can be observed by comparing the rate \( R(0, D_2, \Gamma) \) with rate \( R(0, 0.5, \Gamma) \) corresponding to a distortion level \( D_2 = 0.5 \), which requires no communication to Decoder 2. The smallest distortion \( D_2 \) that does not affect the rate can be found as \( D_2 = \alpha_{2,\text{opt}} + \alpha_{3,\text{opt}} \), where \( \alpha_{2,\text{opt}} \) and \( \alpha_{3,\text{opt}} \) are the optimal values for problem (18) with \( D_2 = 0.5 \) that minimizes \( \alpha_2 + \alpha_3 \).

We now compare the performance between non-causal action observation, as considered above, and strictly causal action observation. The performance in the latter case can be obtained from Proposition 2 and leads to (18) with the more restrictive constraint (13c). Fig. 8 plots the difference between rate-distortion function (18) for the case of non-causal and strictly causal action observation versus \( \delta \) for three values of distortion, namely \( D_2 = 0.1, 0.2, 0.3 \). As shown, irrespective of the value of distortion \( D_2 \), for values of \( \delta = 0 \) and \( \delta = 1 \), the performance with non-causal action observation is equal to that with strictly causal observation. This is due to the facts that: i) for \( \delta = 0 \), the side information \( Y \) is a noiseless measure of the source sequence \( X \).
for both $A = 0$ and $A = 1$ and thus there is no gain in making the actions at Decoder 1 to be dependent of $X$, and thus $\hat{X}_2$; ii) for $\delta = 1$, the side information $Y$ is independent of the source sequence $X$ given both $A = 0$ and $A = 1$, and thus it is without loss of optimality to choose actions at Decoder 1 to be independent of $X$ and $\hat{X}_2$. We can conclude that for both $\delta = 0$ and $\delta = 1$, causal action observation, and in fact even selecting $A$ to be independent of $X$, does not entail any performance loss. Instead, for values $0 < \delta < 1$, it is generally advantageous for Decoder 1 to select actions correlated with the source $X$, and hence some performance loss is observed with strictly causal action observation owing to the more restrictive constraint (13c). This reflects the need to cater to both Decoder 1 and Decoder 2 when selecting actions $A$, which requires description of two different source blocks. Following similar arguments, it is also noted that, as the communication requirements for Decoder 2 become more pronounced, i.e., as $D_2$ decreases, the difference between the rate-distortion function with non-causal and strictly-causal action observation increases. The performance with causal action observation is intermediate between full and strictly causal observation, and it is not shown here.
Figure 8. Difference between the rate-distortion function (18) with non-causal (NC) and strictly causal (SC) action observation versus \( \delta \) for values of distortion \( D_2 = 0.1, 0.2 \) and 0.3.

### III. Encoder-Side Actions for Side Information Acquisition

In the previous section, the actions controlling the quality and availability of the side information were taken by the decoder. In this section, following [1, Sec. III], we consider instead scenarios in which the encoder takes the actions affecting the side information of Decoder 1, as shown in Fig. 2. Specifically, the encoder takes actions \( A^n \in \mathcal{A}^n \), thus influencing the side information available to the Decoder 1 through a discrete memoryless channel \( p(y|x,a) \). Decoder 2 observes the action sequence to obtain the deterministic function \( f(A^n) = (f(A_1),...,f(A_n)) \), or the corresponding causal and strictly causal function, which is used to estimate the source sequence subject to a distortion constraint.

An \((n,R,D_1,D_2,\Gamma)\) code is defined similar to the previous sections with the difference that the action encoder (2) maps directly the source sequence \( X^n \) into the action sequence \( A^n \), i.e.,

\[
h^{(a)}: \mathcal{X}^n \rightarrow \mathcal{A}^n.
\] (19)

As discussed in [1], even in the absence of Decoder 2, the problem at hand is challenging. We thus focus on certain special cases, first the special case in which the side information channel \( p(y|x,a) \) is such that \( Y \) is a deterministic function of \( A \), i.e., \( Y = f_Y(A) \), and \( f(A) = A \). This is
solved in Proposition 4 and generalized by the following remark to the case of all deterministic function \( f \) for which \( H(f_Y(A)|f(A)) = 0 \). Following Proposition 4 is Proposition 5 where the case \( H(f(Y)|f_Y(A)) = 0 \) is solved.

**Proposition 4.** The rate-distortion-cost function \( R(D_1, D_2, \Gamma) \) for the source coding problem with encoder-side actions, non-causal, causal or strictly causal observation of the actions illustrated in Fig. 2 with \( f(A) = A \) and \( Y = f_Y(A) \) is given by

\[
R(D_1, D_2, \Gamma) = \min_{p(u|x), p(\hat{x}_1|u,x), p(\hat{x}_2|u,x), p(a)} \left\{ I(X; \hat{X}_1, U) - H(f_Y(A)) \right\}^+, 
\]

where the information measures are evaluated with respect to the joint pmf

\[
p(x, u, \hat{x}_1, \hat{x}_2, a) = p(x)p(u|x)p(\hat{x}_1|u,x)p(\hat{x}_2|u,x)p(a),
\]

for some pmfs \( p(u|x), p(\hat{x}_1|u,x), p(\hat{x}_2|u,x) \) and \( p(a) \) such that the inequalities

\[
\begin{align*}
E[d_j(X, \hat{X}_j)] & \leq D_j, \text{ for } j = 1, 2, \\
E[\Lambda(A)] & \leq \Gamma, \\
I(X; U) & \leq H(f_Y(A)) \\
\text{and } I(X; \hat{X}_2|U) & \leq H(A|f_Y(A))
\end{align*}
\]

are satisfied. Finally, \( U \) is an auxiliary random variable whose alphabet cardinality can be constrained as \( |U| \leq |X||\hat{X}_1||\hat{X}_2| + 3 \) without loss of optimality.

**Remark 4.** The results above generalizes a number of known single-letter characterizations. Notably, if \( D_2 = D_{2,\text{max}} \), so that the distortion requirements of Decoder 2 are immaterial to the system performance, the result reduces to [1, Theorem 7]. Moreover, in the special case in which \( A = (A_0, A_2), Y = A_0, R = R_1, |A_0| = 2^{R_0}, |A_2| = 2^{R_2} \), the model coincides with the lossy Gray-Wyner problem [23].

As detailed in the proof below, Proposition 4 establishes the optimality of separate source-channel coding for the set-up in Fig. 2 under the stated conditions. In particular, the encoder compresses using a standard successive refinement source code in which \( U \) represents the coarse description and \( \hat{X}_1, \hat{X}_2 \) two independent refinements. The indices of the coarse description \( U \) and

\[\text{2Note that here } 2^{R_0} \text{ and } 2^{R_2} \text{ are constrained to be integers.}\]
of the refined description $\hat{X}_2$ are sent on the degraded (deterministic) broadcast channel with input $A$ and outputs $(A, t(A))$ using superposition coding. Reliable compression and communication is guaranteed by the two bounds (22c)-(22d). A further refined description $\hat{X}_1$ is produced for Decoder 1, and the corresponding index is sent partly over the mentioned broadcast channel and partly over the link of rate $R$, leading to the rate (20). Details of the achievability proof can be found below, while the proof of the converse is given in Appendix C.

Remark 5. Following the discussion above, specializing Proposition 4 to the case $R = 0$ shows the optimality of source-channel coding separation for the lossy transmission of a source over a deterministic degraded broadcast channel (see [3, Chapter 14] for a review of scenarios in which the optimality of separation holds for lossless transmission over a broadcast channel).

Sketch of the achievability proof: As anticipated above, achievability uses the ideas of a source-channel coding separation, successive refinement and superposition coding. We only describe the outline, as the rigorous details can be derived based on standard techniques [3]. We start with the case of non-causal action observation at Decoder 2. Note that the deterministic channel with input $A$ and outputs $A$ (to Decoder 2) and $f_Y(A)$ (to Decoder 1) is not only deterministic but also degraded [3, Chapter 5]. This channel is used to send a common source description of rate $\tilde{R}_1$ to both the decoders and a refined description of rate $\tilde{R}_2$ to Decoder 2 only. To elaborate, fix the pmfs $p(u|x)$, $p(\hat{x}_1|u, x)$, $p(\hat{x}_2|u, x)$ and $p(a)$. Generate a codebook of $2^nI(X:U)$ sequences $U^n$ i.i.d. with the pmf $p(u)$ and, for each $U^n$ sequence, generate a codebook of $2^nI(X;\hat{X}_1|U)$ $\hat{X}_1^n$ sequences i.i.d. with pmf $p(\hat{x}_1|u)$ and a codebook of $2^nI(X;\hat{X}_2|U)$ sequences $\hat{X}_2^n$ i.i.d. with pmf $p(\hat{x}_2|u)$. Given a source sequence $X^n$, the encoder finds a jointly typical $U^n$ codeword, and then a codeword $\hat{X}_1^n$ jointly typical with $(X^n, U^n)$ and similarly for $\hat{X}_2^n$. Using source-channel separation on the broadcast “action” channel described above, the index from the $U$-codebook and a part of the index from the $\hat{X}_1$-codebook, of rate $r$, is described to both decoders, and the index from $\hat{X}_2$-codebook is described to Decoder 2 as its private information. Thus, we have the inequalities

$$\tilde{R}_1 \geq I(X;U) + r$$ (23a)
and
$$\tilde{R}_2 \geq I(X;\hat{X}_2|U).$$ (23b)

The capacity region of the broadcast channel is given by the conditions [3, Chapter 9], $\tilde{R}_1 \leq
$H(f_Y(A))$ and $\tilde{R}_1 + \tilde{R}_2 \leq H(A)$, and thus the following rates are achievable

\begin{align}
\tilde{R}_1 &\leq H(f_Y(A)) \quad (24a) \\
\text{and} \quad \tilde{R}_2 &\leq H(A|f_Y(A)). \quad (24b)
\end{align}

Finally the remaining part of the index of codeword $\hat{X}_1^n$ is sent through the direct rate $R$, leading to the condition

$$R \geq I(X; \hat{X}_1|U) - r. \quad (25)$$

Combining (23), (24) and (25), and using Fourier-Motzkin elimination we obtain

$$R \geq I(X; \hat{X}_1, U) - H(f_Y(A)) \quad (26)$$

and (29c)-(22d). The distortion and cost constraints are handled in a standard manner and hence the details are omitted.

It remains to discuss how to handle the case of causal or strictly causal action observation. Given the converse result in Appendix C, it is enough to show that (20)-(22) is achievable also with strictly causal and causal action observation. This can be simply accomplished by encoding in blocks as per achievability of Proposition 2 and Proposition 3. Specifically, in each block the encoder compresses the source sequence corresponding to the next block. Decoder 2 then operates as above, while Decoder 1 can recover all source blocks at the end of all blocks. $\Box$

**Remark 6.** The scenario solved above is when the action observation is perfect, i.e., $f(A) = A$. The result also carries verbatim for the more general case where $f(A)$ is a generic function as long as $H(f_Y(A)|f(A)) = 0$. The expressions of the rate region remain the same as in the proposition above except that $A$ is replaced by $f(A)$.

Proposition 4 characterizes the optimal performance for the case when Decoder 2 has a better information about the actions taken by the encoder than Decoder 1 in the sense that $H(f_Y(A)|f(A)) = 0$. We note here that a similar characterization can be given also for the dual setting in which $H(f(A)|f_Y(A)) = 0$ so that Decoder 1 has the better observation about the actions.

**Proposition 5.** The rate-distortion-cost function $R(D_1, D_2, \Gamma)$ for the source coding problem with encoder-side actions, non-causal, causal or strictly causal observation of the actions illustrated in Fig. 2 with $H(f(A)|f_Y(A)) = 0$, is given by

May 2, 2014 DRAFT
\[ R(D_1, D_2, \Gamma) = \min_{p(a), p(\hat{x}_1, \hat{x}_2|x)} \{ I(X; \hat{X}_1, \hat{X}_2) - H(f_Y(A)) \}^+, \]  
\[ \text{where the information measures are evaluated with respect to the joint pmf} \]
\[ p(x, \hat{x}_1, \hat{x}_2, a) = p(x)p(\hat{x}_1, \hat{x}_2|x)p(a), \]
\[ \text{such that the following inequalities are satisfied,} \]
\[ E[d_j(X, \hat{X}_j)] \leq D_j, \text{ for } j = 1, 2, \]
\[ E[\Lambda(A)] \leq \Gamma, \]
\[ I(X; \hat{X}_2) \leq H(f(A)). \]

The converse follows similarly as that for Proposition 4 where instead of \( U \) in the converse we use \( \hat{X}_2 \), as knowing \( Y^n = f_Y(A^n) \) implies knowing \( f(A^n) \), due to the assumption \( H(f(A)|f_Y(A)) = 0 \). We just outline the achievability for the non-causal case (the achievability for strictly causal and causal case uses block coding ideas as in Proposition 4). A successive refinement codebook is generated by drawing \( 2^{nI(X; \hat{X}_2)} \) codewords \( \hat{X}_2^n \), and, for each codeword \( \hat{X}_2^n \), a number \( 2^{nI(X; \hat{X}_1|\hat{X}_2)} \) of codewords \( \hat{X}_1^n \). As for Proposition 4 the indices of these two codebooks obtained via standard joint typicality encoding are sent through the degraded broadcast channel \( p(y, b|a) = 1_{\{y = f_Y(A^n), b = f(A^n)\}} \). Splitting the rate for the index of codeword \( \hat{X}_1^n \) so that a rate \( R \) is sent over the direct link to Decoder 1, reliability of compression and communication over the “action” broadcast channel is guaranteed if
\[ I(X; \hat{X}_2) \leq H(f(A)) \]
\[ I(X; \hat{X}_2) + I(X; \hat{X}_1|\hat{X}_2) - R \leq H(f(A), f_Y(A)) = H(f_Y(A)), \]
where the latter inequality implies \( R \geq I(X; \hat{X}_1, \hat{X}_2) - H(f_Y(A)) \). The proof is concluded using the usual steps. \( \square \)

IV. ACTIONS FOR CHANNEL STATE CONTROL AND PROBING

In this section, we consider the impact of information embedding on actions for the set-up of channel coding with actions of [2]. To this end, we consider the model in Fig. 3 in which Decoder 1, based on the observation of a deterministic function of the actions, wishes to retrieve

May 2, 2014
part of the information destined to Decoder 2. Note that for simplicity of notation here the additional encoder that observes the actions is denoted as Decoder 1, rather than Decoder 2 as done above. Also, we emphasize that in the original set-up of [2], Decoder 1 was not present.

A. System Model

The system is defined by the pmfs $p(x)$, $p(y|x,s,a)$, $p(s|a)$, function $f: A \rightarrow B$ and by discrete alphabets $\mathcal{X}, \mathcal{A}, \mathcal{B}, \mathcal{S}$, and $\mathcal{Y}$. Given the messages $(M_1, M_2)$, selected randomly from the set $\mathcal{M}_1 \times \mathcal{M}_2 = [1, 2^{nR_1}] \times [1, 2^{nR_2}]$, an action sequence $A^n \in \mathcal{A}^n$ is selected by the Encoder. Decoder 1 observes the signal $B^n = f(A^n)$ as a deterministic function of the actions, and estimates message $M_1$. Note that the notation here implies a “non-causal” observation of the actions, but it is easy to see that the results below hold also with causal and strictly causal observation of the actions. Moreover, the state sequence $S^n \in \mathcal{S}^n$ is generated as the output of a memoryless channel $p(s|a)$ and we have $p(b^n, s^n|a^n) = \prod_{i=1}^n p(s_i|a_i)1_{\{b_i = f(a_i)\}}$ for an action sequence $A^n = a^n$. The input sequence $X^n \in \mathcal{X}^n$ is selected on the basis of both messages $(M_1, M_2)$ and of the state sequence $S^n$ by the Encoder. The action sequence $A^n$ and the input $X^n$ have to satisfy an average cost constraint defined by a function $\gamma: \mathcal{A} \times \mathcal{X} \rightarrow [0, \infty)$, so that the cost for the input sequences $a^n$ and $x^n$ is given by $\gamma(a^n, x^n) = \frac{1}{n} \sum_{i=1}^n \gamma(a_i, x_i)$. Given $X^n = x^n$, $S^n = s^n$ and $A^n = a^n$, the received signal is distributed as $p(y^n|x^n, s^n, a^n) = \prod_{i=1}^n p(y_i|x_i, s_i, a_i)$.

Decoder 2, having received the signal $Y^n$, estimates both messages $(M_1, M_2)$.

The setting includes the semi-deterministic broadcast channel with degraded message sets [25] (see also [3, Ch. 8]) as a special case by setting $X$ to be constant and $Y = S$, and the channel with action-dependent states studied in [2] for $R_1 = 0$.

**Definition 4.** An $(n, R_0, R_1, \Gamma, \epsilon)$ code for the model in Fig. 3 consists of an action encoder

$$h^{(a)}: \mathcal{M}_1 \times \mathcal{M}_2 \rightarrow \mathcal{A}^n,$$  \hspace{1cm} (32)

which maps message $(M_1, M_2)$ into an action sequence $A^n$; a channel encoder

$$h^{(e)}: \mathcal{M}_1 \times \mathcal{M}_2 \times \mathcal{S}^n \rightarrow \mathcal{X}^n,$$  \hspace{1cm} (33)

which maps message $(M_1, M_2)$ and the state sequence $S^n$ into the sequence $X^n$; two decoding
functions

\[ h_1^{(d)} : B^n \to \mathcal{M}_1, \quad (34) \]

and \( h_2^{(d)} : Y^n \to \mathcal{M}_1 \times \mathcal{M}_2, \quad (35) \)

which map the sequences \( B^n \) and \( Y^n \) into the estimated messages \( \hat{M}_1 \) and \( (\hat{M}_1, \hat{M}_2) \), respectively; such that the probability of error in decoding the messages \( (M_1, M_2) \) is small,

\[
\Pr[h_1^{(d)}(B^n) \neq M_1] \leq \epsilon, \quad (36) \\
\text{and } \Pr[h_2^{(d)}(Y^n) \neq (M_1, M_2)] \leq \epsilon, \quad (37)
\]

and the cost constraint is satisfied, i.e.,

\[
\frac{1}{n} \sum_{i=1}^{n} E[\gamma(A_i, X_i)] \leq \Gamma + \epsilon. \quad (38)
\]

Given a cost \( \Gamma \), a rate pair \( (R_1, R_2) \) is said to be achievable for a cost-constraint \( \Gamma \) if, for any \( \epsilon > 0 \) and sufficiently large \( n \), there a exists a \((n, R_1, R_2, \Gamma, \epsilon)\) code. We are interested in characterizing the capacity-cost region \( C(\Gamma) \), which is the closure of all achievable rate pairs \( (R_1, R_2) \) for the given cost \( \Gamma \).

**B. Capacity-Cost Region**

In this section, a single-letter characterization of the capacity-cost region is derived.

**Proposition 6.** The capacity-cost region \( C(\Gamma) \) for the system in Fig. 3 is given by the union of all rate pairs \((R_1, R_2)\) such that the inequalities

\[
R_1 \leq H(f(A)) \quad (39a)
\]

and \( R_1 + R_2 \leq I(A, U; Y) - I(U; S|A), \quad (39b)\)

are satisfied, where the mutual informations are evaluated with respect to the joint pmf

\[
p(a, s, u, x, y) = p(a)p(s|a)p(u|s, a)1_{(x=g(u, s))}p(y|x, s, a), \quad (40)
\]

for some pmfs \( p(a), p(u|s, a) \) and function \( g : U \times S \to X \) such that

\[
E[\gamma(A, X)] \leq \Gamma. \quad (41)
\]

Finally, we can set \( |U| \leq |X||S||A| + 1 \) without loss of optimality.
The proof of converse is an immediate consequence of cut-set arguments and of the proof of the upper bound obtained in [2, Theorem 1]. Specifically, inequality (39a) follows by considering the cut around Decoder 1, while the inequality (39b) coincides with the bound derived in [2, Theorem 1] on the rate that can be communicated between the Encoder and Decoder 2 with no regards for Decoder 1. The achievability requires rate splitting, superposition coding and the coding strategy proposed in [2, Theorem 1]. A sketch of proof of the achievability is relegated to Appendix D.

C. Probing Capacity

Here we provide an example that illustrates the effect of the communication requirements of the action-cribbing decoder on the system performance. Consider the communication system shown in Fig. 9 where the states is known to Decoder 2. We further assume binary actions, such that, if $A = 1$, the channel encoder observes the state $S$, and if $A = 0$, it does not obtain any information about $S$. We model this problem by defining the state information available at the encoder as $S_e = u(S, A)$, where $u(S, 1) = S$ and $u(S, 0) = \epsilon$, where represents as “erasure” symbol. Following [4], we refer to this problem as having a “probing” encoder.

The channel encoder maps the state information $S^n_e$ and messages $M_1, M_2$ into a codeword $X^n$ (see Fig. 9). Moreover, two cost constraints, namely $\frac{1}{n} \sum_{i=1}^{n} E[\gamma_a(A_i)] \leq \Gamma_A$ and

\footnote{The cardinality constraints follow from [2, Theorem 1]}

Figure 9. Channel coding with actions for channel state probing and with information embedding on actions.
\[ \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\gamma_2(X_i)] \leq \Gamma_X \] are imposed for given action input cost functions \( \gamma_a : \mathcal{A} \rightarrow [0, \Lambda_{a,\text{max}}] \) and \( \gamma_x : \mathcal{X} \rightarrow [0, \Lambda_{x,\text{max}}] \) with \( 0 \leq \Lambda_{a,\text{max}} < \infty \) and \( 0 \leq \Lambda_{x,\text{max}} < \infty \), respectively. In [4], Theorem 1], a correspondence was proved between the set-up of a probing encoder and that of action dependent states. Using [4, Theorem 1] and Proposition 6, we can easily obtain that the capacity-cost region \( C(\Gamma_A, \Gamma_X) \) for the system in Fig. 9 is given by the union of all rate pairs \((R_1, R_2)\) such that the inequalities

\[
R_1 \leq H(A|Q) \tag{42a}
\]

and

\[
R_1 + R_2 \leq I(X;Y|S,Q), \tag{42b}
\]

are satisfied, where the mutual informations are evaluated with respect to the joint pmf

\[
p(q, a, s, s_e, x, y) = p(q)p(a|q)p(s)p(x|s_e, a, q)p(y|x, s), \tag{43}
\]

for some pmfs \( p(q), p(a|q), p(x|s_e, a, q) \) such that \( \mathbb{E}[\gamma_a(A)] \leq \Gamma_A \) and \( \mathbb{E}[\gamma_x(X)] \leq \Gamma_X \).

We now apply (42a)-(42b) to the channel shown in Fig. 9 in which alphabets are binary \( \mathcal{X} = \mathcal{Y} = \mathcal{S} = \{0, 1\} \), \( S \) is a Bern(1 - \( \epsilon \)) variable for \( 0 \leq \epsilon \leq 1 \) and the channel is a binary symmetric with flipping probability 0.5 if \( S = 0 \) (“bad” channel state) and 0 if \( S = 1 \) (“good” channel state).

To evaluate the maximum achievable sum-rate \( R_1 + R_2 \) for a given rate \( R_1 \), we define \( \Pr[A = 1] = \gamma, \Pr[X = 1|S_e = 1, A = 1] = p_1 \) and \( \Pr[X = 1|S_e = e, A = 0] = p_2 \), and we set \( \Pr[X = 1|S_e = 0, A = 1] = 0 \) without loss of optimality. The maximum sum-rate \( R_1 + R_2 \) for a given rate \( R_1 \) is then obtained from (42b) by solving the problem

\[
R_1 + R_2 = \max_{0 \leq p_1, p_2, \gamma \leq 1} \gamma(1 - \epsilon)H(p_1) + (1 - \gamma)(1 - \epsilon)H(p_2), \tag{44}
\]

under the constraint \( \mathbb{E}[X] = p_1 \gamma(1 - \epsilon) + p_2(1 - \gamma) \leq \Gamma_X, \mathbb{E}[A] = \gamma \leq \Gamma_A \) and \( H(A) = H(\gamma) \geq R_1 \). Note that the last constraint imposes that the rate achievable by the Decoder 1 is larger than \( R_1 \) as per (42a).

The sum-rate in (44) is shown in Fig. 10 for \( \epsilon = 0.5, \Gamma_A = 1 \) and different values of \( R_1 \). It can be seen that, for sufficiently small values of the cost constraint \( \Gamma_X \), increasing the communication requirements, i.e., \( R_1 \), of the Decoder 1, reduces the achievable sum-rate \( R_1 + R_2 \). This is due to the fact that increasing \( R_1 \) requires to encode more information in the action sequence, which in turn reduces the portion of the actions that can be set to \( A = 1 \), i.e., \( \Pr[A = 1] \). As a result,
the encoder is less informed about the state sequence and thus bound to waste some power on bad channel states.

Figure 10. Sum-rate $R_1 + R_2$ versus the input cost constraint $\Gamma_X$ for values of $R_1 = 0$, $R_1 = 0.5$ and $R_1 = 0.9$.

**Remark 7.** The communication requirements of Decoder 1 need not necessarily affect the system performance. For instance, consider the example 1 in [4, Sec. V.A], which includes a probing encoder as in Fig. 9 but transmitting over a different channel. There, it turns out that it is sufficient to have $\Pr[A = 1] \gtrsim 0.2$ in order to achieve the same performance that can be achieved with full encoder channel state information. Therefore, the additional constraint on the rate of Decoder 1 (42a), namely $R_1 \geq H(A)$, does not affect the sum-rate achievable in this example for any rate $R_1 \in [0, 1]$.

**V. CONCLUDING REMARKS**

There is a profound interplay between actuation and communication in that both actuation can be instrumental to improve the efficiency of communication, and, vice versa, communication, implicit or explicit, can provide an essential tool to improve control tasks. This work has focused on the first type of interplay, and has investigated the implications of embedding information directly in the actions for the aim of communicating with a separate decoder. The communication
requirements of this decoder are generally in conflict with the goal of improving the efficiency of the given communication link. This performance trade-off has been studied here for both source and channel coding. The results provided in the paper allow to give a quantitative answer to the questions posed in Sec. I-A regarding the impact of the requirements of action information embedding on the system performance. They also shed light into the structure of optimal embedding strategies, which turns out to be related, for the source coding model, with the strategies studied in [20], [24].

The investigation on the theme of information embedding on actions can be further developed in a number of directions, including models with memory [8], [6] and with multiple terminals [9], [14], [11]. We also note that results akin to the ones reported here can be developed assuming causal state information at the decoder for source coding problems or causal state information at the transmitter.

**APPENDIX A: PROOF OF PROPOSITION 1**

Here, we prove the converse part of Proposition 1. For any \((n, R, D_1 + \epsilon, D_2 + \epsilon, \Gamma + \epsilon)\) code, we have

\[
R \geq H(M)
\]

\[
= I(M; X^n, Y^n)
\]

\[
= H(X^n, Y^n) - H(X^n, Y^n|M)
\]

\[
= H(X^n) + H(Y^n|X^n) - H(Y^n|M) - H(X^n|M, Y^n)
\]

\[
\sum_{i=1}^{n} H(X_i) + H(Y_i|Y_i^{-1}, X^n) - H(Y_i|Y_i^{-1}, M) - H(X_i|X_i^{-1}, M, Y^n)
\]

\[
\geq \sum_{i=1}^{n} H(X_i) + H(Y_i|Y_i^{-1}, X^n, A^n) - H(Y_i|Y_i^{-1}, M, A^n) - H(X_i|X_i^{-1}, M, Y^n, A^n)
\]  

\[
\geq \sum_{i=1}^{n} H(X_i) + H(Y_i|Y_i^{-1}, X^n, A^n, \hat{X}_{2i}) - H(Y_i|Y_i^{-1}, M, A^n, \hat{X}_{2i})
\]

\[
\geq \sum_{i=1}^{n} H(X_i) + H(Y_i|X_i, A_i, \hat{X}_{2i}) - H(Y_i|A_i, \hat{X}_{2i}) - H(X_i|U_i, Y_i, A_i, \hat{X}_{2i})
\]  

May 2, 2014 DRAFT
where \((a)\) because \(A^n\) is a function of \(M\) and since conditioning reduces entropy; \((b)\) follows since \(\hat{X}_{2i}\) is a function of \(A^n\); and \((c)\) follows because we have the Markov relation \(Y_i = (X_i, A_i, \hat{X}_{2i}) = (X_n^{n:i}, A^{n:i})\), by defining \(U_i = (M, X_i^{i-1}, Y_n^{n:i}, A^{i-1})\) and since conditioning reduces entropy.

Defining \(Q\) to be a random variable uniformly distributed over \([1, n]\) and independent of all the other random variables and with \(X \triangleq X_Q\), \(Y \triangleq Y_Q\), \(A \triangleq A_Q\), \(\hat{X}_1 \triangleq \hat{X}_{1Q}\), \(\hat{X}_2 \triangleq \hat{X}_{2Q}\) and \(U \triangleq (U_Q, Q)\), from \((47)\) we have

\[
R \geq H(X|Q) + H(Y|X, A, \hat{X}_2, Q) - H(Y|A, \hat{X}_2, Q) - H(X|U, Y, A, \hat{X}_2, Q)
\]

\[
\geq H(X) + H(Y|X, A, \hat{X}_2) - H(Y|A, \hat{X}_2) - H(X|U, Y, A, \hat{X}_2)
\]

\[
= I(X; U, Y, A, \hat{X}_2) - I(Y; X|A, \hat{X}_2)
\]

\[
= I(X; A, \hat{X}_2) + I(X; U|Y, A, \hat{X}_2),
\]

where in \((a)\) we have used the fact that \(X^n\) is i.i.d., conditioning reduces entropy and by the problem definition. Moreover, we have the following chain of inequalities

\[
H(f(A^n)) \leq \sum_{i=1}^{n} H(f(A_i)) = nH(f(A)|Q) \leq nH(f(A)),
\]

where the last inequality follows since conditioning reduces entropy, and

\[
H(f(A^n)) \geq I(f(A^n); X^n)
\]

\[
= \sum_{i=1}^{n} I(f(A^n); X_i|X^{i-1})
\]

\[
= \sum_{i=1}^{n} I(f(A^n), \hat{X}_{2i}; X_i|X^{i-1})
\]

\[
= \sum_{i=1}^{n} I(f(A^n), \hat{X}_{2i}, X^{i-1}; X_i)
\]

\[
\geq \sum_{i=1}^{n} I(f(A_i), \hat{X}_{2i}; X_i)
\]

\[
= n(H(X|Q) - H(X|f(A), \hat{X}_2, Q))
\]

\[
\geq n(H(X) - H(X|f(A), \hat{X}_2))
\]

\[
= n(I(X; f(A), \hat{X}_2)),
\]

\((49)\)
where (a) follows by the chain for mutual information and since mutual information is non-negative; and (b) follows since $X^n$ is i.i.d. and due to the fact that conditioning decreases entropy. Combining (49) and (48), we obtain the inequality

$$I(X; f(A), \hat{X}_2) \leq H(f(A)).$$  \tag{50}$$

We note that the defined random variables factorizes as (8) since we have the Markov chain relationship $(\hat{X}_1, \hat{X}_2, U)-(A, X)-Y$ by the problem definition and that $\hat{X}_2$ is a function $g(U, Y)$ of $U$ and $Y$ by the definition of $U$. Moreover, from cost and distortion constraints (5)-(6), we have

$$D_j + \epsilon \geq \frac{1}{n} \sum_{i=1}^{n} E[d_j(X_i, \hat{X}_{ji})] = E[d_j(X, \hat{X}_j)], \text{ for } j = 1, 2, \tag{51a}$$

and

$$\Gamma + \epsilon \geq \frac{1}{n} \sum_{i=1}^{n} E[\Lambda(A_i)] = E[\Lambda(A)]. \tag{51b}$$

The cardinality constraint on the auxiliary random variable $U$ is obtained as follows using Caratheodory’s theorem as in [3, Appendix C]. Note that we can write $I(X; \hat{X}_2, A) + I(X; U|\hat{X}_2, A, Y) = H(X) - H(X|\hat{X}_2, A) + H(X|\hat{X}_2, A, Y) - H(X|\hat{X}_2, A, Y, U)$.

Now, to preserve the joint distribution of variables $(X, \hat{X}_2, A)$, and thus the distribution of all variables $(X, \hat{X}_2, A, Y)$ and the terms $H(X), H(X|\hat{X}_2, A)$ and $H(X|\hat{X}_2, A, Y)$ (since $p(y|x, a)$ is fixed), the set $U$ should have $|\mathcal{X}||\hat{X}_2||A| - 1$ elements; moreover, we need one further element to preserve the conditional entropy $H(X|\hat{X}_2, A, Y, U)$ and one for the distortion $E[d_1(X, \hat{X}_1)]$.

**APPENDIX B: PROOF OF PROPOSITION 2 AND PROPOSITION 3**

Here, we first prove the converse part of Proposition 2 and then describe the different steps needed to prove Proposition 3. The first part of the converse follows the same steps as in Appendix A. However, we note that in (45) and (46), we can write $A^i$ instead of $A^n$, without changing the following steps. This is due to the strictly causal dependence of $\hat{X}_2$, on the action sequence which is used in (46). This allows to validate the claim in Remark 2. To prove the constraint in (13c), we have the following chain of inequalities

$$H(f(A^n)) = \sum_{i=1}^{n} H(f(A_i)|f(A^{i-1})) = \sum_{i=1}^{n} H(f(A_i)|\hat{X}_{2i}) \overset{(a)}{=} nH(f(A)|\hat{X}_2). \tag{52}$$
Moreover, we can write

\[
H(f(A^n)) \geq I(f(A^n); X^n) \\
= \sum_{i=1}^{n} I(f(A^n); X_i | X^{i-1}) \\
= \sum_{i=1}^{n} I(f(A^n), \hat{X}_{2i}; X_i | X^{i-1}) \\
= \sum_{i=1}^{n} I(f(A^n), \hat{X}_{2i}, X^{i-1}; X_i) \\
\overset{(a)}{=} \sum_{i=1}^{n} I(f(A_i), \hat{X}_{2i}; X_i) \\
= n(H(X) - H(X|f(A), \hat{X}_2, Q)) \\
\overset{(b)}{=} n(H(X) - H(X|f(A), \hat{X}_2)) \\
= n(I(X; f(A), \hat{X}_2)),
\]

where \((a)\) follows by the chain for mutual information and since mutual information is non-negative; and \((b)\) follows since \(X^n\) is i.i.d. and due to the fact that conditioning decreases entropy. Combining (53) and (52), we obtain the inequality (13c). We note that the joint pmf of the defined random variables factorizes as (12) since we have the Markov chain relationship \((\hat{X}_1, \hat{X}_2, U) \rightarrow (A, X) \rightarrow Y\) by the problem definition and that \(\hat{X}_1\) is a function \(g(U, Y)\) of \(U\) and \(Y\) by the definition of \(U\) as in Appendix A. The distortion, cost and cardinality constraint are obtained as in Appendix A.

The converse for Proposition 3 follows from similar steps by defining \(V_i = f(A^{i-1})\) and noting that \(\hat{X}_{2i}\) is a function of \(V_i\) and \(f(A_i)\).

We bound the cardinality of the auxiliary random variables \(U\) and \(V\) for Proposition 3 using [3, Appendix C]. The bounds for \(U\) in Proposition 2 follow in the same way. Note that we can write \(I(X; V, A) + I(X; U|V, A, Y) = H(X) - H(X|V, A) + H(X|V, A, Y) - H(X|V, A, Y, U)\). Starting with \(V\), the alphabet \(\mathcal{V}\) should have \(|\mathcal{X}| - 1\) elements to preserve the distribution \(p(x)\) and hence \(H(X)\), one element to preserve \(-H(X|V, A) + H(X|V, A, Y)\), two elements to preserve the distortion constraints and and one more to preserve the condition \(I(X; V, f(A)) \leq H(f(A)|V)\). As for \(U\), just as in Appendix A, \(U\) should have \(|\mathcal{X}||\mathcal{Y}||A| - 1\) elements to preserve the joint distribution \(p(x, v, a)\) (which preserves the joint distribution \(p(x, a, v, y)\) and hence
\[ H(X), H(X|V, A), H(X|V, A, Y), \] one element to preserve \( H(X|V, A, Y, U) \) and one more to preserve the distortion constraint of Decoder 1.

**Appendix C: Proof of Proposition 4**

Here, we prove the converse part of Proposition 4. To establish the converse, it is sufficient to consider the case of non-causal action observation, as done in the following. For any \((n, R, D_1 + \epsilon, D_2 + \epsilon, \Gamma + \epsilon)\) code, define the auxiliary variable \( U_i = (Y^n, X^{i-1}) \) and \( Q \) as a random time sharing variable uniformly distributed in the interval \([1, n]\) independent of \((X, U, \hat{X}_1, \hat{X}_2, A, Y)\). We then have,

\[
H(Y^n) = \sum_{i=1}^{n} H(Y_i|Y^{i-1}) \\
\leq \sum_{i=1}^{n} H(Y_i) \\
= \sum_{i=1}^{n} H(f_Y(A_i)) \\
= nH(f_Y(A_Q)|Q) \\
\leq H(f_Y(A_Q)). \hspace{1cm} (54)
\]

Also, we can write

\[
H(Y^n) \geq I(X^n; Y^n) \\
= \sum_{i=1}^{n} I(X_i; Y^n|X^{i-1}) \\
\overset{(a)}{=} \sum_{i=1}^{n} I(X_i; Y^n, X^{i-1}) \\
= \sum_{i=1}^{n} I(X_i; U_i) \\
= nI(X_Q; U_Q|Q) \\
\overset{(b)}{=} I(X_Q; U_Q, Q), \hspace{1cm} (55)
\]

May 2, 2014 DRAFT
along with

\[
H(A^n|Y^n) = \sum_{i=1}^{n} H(A_i|Y^n, A^{i-1})
\]
\[
\leq \sum_{i=1}^{n} H(A_i|Y_i)
\]
\[
= \sum_{i=1}^{n} H(A_i|f_Y(A_i))
\]
\[
= n(H(A_Q)|f_Y(A_Q), Q)
\]
\[
\leq H(A_Q|f_Y(A_Q)), \quad (56)
\]

and

\[
H(A^n|Y^n) \geq I(X^n; A^n|Y^n)
\]
\[
\overset{(c)}{=} I(X^n; A^n, \hat{X}_2^n|Y^n)
\]
\[
\geq \sum_{i=1}^{n} I(X_i; \hat{X}_{2,i}|Y^n, X^{i-1})
\]
\[
= \sum_{i=1}^{n} I(X_i; \hat{X}_{2,i}|U_i)
\]
\[
= nI(X_Q; \hat{X}_{2,Q}|U_Q, Q). \quad (57)
\]

Furthermore, we have

\[
H(Y^n, M) \leq \sum_{i=1}^{n} H(Y_i) + nR
\]
\[
\leq nH(Y_Q) + nR, \quad (58)
\]

and

\[
H(Y^n, M) \geq I(X^n; Y^n, M)
\]
\[
\overset{(d)}{=} I(X^n; \hat{X}_1^n, Y^n, M)
\]
\[
\geq \sum_{i=1}^{n} I(X_i; \hat{X}_{1,i}, Y^n|X^{i-1})
\]
\[
= \sum_{i=1}^{n} I(X_i; \hat{X}_{1,i}, Y^n, X^{i-1})
\]
\( \sum_{i=1}^{n} I(X_i; \hat{X}_{1,i}, U_i) = n(X_Q; \hat{X}_{1,Q}, U_Q, Q) \) \tag{59}

where (a) follows from the independence of \( X_i \) and \( X^{i-1} \); (b) follows from the independence of \( Q \) from all other random variables; (c) follows from the fact that \( \hat{X}_2^n \) is a function of \( A^n \); and (d) follows from the fact that \( \hat{X}_1^n \) is a function of \( (M, Y^n) \). Defining \( U \triangleq (U_Q, Q) \) along with \( X \triangleq X_Q, Y \triangleq Y_Q, A \triangleq A_Q, \hat{X}_1 \triangleq \hat{X}_{1Q}, \hat{X}_2 \triangleq \hat{X}_{2Q} \) and combining (54), (55), (56), (57), (58) and (59), we obtain the rate region inequalities as mentioned in the proposition. Note that the joint distribution of the random variables \( (X, Y, A, \hat{X}_1, \hat{X}_2) \) established above factorizes as

\[
p(x)p(u, \hat{x}_1, \hat{x}_2, a|x) \]

but can be restricted only to pmfs factorizing as in (21). This is because the information measures in (20)-(22) only depends on the marginals \( p(x, u, \hat{x}_1) \), \( p(a) \) and \( p(x, u, \hat{x}_2) \). Distortion and cost constraints are handled in the standard manner [3].

We bound the cardinality of the auxiliary random variables \( U \) using [3, Appendix C]. The set \( U \) should have \( |X||\hat{X}_1||\hat{X}_2| - 1 \) elements to preserve the joint distribution \( p(x, \hat{x}_1, \hat{x}_2) \), one element to preserve the Markov chain \( \hat{X}_1 - U - \hat{X}_2 \), and three elements to preserve \( H(X|\hat{X}_1, U) \), \( H(X|\hat{X}_2, U) \) and \( H(X|U) \).

**APPENDIX D: SKETCH OF PROOF OF ACHIEVABILITY FOR PROPOSITION 6**

We will prove below that the following rate region is achievable

\[
R_1 \leq H(f(A)) \tag{60a}
\]

\[
R_1 + R_2 \leq I(A, U; Y) - I(U; S|A), \tag{60b}
\]

\[
R_2 \leq I(A; Y|f(A)) + I(A, U; Y|A) - I(U; S|A), \tag{60c}
\]

for a given joint distribution as in (40). Assuming now that this rate region is achievable, we show that the rate region (39) is also achievable. Region (39) is larger than (60) owing to the absence of the inequality (60c). The two regions are illustrated in Fig. 11 for a given choice of the distribution (40), with region (60) in solid lines and (39) in dashed lines. We now argue that the achievability of region (60) (solid lines) implies the achievability of region (39) (dashed lines) as well, by following the same arguments as in [25]. Specifically, we observe that, if \((R_1, R_2)\) is achievable with some scheme, then \((R_1 - t, R_2 + t)\) is also achievable for all \(0 \leq t \leq R_1\).
This is due to the fact that, if the rate pair \((R_1, R_2)\) is achievable, then some of the rate of the common message \(M_1\) can always be transferred to the private message \(M_2\) for Decoder 2 to achieve \((R_1 - t, R_2 + t)\) if \(0 \leq t \leq R_1\). It follows immediately that all the points on the dashed line in Fig. 11 are also achievable.

The discussion above allows us to conclude that concludes that, if region (60) is achievable, then the desired rate region (39) is also achievable. We now focus on proving the achievability of (60). To this end, we combine superposition coding and the technique proposed in [2]. Fix the joint distribution as in (40). We first generate the codebook \(b^n(m_1), m_1 \in [1 : 2^{nR_1}]\), i.i.d. with pmf \(p(b)\). Next, we generate a superimposed codebook for each \(b^n\) of \(a^n(m_1, m_2)\) codewords, \(m_2 \in [1 : 2^{nR_2}]\), i.i.d. with pmf \(p(a|b)\). For every \(a^n\) sequence, a codebook of \(u^n(m_1, m_2, j)\) sequences is generated, \(j \in [1 : 2^{n\tilde{R}}]\), i.i.d. with pmf \(p(u|a)\).

To encode messages \((m_1, m_2)\), Encoder selects the codeword \(a^n(m_1, m_2)\), and chooses a \(u^n\) codeword jointly typical with action and state sequence, which requires \(\tilde{R} \geq I(U; S|A)\). Then \(x_i = g(u_i, s_i)\) is then sent through the channel. Decoder 1 decodes the message \(m_1\) correctly if \(R_1 \leq H(B)\). Decoder 2 looks for the unique pair of messages \((m_1, m_2)\) such that the tuple

\[ (R_1, R_2) \]
\((y^n, b^n(m_1), a^n(m_1, m_2), u^n(m_1, m_2, j))\) is jointly typical for some \(j \in [1 : 2^n\hat{R}]\). This step is reliable if \(R_1 + R_2 + \hat{R} \leq I(A, U; Y)\) and \(R_2 + \hat{R} \leq I(U, A; Y | B) = I(A; Y | B) + I(U; Y | A)\). Using Fourier-Motzkin elimination to eliminate rate \(\hat{R}\) leads to the bounds \((60)\).

REFERENCES

[1] H. Permuter and T. Weissman, “Source coding with a side information “vending machine”,” IEEE Trans. Inf. Theory, vol. 57, pp. 4530–4544, Jul 2011.

[2] T. Weissman, “Capacity of channels with action-dependent states,” IEEE Trans. Inform. Theory, vol. 56, no. 11, pp. 5396–5411, Nov. 2010.

[3] A. El Gamal and Y. Kim, Network Information Theory, Cambridge University Press, Dec 2011.

[4] H. Asnani, H. Permuter, and T. Weissman, “Probing capacity,” IEEE Trans. Inform. Theory, vol. 57, no. 11, pp. 7317–7332, Nov. 2011.

[5] K. Kittichokechhai, T. J. Oechtering, and M. Skoglund, “Source coding with common reconstruction and action-dependent side information,” in Proc. IEEE Information Theory Workshop (ITW 2010), pp.1-5, Aug. 30 2010-Sept. 3 2010.

[6] H. Asnani and T. Weissman, “On real time coding with limited lookahead,” submitted [arXiv:1105.3755].

[7] C. Choudhuri and U. Mitra, “How useful is adaptive action?,” in Proc. IEEE Globecom 2012.

[8] H. Asnani, H. Permuter, and T. Weissman, “To feed or not to feed back,” submitted [arXiv:1011.1607].

[9] Y. Chia, H. Asnani, and T. Weissman, ”Multi-terminal source coding with action dependent side information,” submitted [arXiv:1110.6916].

[10] B. Ahmadi and O. Simeone, “Distributed and Cascade Lossy Source Coding with a Side Information "Vending Machine"," in Proc. IEEE International Symposium on Information Theory (ISIT 2012), Cambridge, MA, USA, July 1-6, 2012.

[11] L. Zhao, Y. K. Chia, and T. Weissman, “Compression with actions,” submitted [arXiv:1204.2331].

[12] Y. Steinberg and T. Weissman, “The degraded broadcast channel with action-dependent states,” in Proc. IEEE International Symposium on Information Theory (ISIT 2012), Cambridge, MA, USA, July 1-6, 2012.

[13] B. Ahmadi and O. Simeone, “On channels with action-dependent states,” arXiv:1202.4438

[14] B. Ahmadi, O. Simeone, C. Choudhuri, and U. Mitra, “On cascade source coding with a side information "Vending Machine"," arXiv:1204.1548.

[15] N. Elia, “When Bode meets Shannon: Control-oriented feedback communication schemes,” IEEE Trans. Autom. Contr., vol. 49, no. 9, pp. 1477–1488, Sep. 2004.

[16] T. P. Coleman, “A stochastic control viewpoint on ‘Posterior Matching’-style feedback communication schemes,” in Proc. IEEE International Symposium on Information Theory (ISIT 2009), pp. 1520-1524, 2009.

[17] S. Tatikonda and S. Mitter, “Control under communication constraints,” IEEE Trans. Automatic Control, vol. 49, no. 7, pp. 1056–1068, July 2004.

[18] A. Sahai and P. Grover, “Demystifying the Witsenhausen counterexample,” IEEE Control System Magazine, vol. 30, no. 6, pp. 20–24, Dec. 2010.

[19] P. Moulin and J. A. O’Sullivan, “Information-theoretic analysis of information hiding,” IEEE Trans. Inform. Theory, vol. 49, no. 3, pp. 563–593, Mar. 2003.

[20] H. Asnani, H. Permuter and T. Weissman, “Successive refinement with decoder cooperation and its channel coding duals,” submitted (arxiv.org/pdf/1203.4865).
[21] S. S. Pradhan, “On the role of feedforward in Gaussian sources: point-to-point source coding and multiple description source coding,” *IEEE Trans. Inform. Theory*, vol. 53, no. 1, pp. 331-349, Jan. 2007.

[22] F. Willems, “The discrete memoryless multiple access channel with partially cooperating encoders (Corresp.),” *IEEE Trans. Inform. Theory*, vol. 29, no. 3, pp. 441-445, May 1983.

[23] R. M. Gray and A. D. Wyner, “Source coding for a simple network,” *Bell System Tech. J.*, vol. 53, pp. 1681–1721, 1974.

[24] P. Cuff and L. Zhao, “Coordination using implicit communication,” in *Proc. IEEE Information Theory Workshop* (ITW 2011), pp. 467-471, Oct. 2011.

[25] J. Korner and K. Marton, “General broadcast channels with degraded message sets,” *IEEE Trans. on Information Theory*, vol.23, no.1, pp. 60-64, Jan 1977.