Coordination Coding with Causal Decoder for Vector-valued Witsenhausen Counterexample Setups

Tobias J. Oechtering
KTH Royal Institute of Technology
EECS, Div. Inform. Science and Engineering
10044 Stockholm, Sweden

Maël Le Treust
ETIS UMR 8051, Université Paris Seine,
Université Cergy-Pontoise, ENSEA, CNRS
95014 Cergy-Pontoise, France

***Abstract*—The vector-valued extension of the famous Witsenhausen counter-example setup is studied where the first decision maker (DM1) non-causally knows and encodes the iid state sequence and the second decision maker (DM2) causally estimates the interim state. The coding scheme is transferred from the finite alphabet coordination problem for which it is proved to be optimal. The extension to the Gaussian setup is based on a non-standard weak typicality approach and requires a careful average estimation error analysis since the interim state is estimated by the decoder. Next, we provide a choice of auxiliary random variables that outperforms any linear scheme. The optimal scheme remains unknown.**

I. Introduction

In 1968, Witsenhausen introduced in [1] his famous counterexample that showed that the best affine policy is outperformed by non-linear policies. Since then the example serves as study object illustrating the importance of the information pattern in distributed decision making, see [2] for a comprehensive discussion.

The first vector-valued extension considering a non-causal setup was studied by Grover and Sahai in 2008 followed by a series of works, e.g. [3]. A comprehensive overview on the corresponding results is provided in [4], where we also discuss its close relation to the coordination problem. Optimal coding schemes for relevant setups are derived in [5], which also provides a review on the related literature. Most of the results were derived for finite alphabet setup where the concept of strong typicality provides the Markov Lemma. A rigorous extension to the Gaussian case has been done by Grover and Sahai in [3]. The main result was proved to be optimal in [6].

In this work we extend the finite alphabet coding scheme based on the concept of weak typicality [7] that we extend so that the need of the Markov Lemma can be avoided. Conceptually, the extension is similar to an extension as done in [8] where also Wyner's approach on how to analyze the average estimation error has been used [9]. We extend this approach in this work since the second decision maker estimates the interim state and not the iid source.

In the following we show that also in this vector-valued setup the best affine policies can be outperformed by non-linear policies. In more detail, there exists parameter configurations where our coordination coding outperforms a simple *amplification* strategy.

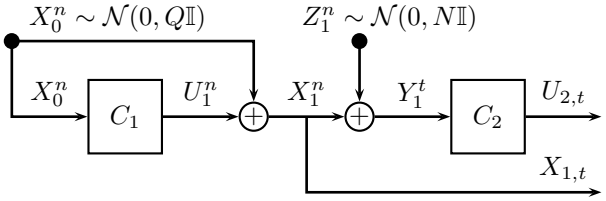


Fig. 1. The state and the channel noise are drawn according to the i.i.d. Gaussian distributions $X_0^n \sim \mathcal{N}(0, Q\mathbb{I})$ and $Z_1^n \sim \mathcal{N}(0, N\mathbb{I})$.

II. System Model

We consider the vector-valued Witsenhausen setup in which the sequences of states and channel noises are drawn independently according to the i.i.d. Gaussian distributions $X_0^n \sim \mathcal{N}(0, Q\mathbb{I})$ and $Z_1^n \sim \mathcal{N}(0, N\mathbb{I})$. We denote by X_1 the *interim* state and Y_1 the output of the noisy channel.

$$X_1 = X_0 + U_1, \tag{1}$$

$$Y_1 = X_1 + Z_1 = X_0 + U_1 + Z_1 \text{ with } Z_1 \sim \mathcal{N}(0, N). \tag{2}$$

We denote by $\mathcal{P}_{X_0} = \mathcal{N}(0, Q)$ the Gaussian state distribution and by $\mathcal{P}_{X_1 Y_1 | X_0 U_1}$ the conditional probability distribution corresponding to equations (1)-(2).

Definition 1. *A "control design" with non-causal encoder and causal decoder is a tuple of stochastic functions $c = (f, \{g_i\}_{i\in\{1,\dots,n\}})$ defined by:*

$$f : \mathcal{X}_0^n \longrightarrow \mathcal{U}_1^n, \quad g_i : \mathcal{Y}_1^i \longrightarrow \mathcal{U}_2, \ \forall i \in \{1, \dots, n\}. \tag{3}$$

We denote by $\mathcal{C}_{\mathsf{d}}(n)$ the set of control designs with non-causal encoder and causal decoder. This code induces a probability distribution over the sequences given by:

$$\prod_{i=1}^{n} \mathcal{P}_{X_{0,i}} f_{U_1^n | X_0^n} \prod_{i=1}^{n} \mathcal{P}_{X_{1,i} Y_{1,i} | X_{0,i} U_{1,i}} \prod_{i=1}^{n} g_{U_{2,i} | Y_1^i}. \tag{4}$$

Definition 2. *We define the two long-run costs functions $c_P(u_1^n) = \frac{1}{n}\sum_{t=1}^{n} u_{1,t}^2$ and $c_S(x_1^n, u_2^n) = \frac{1}{n}\sum_{t=1}^{n}(x_{1,t} - u_{2,t})^2$. The pair of costs (P, S) are achievable if for all $\varepsilon > 0$, there exists $\bar{n} \in \mathbb{N}^*$, for all $n \geq \bar{n}$, there exists a "control design" $c \in \mathcal{C}_{\mathsf{d}}(n)$ such that:*

$$\left| P - \mathbb{E}\Big[c_P(U_1^n)\Big] \right| \leq \varepsilon, \quad \left| S - \mathbb{E}\Big[c_S(X_1^n, U_2^n)\Big] \right| \leq \varepsilon. \tag{5}$$

Theorem 3 (Main result)**.** *The pair of Witsenhausen costs* (P, S) *are achievable* if and only if *there exists a joint probability distribution*

$$\mathcal{Q} = \mathcal{P}_{X_0}\mathcal{Q}_{U_1W_1W_2|X_0}\mathcal{P}_{X_1Y_1|X_0U_1}\mathcal{Q}_{U_2|W_2Y_1}, \tag{6}$$

involving two auxiliary random variables (W_1, W_2)*, such that* $I(W_1; Y_1|W_2) - I(W_1, W_2; X_0) \geq 0$ *and*

$$P = \mathbb{E}_{\mathcal{Q}}\big[U_1^2\big], \qquad S = \mathbb{E}_{\mathcal{Q}}\big[(X_1 - U_2)^2\big]. \tag{7}$$

Remark 4. *The probability distribution of* (6) *satisfies:*

$$\begin{cases} (X_1, Y_1) \multimap (X_0, U_1) \multimap (W_1, W_2), \\ U_2 \multimap (Y_1, W_2) \multimap (X_0, X_1, U_1, W_1), \\ X_2 \multimap (X_1, U_2) \multimap (X_0, U_1, Y_1, W_1, W_2). \end{cases} \tag{8}$$

The causality condition prevents the controller C_2 *to recover* W_1*. This corresponds to the second Markov chain of* (8)*.*

The achievability proof of Theorem 3 is in Appendix A and the converse proof follows from [4, Theorem II.5]. In order to characterize the set of achievable Witsenhausen costs (P, S), we consider a fixed parameter $P \geq 0$ and we minimize $S = \mathbb{E}_{\mathcal{Q}}\big[(X_1 - U_2)^2\big]$. The optimization problem writes

$$\underset{\mathcal{Q}\in\mathbb{Q}(P)}{\text{minimize}} \ \mathbb{E}_{\mathcal{Q}}\big[(X_1 - U_2)^2\big], \quad \text{where} \tag{9}$$

$$\mathbb{Q}(P) = \Big\{ \big(\mathcal{Q}_{U_1W_1W_2|X_0}, \mathcal{Q}_{U_2|W_2Y_1}\big) \ \text{ s.t. } \ P = E_{\mathcal{Q}}\big[U_1^2\big] \\ \text{and } I(W_1; Y_1|W_2) - I(W_1, W_2; X_0) \geq 0 \Big\}. \tag{10}$$

Although the set $\mathbb{Q}(P)$ is convex, the characterization of the optimal pairs of probability distributions $\big(\mathcal{Q}_{U_1W_1W_2|X_0}, \mathcal{Q}_{U_2|W_2Y_1}\big) \in \mathbb{Q}$ seems difficult. We investigate two schemes referred to as "state amplification" in section III-A and "coordination coding" in section III-B and we compare there performances in section III-C.

III. Numerical Results

A. Best Affine Policy

For a given parameter $P \geq 0$, the first controller zero-forces the state by using $U_1 = -\sqrt{\frac{P}{Q}} \cdot X_0$. The best affine policy [12], also referred to as "state amplification", does not involve coding so there is no need of (W_1, W_2).

Proposition 5. *When the controllers implement* $U_1 = -\sqrt{\frac{P}{Q}} \cdot X_0$ *and* $U_2 = \mathbb{E}\big[X_1|Y_1\big]$*, we have:*

$$S_{\mathsf{sa}} = \mathbb{E}\big[(X_1 - U_2)^2\big] = \frac{\big(\sqrt{Q} - \sqrt{P}\big)^2 \cdot N}{\big(\sqrt{Q} - \sqrt{P}\big)^2 + N}. \tag{11}$$

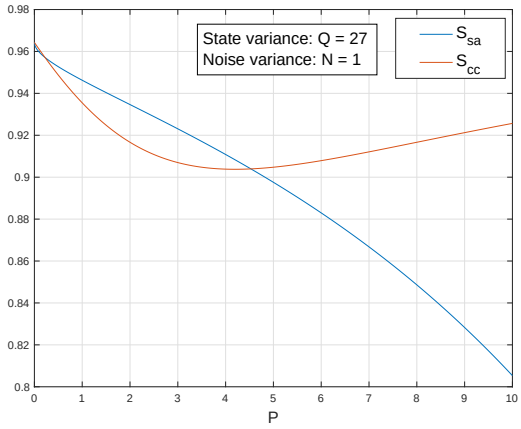


Fig. 2. The functions S_{sa} and S_{cc} depending on P. Coordination coding with Costa's choice of auxiliary random variables outperforms the best affine strategy for a range of low powers.

B. Coordination Coding

As in [5], we combine source coding with Costa's coding.

$$I(W_1; Y_1|W_2) - I(W_1, W_2; X_0) \geq 0 \tag{12}$$

$$\iff I(W_2; X_0) \leq I(W_1; Y_1, W_2) - I(W_1; X_0, W_2). \tag{13}$$

Given two positive correlations parameters (δ, D), we define the test channel $X_0 = \delta \cdot W_2 + Z_0$, where the random variables $Z_0 \sim \mathcal{N}(0, D)$ and $W_2 \sim \mathcal{N}\big(0, \frac{1}{\delta^2}(Q - D)\big)$ are independent. The correlation matrix of $(X_0, W_2) \sim \mathcal{N}(0, K)$ is given by:

$$K = \begin{bmatrix} Q & \frac{1}{\delta}(Q - D) \\ \frac{1}{\delta}(Q - D) & \frac{1}{\delta^2}(Q - D) \end{bmatrix} \tag{14}$$

As in Costa's scheme [11], U_1 is independent of (X_0, W_2) and $W_1 = U_1 + \alpha X_0 + \beta W_2$.

Lemma 6. *The parameters* (δ, β) *have no impact, moreover*

$$I(W_1; Y_1|W_2) - I(W_1, W_2; X_0) \tag{15}$$

$$= \frac{1}{2}\log_2\left(\frac{P \cdot D \cdot (P + D + N)}{Q \cdot \big(N \cdot (P + \alpha^2 D) + (1 - \alpha)^2 \cdot P \cdot D\big)}\right). \tag{16}$$

Lemma 7. *By using Costa's optimal* $\alpha^\star = \frac{P}{P+N}$*, we have*

$$I(W_1; Y_1|W_2) - I(W_1, W_2; X_0) = \frac{1}{2}\log_2\left(\frac{D \cdot (P + N)}{Q \cdot N}\right). \tag{17}$$

The minimal D *such that* (17) *is positive, is* $D^\star = \frac{Q \cdot N}{P+N}$*.*

Proposition 8. *For the coordination coding scheme with* $W_1 = U_1 + \alpha^\star X_0$*,* $W_2 = X_0 - Z_0$*,* $Z_0 \sim \mathcal{N}(0, D^\star)$ *and* $U_2 = \mathbb{E}\big[X_1|Y_1, W_2\big]$*, we have*

$$S_{\mathsf{cc}} = \mathbb{E}\big[(X_1 - U_2)^2\big] = \frac{N \cdot \big(P \cdot (P + N) + Q \cdot N\big)}{(P + N)^2 + Q \cdot N}. \tag{18}$$

C. Performance Comparison

In Fig. 2, we compare the S_{sa} and S_{cc} as functions of P. For the parameters $(P, Q, N) = (3, 27, 1)$ the "coordination coding" outperforms the "state amplification" scheme.

Proposition 9. *For $N > 0$ and $Q > 0$ we have:*

$$S_{cc} - S_{sa} \geq 0 \Longleftrightarrow 2(P+N) - \sqrt{Q \cdot P} \geq 0. \tag{19}$$

Appendix A
Proof of the Main Result

The achievability proof uses the block-Markov coding scheme with B blocks each of length n using backward encoding at DM 1 and forward decoding at DM 2. The coding scheme follows the empirical coordination scheme with non-causal encoding and causal decoding [5]. Before the regular transmission will be a initialisation phase of length n'. The *'error'* analysis is based on the concept of weak typicality with an extension that circumvents the need of the Markov Lemma available for strong typicality. A similar approach has been taken in [8].

Preliminaries: Given an arbitrary but fixed $\varepsilon > 0$. Further, assume $(X_0^n, X_1^n, U_1^n, U_2^n, W_1^n, W_2^n, Y_1^n)$ is generated iid $\sim \mathcal{Q}_{X_0X_1U_1U_2W_1W_2Y_1}$ with $P = \mathbb{E}[U_1^2]$ and $S = \mathbb{E}[(X_1 - U_2)^2]$. Then let $\psi^{(n)} : \mathcal{X}_0^n \times \mathcal{X}_1^n \times \mathcal{U}_1^n \times \mathcal{U}_2^n \times \mathcal{W}_1^n \times \mathcal{W}_2^n \times \mathcal{Y}_1^n \to \{0, 1\}$ denote an indicator function for sequences of length n with

$$\psi^{(n)}(x_0^n, x_1^n, u_1^n, u_2^n, w_1^n, w_2^n, y_1^n) = \begin{cases} 1 & \text{if } |c_P(u_1^n) - P| \geq \frac{1}{2}\varepsilon \text{ or } |c_S(x_1^n, u_2^n) - S| \geq \frac{1}{12}\varepsilon \\ & \quad \text{or } (w_1^n, w_2^n, y_1^n) \notin \mathcal{A}_\varepsilon^{(n)}(W_1, W_2, Y_1) \\ 0 & \text{otherwise} \end{cases}$$

Using the weak law of large numbers and the union bound we have

$$\delta_n = \mathbb{E}[\psi^{(n)}(X_0^n, X_1^n, U_1^n, U_2^n, W_1^n, W_2^n, Y_1^n)] \overset{n\to\infty}{\longrightarrow} 0.$$

Define

$$\mathcal{S}_\varepsilon^{(n)} = \{(x_0^n, w_1^n, w_2^n) \,|\, \eta^{(n)}(x_0^n, w_1^n, w_2^n) \leq \sqrt{\delta_n}\}$$

with $\eta^{(n)}(x_0^n, w_1^n, w_2^n) = \mathbb{E}[\psi^{(n)}(x_0^n, X_1^n, U_1^n, U_2^n, w_1^n, w_2^n, Y_1^n) | X_0^n = x_0^n, W_1^n = w_1^n, W_2^n = w_2^n]$. Then from the Markov inequality we obtain

$$\mathbb{P}\{(X_0^n, W_1^n, W_2^n) \notin \mathcal{S}_\varepsilon^{(n)}\} \leq \frac{\mathbb{E}[\psi^{(n)}(X_0^n, X_1^n, U_1^n, U_2^n, W_1^n, W_2^n, Y_1^n)]}{\sqrt{\delta_n}} \leq \sqrt{\delta_n}$$

We finally define the set

$$\mathcal{B}_\varepsilon^{(n)} = \mathcal{A}_\varepsilon^{(n)}(X_0, W_1, W_2) \cap \mathcal{S}_\varepsilon^{(n)},$$

which denotes the set of jointly typical pairs that also satisfy the cost constraints. Note that for (X_0^n, W_1^n, W_2^n) iid $\sim \mathcal{Q}_{X_0W_1W_2}$ we have $\mathbb{P}\{(X_0^n, W_1^n, W_2^n) \in \mathcal{B}_\varepsilon^{(n)}\} \to 1$ as $n \to \infty$. Furthermore, we have the following lemma, which can be similarly shown as Lemma 1 in the journal draft [8].

Lemma 10. *Let X_0^n iid$\sim$ $\mathcal{Q}_{X_0}$. For $M = 2^{nR} \geq 2^{n(I(X_0;W_2)+3\varepsilon)}$ codewords $w_2^n(m)$ iid$\sim$ $\mathcal{Q}_{W_2}$, $1 \leq m \leq M$ and $L = 2^{nR_L} \geq 2^{n(I(W_1;X_0,W_2)+4\varepsilon)}$ codewords $w_1^n(\ell, m)$ iid$\sim$ $\mathcal{Q}_{W_1}$, $1 \leq \ell \leq L$ and $\varepsilon > 0$, we have*

$$\mathbb{P}\{(X_0^n, W_1^n(1, \ell), W_2^n(m)) \notin \mathcal{B}_\varepsilon^{(n)} \,\forall\, m, \ell\} \to 0 \text{ as } n \to \infty.$$

Proof. The proof follows the same arguments as the proof of Lemma 1 in the journal draft [8] with X, U, and V replaced by X_0, W_1, and W_2 as well as $p_{V|U}$ replaced by $\mathcal{Q}_{W_1}$ so that (155) changes as follows

$$\frac{\mathcal{Q}_{W_1}^{\otimes n}(w_1^n)}{\mathcal{Q}_{W_1|X_0W_2}^{\otimes n}(w_1^n|x_0^n, w_2^n)} \geq \frac{2^{-n(h(W_1)+2\varepsilon)}}{2^{-n(h(W_1|X_0,W_2)-2\varepsilon)}} = 2^{-n(I(W_1;X_0,W_2)+4\varepsilon}. \qquad \square$$

To ensure that the second cost constraint remains bounded even when a coding error happens, DM 2 is going to quantise its output. Since we assume a joint distribution with $\mathbb{E}[(X_1 - U_2)^2] = S$, for any $\hat{\delta} > 0$ there exists a quantisation $q_{U_2} : \mathcal{U}_2 \to \{\hat{u}_{2,k}\}_{k=1}^{N_{U_2}}$ such that

$$\hat{S} = \mathbb{E}[(X_1 - q_{U_2}(U_2))^2] \leq (1 + \hat{\delta})S,$$

in particular such that $\hat{\delta}S < \frac{1}{4}\varepsilon$.

With those preliminaries we are now ready to provide the coding scheme.

Random codebook: For rate $R \geq I(X_0; W_2) + 3\varepsilon$ and rate $R_L \geq I(W_1; W_2; X_0) + 4\varepsilon$, generate 2^{nR} codewords $w_2^n(m)$ iid $\sim$ $\mathcal{Q}_{W_2}$ and $2^{n(R+R_L)}$ codewords $w_1^n(m, \ell)$ iid $\sim$ $\mathcal{Q}_{W_1}$ with indices $m \in [1 : 2^{nR}]$ and $\ell \in [1 : 2^{nR_L}]$.

Backward encoding at DM 1: Let m_b and $x_{0,b}^n$ denote the message and processed source sequence of length n of block b, $1 \leq b \leq B$. Due to non-causal knowledge, the encoder performs backward encoding, i.e., the encoder starts with block $b = B$ with initialisation $m_{B+1} = 1$ and subsequently encodes the previous blocks. In block b, the encoder takes sequence $x_{0,b}^n$ and message m_{b+1} and looks for ℓ_b and m_b such that

$$(x_{0,b}^n, w_1^n(m_{b+1}, \ell_b), w_2^n(m_b)) \in \mathcal{B}_\varepsilon^{(n)}$$

If there are none or more than one pair, then the encoder randomly picks one. Let $w_{1,b}^n = w_1^n(m_{b+1}, \ell_b)$ and $w_{2,b}^n = w_2^n(m_b)$ denote the choice. Next, we generate $u_{1,b}^n \sim \mathcal{Q}_{U_1|W_1W_2X_0}^{\otimes n}(w_{1,b}^n, w_{2,b}^n, x_0^n)$.

Forward transmission of DM 1: In block b, $1 \leq b \leq B$, if $|c_P(u_{1,b}^n) - P| < \frac{1}{4}\varepsilon$ then DM 1 transmits $u_{1,b}^n$ synchronously with $x_{0,b}^n$, otherwise DM1 transmits the all zero codeword. Channel $P_{X_1,Y_1|X_0,U_1}^{\otimes n}$ produces channel outputs $x_{1,b}^n$ and $y_{1,b}^n$.

Forward decoding at DM 2: Let $\tilde{w}_{2,b}^n$ be an abbreviation for $w_{2,b}^n(\tilde{m}_b)$ for block b, $1 \leq b \leq B$, where $\tilde{m}_b$ denotes the index decided on in the previous block $b-1$. Note that message $\tilde{m}_1$ will have been obtained from the initialisation phase. Upon receiving $y_{1,b}^n$, DM 2 looks for $\tilde{\ell}_b$ and $\tilde{m}_{b+1}^n$ such that

$$(y_{1,b}^n, w_1^n(\tilde{m}_{b+1}, \tilde{\ell}_b), \tilde{w}_{2,b}) \in \mathcal{A}_\varepsilon^{(n)}(Y_1, W_1, W_2).$$

If there are none or more than one pair, then the decoder randomly picks one.

Forward transmission of DM 2: In block b, $1 \le b \le B$, DM 2 generates $u_{2,b}^n \sim \mathcal{Q}_{U_2|W_2Y_1}^{\otimes n}(\tilde{w}_{2,b}^n, y_1^n)$. DM 2 transmits the quantised sequence $\hat{u}_{2,b}^n$ with elements $\hat{u}_{2,i,b} = q_{U_2}(u_{2,i,b})$ synchronously with $y_{1,b}^n$.

Sketch for initialisation phase: Before the first block, message m_1 is communicated from DM 1 to DM 2 using a Gel'fand Pinsker coding scheme treating $X_0^{n'}$ as non-causal channel state knowledge. The auxiliary random variable is picked according to Costa [11] with transmit power P so that the rate $R_{GP} = \frac{1}{2}\log(1+\frac{P}{N})$ is achievable. The block length of the initial phase $n' = \alpha n$ is chosen such that message m_1 with rate R can be communicated with an arbitrary small error, i.e., we pick a finite $\alpha > 0$ such that $\alpha > R/R_{GP}$. Beside decoding message m_1, similarly as in [12] where the channel state sequence is estimated, DM 2 will estimate the evolved state sequence $X_1^{n'}$ using the MMSE estimator

$$U_{2,i} = \frac{P+Q}{P+Q+N}\hat{Y}_{1,i}$$

for $1 \le i \le n'$. The corresponding mean-squared state estimation error is given by

$$S' = \mathbb{E}\left[\frac{1}{n'}\left\|X_1^{n'} - U_2^{n'}\right\|_2^2\right] = \frac{(P+Q)N}{P+Q+N}$$

In the following error analysis, the initialisation phase will be denoted as block $b = 0$.

Error analysis per block: Let E^e and $E_b^e(m_{b+1})$ denote the events of a failed encoding process and failed encoding in block b given m_{b+1}, i.e., $E_b^e(m_{b+1}) = E_b^{e,1}(m_{b+1}) \cup E_b^{e,2}$ with $E_b^{e,1}(m_{b+1}) = \{(X_{0,b}^n, W_1^n(m_{b+1}, \ell_b), W_2^n(m_b)) \notin \mathcal{B}_\varepsilon^{(n)}\ \forall(\ell_b, m_b)\}$ and $E_b^{e,2} = \{|c_P(U_{1,b}^n) - P| \ge \frac{1}{4}\varepsilon\}$. Due to the independence between codewords, the probability of an encoding error in block b given no encoding error in previous blocks does not depend on previous blocks. Accordingly, it is sufficient to analyze the case $m_{b+1} = 1$. Thus,

$$\begin{aligned}\mathbb{P}\{E_b^e(M_{b+1}) \,|\, \bigcup_{\beta=b+1}^{B} \bar{E}_\beta^e(M_{\beta+1})\} \\ = \mathbb{P}\{E_b^e(1)\} \le \mathbb{P}\{E_b^{e,1}(1)\} + \mathbb{P}\{E_b^{e,2} \,|\, \bar{E}_b^{e,1}(1)\}\end{aligned}$$

where the bar in $\bar{E}^e$ denotes the complementary event of E^e. If $R \ge I(X_0; W_2) + 3\varepsilon$ and $R_L \ge I(W_1; W_2, X_0) + 4\varepsilon$ following Lemma 10, we have $\mathbb{P}\{E_b^{e,1}(1)\} = \mathbb{P}\{(X_{0,b}^n, W_1^n(1,\ell), W_2(m)) \notin \mathcal{B}_\varepsilon^{(n)}\, \forall\, m, \ell\} \to 0$ as $n \to \infty$. Further, we have $\mathbb{P}\{E_b^{e,2} | \bar{E}_b^{e,1}(1)\} = \mathbb{P}\{|c_P(U_{1,b}^n) - P| \ge \frac{1}{4}\varepsilon \,|\, (X_{0,b}^n, W_1^n(1, L_b), W_2(m)) \in \mathcal{B}_\varepsilon^{(n)}\} \le \sqrt{\delta_n} \to 0$ as $n \to \infty$ due to the law of large numbers.

For the initialisation phase, i.e. block $b = 0$, the encoding and decoding is successful if the message $m_1 \in [1 : 2^{nR}]$ can be successfully send in the initialisation block. This can be done with arbitrary small, but positive probability of error with a sufficiently long block length $n' = \alpha n$ since α has been chosen such that $nR < n'R(1) = \alpha n R_{GP}$ holds. Thus, we have $\mathbb{P}\{E_0^e(M_1) \,|\, \bigcup_{\beta=1}^{B} \bar{E}_\beta^e(M_{\beta+1})\} \to 0$ as well as $\mathbb{P}\{E_0^d\} \to 0$ as $n \to \infty$. It follows that $\mathbb{P}\{E^e\} \to 0$ as $n \to \infty$.

Next, we analyze the decoding error in block b, $1 \le b \le B$. Let $Y_{1,b}^n$ denote the received sequences at DM 2 in block b. Further, let E_b^t denote the event that sequence $Y_{1,b}^n$ is not jointly typical, i.e., $E_b^t = \{(Y_{1,b}^n, W_1^n(M_{b+1}, L_b), \tilde{W}_{2,b}^n) \notin \mathcal{A}_\varepsilon^{(n)}(Y_1, W_1, W_2)\}$. Then the decoding error probability $\mathbb{P}\{E_b^d \,|\, \bigcup_{\beta=0}^{b-1} \bar{E}_\beta^d \cup \bar{E}^e\}$ can be upper bounded by

$$\mathbb{P}\{E_b^d \,|\, \bigcup_{\beta=0}^{b-1} \bar{E}_\beta^d \cup \bar{E}^e \cup \bar{E}_b^t\} + \mathbb{P}\{E_b^t \,|\, \bigcup_{\beta=0}^{b-1} \bar{E}_\beta^d \cup \bar{E}^e\}$$

using the union bound. Using the definition of $\mathcal{B}_\varepsilon^{(n)}$ we obtain the following upper bound for the second term

$$\begin{aligned}&\mathbb{P}\{E_b^t \,|\, \bigcup_{\beta=0}^{b-1} \bar{E}_\beta^d \cup \bar{E}^e\} \\ &= \mathbb{P}\{Y_{1,b}^n \notin \mathcal{A}_\varepsilon^{(n)}(Y_1 \,|\, W_{1,b}^n, W_{2,b}^n) \,|\, (X_{0,b}, W_{1,b}^n, W_{2,b}^n) \in \mathcal{B}_\varepsilon^{(n)}\} \\ &\le \max_{(x_0^n, w_1^n, w_2^n) \in \mathcal{B}_\varepsilon^{(n)}} \eta^{(n)}(x_0^n, w_1^n, w_2^n) \le \sqrt{\delta_n} \to 0 \text{ as } n \to \infty,\end{aligned}$$

which also ensures that $W_1^n(M_{b+1}, L_b)$ will be jointly typical with $Y_{1,b}^n$ and $W_{2,b}^n$. For the correct decoding in block b, $1 \le b \le B$ we have

$$\begin{aligned}&\mathbb{P}\{E_b^d \,|\, \bigcup_{\beta=0}^{b-1} \bar{E}_\beta^d \cup \bar{E}^e \cup \bar{E}_b^t\} \le \mathbb{P}\{\exists \tilde{\ell}_b, \tilde{m}_{b+1} \ne M_{b+1} : \\ &\qquad W_1^n(\tilde{m}_{b+1}, \tilde{\ell}_{b+1}) \in \mathcal{A}_\varepsilon^{(n)}(W_1 | Y_{1,b}^n, W_{2,b}^n)\} \\ &\le \sum_{\substack{\ell_b, \tilde{m}_{b+1}: \\ \tilde{m}_{b+1} \ne M_{b+1}}} \max_{(y_1^n, w_2^n) \in \mathcal{A}_\varepsilon^{(n)}} \mathbb{P}\{W_1^n(\tilde{m}_{b+1}, \tilde{\ell}_{b+1}) \in \mathcal{A}_\varepsilon^{(n)}(W_1 | y_1^n, w_2^n)\} \\ &\le 2^{nR} 2^{nR_L} 2^{-n(I(W_1;Y_1,W_2)-3\varepsilon)} = 2^{n(R+R_L-I(W_1;Y_1,W_2)+3\varepsilon)},\end{aligned}$$

which goes to 0 as $n \to \infty$ if $R + R_L < I(W_1; Y_1, W_2) - 3\varepsilon$. It follows that $\mathbb{P}\{E^d\} \to 0$ as $n \to \infty$.

Witsenhausen cost analysis: We first analyze the cost of control. Let $\psi_b = \psi^{(n)}(X_{0,b}^n, X_{1,b}^n, U_{1,b}^n, U_{2,b}^n, W_{1,b}^n, W_{2,b}^n, Y_{1,b}^n)$ indicate an *error* in block b. From the previous we have $\mathbb{E}[\psi_b] \to 0$ as $n \to \infty$. If $\psi_b = 1$, either for the generated input sequence $u_{1,b}^n$ we have $|c_P(u_{1,b}^n) - P| \ge \frac{1}{2}\varepsilon$, or the first cost constraint is satisfied but the second cost constraint or the jointly typicality condition are not satisfied. If the first constraint is not satisfied, then the encoder sets $u_{1,b}^n$ to the all zero codeword, i.e., bounded error for $\psi_b = 1$ so that $\mathbb{E}[|c_P(U_{1,b}^n) - P|] < \varepsilon$ can be shown for n sufficiently large.

Next, for the estimation error cost we extend the distortion analysis approach by Wyner [9]. Let $\chi_{E,b}$ be an indicator function of the event of an encoding or decoding error in block b. From the previous error analysis we have $\mathbb{P}\{\chi_{E,b}\} \to 0$ as $n \to \infty$. Define $\phi_b = (1-\psi_b)(1-\chi_{E,b})$ indicating the *event of desired sequences that satisfy cost and joint typicality constraints* AND *no coding error event* in block b. From

the previous we have $\mathbb{E}[\phi_b] \to 1$ as $n \to \infty$. In particular, if $\phi_b = 1$, then we have $\mathbb{E}[|c_S(X_{1,b}^n, \hat{U}_{2,b}^n) - \hat{S}|] < \frac{1}{12}\varepsilon$. Therewith, we obtain

$$\begin{aligned}&\mathbb{E}[|c_S(X_{1,b}^n, \hat{U}_{2,b}^n) - \hat{S}]\\ &= \mathbb{E}[\phi_b|c_S(X_{1,b}^n, \hat{U}_{2,b}^n) - \hat{S}|] + \mathbb{E}[\bar{\phi}_b|c_S(X_{1,b}^n, \hat{U}_{2,b}^n) - \hat{S}|]\\ &\leq \tfrac{1}{12}\varepsilon + \mathbb{E}[\bar{\phi}_b\hat{S}] + \mathbb{E}[\bar{\phi}_b c_S(X_{1,b}^n, \hat{U}_{2,b}^n)].\end{aligned}$$

For n sufficiently large we have $\mathbb{E}[\bar{\phi}_b\hat{S}] \leq \frac{1}{12}\varepsilon$ since $\hat{S} < \infty$. The last term can be bounded following Wyner's trick as done in [8], which we however need to extend because the internal state X_1 instead of source X_0 is estimated.

First note that using Cauchy-Schwartz inequality, we have $\sum_{i=1}^n (a_i + b_i)^2 \leq \sum_{i=1}^n a_i^2 + b_i^2 + 2\sqrt{(\sum_{i=1}^n a_i^2)(\sum_{i=1}^n b_i^2)}$, for any $a_i, b_i \in \mathbb{R}$. Since $X_1 = X_0 + U_1$, we have $c_S(X_{1,b}^n, \hat{U}_{2,b}^n) = \frac{1}{n}\sum_{i=1}^n (X_{0i,b} + U_{1i,b} - \hat{U}_{2i,b})^2$. Associating $U_{1i,b}$ as a_i and $X_{0i,b} - \hat{U}_{2i,b}$ as b_i, we obtain the following inequality

$$\begin{aligned}c_S(X_{1,b}^n, \hat{U}_{2,b}^n) \leq c_P(U_{1,b}^n) + c_S(X_{0,b}^n, \hat{U}_{2,b}^n)\\ + 2\sqrt{c_P(U_{1,b}^n)c_S(X_{0,b}^n, \hat{U}_{2,b}^n)}.\end{aligned}$$

Further, the encoding ensures that we always have $c_P(U_{1,b}^n) \leq P + \varepsilon$. Since $\sqrt{\cdot}$ is concave, using Jensen inequality we have

$$\begin{aligned}\mathbb{E}[\bar{\phi}_b c_S(X_{1,b}^n, \hat{U}_{2,b}^n)] \leq \mathbb{E}\{\bar{\phi}_b(P + \varepsilon + c_S(X_{0,b}^n, \hat{U}_{2,b}^n))\}\\ + 2\sqrt{\mathbb{E}[\bar{\phi}_b(P + \varepsilon)c_S(X_{0,b}^n, \hat{U}_{2,b}^n)]}.\end{aligned}$$

Now, we can argue following Wyner's trick exploiting the discretisation of U_2 as follows

$$\begin{aligned}\mathbb{E}[\bar{\phi}_b c_S(X_{0,b}^n, \hat{U}_{2,b}^n)] &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\bar{\phi}_b c_S(X_{0i,b}, \hat{U}_{2,i,b})]\\ &\leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\bar{\phi}_b D(X_{0,i,b})]\end{aligned}$$

with $D(X_{0,i,b}) = \max_{\hat{u}_{2,k}} c_S(X_{0,i,b}, \hat{u}_{2,k})$. The random variables $\{D(X_{0,i,b})\}_i$ are iid and integrable since $c_S(\cdot,\cdot)$ is a squared distance measure and $X_{0,i,b}$ is Gaussian distributed. Next, let $\chi_{\{D(X_{0i,b})>d\}}$ denote an indicator function which is one if $D(X_{0i,b}) > d$. Then we have

$$\begin{aligned}&\mathbb{E}[\bar{\phi}_b(P + \varepsilon + c_S(X_{0,b}^n, \hat{U}_{2,b}^n))]\\ &\quad\leq (P + \varepsilon + d)\,\mathbb{E}[\bar{\phi}_b] + \mathbb{E}[D(X_{0i,b})\chi_{\{D(X_{0i,b})>d\}}]\end{aligned}$$

as well as

$$\begin{aligned}&2\sqrt{\mathbb{E}[\bar{\phi}_b(P + \varepsilon)c_S(X_{0,b}^n, \hat{U}_{2,b}^n)}\\ &\leq 2\sqrt{(P + \varepsilon)(\,\mathbb{E}[\bar{\phi}_b]d + \mathbb{E}[D(X_{0i,b})\chi_{\{D(X_{0i,b})>d\}}])}\end{aligned}$$

Since $D(X_{0i,b})$ is integrable, for any $\varepsilon_d > 0$ there must exist a d_0 such that $\mathbb{E}[D(X_{0i,b})\chi_{\{D(X_{0i,b})>d\}}] < \varepsilon_d$ for all $d > d_0$ due to the monotone convergence theorem. Thus for a sufficiently small ε_d and a sufficiently large n both right hand sides can be upper bounded by $\frac{1}{24}\varepsilon$ so that

$$\mathbb{E}[\bar{\phi}_b c_S(X_{1,b}^n, \hat{U}_{2,b}^n)] \leq \tfrac{1}{12}\varepsilon.$$

Thus, for the costs of block b we have

$$\begin{aligned}\mathbb{E}[|c_S(X_{1,b}^n, \hat{U}_{2,b}^n) - S|] &\leq |\hat{S} - S| + \mathbb{E}[|c_S(X_{1,b}^n, \hat{U}_{2,b}^n) - \hat{S}|]\\ &\leq \tfrac{1}{4}\varepsilon + \mathbb{E}[|c_S(X_{1,b}^n, \hat{U}_{2,b}^n) - \hat{S}|] \leq \tfrac{1}{2}\varepsilon\end{aligned}$$

and $\mathbb{E}\{|c(U_{1,b}^n, X_{1,b}^n, U_{2,b}^n) - P - S|\} \leq \frac{1}{2}\varepsilon$.

Lastly, we have to include the cost of the initialisation block $b = 0$. Since the average transmit power in the initial phase is also set to P, we have

$$\begin{aligned}\mathbb{E}[|c_P(U_1^{Bn+n'}) - P|] \leq \frac{\alpha n}{(B+\alpha)n}\mathbb{E}[|c_P(U_{1,0}^{n'}) - P|]\\ + \frac{n}{(B+\alpha)n}\sum_{b=1}^B \mathbb{E}[|c_P(U_{1,b}^n) - P|] \leq \varepsilon\end{aligned}$$

For the estimation error, the initial phase results in a larger but bounded error average error $S' < \infty$. The impact however can be made arbitrary small with a sufficiently large number of blocks B as follows

$$\begin{aligned}\mathbb{E}[|c_S(X_1^{Bn+n'}, U_2^{Bn+n'}) - S|] \leq \frac{\alpha n}{(B+\alpha)n}\mathbb{E}[|c_S(X_1^{n'}, U_2^{n'})\\ - S|] + \frac{n}{(B+\alpha)n}\sum_{b=1}^B \mathbb{E}[|c_S(X_{1,b}^n, U_{2,b}^n) - S|] \leq \varepsilon\end{aligned}$$

for n and B sufficiently large.

Lastly, the existence of a coordination scheme follows from the extension of the random coding argument as in the proof of [10, Lemma 2.2].

Closedness: The previous holds if the rate constraint holds with strict inequality. For equality, we can argue as in [5, Appendix C], i.e., since $N < \infty$, we can always find an approximation of the random variables W_1, W_2, U_1 and U_2 that result in an arbitrary small increase of the costs, but satisfy the rate constraint with strict inequality. □

REFERENCES

[1] H. Witsenhausen, "A counterexample in stochastic optimum control," *SIAM Journal on Control*, vol. 6, no. 1, pp. 131–147, 1968.

[2] S. Yuksel and T. Basar, *Stochastic Networked Control Systems: Stabilization and Optimization under Information Constraints*, ser. Systems & Control Foundations & Applications. New York, NY: Springer, 2013.

[3] P. Grover and A. Sahai, "Witsenhausen's counterexample as assisted interference suppression," in *Int. J. Syst., Control Commun.*, 2010.

[4] M. Le Treust and T. J. Oechtering, "Optimal Control Designs for Vector-valued Witsenhausen Counterexample Setups," *IEEE 56th Annual Allerton Conf. on Commun., Control, and Comp.*, Sept. 2018.

[5] M. Le. Treust, "Joint Empirical Coordination of Source and Channel," *IEEE Trans. Inf. Theory*, vol. 63, no. 8, pp. 5087–5114, Aug. 2017.

[6] C. Choudhuri and U. Mitra, "On Witsenhausen's counterexample: the asymptotic vector case," *2012 IEEE ITW*, pp. 162–166, July 2012.

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Wiley & Sons, 2006.

[8] M. T. Vu, T. J. Oechtering, and M. Skoglund, "Gaussian Hierarchical Identification with Pre-processing," in *Proc. IEEE Data Compression Conf.*, Mar. 2018. [Submitted journal version]. Available: https://people.kth.se/$\sim$oech/GaussIdent.pdf

[9] A. D. Wyner, "The rate-distortion function for source coding with side information at the decoder-ii: General sources," *Inform. & control,* 1978.

[10] M. Bloch and J. Barros, *Physical-layer Security: From Information Theory to Security Engineering*. Cambridge University Press, 2011.

[11] M. H. M. Costa, "Writing on dirty paper," *IEEE Trans. on Inform. Theory*, vol. 29, no. 3, pp. 439–441, 1983.

[12] A. Sutivong, M. Chiang, T. M. Cover, and Y.-H. Kim, "Channel capacity and state estimation for state-dependent Gaussian channels," *IEEE Trans. Inf. Theory,* vol. 51, no. 4, pp. 1486?1495, Apr. 2005.