Lower bounds on the estimation error in problems of distributed computation

Giacomo Como
Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
77 Massachusetts Avenue, Cambridge, MA
Email: giacomo@mit.edu

Munther Dahleh
Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
77 Massachusetts Avenue, Cambridge, MA
Email: dahleh@mit.edu

Abstract—**Information-theoretic lower bounds on the estimation error are derived for problems of distributed computation. These bounds hold for a network attempting to compute a real-vector-valued function of the global information, when the nodes have access to partial information and can communicate through noisy transmission channels. The presented bounds are algorithm-independent, and improve on recent results by Ayaso et al., where the exponential decay rate of the mean square error was upper-bounded by the minimum normalized cut-set capacity. We show that, if the transmission channels are stochastic, the highest achievable exponential decay rate of the mean square error is in general strictly smaller than the minimum normalized cut-set capacity of the network. This is due to atypical channel realizations, which, despite their asymptotically vanishing probability, affect the error exponent.**

I. Introduction

As large-scale networks have emerged –characterized by the lack of centralized access to information, and possibly time-varying topologies–, problems of distributed computation have received an increasing amount of attention by the research community in the last few years. In these scenarios, large collections of agents –each having access to some partial information– aim at computing an application-specific function of the global information. The computation must be completely distributed, i.e. each agent can rely only on local observations, while iteratively processing the available information and communicating with the other agents. The main challenge in the design of such distributed computation systems is posed by the scarce energetic autonomy of the agents, which severely constrains both their computational and communication capabilities. In the present paper we shall focus on the latter and investigate the fundamental performance limitations of distributed computation algorithms on networks with noisy communication channels.

Different models for problems of distributed computation over networks have been proposed in the information-theoretic literature: the reader is referred to [1] for an overview of the main research lines which have been developed. In this paper, we shall study the case of a network attempting to evaluate a real-vector-valued function of the global information with increasing precision. The motivations for considering such a model mainly come from applications to distributed inference and control, as well as to opinion dynamics, where the quantities of interest are very often continuous- rather than discrete-valued. As an example, one can consider the average consensus problem, which has been the object of recent extensive research: here, each node of the network has access to a real number –or possibly a vector– representing a noisy measurement of the same physical quantity, and the goal is to evaluate the arithmetic mean of all the measurements.

The recent literature on distributed control and estimation problems with communication constraints has highlighted the centrality of delay. In fact, large delays can be detrimental for the overall system performance. For this reason, one of the main performance measures of distributed computation algorithms is the speed of convergence to zero of the estimation error, i.e. of the distance between the value of the function to be evaluated and the estimate each node of the network has of it. In the recent work [2], which considers a framework very similar to the one studied here, it was shown that the mean square error of the nodes' estimates of the global function cannot decrease to zero at an exponential rate faster than the normalized capacity of the worst cut-set of the network.

In the present paper, upper bounds will be proved for the exponential decay rate of the tails of the probability distribution of the error made by any node in the network in estimating a function of the global information. As a corollary, upper bounds on the exponential decay rate of arbitrary moments of the estimation error will be obtained. In particular, it will be shown that, for non-deterministic channels, the exponential decay rate of any moment of the error is bounded away from worst normalized cut-set capacity. The insufficiency of the Shannon capacity as a measure of the achievable performance stems from the atypical channel realizations which, despite their asymptotically vanishing probability, strongly impact the error rate. This observation is coherent with some of the available results in the literature on control and estimation with communication constraints [3], [4].

Our approach draws on techniques developed for upper bounds on the error exponent of fixed-length block-codes on discrete memoryless channels with feedback [5], [6], combined with a novel inequality playing the role of Fano's inequality in Euclidean spaces. Our arguments involve three main steps. First, an upper bound on the probability that two real-vector-valued random variables are within a certain distance is derived in terms of their conditional entropy. Second, using network-information-theoretic techniques, the conditional entropy between a function of the global information and the estimate any node of the network can have of it is bounded in terms of the mutual information across a cut-set of the network. Finally, a change of probability measure argument is used in order to capture the large deviations of the channel behaviour.

The remainder of this paper is organized as follows. In Sect. II a the problem is formally stated and the main results of the paper are presented. Sect. III contains two of the aforementioned technical results: Sect. III-A presents a Fano-like inequality in Euclidean spaces, while Sect. III-B we discuss bounds for the conditional entropy across a cut-set. A change of probability measure argument is developed in Sect. IV-A, and subsequently applied in Sect. IV-B and Sect. IV-C in order to prove the main results.

II. Problem statement and main results

In this section, we shall present a formal statement of the problem and anticipate the main results of the paper, to be proved in the following sections.

We start by introducing a few notational conventions. The set of the first n naturals will be denoted by $[n] := \{1, \dots, n\}$. For subscript-indexed (respectively superscript-indexed) vector $v = (v_i)_{i\in\mathcal{I}}$ ($v = (v^{(i)})_{i\in\mathcal{I}}$), and a subset of indices $\mathcal{S} \subseteq \mathcal{I}$, $v_\mathcal{S} := (v_i)_{i\in\mathcal{S}}$ ($v^{(\mathcal{S})} = (v^{(i)})_{i\in\mathcal{S}}$) will denote the restriction of v to $\mathcal{S}$. For two finite-valued random variables (r.v.) V, W, the entropy of V, the conditional entropy of V given W and their mutual information will be denoted by $\mathrm{H}(V)$, $\mathrm{H}(V|W)$ and $I(V;W)$, respectively. The same notation will be used for continuous-valued random variables to denote their differential entropy, conditional entropy and mutual information. With a common abuse of notation, for a probability measure μ on $\mathbb{R}^d$ $\mathrm{H}(\mu)$ will denote its differential entropy (whenever it exists); for $x \in [0,1]$, $\mathrm{H}(x)$ will denote the binary entropy of x.

We shall consider a network consisting of a finite set of nodes $\mathcal{V}$. Each node v has access to some local information, by observing a r.v. W_v; the complete vector of observations will be denoted by $W = (W_v)_{v\in\mathcal{V}}$. The goal of the network is to evaluate a function $Z = f(W)$ of the global information in a distributed way, through successive rounds of computation/communication. At each time $t \in \mathbb{N}$, every node $v \in \mathcal{V}$ transmits a signal $X_t^{(v)}$, and receives a signal $Y_t^{(v)}$; $X_t = (X_t^{(v)})_{v\in\mathcal{V}}$ and $Y_t = (Y_t^{(v)})_{v\in\mathcal{V}}$ will denote the complete vectors of transmitted and received signals, respectively. The communication channel is represented by a stochastic kernel $P(y|x)$ describing the probability that $Y_t = y$ is received given that $X_t = x$ has been transmitted. The channel is assumed to be memoryless, i.e. Y_t is conditionally independent from $W, X_{[t-1]}, Y_{[t-1]}$ given X_t. Distributedness of the algorithm is then ensured by requiring that $X_t^{(v)}$ depends only on the local information $(W_v, Y^{(v)})_{[t-1]})$ available at node v at the beginning of the t-th round of communication. Finally, at time t, each node v makes an estimate $\hat{Z}_t^{(v)}$ of Z based on the local information $(W_v, Y_{[t]}^{(v)})$ available at the end of the t-th round of communication. The performance of the distributed computation algorithm is measured in terms of the decay rate of the estimation errors of the nodes

$$\Delta_t^{(v)} := \left|\left|Z_t^{(v)} - Z\right|\right|, \qquad v \in \mathcal{V}, \tag{1}$$

where $||z||$ denotes the Euclidean norm of a vector z.

More formally, we shall assume that the r.v. observed by node v, W_v , takes values in some measurable space $\mathcal{W}_v$.[1] The a priori distribution of the complete observation vector W is described by an arbitrary probability measure μ_W on the product space $\mathcal{W} := \prod_{v\in\mathcal{V}} \mathcal{W}_v$. The measure μ_W need not have a product structure, so that the proposed model is able to handle the case of correlated observations. The function

$$f : \mathcal{W} \to \mathbb{R}^d$$

is assumed to be measurable. The transmitted (respectively received) signals $X_t^{(v)}$ ($Y_t^{(v)}$) take values in a finite alphabet $\mathcal{X}_v$ ($\mathcal{Y}_v$); the complete channel input (output) alphabet will be denoted by $\mathcal{X} := \prod_{v\in\mathcal{V}} \mathcal{X}_v$ ($\mathcal{Y} := \prod_{v\in\mathcal{V}} \mathcal{Y}_v$). The distributed algorithm consists of a sequence of encoders $\Phi = (\phi_t^{(v)})$ and a sequence of decoders $\Psi = (\psi_t^{(v)})$, where

$$\phi_t^{(v)} : \mathcal{W}_v \times \mathcal{Y}_v^{t-1} \to \mathcal{X}_v\,, \qquad \psi_t^{(v)} : \mathcal{W}_v \times \mathcal{Y}_v^t \to \mathbb{R}^d\,,$$

are measurable functions, such that

$$X_t^{(v)} = \phi_t^{(v)}\Big(W_v, Y_{[t-1]}^{(v)}\Big)\,, \qquad Z_t^{(v)} = \psi_t^{(v)}\Big(W_v, Y_{[t]}^{(v)}\Big)\,. \tag{2}$$

Observe that the a priori measure μ_W, the encoders' sequence Φ and the channel P naturally define a joint probability measure $\mathbb{P}$ on the space $\Omega := \mathcal{W} \times \mathcal{Y}^\mathbb{N}$, equipped with its standard product sigma-field $\mathcal{A}$. All the r.v.s of interest can be though of as defined over $(\Omega, \mathcal{A}, \mathbb{P})$. Throughout the paper the symbol $\mathbb{E}$ will denote the expectation operator with respect to this probability space. We shall make the following assumption on μ_W and f.

Assumption 1. (a) $\mathrm{H}(Z|W_\mathcal{S}) < +\infty$ *for all* $\mathcal{S} \subsetneq \mathcal{V}$;
(b) $m := \mathbb{E}[||Z||^2] < +\infty$.

In the rest of the paper bounds on the estimation error will be derived, which depend on the channel P, the a priori measure μ_W, as well as the function f, and hold for any distributed algorithm (Φ, Ψ). Although some of the arguments which will be presented hold true for general memoryless channels $P(\,\cdot\,|\,\cdot\,)$, we shall confine our discussion to channels which are adapted to some graph topology. More precisely, we shall consider a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{E} \subseteq \mathcal{V}^2 \setminus \{(v,v)|v \in \mathcal{V}\}$ is a set of directed edges. To each edge $e \in \mathcal{E}$ a discrete memoryless channel is associated, having finite input $\mathcal{X}_e$, output $\mathcal{Y}_e$, and transition probabilities $P_e(y|x)$. Transmission is assumed to be independent among the different edges, so that

$$\mathcal{X} = \prod_{e\in\mathcal{E}} \mathcal{X}_e\,, \quad \mathcal{Y} = \prod_{e\in\mathcal{E}} \mathcal{Y}_e\,, \quad P(y|x) = \prod_{e\in\mathcal{E}} P_e(y_e|x_e)\,.$$

[1] For concreteness the reader may assume that $\mathcal{W}_v = \mathbb{R}^n$ for some $n \geq d$, though this assumption is not needed. Keeping this abstract setting allows to treat many different cases of relevant interest at once.

The bounds presented in this paper involve cut-set arguments. Given a proper subset of nodes, $\emptyset \neq \mathcal{S} \subsetneq \mathcal{V}$, we imagine to have cut the graph $\mathcal{G}$ by an hypothetic boundary leaving nodes in $\mathcal{S}$ on left-hand side and nodes on $\mathcal{S}^c$ on the right-hand side. Consider the cut-set $\mathcal{E}_\mathcal{S} := \mathcal{S} \times \mathcal{S}^c \cap \mathcal{E}$ of edges crossing this boundary from left to right, and the associated memoryless channel, having input, output and transition probabilities respectively given by

$$\mathcal{X}_\mathcal{S} := \prod_{e\in\mathcal{E}_\mathcal{S}} \mathcal{X}_e\,, \quad \mathcal{Y}_\mathcal{S} := \prod_{e\in\mathcal{E}_\mathcal{S}} \mathcal{Y}_e\,, \quad P_\mathcal{S}(y|x) = \prod_{e\in\mathcal{E}_\mathcal{S}} P_e(y_e|x_e)\,.$$

Let $\mathcal{Q}_\mathcal{S}$ be the class of all stochastic kernels with input $\mathcal{X}_\mathcal{S}$ and output $\mathcal{Y}_\mathcal{S}$. For $Q \in \mathcal{Q}_\mathcal{S}$, we shall denote by

$$C_Q := \max I\left(X_\mathcal{S}, Y_\mathcal{S}\right)$$

its Shannon capacity, and by

$$D\left(Q||P_\mathcal{S}\right) := \max_{x\in\mathcal{X}_\mathcal{S}} D\left(Q(\,\cdot\,|x)||P_\mathcal{S}(\,\cdot\,|x)\right)$$

the maximal Kullback-Leiber divergence between the output distributions of Q and $P_\mathcal{S}$.

The main result of this paper consists in an upper bound on the exponential error decay of the estimation error. Define

$$E_\mathcal{S}(R) := \min\left\{D\left(Q||P_\mathcal{S}\right) \,|\, Q \in \mathcal{Q}_\mathcal{S} : C_Q \leq R\right\}\,; \qquad (3)$$

The quantity $E_\mathcal{S}(R)$ coincides with the Dobrushin-Haroutunian's bound on the error exponent of rate-R fixed-length block-codes with feedback on the channel $P_\mathcal{S}$ [6]. Let

$$\Gamma_t^{(v)} := -\frac{1}{t}\log\Delta_t^{(v)}\,, \qquad v \in \mathcal{V}, t \in \mathbb{N}\,. \qquad (4)$$

The following statement is proved in Sect. IV-B.

Theorem 1. *If Assumption 1 holds,*

$$\limsup_t \left\{-\frac{1}{t}\log\mathbb{P}\left(\Gamma_t^{(v)} \geq R\right)\right\} \leq \min_{\substack{\emptyset\neq\mathcal{S}\subsetneq\mathcal{V}:\\ v\in\mathcal{S}^c}} E_\mathcal{S}(Rd)\,, \qquad (5)$$

for every node $v \in \mathcal{V}$.

As a corollary of Theorem 1, it is possible to get, for all $\eta > 0$, an upper bound on the exponential decay rate of the average η-moment of the error:

$$\Lambda_t^{(\eta)} := \left(\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}\left[\Delta_t^{(v)}\right]^\eta\right)^{1/\eta}\,. \qquad (6)$$

Define

$$\beta_\eta^{(\mathcal{S})} := \min_{Q\in\mathcal{Q}_\mathcal{S}}\left\{\tfrac{1}{d}C_Q + \tfrac{1}{\eta}D(Q||P_\mathcal{S})\right\}, \quad \beta_\eta := \min_{\emptyset\neq\mathcal{S}\subsetneq\mathcal{V}}\left\{\beta_\eta^{(\mathcal{S})}\right\}. \qquad (7)$$

The following result is proved in Sect. IV-C.

Corollary 1. *If Assumption 1 holds,*

$$\limsup_t -\frac{1}{t}\log\Lambda_t^{(\eta)} \leq \beta_\eta\,. \qquad (8)$$

A few comments are in order. First, observe that Assumption 1(a) captures a fundamental feature of the distributed computation problem, namely that no proper subset of the nodes has enough information in order to compute $Z = f(W)$. On the other end, Assumption 1(b) is more of a technical nature: for instance it guarantees that $\mathrm{H}(Z)$ exists and is bounded from above by some finite constant (see Lemma 1).

Second, observe that $\beta_\eta^{(\mathcal{S})} \leq \frac{1}{d}C_{P_\mathcal{S}}$, as can be easily seen by choosing $Q = P_\mathcal{S}$ in (7). In particular, $\beta_\eta^{(\mathcal{S})} = \frac{1}{d}C_{P_\mathcal{S}}$ whenever $P_\mathcal{S}$ is a deterministic channel, i.e. when, for all $x \in \mathcal{X}_\mathcal{S}$, $P_\mathcal{S}(\,\cdot\,|x) = \delta_{y_x}$ for some $y_x \in \mathcal{Y}_\mathcal{S}$. Indeed, in this case, the only stochastic kernel $Q \in \mathcal{Q}_\mathcal{S}$ such that $D(Q||P_\mathcal{S}) < +\infty$ is $P_\mathcal{S}$ itself. Hence, for deterministic channels, Corollary 1 states that the exponential rate of the mean square error is upper-bonded by $1/d$ times the capacity of the worst cut-set in the network. However, for channels that are not deterministic, it can be shown that $\beta_\eta^{(\mathcal{S})} < \frac{1}{d}C_\mathcal{S}$, i.e. the achievable exponential decay rate of the mean square error is strictly smaller than the normalized capacity of the worst cut-set in the network. In particular, for any non-deterministic channel $P_\mathcal{S}$, it is not hard to see that

$$\lim_{\eta\to+\infty}\beta_\eta^{(\mathcal{S})} = 0\,. \qquad (9)$$

Equation (9) has to be interpreted as follows: the higher η, the more detrimental atypical channel realizations are for the system performance.

III. A first bound based on the cut-set capacity

A. A Fano-like inequality in Euclidean spaces

We shall obtain a result which may be interpreted as a geometric analogous of Fano's inequality for real-vector-valued r.v.s.

Recall that Fano's inequality states that for two r.v.s $Z, \hat{Z}$, taking values in a finite set $\mathcal{Z}$, the probability p that $\hat{Z} = Z$ can be estimated in terms of the conditional entropy $\mathrm{H}(Z|\hat{Z})$ as follows:

$$(1-p)\log(|\mathcal{Z}|-1) + \mathrm{H}(p) \leq \mathrm{H}(Z|\hat{Z})\,. \qquad (10)$$

The proof of (10) relies on two basic properties of the discrete entropy function: its grouping property, and the fact that the entropy of a probability measure over a finite set is upper-bounded by that of a uniform measure on that set.

In the what follows, we wish to prove a similar result for two r.v.s $W, \hat{W}$ taking values in the d-dimensional Euclidean space $\mathbb{R}^d$. Rather than estimating the probability that W and $\hat{W}$ coincide, [2] we shall look at the probability that the distance between W and $\hat{W}$ does not exceed some positive constant r. We shall estimate this probability in terms of the logarithm of the volume of a ball of radius r in $\mathbb{R}^d$, and of the conditional entropy associated to the joint law of $\hat{Z}$ and Z. Beside the grouping property of the entropy functional, our proof relies on some variational properties of the entropy which are recalled in the following lemma.

Lemma 1. *Consider $\mu \in \mathcal{P}(\mathbb{R}^d)$. Then:*

(a) *if μ is supported in some compact subset $A \subseteq \mathbb{R}^d$,*

$$\mathrm{H}(\mu) \leq \log\lambda_d(A)\,,$$

[2] However, see the remark following Lemma 2.

with equality if and only if μ is the uniform measure over A.

(b) *if $\int_{\mathbb{R}^d} ||z||^2 \mathrm{d}\mu(z) \leq md$ for some $m > 0$,*

$$\mathrm{H}(\mu) \leq \frac{d}{2} \log(2\pi e m) \,,$$

with equality if and only if μ is a homogeneous, zero-mean, d-dimensional Gaussian measure.

Lemma 2. *Let Z and $\hat{Z}$ be two $\mathbb{R}^d$-valued r.v.s, with joint probability law $\mu_{Z,\hat{Z}}$, and such that*

$$m := \mathbb{E}\left[||Z||^2\right] < +\infty \,.$$

For any $r > 0$, let $A_r := \{(z,\hat{z}) : ||z - \hat{z}|| \leq r\} \subseteq \mathbb{R}^d \times \mathbb{R}^d$. Then,

$$\mu(A_r) \log r^d + \frac{d}{2} \log(J_d m) \geq \mathrm{H}\left(Z|\hat{Z}\right) . \qquad (11)$$

where $J_d := \frac{4\pi e}{d}(2K_d)^{2/d}$, with $K_d := \frac{\pi^{d/2}}{\Gamma(d/2+1)}$ denoting the volume of a unitary ball in $\mathbb{R}^d$. [3]

Remark: Notice that the assumption $m < +\infty$ implies that $\mathrm{H}(Z) < +\infty$, and, a fortiori, $\mathrm{H}(Z|\hat{Z}) < +\infty$. Observe that (11) holds true also for $r = 0$: in this case, it implies that either $\mu(A_r) = 0$ or $\mathrm{H}(Z|\hat{Z}) = -\infty$.

Proof: Let $\mu_{\hat{Z}}$ and $\mu_{Z|\hat{Z}}(\,\cdot\,|Z)$ be the marginal law of $\hat{Z}$ and the conditional law of Z given $\hat{Z}$, respectively. [4] Since $\mathrm{H}(Z|\hat{Z}) < +\infty$, necessarily

$$\mathrm{H}\left(\mu_{Z|\hat{Z}}(\,\cdot\,|\hat{Z})\right) < +\infty \,, \quad \mu_{\hat{Z}} - a.s.$$

For $\hat{z} \in \mathbb{R}^d$, let us denote by $B_{\hat{z}} := \{z : ||z - \hat{z}|| \leq r\} \subset \mathbb{R}^d$ the closed ball centered in $\hat{z}$ of radius r, and let $B^c_{\hat{z}} := \mathbb{R}^d \backslash B_{\hat{z}}$. For $\hat{z} \in \mathbb{R}^d$, let

$$p_{\hat{z}} := \mu_{Z|\hat{Z}}(B_{\hat{z}}|\hat{z}) \,, \qquad q_{\hat{z}} := \mu_{Z|\hat{Z}}(B^c_{\hat{z}}|\hat{z}) = 1 - p_{\hat{z}} \,,$$

and define the probability measures $\nu_{\hat{v}}, \gamma_{\hat{v}} \in \mathcal{P}(\mathbb{R}^d)$ by

$$\nu_{\hat{z}}(A) := \frac{1}{p_{\hat{z}}} \mu_{Z|\hat{Z}}(A \cap B_{\hat{z}}|\hat{z}) \,, \quad \gamma_{\hat{z}}(A) := \frac{1}{q_{\hat{z}}} \mu_{Z|\hat{Z}}(A \cap B^c_{\hat{z}}|\hat{z}) \,,$$

for all Borel set $A \subseteq \mathbb{R}^d$. By the grouping property of differential entropy, we have that

$$\mathrm{H}(p_{\hat{z}}) + p_{\hat{z}} \mathrm{H}(\nu_{\hat{z}}) + q_{\hat{z}} \mathrm{H}(\gamma_{\hat{z}}) \,. \qquad (12)$$

Since $\nu_{\hat{z}}$ is supported on $B_{\hat{z}}$, Lemma 1 (a) allows one to estimate its entropy by that of a uniform measure on $B_{\hat{z}}$:

$$\mathrm{H}(\nu_{\hat{z}}) \leq \log \mathrm{Vol}(B_{\hat{z}}) = \log\left(K_d r^d\right) \,. \qquad (13)$$

On the other hand, Lemma 1 (b) allows one to estimate the entropy of $\gamma_{\hat{z}}$ with that of a zero-mean homogeneous Gaussian measure with the same second moment, obataining

$$\mathrm{H}(\gamma_{\hat{z}}) \leq \frac{d}{2} \log\left(\frac{2\pi e}{d} \int_{\mathbb{R}^d} ||z||^2 \mathrm{d}\gamma_{\hat{z}}(z)\right) \,. \qquad (14)$$

Now, observe that

$$\begin{aligned} q_{\hat{z}} \int_{\mathbb{R}^d} ||z||^2 \mathrm{d}\gamma_{\hat{z}}(z) &= \int_{B^c_{\hat{z}}} ||z||^2 \mathrm{d}\mu_{Z|\hat{Z}}(z|\hat{z}) \\ &\leq \int_{\mathbb{R}^d} ||z||^2 \mathrm{d}\mu_{Z|\hat{Z}}(z|\hat{z}) \quad =: \; m_{\hat{z}} \,. \end{aligned} \qquad (15)$$

By combining formula (12) with the inequalities (13), (14) and (15), and using the fact that

$$-x \log x \leq \mathrm{H}(x) \leq \log 2 \,, \qquad \forall\, 0 \leq x \leq 1 \,,$$

we get that, $\mu_{\hat{Z}}$-almost surely,

$$\begin{aligned} \mathrm{H}(\mu_{Z|\hat{Z}}(\,\cdot\,|\hat{Z})) &= p_{\hat{Z}} \mathrm{H}(\nu_{\hat{Z}}) + q_{\hat{Z}} \mathrm{H}(\gamma_{\hat{Z}}) + \mathrm{H}(p_{\hat{Z}}) \\ &\leq p_{\hat{Z}} \log(K_d r^d) + q_{\hat{Z}} \tfrac{d}{2} \log \tfrac{2\pi e m_{\hat{Z}}}{d q_{\hat{Z}}} + \log 2 \\ &\leq p_{\hat{Z}} \log r^d + \tfrac{d}{2} \log\left(J_d m_{\hat{Z}}\right) \,. \end{aligned}$$

Hence, Jensen's inequality implies that

$$\begin{aligned} \mathrm{H}\left(Z|\hat{Z}\right) &= \mathbb{E}\left[\mathrm{H}\left(\mu_{Z|\hat{Z}}(\,\cdot\,|\hat{Z})\right)\right] \\ &\leq \mathbb{E}\left[p_{\hat{Z}}\right] \log r^d + \tfrac{d}{2} \mathbb{E}\left[\log\left(J_d m_{\hat{Z}}\right)\right] \\ &\leq \mu(A_r) \log r^d + \tfrac{d}{2} \log(J_d m) \,, \end{aligned}$$

and the claim follows. ■

B. Bounding the conditional entropy through a cut-set

Consider a non-trivial cut-set $\mathcal{E}_{\mathcal{S}}$. For an arbitrary node on the right-hand side, $v \in \mathcal{S}^c$, Lemma 2 can be applied in order to upper bound the left tails of the estimation error $\Delta^{(v)}_t$ in terms of the conditional entropy $\mathrm{H}(Z|\hat{Z}^{(v)}_t)$. The next natural step consists in deriving a lower bound on $\mathrm{H}(Z|\hat{Z}^{(v)}_t)$, a task which is accomplished below. The key idea, borrowed from standard cut-set arguments in network information theory [7, pagg. 587-594], consists in relaxing the problem, by assuming that all the nodes on the left-hand side of the cut can share instantaneous information among themselves in order to establish communication in the most efficient way with the nodes on the right-hand side, which in turn are able to distribute the received information instantaneously among themselves. These arguments lead to the proof of the following result.

Lemma 3. *Let $\mathcal{E}_{\mathcal{S}}$ be non-trivial cut-set. Then,*

$$\mathrm{H}\left(Z|\hat{Z}^{(v)}_t\right) \geq \mathrm{H}(Z|W_{\mathcal{S}^c}) - \sum_{1 \leq j \leq t} I\left(X^{(\mathcal{S})}_j, \mathcal{Y}^{(\mathcal{S}^c)}_j | X^{(\mathcal{S}^c)}_j\right) \,, \qquad (16)$$

for every node $v \in \mathcal{S}^c$, and all $t \in \mathbb{N}$.

Proof: It is an immediate consequence of the assumption (2) that the vector $X^{(\mathcal{S}^c)}_t$ of the signals transmitted by all the nodes on the left-hand side of the cut, is a function of the total information available to them $\left(W_{\mathcal{S}^c}, Y^{(\mathcal{S})}_{[t-1]}\right)$. Again from (2), it follows that the estimation $\hat{Z}^{(v)}_t$ is a function of the total information $\left(W_{\mathcal{S}^c}, Y^{(\mathcal{S}^c)}_{[t]}\right)$ available at the right-hand side of

[3] Here $\Gamma(\,\cdot\,)$ denotes Euler's Gamma function.

[4] Recall that $\mu_{Z|\hat{Z}}(\,\cdot\,|Z)$ is a random probability measure on $\mathbb{R}^d$, which is well defined $\mu_{\hat{Z}}$-almost surely.

the cut. As a consequence, we have the following chain of (in)equalities:

$$\begin{aligned}
& \mathrm{H}\left(Z|\hat{Z}_t^{(v)}\right) \\
\overset{(a)}{\geq}\ & \mathrm{H}\left(Z|\hat{Z}_t^{(v)}, W_{\mathcal{S}^c}\right) \\
=\ & \mathrm{H}(Z|W_{\mathcal{S}^c}) - I\left(Z;\hat{Z}_t^{(v)}|W_{\mathcal{S}^c}\right) \\
\overset{(b)}{\geq}\ & \mathrm{H}\left(Z|W_{\mathcal{S}^c}\right) - I\left(W; W_{\mathcal{S}^c}, Y_{[t]}^{(\mathcal{S}^c)}|W_{\mathcal{S}^c}\right) \\
=\ & \mathrm{H}\left(Z|W_{\mathcal{S}^c}\right) - I\left(W_{\mathcal{S}}; Y_{[t]}^{(\mathcal{S}^c)}|W_{\mathcal{S}^c}\right) \\
\overset{(c)}{=}\ & \mathrm{H}\left(Z|W_{\mathcal{S}^c}\right) - \sum_{1\leq j\leq t} I\left(W_{\mathcal{S}}; Y_j^{(\mathcal{S}^c)}|W_{\mathcal{S}^c}, Y_{[j-1]}^{(\mathcal{S}^c)}\right) ,
\end{aligned} \tag{17}$$

where: inequality (a) follows since conditioning does not increase entropy; inequality (b) is a consequence of the data processing inequality and the fact that $Z = f(W)$ and $\hat{Z}_t^{(v)}$ is a function of $W_{\mathcal{S}^c}$ and $Y_{[t]}^{(\mathcal{S}^c)}$; equality (c) follows from the chain rule for mutual information. Now, for all $1 \leq j \leq t$, we have that

$$\begin{aligned}
& I\left(X_j^{(\mathcal{S})}; Y_j^{(\mathcal{S}^c)}|X_j^{(\mathcal{S}^c)}\right) \\
\overset{(d)}{=}\ & I\left(Y_j^{(\mathcal{S}^c)}; X_j^{(\mathcal{S})}|W_{\mathcal{S}^c}, Y_{[j-1]}^{(\mathcal{S}^c)}, X_j^{(\mathcal{S}^c)}\right) \\
=\ & \mathrm{H}\left(Y_j^{(\mathcal{S}^c)}|W_{\mathcal{S}^c}, Y_{[j-1]}^{(\mathcal{S}^c)}, X_j^{(\mathcal{S}^c)}\right) - \mathrm{H}\left(Y_j^{(\mathcal{S}^c)}|W, Y_{[j-1]}^{(\mathcal{S}^c)}, X_j\right) \\
\overset{(e)}{\geq}\ & \mathrm{H}\left(Y_j^{(\mathcal{S}^c)}|W_{\mathcal{S}^c}, Y_{[j-1]}^{(\mathcal{S}^c)}\right) - \mathrm{H}\left(Y_j^{(\mathcal{S}^c)}|W, Y_{[j-1]}^{(\mathcal{S}^c)}\right) \\
=\ & I\left(W_{\mathcal{S}}; Y_j^{(\mathcal{S}^c)}|W_{\mathcal{S}^c}, Y_{[j-1]}^{(\mathcal{S}^c)}\right) ,
\end{aligned} \tag{18}$$

where: equality (d) follows from the fact that, due to the assumptions of causality of the encoders and memorylessness of the channel, $Y_j^{(\mathcal{S}^c)}$ is conditionally independent from W and $Y_{1,\ldots,j-1}^{(\mathcal{S}^c)}$ given X_j; inequality (e) follows form the fact that, as observed, $X_j^{(\mathcal{S}^c)}$ is a function of $W_{\mathcal{S}^c}$ and $Y_{[j-1]}^{(\mathcal{S}^c)}$, whereas removing the conditioning does not increase the entropy in the second term. Therefore, by combining (17) with (18), the claim (16) follows. ■

Observe that Lemma 3 holds for every memoryless channel $P(\,\cdot\,|\,\cdot\,)$. Imposing the further constraint that P is adapted to some graph topology $\mathcal{G}$ allows one to bound the conditional mutual information terms $I\left(X_j^{(\mathcal{S})}; Y_j^{(\mathcal{S}^c)}|X_j^{(\mathcal{S}^c)}\right)$, as in the following statement.

Proposition 1. *Let $(\mathcal{S}, \mathcal{S}^c)$ be a non-trivial cut, and $v \in \mathcal{S}^c$. Then, for every $t \in \mathbb{N}$ and $r > 0$,*

$$-\mathbb{P}\left(\Delta_t^{(v)} \leq r\right) \log r^d \leq tC_{\mathcal{S}} + \frac{d}{2}\log\left(J_d m\right) - \mathrm{H}(Z|W_{\mathcal{S}^c}) , \tag{19}$$

where $C_{P_{\mathcal{S}}} := \sum_{e\in\mathcal{E}_{\mathcal{S}}} C_e$ is the capacity of the cut-set $\mathcal{E}_{\mathcal{S}}$.

Proof: By applying Lemmas 2 and 3, one gets, for every node $v \in \mathcal{S}^c$,

$$-\mathbb{P}\left(\Delta_t^{(v)} \leq r\right) \log r^d \leq \sum_{1\leq j\leq t} I\left(X_j^{(\mathcal{S})}, \mathcal{Y}_j^{(\mathcal{S}^c)}|X_j^{(\mathcal{S}^c)}\right) + K ,$$

where $K := -\mathrm{H}(Z|W_{\mathcal{S}^c}) + \frac{d}{2}\log(J_d m)$. Then, observe that, since $\mathcal{Y}_j^{(\mathcal{S}^c)}$ is conditionally independent from $X_j^{(\mathcal{S}^c)}$ given $X_j^{(\mathcal{S})}$,

$$I\left(X_j^{(\mathcal{S})}, \mathcal{Y}_j^{(\mathcal{S}^c)}|X_j^{(\mathcal{S}^c)}\right) = I\left(X_j^{(\mathcal{S})}, \mathcal{Y}_j^{(\mathcal{S}^c)}\right) \leq C_{\mathcal{S}} ,$$

the last inequality above following from the definition of cut-set capacity as maximal mutual information between the input and output of the channels crossing the cut. ■

IV. Upper bounds on the error exponent

A. A change of measure argument

We shall now develop some arguments based on a change of measure. Recall that, a memoryless channel with input $\mathcal{X}$, output $\mathcal{Y}$, and transition probabilities P, and a sequence of encoders $\Phi = \{\phi_t^{(v)}\}$ induce a probability measure $\mathbb{P}$ on $\Omega = \mathcal{W} \times \mathcal{Y}^{\mathbb{N}}$. Now consider a stochastic kernel $Q(\,\cdot\,|\,\cdot\,)$, having the same input $\mathcal{X}$ and output $\mathcal{Y}$. The stochastic kernel Q, together with the encoder sequence of encoders Φ, induces another probability measure on Ω, to be denoted by $\mathbb{Q}$.

The core idea consists in finding a relationship between the probability of an event A measured by $\mathbb{P}$ and that measured by $\mathbb{Q}$, by proving a large deviations bound on the channel behavior. In doing that, the stochastic kernel Q should be interpreted as a conditional empirical distribution of the channel output sequence (Y_t) given (X_t).

Let us assume that, for all input symbols $x \in \mathcal{X}$, $Q(\,\cdot\,|x)$ is absolutely continuous with respect to $P(\,\cdot\,|x)$, so that

$$\lambda_Q := \max\left\{\left|\log\frac{Q(y|x)}{P(y|x)}\right| \,\middle|\, x, y : P(y|x) > 0\right\} < +\infty , \tag{20}$$

and, a fortiori,

$$D(Q||P) := \max_{x\in\mathcal{X}} \sum_y Q(y|x) \log\frac{Q(y|x)}{P(y|x)} < +\infty .$$

Lemma 4. *For $t \in \mathbb{N}$, let $A \in \mathcal{A}$ be an event measurable with respect to (W, Y_1^t). Then, for all $\alpha > 1$ and $\varepsilon > \sqrt{(\alpha-1)8\lambda_Q^3}$, it holds*

$$\mathbb{P}(A) \geq 2^{1/1-\alpha}\mathbb{Q}(A)^{\overline{\alpha}} \exp\left(-t[D(Q||P) + \varepsilon]\right) ,$$

where $\overline{\alpha}$ is such that $\frac{1}{\alpha} + \frac{1}{\overline{\alpha}} = 1$.

Proof: Let us consider the r.v.

$$\Upsilon_t := \frac{\mathbb{Q}_{Y_{[t]}|W}(Y_{[t]}|W)}{\mathbb{P}_{Y_{[t]}|W}(Y_{[t]}|W)} .$$

From Hölder's inequality, it follows that

$$\begin{aligned}
\mathbb{Q}(A) &= \mathbb{E}_{\mathbb{Q}}\left[\mathbb{1}_A\right] \\
&= \mathbb{E}\left[\mathbb{1}_A \Upsilon_t\right] \\
&\leq \mathbb{E}\left[\Upsilon_t^{\alpha}\right]^{1/\alpha} \mathbb{E}\left[\mathbb{1}_A^{\overline{\alpha}}\right]^{1/\overline{\alpha}} \\
&= \mathbb{E}\left[\Upsilon_t^{\alpha}\right]^{1/\alpha} \mathbb{P}(A)^{1/\overline{\alpha}} .
\end{aligned} \tag{21}$$

We now look for an upper bound on $\mathbb{E}\left[\Upsilon_t^\alpha\right]$. To this end, observe that

$$\mathbb{E}\left[\Upsilon_t^\alpha\right] = \mathbb{E}_{\mathbb{Q}}\left[\Upsilon_t^{\alpha-1}\right] ,$$

where $\mathbb{E}_{\mathbb{Q}}$ denotes the expectation operator on the probability space $(\Omega, \mathcal{A}, \mathbb{Q})$. For all $1 \leq j \leq t$, we consider the σ-field $\mathcal{A}_j := \sigma(W, Y_{[j]})$, and the r.v.s

$$\Xi_j := D\left(Q(\,\cdot\,|X_j)||P(\,\cdot\,|X_j)\right) ,$$

$$M_j := \log\frac{\mathbb{Q}(Y_{[j]}|W)}{\mathbb{P}(Y_{[j]}|W)} - \sum_{i=1}^{j}\Xi_i \,.$$

Let us also define $\mathcal{A}_0 := \sigma(W)$, and $M_0 \equiv 0$. Then, $(M_j)_{j\geq 0}$ is a martingale on the filtrated probability space $(\Omega, (\mathcal{A}_j), \mathbb{Q})$. Indeed, it is easily verified that, for all $s \geq 0$, M_j is $\mathcal{A}_j$-measurable, and that

$$\mathbb{E}_{\mathbb{Q}}[M_{j+1}|\mathcal{A}_j]-M_j = \mathbb{E}_{\mathbb{Q}}\left[\log\tfrac{\mathbb{Q}(Y_{j+1}|W,Y_{[j]})}{\mathbb{P}(Y_{j+1}|W,Y_{[j]})}\middle|\mathcal{A}_j\right]-\Xi_{j+1} = 0\,.$$

Moreover, since the channel transition probabilities $Q(\,\cdot\,|x)$ are absolutely continuous with respect to $P(\,\cdot\,|x)$ for all $x \in \mathcal{X}$, we have that

$$|M_j - M_{j-1}| \leq \left|\log\tfrac{Q(Y_j|X_j)}{P(Y_j|X_j)}\right| + \Xi_j \leq 2\lambda_Q\,,$$

so that (M_j) has uniformly bounded increments. Hence, we can apply the Hoeffding-Azuma inequality [8] obtaining that

$$\mathbb{Q}\left(M_t \geq \varepsilon t|\,W\right) = \mathbb{Q}\left(M_t \geq M_0 + \varepsilon t|\,W\right) \leq \exp\left(-t\frac{\varepsilon^2}{8\lambda_Q^2}\right) .$$

Now, observe that, since

$$\sum_{1\leq j\leq t}\Xi_j \leq tD(Q||P)\,,$$

the $\mathbb{Q}$-probability of the event

$$E := \left\{\Upsilon_t \geq \exp\left(t[D(Q||P)+\varepsilon]\right)\right\}$$

can be estimated as follows:

$$\mathbb{Q}\left(E\right) \leq \mathbb{Q}\left(M_t \geq \varepsilon t\right) \leq \exp\left(-t\frac{\varepsilon^2}{8\lambda_Q^2}\right) .$$

Hence, we obtain, for $\beta := \alpha - 1$,

$$\begin{aligned}
\mathbb{E}\left[\Upsilon_t^\alpha\right] &= \mathbb{E}_{\mathbb{Q}}\left[\Upsilon_t^\beta\right]\\
&= \mathbb{E}_{\mathbb{Q}}\left[\Upsilon_t^\beta\mathbb{1}_E\right] + \mathbb{E}_{\mathbb{Q}}\left[\Upsilon_t^\beta\mathbb{1}_{E^c}\right]\\
&\leq \exp\left(\beta t\lambda_Q\right)\mathbb{Q}(E) + \exp\left(\beta t[D(Q||P)+\varepsilon]\right))\mathbb{Q}(E^c)\\
&\leq \exp\left(-t\left[\tfrac{\varepsilon^2}{8\lambda_Q^2} - \beta\lambda_Q\right]\right) + \exp\left(\beta t[D(Q||P)+\varepsilon]\right)\\
&\leq \exp\left(\beta t[D(Q||P)+\varepsilon]\right)\left[1+\exp\left(-t\left[\tfrac{\varepsilon^2-8\beta\lambda_Q^3}{8\lambda_Q^2}\right]\right)\right]\\
&\leq \exp\left(\beta t[D(Q||P)+\varepsilon]\right)2\,,
\end{aligned} \tag{22}$$

the last inequality following since $\frac{\varepsilon^2}{8\lambda_Q^2} > \beta\lambda_Q$. Then, the claim follows by substituting (22) into (21). ■

B. Proof of Theorem 1

We are now ready to prove the main result. Given a node $v \in \mathcal{V}$, and a non-trivial cut-set $\mathcal{E}_\mathcal{S}$ such that $v \in \mathcal{S}^c$, applying Proposition 1 to some stochastic kernel $Q \in \mathcal{Q}_\mathcal{S}$ leads to an upper bound on the left $\mathbb{Q}$-tail of the estimation error $\Delta_t^{(v)}$. Then, Lemma 4 allows one to recover an upper bound on the left $\mathbb{P}$-tail of $\Delta_t^{(v)}$, which is stated below.

Proposition 2. *For a node $v \in \mathcal{V}$, consider a non-trivial cut-set $\mathcal{E}_\mathcal{S}$ such that $v \in \mathcal{S}^c$, and a stochastic kernel $Q \in \mathcal{Q}_\mathcal{S}$. Then, for every $0 < r < 1$, $\alpha > 1$ and $\varepsilon > \sqrt{(\alpha-1)8\lambda_Q^3}$, it holds*

$$\mathbb{P}\left(\Delta_t^{(v)} > r\right) \geq \theta_t^{(\alpha)}\exp\left(-t[D(Q||P)+\varepsilon]\right) ,$$

where

$$\theta_t^{(\alpha)} := 2^{\frac{1}{1-\alpha}}\left(1 - \frac{tC_Q - \mathrm{H}(Z|W_{\mathcal{S}^c}) + \frac{d}{2}\log\left(J_d m\right)}{-\log r^d}\right)^{\overline{\alpha}}$$

We can now prove Theorem 1. For a given $R > 0$, fix $\delta > 0$ and choose a stochastic kernel $Q \in \mathcal{Q}_\mathcal{S}$ such that $C_Q \leq d(R-\delta)$. Clearly, for any $\alpha > 1$,

$$\liminf_t\left\{\theta_t^{(\alpha)}\right\} \geq \frac{\delta 2^{\frac{1}{1-\alpha}}}{C_Q+\delta} > 0\,.$$

Then, Proposition 2 implies that, for all $\varepsilon > \sqrt{(\alpha-1)\lambda_Q^3 8}$,

$$\begin{aligned}
D(Q||P)+\varepsilon &= D(Q||P)+\varepsilon+\limsup_t\left\{-\tfrac{1}{t}\log\theta_t^{(\alpha)}\right\}\\
&\geq \limsup_t\left\{-\tfrac{1}{t}\log\mathbb{P}\left(\Gamma_t^{(v)} < R\right)\right\}.
\end{aligned} \tag{23}$$

From the arbitrariness of the choice of the constants α and ε, and of the stochastic kernel $Q \in \mathcal{Q}_\mathcal{S}$, it follows that

$$\limsup_t\left\{-\frac{1}{t}\log\mathbb{P}\left(\Gamma_t^{(v)} < R)\right)\right\} \leq E_\mathcal{S}(R-\delta)\,.$$

Finally, (5) follows from the arbitrariness of $\delta > 0$ and the continuity of the exponent $E_\mathcal{S}(R)$ as a function of R.

C. Proof of Corollary 1

Let $\mathcal{S}$ and $Q \in \mathcal{Q}_\mathcal{S}$ be, respectively, the minimizing cut-set and stochastic kernel in (7). For $\delta > 0$, we have that

$$\begin{aligned}
(\Lambda_t^{(\eta)})^\eta &= \tfrac{1}{|\mathcal{V}|}\sum_{u\in\mathcal{V}}\mathbb{E}\left[(\Delta_t^{(u)})^\eta\right]\\
&\geq \tfrac{1}{|\mathcal{V}|}\mathbb{E}\left[(\Delta_t^{(u)})^\eta\right]\\
&\geq \mathbb{P}\left(\Gamma_t^{(u)} < \tfrac{1}{d}(C_Q+\delta)\right)\exp(-t\tfrac{\eta}{d}(C_Q+\delta))\,.
\end{aligned} \tag{24}$$

As in (23) one gets that

$$\limsup_t\left\{-\tfrac{1}{t}\log\mathbb{P}\left(\Gamma_t^{(u)} < \tfrac{1}{d}(C_Q+\delta)\right)\right\} \leq D(Q||P)+\varepsilon\,. \tag{25}$$

Then, (24) and (25) imply that

$$\limsup_t\left\{-\tfrac{1}{t}\log\Lambda_t^{(\eta)}\right\} \leq \tfrac{1}{d}(C_Q+\delta) + \tfrac{1}{\eta}\left(D(Q||P)+\varepsilon\right) .$$

Finally, (8) follows from the arbitrariness of the choices of $\delta > 0$, $\alpha > 0$, and $\varepsilon > \sqrt{(\alpha-1)\lambda_Q^3 8}$.

V. Conclusion

Upper bounds on the error exponent have been presented for problems of distributed computation of a real-vector-valued function on a network with noisy communication channels. It has been shown that, on non-deterministic channels, the exponential decay rate of any moment of the estimation error is strictly smaller than the capacity of the worst cut-set capacity.

Current research includes understanding how these bounds affect scaling limits of large networks, and proving tighter bounds for cases when the system dynamics cannot be fully designed but rather it is partially given.

Acknowledgements

The authors acknowledge support from the NSF under the EFRI-ARESCI project "Foundations for reconfigurable and autonomous cyber-physical systems: cyber-cities and cyber-universities".

References

[1] A. Giridhar and P. Kumar, "Towards a theory of in-network computation in wireless sensor networks," *IEEE Commun. Mag.*, vol. 44, no. 4, pp. 98–107, April 2006.

[2] O. Ayaso, D. Shah, and M. Dahleh, "Information theoretic bounds on distributed computation," *submitted to IEEE Trans. Inf. Theory*, 2008. [Online]. Available: http://mit.edu/dahleh/www/pubs/Ayaso2008.pdf

[3] A. Sahai and S. Mitter, "The necessety and sufficiency of anytime capacity for stabilization of a linear system over a noisy communication link –part i:scalar systems," *IEEE Trans. Inf. Theory*, vol. 52, no. 8, pp. 3369–3395, August 2006.

[4] G. Como, F. Fagnani, and S. Zampieri, "Anytime reiable tansmission of real-valued information through digital noisy channels," *submitted to SIAM J. Control Optim.*, 2009. [Online]. Available: http://mit.edu/giacomo/www/material/broadcast31.pdf

[5] R. Dobrushin, "An asymptotic bound for the probability error of information transmission through a channel without memory using the feedback," *Probl. Kibern.*, vol. 8, pp. 161–168, 1962.

[6] E. Haroutunian, "Lower bound for error probability in channels with feedback," *Probl. Pered. Inform.*, vol. 13, pp. 36–44, 1977.

[7] T. Cover and J. Thomas, *Elements of Information Theory, 2nd Edition*. John Wiley and Sons, New York, 2006.

[8] C. McDiarmid, *Concentration*, ser. Probabilistic Methods in Algorithmic Discrete Mathematics. Springer-Verlag, 1998.