Joint Source Coding, Routing and Power Allocation in Wireless Sensor Networks

Jun Yuan, *Student Member, IEEE*, and Wei Yu, *Member, IEEE*

***Abstract*—This paper proposes a cross-layer optimization framework for the wireless sensor networks. In a wireless sensor network, each sensor makes a local observation of the underlying physical phenomenon and sends a quantized version of the observation to a central location via wireless links. As the sensor observations are often partial and correlated, the network performance is a complicated and nonseparable function of individual data rates at each sensor. In addition, due to the shared nature of wireless medium, nearby transmissions often interfere with each other. Thus, the traditional "bit-pipe" model for network link capacity no longer holds. This paper deals with the joint optimization of source quantization, routing, and power control in a wireless sensor network. We follow a separate source and channel coding approach and show that the overall network optimization problem can be naturally decomposed into a source coding subproblem at the application layer and a wireless power control subproblem at the physical layer. The interfaces between the layers are precisely the dual optimization variables. In addition, we introduce a novel source coding model at the application layer, which allows the efficient design of practical source quantization schemes at each sensor. Finally, we propose a dual algorithm for the overall network optimization problem. The dual algorithm, when combined with a column-generation method, allows an efficient solution for the overall network optimization problem.**

***Index Terms*—Channel capacity region, convex optimization, dual decomposition, power control, rate-distortion region, routing, sensor networks, source coding.**

I. Introduction

SENSOR networks have emerged as a promising application for future wireless networks. Wireless sensor networks are envisioned to consist of a large number of low-cost low-power sensors deployed over a large area. Each sensor is capable of not only making local observations of the underlying physical phenomenon, but also transmitting the observed information to and relaying the information for its neighbors using wireless links. Information from all the sensors is eventually collected at a central processing unit (sometime called a central estimation officer, or a CEO), which produces a global picture of the physical phenomenon. Potential applications for sensor networks include military sensing, security monitoring, traffic control and environment monitoring.

The design objective of the sensor network is to reconstruct the underlying physical phenomenon as accurately as possible at the central processing unit. Thus, it is natural to define the network utility optimization problem for the sensor network as that of minimizing the overall estimation distortion. However, because each sensor makes only a partial observation of the underlying physical phenomenon, the overall estimation at the CEO depends on the individual data rates from all sensors in a complicated fashion. In particular, using distributed source coding techniques, it is possible to tradeoff transmission rate in one sensor with transmission rate in another sensor by rate adaptation. In this case, the network utility function is coupled across all individual rates and is therefore nonseparable.

Further, due to the shared nature of wireless medium, geographically close transmissions often interfere with each other. Because of the interference, the traditional 'bit-pipe' assumption for link capacity no longer holds. Instead, it is possible to tradeoff the capacity of one link with the capacity of another by adaptive power control.

Finally, as a large number of sensors are deployed in a field, they form an ad-hoc network. The routing problem in a sensor network becomes that of selecting the best multihop paths to the CEO. This routing problem is coupled with the source coding and link capacity problems, further complicating the overall network design.

In this paper, we address the overall joint source coding, routing and power allocation problem for the sensor network. We focus on a complete *digital* approach, where each sensor compresses its local observation into bits and transmits/relays the data digitally. Such a design choice follows a source-channel separation approach. Although source-channel separation may not necessarily lead to an information theoretical optimal design, there are clear practical benefits for separation. As the main result of this paper shows, the separation assumption naturally leads to a layered structure in network optimization, which makes the overall problem more amendable to analysis and algorithmic solutions.

The highlights of this paper are as follows:

- *Framework*: We provide an optimization framework that incorporates design requirements for different layers in a sensor network. We use a dual decomposition approach to decompose the overall network optimization problem into two disjoint subproblems: a source coding subproblem at the application layer and a power control subproblem at

Paper approved by M. Skoglund, the Editor for Source/Channel Coding of the IEEE Communications Society. Manuscript received May 5, 2006; revised December 27, 2006 and May 4, 2007. This paper was presented in part at the IEEE International Conference on Communications (ICC), Seoul, Korea, May 2005, and in part at the 23rd Queen's Biennial Symposium on Communications, Kingston, ON, Canada, May 2006. This work was supported by a discovery grant from the Natural Science and Engineering Research Council (NSERC) of Canada, and by the Canada Research Chairs program. Correspondence should be addressed to W. Yu.

J. Yuan was with The Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, 10 King's College Road, Toronto, Ontario M5S 3G4, Canada. He is now with the Wireless Technology Lab, Nortel, 3500 Carling Ave., Ottawa, Ontario K2H 8E9, Canada (e-mail: junyu@nortel.com).

W. Yu is with The Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, 10 King's College Road, Toronto, Ontario M5S 3G4, Canada (e-mail: weiyu@comm.utoronto.ca).

Digital Object Identifier 10.1109/TCOMM.2008.060237.

the physical layer. A set of Lagrangian dual variables play the role of coordinating the cross-layer interface.

- *Model*: We define a novel utility function for the sensor network that characterizes the tradeoff between estimation accuracy and power consumption in a sensor network application. In addition, we present a novel simplified source coding model at the application layer to capture the rate-distortion optimization in the source coding subproblem. Similarly, we address the power control and interference management problem by setting up a coupled link capacity model at the physical layer.
- *Algorithm*: We present a dual algorithm to solve the joint optimization problem at the system level. The algorithm consists of a subgradient update of the Lagrangian dual variables that coordinate the application layer and the physical layer interface, an efficient and iterative solution of the two subproblems in the two layers, and a column-generation method that ensures convexity and convergence.

The general framework proposed in this paper represents a cross-layer optimization strategy. It balances the bit-rate *demand* by the source coding module at the application layer and the bit-rate *supply* by routing and channel coding modules at the network and physical layers. Modeling and solution algorithms for each subproblem can be easily tuned when new networking techniques or new optimization tools become available.

This paper focuses on sensor networks or ad-hoc networks for environment monitoring or video surveillance applications in which continuous data streams are generated at each sensor. The models considered in this paper are distinct from detection-based sensor networks in which sensors are put into a sleep mode most of the time and wake up only upon intrusion detection. Further, this paper considers a static network in which sensor locations are fixed once deployed, and where centralized network optimization is feasible.

A. Related Work

The cross-layer optimization of networks using dual decomposition has been studied by many authors in recent literature. In their seminal paper, Kelly, Maulloo, and Tan [1] showed that the network optimization problem may be cast in a primal or a dual form. The authors further used the dual decomposition approach to decompose the utility maximization problem into a user subproblem and a network subproblem. In a related work, Low and Lapsley [2] investigated the flow control problem by solving its dual using a gradient projection approach. More recently, Xiao, Johansson and Boyd [3] used the dual decomposition approach to perform a joint optimization of routing and physical-layer resource allocation in a wireless network. Dual decomposition is also used by Chiang [4] in a joint optimization of the Internet Transmission Control Protocol (TCP) window size and power levels for wireless networks, where a set of dual variables are used as a means of cross-layer optimization. Further, Wang, Li, Low and Doyle [5] investigated a joint TCP and Internet Protocol (IP) routing problem, where congestion parameters are interpreted as primal and dual optimization variables. Recently, Lin and Shroff [6] employed a dual decomposition technique to study the impact of imperfect scheduling on cross-layer rate control. The present paper is inspired by these work in the literature. Our main contribution is to illustrate that the dual decomposition approach is applicable not only to the joint optimization of the physical layer and the network layer, but also to the joint optimization of the application layer. We choose the sensor network as a natural application as the network utility function of a sensor network typically depends on the application-layer data rates in a tightly coupled manner.

A key requirement for the use of the dual decomposition approach is the underlying convexity structure of the overall optimization problem. Convexity does not necessarily hold in wireless networks where nodes interfere with each other. In this paper, we show that while convexity is indeed crucial, convexity can always be assured if time-sharing is allowed. We use a technique called the column-generation method to ensure convexity. Column generation was first used for the channel capacity problem by Johansson and Xiao [7]. In this paper, we show that the column-generation method is applicable to both rate-distortion and channel capacity problems.

Our work is related to the work of Kim, Rajeswaran and Negi [8], where the joint power adaptation, scheduling and routing problem for wireless ad hoc networks in a ultra-wideband (UWB) context is investigated. This paper is also related to the work of Xiao, Cui, Luo, and Goldsmith [9], which studied an optimal power scheduling of decentralized estimation in a sensor network with energy constraints, and the work of Xiao and Luo [10], which addressed a multiterminal source-channel communication problem under orthogonal multiple access. The architectural issue and cross-layer design for wireless sensor networks were also explored in the work of Scaglione and Servetto [11], where the interaction of distributed source coding problem and routing was considered. The network optimization problem treated in this paper can be considered as an extension or generalization of the above work in the sense that source coding, routing and power control are jointly taken into account in an overall design of a sensor network.

B. Organization

The remainder of this paper is organized as follows. In Section II, we provide a joint optimization framework and propose a dual algorithm for the sensor network. In Section III, we treat the source coding subproblem and the power control subproblem in detail. The modeling aspect and the structure of the subproblems are discussed. In Section IV, we present an iterative column-generation-based approach to efficiently solve the two subproblems in each layer and show how the solutions may be incorporated in the overall dual algorithm. The interaction between the two subproblems and the convergence property of the dual algorithm are illustrated by simulation examples in Section V. Finally, we conclude the paper in Section VI.

II. Joint Optimization Framework

Consider a wireless sensor network with many sensors deployed over a field. Each sensor makes a local observation of some underlying physical phenomenon, quantizes

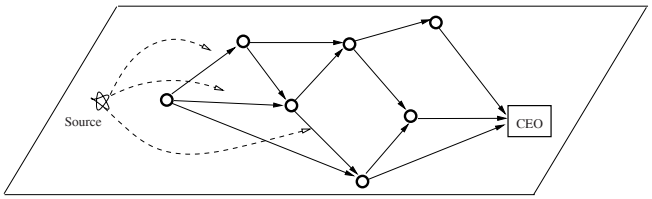


Fig. 1. Sensor network.

its observation, and transfers the quantized data to a CEO through a wireless network. The wireless network is typically infrastructureless, thus each sensor may also act as a relay for its neighbors. Fig. 1 illustrates a typical sensor network.

For convenience and without loss of generality, this paper considers a discrete-time model, in which the physical phenomenon is modeled as an i.i.d. discrete-time random vector (with spatial correlation.)

This paper considers a separate source and channel coding approach in which the distributed source coding and the information transmission problems are considered in two different layers. Such a separation approach, although not necessarily optimal, is however architecturally appealing. Recent work such as [12] considered uncoded transmission with joint source and channel coding strategies. Although joint source and channel coding may lead to a better overall performance in an information theoretical sense, the layered approach advocated in this paper is much easier to design in practice. As an alternative to the sensing-and-routing approach considered in this paper, a tree-based architecture [13], where each sensor may combine local observation with quantized observation from neighboring sensors, is also possible. Such a tree-based architecture is not considered in the present framework.

A. General Framework

The design objective of a sensor network is to reconstruct the underlying physical phenomenon as accurately as possible, while making the most efficient use of network resources. In this section, we propose an optimization framework to achieve this objective. The formulation of the joint source coding, routing and power control problem is based on the following assumptions. First, the overall estimation distortion is determined by the source coding rates of all sensors. Second, the source coding rate in each sensor is supported by the network via different routes from the sensors to the CEO. Third, at each transmission link, the aggregated flow rate cannot exceed the link capacity. Fourth, the achievable link capacity is determined by the signal-to-interference-and-noise ratio (SINR) of the link, which in turn is determined by the power levels at all link transmitters.

Let $\mathcal{G} = (V, E)$ be the network topology, where V and E are the sets of vertices and edges respectively. Let $\mathbf{d}$ be the distortion vector and $\mathbf{p} = [p_1, ..., p_L]^T$ be the link transmit power vector, where L is the number of total links. Let $\mathbf{s} = [s_1, ..., s_N]^T$ be a vector of source coding rates, where N is the number of total nodes. Let $\mathbf{c} = [c_1, ..., c_L]^T$ be a vector of link capacities.

In this paper, source rates are expressed in bits per source sample. Link capacities are expressed in bits per channel sample. Let T_s be the source sampling period, T_c be the channel sampling period. The source rates and link capacities in bits per second are then $\mathbf{s}/T_s$ and $\mathbf{c}/T_c$, respectively. For convenience and without loss of generality, we assume $T_s = T_c = 1$ in the rest of the paper.

The fundamental concept in source coding is the *rate-distortion region* $\mathcal{R}$, which is the closure of the set of achievable rate distortion pairs $(\mathbf{s}, \mathbf{d})$ such that source rates $\mathbf{s}$ are sufficient to allow reconstruction of the underlying phenomenon with distortion $\mathbf{d}$. Rate-distortion region $\mathcal{R}$ characterizes the tradeoff between source rates and distortion.

The fundamental concept in channel coding is the *power-capacity region* $\mathcal{C}$. Power-capacity region $\mathcal{C}$ characterizes the tradeoff between different link capacities and power constraints such that power $\mathbf{p}$ is sufficient to support link capacities $\mathbf{c}$. Such a tradeoff exists because links mutually interfere with each other.

Source rate is a node-based concept. Link capacity is a link-based concept. The network routing problem is that of supporting the distortion-achievable rate by link capacity. This problem can be formulated using a multicommodity flow model [14] [3], where quantized observations at different sensors are represented by different commodity types. The multicommodity flow model ensures that flow conservation is obeyed in each node in the network. Mathematically, the flow conservation constraint can be represented by a so-called node-link incidence matrix. Let A be an $N \times L$ matrix in a network with N nodes and L links. Define[1]

$$a_{il} = \begin{cases} 1 & \text{if } i \text{ is the start node for link } l \\ -1 & \text{if } i \text{ is the end node for link } l \\ 0 & \text{else} \end{cases}$$

Then, the condition $A\mathbf{c} \geq \mathbf{s}$ ensures that the source rates can be supported by link capacities via a suitable routing scheme. Note that the source data may be routed to the CEO through multiple paths.

The fundamental objective in a sensor network is to minimize the distortion $\mathbf{d}$ incurred in the estimation process. The fundamental resource constraint in the network is the transmit power $\mathbf{p}$. The goal of the network optimization problem is to characterize a tradeoff between the two. Such a trade-off can be parameterized by weighting vectors α and β in an overall network optimization problem as follows:

$$\begin{aligned} \text{minimize} \quad & \alpha^T \mathbf{d} + \beta^T \mathbf{p} \\ \text{subject to} \quad & \begin{pmatrix} \mathbf{s} \\ \mathbf{d} \end{pmatrix} \in \mathcal{R} \\ & \begin{pmatrix} \mathbf{c} \\ \mathbf{p} \end{pmatrix} \in \mathcal{C} \\ & A\mathbf{c} \geq \mathbf{s} \end{aligned} \tag{1}$$

where α and β represent the relative emphasis on the distortion $\mathbf{d}$ and power consumption $\mathbf{p}$. The constraint $\begin{pmatrix} \mathbf{s} \\ \mathbf{d} \end{pmatrix} \in \mathcal{R}$ models interdependence of the distortion $\mathbf{d}$ on source coding

[1]For simplicity, we consider a network with only one CEO and set its corresponding node-link row as the last row in A. In the formulation of the optimization problem, this last row may be deleted without loss of generality because it is linearly dependent of previous rows.

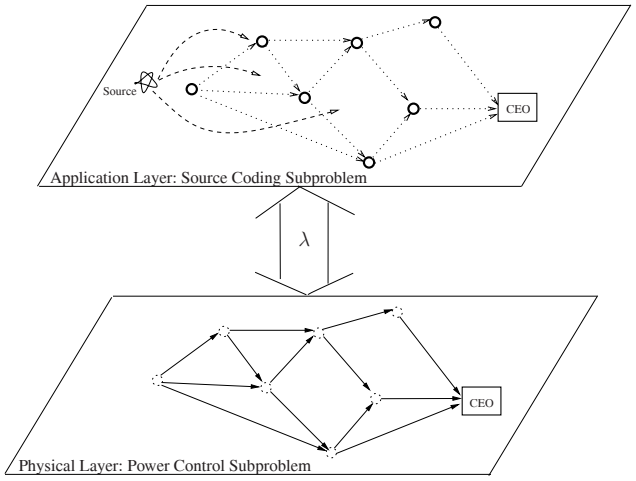


Fig. 2. Layering in sensor networks.

rates $\mathbf{s}$. The constraint $\begin{pmatrix} \mathbf{c} \\ \mathbf{p} \end{pmatrix} \in \mathcal{C}$ models the interdependence of link capacity vector on the power constraint. The last inequality $A\mathbf{c} \geq \mathbf{s}$ reflects the fact that the source rate at each node is upper bounded by the link capacity. In the remaining of this section, we focus on an optimization framework for solving (1), while assuming for now that the rate-distortion and power-capacity regions are given. Detailed discussions on the structures of these regions are deferred to Section III.

B. Dual Decomposing

The solution to the joint optimization problem (1) can be greatly simplified using a dual decomposition technique. The idea is to decompose the original problem into two subproblems by relaxing the inequality constraint $A\mathbf{c} \geq \mathbf{s}$ using a set of dual variables (or Lagrange multipliers). Such a decomposition makes efficient and distributed algorithm design possible. We start by writing down the Lagrangian:

$$L(\lambda, \mathbf{s}, \mathbf{d}, \mathbf{c}, \mathbf{p}) \quad = \quad \alpha^T \mathbf{d} + \beta^T \mathbf{p} + \lambda^T (\mathbf{s} - A\mathbf{c}), \quad (2)$$

where $\lambda = [\lambda_1, \ldots, \lambda_N]^T$ is the Lagrange multiplier. The minimization of the Lagrangian (2) consists of two sets of variables: $(\mathbf{s}, \mathbf{d})$ and $(\mathbf{c}, \mathbf{p})$. In this paper, the rate and distortion variables $(\mathbf{s}, \mathbf{d})$ are referred to as *application-layer* variables; the capacity and power variables $(\mathbf{c}, \mathbf{p})$ are referred to as *physical-layer* variables. Moreover, the Lagrangian optimization problem is now decoupled into two disjoint parts. The application-layer part is a pure source coding subproblem

$$\begin{aligned} \text{minimize} \quad & \alpha^T \mathbf{d} + \lambda^T \mathbf{s} \\ \text{subject to} \quad & \begin{pmatrix} \mathbf{s} \\ \mathbf{d} \end{pmatrix} \in \mathcal{R} \end{aligned} \quad (3)$$

and the physical-layer part is a pure power control subproblem

$$\begin{aligned} \text{maximize} \quad & \mu^T \mathbf{c} - \beta^T \mathbf{p} \\ \text{subject to} \quad & \begin{pmatrix} \mathbf{c} \\ \mathbf{p} \end{pmatrix} \in \mathcal{C} \end{aligned} \quad (4)$$

where $\mu = [\mu_1, \ldots, \mu_L]^T$ is related to the dual variable λ by the link price consistency equation

$$\mu^T = \lambda^T A. \quad (5)$$

Thus, the optimization framework naturally provides a layered structure to the joint source coding, routing, and power control problem as shown in Fig. 2. The solution to the overall optimization problem (1) may be obtained by solving the two subproblems individually and by suitably updating the dual variables.

The Lagrange multipliers λ and μ have the interpretation of being the shadow prices coordinating the application-layer demand and physical-layer supply. A larger value of λ_i for node i signals to the application layer that transporting data for that node is costly. This has the effect of decreasing the source rate at that node. A larger value of μ_l for link l signals to the physical layer to give that particular link a higher priority. This has the effect of increasing the transport capacity on that link. The optimization process can be thought of as a process of matching the supply and demand by iteratively finding an intersection of the distortion-rate and power-capacity regions. Note that routing is implicitly taken into account in the multicommodity flow model. The Lagrangian dual variables λ and μ are related via the node-link incidence matrix A.

C. Dual Algorithm

The key requirement that allows the efficient and optimal solution of the joint optimization problem is the convexity of rate-distortion region and power-capacity region. Assuming the *strict* convexity of the regions, we propose the following algorithm that solves the entire network optimization problem (1).

Algorithm 1: Basic Algorithm

1) Set $t = 0$. Initialize $\lambda^{(0)}$ and set $\mu^{(0)} = A^T \lambda^{(0)}$.
2) In the primal domain, solve the following subproblems:

$$\left\{ \min_{\mathbf{s},\mathbf{d}} \ \alpha^T \mathbf{d} + (\lambda^{(t)})^T \mathbf{s} \ \middle| \ \begin{pmatrix} \mathbf{s} \\ \mathbf{d} \end{pmatrix} \in \mathcal{R} \right\} \quad (6)$$

$$\left\{ \max_{\mathbf{c},\mathbf{p}} \ (\mu^{(t)})^T \mathbf{c} - \beta^T \mathbf{p} \ \middle| \ \begin{pmatrix} \mathbf{c} \\ \mathbf{p} \end{pmatrix} \in \mathcal{C} \right\} \quad (7)$$

3) In the dual domain, update dual variables as follows:

$$\lambda^{(t+1)} \quad = \quad \left[\lambda^{(t)} + \epsilon^{(t)} (\mathbf{s} - A\mathbf{c}) \right]^+ \quad (8)$$

$$\mu^{(t+1)} \quad = \quad A^T \lambda^{(t+1)} \quad (9)$$

where $[\cdot]^+$ denotes $\max(0, \cdot)$.
4) Set $t = t + 1$. Goto Step 2 until convergence.

Theorem 1: Algorithm 1 always converges to the global optimum of the overall network optimization problem (1), provided that the rate-distortion region and power-capacity region are strictly convex and that the step sizes $\epsilon^{(t)}$ is appropriately chosen.

Proof: The crucial ingredient that makes the algorithm work is convexity. Given the convexity of $\mathcal{R}$ and $\mathcal{C}$, the network optimization problem (1) is convex. Define the dual

objective function

$$g(\lambda) = \min_{\mathbf{s},\mathbf{d},\mathbf{c},\mathbf{p}} L(\mathbf{s},\mathbf{d},\mathbf{c},\mathbf{p},\lambda). \tag{10}$$

As mentioned earlier, the above minimization decomposes into two subproblems: (6) and (7). The solutions of the two subproblems allow $g(\lambda)$ to be evaluated.

By strong duality [15], the overall network optimization problem (1) is solved by the following dual maximization problem:

$$\max_{\lambda} g(\lambda) \tag{11}$$

It remains to show that the update step (8) solves the dual maximization problem. This is due to the fact that the update step is a subgradient update for λ. Let $(\mathbf{s}^*, \mathbf{d}^*)$ and $(\mathbf{c}^*, \mathbf{p}^*)$ be the optimal solutions of (6) and (7) given λ. Then,

$$g(\lambda) = \alpha^T \mathbf{d}^* + \beta^T \mathbf{p}^* + \lambda^T \mathbf{s}^* - \lambda^T A \mathbf{c}^*,$$

and $\forall \lambda'$

$$g(\lambda') \leq \alpha^T \mathbf{d}^* + \beta^T \mathbf{p}^* + \lambda'^T \mathbf{s}^* - \lambda'^T A \mathbf{c}^*.$$

Thus,

$$g(\lambda) - g(\lambda') \geq (\mathbf{s}^* - A\mathbf{c}^*)^T (\lambda - \lambda'), \quad \forall \lambda'.$$

By the definition of subgradient [16] [17], $(\mathbf{s}^* - A\mathbf{c}^*)$ is a subgradient of $g(\lambda)$. Thus, as long as step sizes $\epsilon^{(t)}$ are chosen appropriately, the subgradient update eventually converges to the optimal dual variable. For example, one appropriate choice for the step sizes is a square summable but not absolutely summable sequence [16] [17], such as

$$\epsilon^{(t)} = \frac{a}{b+t}, \ a > 0, \ b \geq 0.$$

Once the optimal dual variables are found, the corresponding primal variables can be recovered by explicitly solving (6) and (7) given the optimal λ and μ. Since the regions are strictly convex, the recovered primal variables are unique. Further, they are primal feasible because the original problem (1) always has a feasible solution[2]. Due to strong duality, the primal variables must be global optimum. ■

To summarize, this section provides a general optimization framework for modeling and solving the joint source coding, routing and power allocation problem in wireless sensor networks. In this framework, the joint optimization problem is decomposed into the application-layer and the physical-layer subproblems. As mentioned earlier, the key attribute that is needed for this dual decomposition approach is the convexity of rate-distortion and power-capacity regions. In the following, we first provide concrete examples of rate-distortion and power-capacity regions, then illustrate a method to ensure convexity.

III. Source Coding and Power Control Subproblems

To make the optimization framework concrete, this section presents examples of specific models for each of the subproblems. We choose an information-theoretical characterization of the rate-distortion region and power-capacity region, which allows us to gain insights into the fundamental tradeoffs between performance (e.g., distortion) and resource (e.g., power) in a sensor network. The information-theoretical models also provide useful bounds for practical system design.

[2]For example, an obvious feasible solution is the vector of zero source rates, zero link capacities and zero power consumption with the distortion being the variance of the underlying source itself.

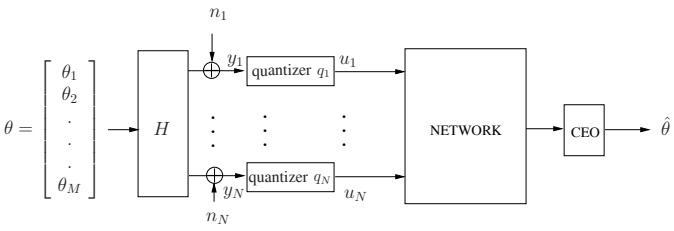


Fig. 3. Source coding.

A. Application Layer: Source Coding Subproblem

Consider an environment sensing application depicted in Fig. 3. The underlying physical phenomenon is modeled by a vector random variable $\theta = (\theta_1, \cdots, \theta_M)^T$, where M is the vector size. A total number of N sensors are deployed in the field, each making a local (and possibly partial) observation of θ, while being corrupted by sensor noise $\mathbf{n}$. The observation channel from the physical phenomenon to the sensors is characterized by a matrix H, which depends on the geographic deployment of sensors. Since sensors are static once deployed, it is reasonable to assume that H does not change over time.

From a rate-distortion theory point of view, a generic quantization scheme can be described as follows. At each sensor $i = 1 \cdots N$, the noisy observation y_i is quantized into a codeword u_i (possibly using a vector quantizer.) The effect of quantization process at each sensor can be modeled by a quantization noise random variable q_i [18][19]:

$$\mathbf{u} = \mathbf{y} + \mathbf{q} = H\theta + \mathbf{n} + \mathbf{q} \tag{12}$$

The quantized information from all sensors is transmitted back through the network to the CEO at source rates $(s_1, ..., s_N)$. At the CEO, the decoder first jointly decodes the codewords $\mathbf{u} = [u_1, ..., u_N]^T$, then estimates the source. The source estimation is denoted as $\hat{\theta}$. The distortion criterion is denoted as $\mathbf{d}(\hat{\theta}, \theta)$. The task of source coding optimization subproblem is to design good joint quantizers $(q_1, \cdots, q_N)$ to minimize some weighted objective function $\alpha^T \mathbf{d} + \lambda^T \mathbf{s}$.

From the information-theory literature, the largest achievable rate region subject to a distortion constraint is known as the Berger-Tung region [20], which can be summarized as follows

$$\sum_{i \in \phi} s_i \geq I(\mathbf{y}_\phi; \mathbf{u}_\phi | \mathbf{u}_{\phi^c}), \quad \forall \ \phi \subseteq \mathcal{P}(\{1, 2, ...N\}) \tag{13}$$

where $I(\cdot;\cdot)$ denotes mutual information, $\mathcal{P}(S)$ is used to denote the power set of S, $\mathbf{u}$ is an auxiliary random vector with a distribution $\Pi_{i=1}^N p(u_i|y_i)$, and $\mathbf{u}_\phi = \{u_i | i \in \phi\}$. The joint distribution of $\mathbf{u}$ is such that the best estimate of θ given $\mathbf{u}$ satisfies the given distortion constraint. The idea behind the Berger-Tung region can be understood in terms of the quantization process as follows. The observation y_i at

each sensor is quantized into a codeword u_i, which is to be transmitted to the CEO. The estimation of θ at the CEO is based on $(u_1, \cdots, u_N)$. Since the observations $(y_1, ..., y_N)$ are correlated, the codewords $(u_1, ..., u_N)$ are also correlated. Slepian-Wolf coding [21] can then be applied to $(u_1, \cdots, u_N)$ to achieve a rate bound as expressed in the conditional mutual information in (13).

In general, the computation of the Berger-Tung achievable rate region for distributed source coding is difficult because it involves an optimization of conditional mutual information over the conditional distribution $\Pi_{i=1}^{N} p(u_i|y_i)$. In this paper, we make several simplifying assumptions to arrive at a suboptimal but computationally feasible rate-distortion region. The main idea is to ignore the possibility of Slepian-Wolf coding when transmitting correlated u_i back to the CEO. More specifically, the following rates are sufficient for achieving the distortion constraint

$$s_i \geq I(y_i; u_i), \quad \forall\, i = 1, \cdots, N \tag{14}$$

Here, each u_i again represents a quantized version of y_i. However, the set of $(u_1, \cdots, u_N)$ is transmitted to the CEO without taking advantage of the fact that they are correlated, while the estimation of θ at the CEO is still based on all of $(u_1, \cdots, u_N)$. The rate region (14) is simpler to evaluate than (13) because no conditional mutual information is involved.

We now consider the special Gaussian case to illustrate this simplified Berger-Tung region. Let us assume that θ is an independent Gaussian vector with zero mean and a diagonal covariance matrix Σ_θ. Let the distortion measure be the mean-squared-error (MSE) measure, i.e. $\mathbf{d} = \mathrm{diag}(\mathbb{E}[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T])$. Let the observation noise $\mathbf{n}$ be independent Gaussian with zero mean and covariance matrix Σ_n. Let us further assume that an ideal Gaussian rate-distortion vector quantizer is used, so that $\mathbf{q}$ can be modeled as a zero-mean Gaussian random vector with a diagonal covariance matrix Σ_q.

Under the Gaussian assumption, the achievable rate for the simplified Berger-Tung region can be computed as

$$s_i \quad \geq \quad I(y_i; u_i) = \frac{1}{2} \log \left(\frac{\mathbf{h}_i^T \Sigma_\theta \mathbf{h}_i + \sigma_{ni}^2 + \sigma_{qi}^2}{\sigma_{qi}^2} \right) \tag{15}$$

where $\mathbf{h}_i^T$ is the ith row of channel matrix H, σ_{ni}^2 and σ_{qi}^2 are the ith diagonal element of Σ_n and Σ_q, which represent the noise variance and quantization variance respectively, and Σ_θ is the covariance matrix of the underlying source vector θ. In this paper, we assume that the source and the noise are stationary, and assume that Σ_θ and Σ_n are known *a priori*.

The corresponding distortion can likewise be computed using the minimum mean-squared-error (MMSE) estimation theory. At the CEO, the optimal estimation $\hat{\theta}$ and the corresponding estimation error covariance matrix K_0 are

$$\begin{aligned} \hat{\theta} &= \mathbb{E}[\theta|\mathbf{u}] = \Sigma_{\theta u} \Sigma_{uu}^{-1} \mathbf{u} \\ K_0 &= \mathbb{E}\left[\left(\theta - \hat{\theta}\right) \left(\theta - \hat{\theta}\right)^T \right] = \Sigma_\theta - \Sigma_{\theta u} \Sigma_{uu}^{-1} \Sigma_{u\theta} \\ &= \Sigma_\theta - \Sigma_\theta H^T (H \Sigma_\theta H^T + \Sigma_n + \Sigma_q)^{-1} H \Sigma_\theta^T \end{aligned}$$

The distortion vector $\mathbf{d}$ is then a vector of the diagonal elements of K_0, i.e. $\mathbf{d} = \mathrm{diag}(K_0)$. Note that the distortion implicitly depends on all quantization levels. This distortion expression characterizes the coupled relation among the sensors.

The optimization problem is that of minimizing $\alpha^T \mathbf{d} + \lambda^T \mathbf{s}$. For simplicity, the rest of the paper considers sum distortion only, i.e. $\alpha^T = \mathbf{1}^T$. In this case, $\alpha^T \mathbf{d} = \mathrm{tr}(K_0)$. Also, for reasons that will be clear later, it is convenient to define a new variable $w_i = 1/(\sigma_{ni}^2 + \sigma_{qi}^2)$. We call $\mathbf{w}$ *quantization effort*. Note that quantization effort $\mathbf{w}$ has a one-to-one relation with the quantization variance σ_q^2, i.e. $\sigma_{qi}^2 = \frac{1}{w_i} - \sigma_{ni}^2, \Sigma_w^{-1} = \Sigma_n + \Sigma_q$. Further, denote $\sigma_{si}^2 = \mathbf{h}_i^T \Sigma_\theta \mathbf{h}_i$.

The source coding subproblem (3) can now be reformulated as follows:

$$\begin{aligned} \text{minimize} \quad & \mathrm{tr}\,(\Sigma_\theta) - \mathrm{tr}\left(\Sigma_\theta H^T (H \Sigma_\theta H^T + \Sigma_w^{-1})^{-1} H \Sigma_\theta^T\right) \\ & + \sum_{i=1}^{N} \lambda_i \log \left(\frac{1 + \sigma_{si}^2 w_i}{1 - \sigma_{ni}^2 w_i} \right) \\ \text{subject to} \quad & 0 \leq w_i \leq \frac{1}{\sigma_{ni}^2}, \quad \sigma_{si}^2 = \mathbf{h}_i^T \Sigma_\theta \mathbf{h}_i \quad i = 1, ..., N \end{aligned} \tag{16}$$

where Σ_w is a diagonal matrix with w_i as the ith diagonal element. The first two terms in (16) are the MMSE distortion estimation. We absorb $\frac{1}{2}$ term of (15) into λ in the third term of (16). For notational convenience, the objective function of (16) is referred to as Q^{APP} later in the paper.

Note that $\mathbf{s}$ and $\mathbf{d}$ are determined given $\mathbf{w}$. Thus, the optimization is over the variable $\mathbf{w}$. The parameterization of the optimization problem over $\mathbf{w}$ is useful, because under certain conditions, the optimization problem is convex over $\mathbf{w}$, which greatly facilitates the task of finding the global optimal solution.

Definition 1: The source coding optimality condition holds, if either of the following is true:

- $\sigma_{ni}^2 \geq \sigma_{si}^2, \quad \forall i$ (17)
- $\sigma_{ni}^2 < \sigma_{si}^2 \quad \text{and} \quad \sigma_{qi}^2 \in \left[0, \frac{\sigma_{si}^2 + \sigma_{ni}^2}{\sigma_{si}^2 - \sigma_{ni}^2} \sigma_{ni}^2\right], \ \forall i$ (18)

where $\sigma_{si}^2 = \mathbf{h}_i^T \Sigma_\theta \mathbf{h}_i$.

Theorem 2: The source coding subproblem (16) is a convex optimization problem under the source coding optimality condition.

A detailed proof is not presented here due to space constraint. The proof consists of showing that the objective function of (16) is convex in $\mathbf{w}$ by checking that its Hessian is positive semidefinite under the source coding optimality condition. A complete proof can be found in [22].

The conditions under which convexity holds have the following interpretation. Convexity holds whenever the sensor noise variance is larger than the source variance. When the sensor noise variance is smaller than the source variance, convexity holds when the quantization noise (or equivalently quantizer resolution) is smaller than the sensor noise variance times a constant, which is larger than 1. In most practical situations, the quantizers are designed so that the quantization noise is below the sensor noise. Thus, convexity usually holds in a well-designed sensor network.

B. Physical Layer: Power Control Subproblem

We now turn to the physical-layer subproblem and present a capacity region formulation for a wireless network. The fundamental issues are power control and interference management. We use the notion of power-capacity region (more rigorously, the achievable rate region) to characterize a tradeoff between achievable rates at different links and the transmit power levels. We assume each link treats the interference from neighboring links as additive noise. In this case, the power control subproblem (4) can be stated as follows:

$$\begin{aligned} \text{maximize} \quad & \sum_l \mu_l c_l - \sum_l \beta_l p_l && (19)\\ \text{subject to} \quad & c_l = \frac{1}{2}\log\left(1+\frac{1}{\Gamma}\text{SINR}_l\right) && l=1,...,L\\ & \text{SINR}_l = \frac{G_{ll}p_l}{\sum_{j\neq l} G_{lj}p_j+\sigma_l^2} && l=1,...,L\\ & 0 \le p_l \le p_{l,\max} && l=1,...,L \end{aligned}$$

where c_l is the capacity of link l. An SINR gap term Γ is used in the rate expression to take into account practical effects of modulation scheme, bit error rate (BER), and error correct code. For example, for uncoded QAM constellations, at bit error rate $\bar{P}_e = 10^{-6}$, the gap is 8.8 dB. Here, SINR_l is the signal-to-interference-and-noise ratio of link l, p_l and σ_l^2 are power and noise respectively, G_{ll} denotes the effective channel gain between the transmitter and receiver of link l, and G_{lj} is the interference coefficient from link j to link l. Each link has a power constraint $p_{l,\max}$. Again, for notational convenience, the objective function of (19) is referred to as Q^{PHY} later in the paper.

As mentioned earlier, the optimization framework presented in this paper is most applicable to either static networks in which channel gains are fixed or slow-fading networks in which fading coefficients can be estimated and fedback to the CEO for network optimization. The interference term in the SINR expression makes (19) a nonconvex function of $\mathbf{p} = [p_1, ..., p_L]^T$. Consequently, the power control problem is not a convex optimization problem and is difficult to solve.

Fortunately, in the physical layer, it is often possible to operate in time-sharing points of different directly achievable rate points, which ensure the convexity of the overall power-capacity region. Time-sharing can be implemented either via time-division multiplexing if the network is synchronized, or via frequency-division multiplexing, if independent transmissions may be implemented in multiple frequency bands. In the next section, we take advantage of this observation to convexify the problem.

IV. Algorithms

When the rate-distortion or the power-capacity region optimization problem is nonconvex, the optimal solutions to the source-coding and power control subproblems become numerically difficult to find. Nevertheless, local search strategies can often be used to identify efficiently computable, yet suboptimal, solutions. However, the lack of convexity also makes the convergence of the dual algorithm a difficult issue. In this section, we propose the use of a column-generation method, which, when combined with a coordinate-descent algorithm for the subproblems, guarantees the convergence of the overall network optimization problem.

A. Coordinate-Descent Algorithms for Subproblems

A local optimum of nonconvex optimization problems can be found using many different methods. A particularly effective local search is the coordinate-descent method, in which a multivariable function is optimized over each variable iteratively. When applied to the source coding subproblem (16) and the power control subproblem (19), the algorithm works as follows:

Algorithm 2: Source Coding Algorithm

1) Initialize $(w_1^{(0)}, \cdots, w_N^{(0)})$.
2) At each round, sequentially update each sensor's quantization effort w_i as follows:
$w_i^{(\tau+1)}$
$= \text{argmin}_{w_i} Q^{\text{APP}}\left(w_1^{(\tau+1)}, \cdots, w_{i-1}^{(\tau+1)}, w_i, w_{i+1}^{(\tau)}, \cdots, w_N^{(\tau)}\right)$
$i = 1, \cdots, N$
3) Set $\tau = \tau + 1$. Repeat 2) until convergence.

Algorithm 3: Power Control Algorithm

1) Initialize $(p_1^{(0)}, \cdots, p_L^{(0)})$.
2) At each round, sequentially update each link's power allocation p_l as follows.
$p_l^{(\tau+1)}$
$= \text{argmax}_{p_l} Q^{\text{PHY}}\left(p_1^{(\tau+1)}, \cdots, p_{l-1}^{(\tau+1)}, p_l, p_{l+1}^{(\tau)}, \cdots, p_L^{(\tau)}\right)$
$l = 1, \cdots, L$
3) Set $\tau = \tau + 1$. Repeat 2) until convergence.

The updates of w_i and p_l in Step 2 of Algorithm 2 and Algorithm 3 are based on individual optimization of each w_i and p_l assuming that all other variables $w_{j,j\neq i}$ and $p_{k,k\neq l}$ are held fixed. Such an optimization step can be done centrally via a one-dimensional search, which is computationally feasible.

The convergence of the source coding algorithm and the power control algorithm can be easily established based on the monotonicity of the optimization steps (see e.g., [23, Prop 2.7.1])[3]. Further, since the objective functions Q^{APP} and Q^{PHY} are continuously differentiable in $\mathbf{w}$ and $\mathbf{p}$, the coordinate-descent-based Algorithm 2 and Algorithm 3 would converge to stationary points, which are local optima of the source coding and power control subproblems. Finally, when the problem is convex (e.g., under the source coding optimality condition (17) or (18) for the source coding problem), coordinate descent reaches a global optimum.

B. Column-Generation Method

Although Algorithm 2 and Algorithm 3 can be used to compute locally optimal points for the two subproblems efficiently, because the objective functions are nonconvex, the local optimal solutions are not necessarily globally optimal. The main purpose of this section is to propose the use of a column-generation technique that allows efficient representations of

[3]Technically, the convergence of $\mathbf{w}^{(\tau)}$ and $\mathbf{p}^{(\tau)}$ also requires a condition that the optimum in each step is uniquely attained. However, this condition can be circumvented by adding a small quadrature term in the iterative process [23].

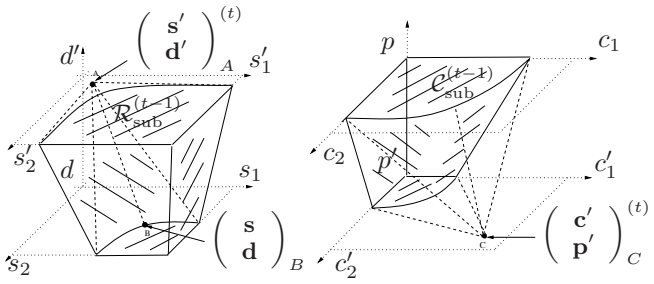


Fig. 4. (a) Convexifying the rate-distortion subregion (b) Convexifying the power-capacity subregion.

rate-distortion and power-capacity subregions based on locally optimal solutions.

The column-generation method basically collects different solution points from the iterative algorithm, and convexifies the region by time sharing among these points. Let $\mathrm{Co}\{\}$ denote the convex hull operator. Then, each time a new rate-distortion pair or a new power-capacity pair is found, we update the respective regions by

$$\begin{aligned}\mathcal{R}_{\rm sub}^{(t)} &= \mathrm{Co}\left\{\mathcal{R}_{\rm sub}^{(t-1)} \bigcup \begin{pmatrix}\mathbf{s}'\\ \mathbf{d}'\end{pmatrix}^{(t)}\right\}\\ \mathcal{C}_{\rm sub}^{(t)} &= \mathrm{Co}\left\{\mathcal{C}_{\rm sub}^{(t-1)} \bigcup \begin{pmatrix}\mathbf{c}'\\ \mathbf{p}'\end{pmatrix}^{(t)}\right\}\end{aligned}$$

where $\mathcal{R}_{\rm sub}^{(t)}$ and $\mathcal{C}_{\rm sub}^{(t)}$ are the convexified subregions up to time t, and $\begin{pmatrix}\mathbf{s}'\\ \mathbf{d}'\end{pmatrix}^{(t)}$, $\begin{pmatrix}\mathbf{c}'\\ \mathbf{p}'\end{pmatrix}^{(t)}$ are the solutions found using Algorithm 2 and Algorithm 3 respectively at time t. These newly found solutions are essentially new modes of operations for the network. The column-generation method allows the network to operate at time-sharing points among these modes.

This convexification process can be incorporated into the overall dual iterative algorithm with different rate-distortion and power-capacity points generated in each iteration step. Clearly, the result of this convexification process does not necessarily converge to the full rate-distortion or power-capacity regions, since only local optimal solutions are obtained in each step. Nevertheless, as it is in general difficult to enumerate all extreme points of the full region explicitly, the subregions $\mathcal{R}_{\rm sub}^{(t)}$ and $\mathcal{C}_{\rm sub}^{(t)}$ are reasonable alternatives. The subregions are upper bounded by the full-regions and they are always convex.

The convexifying process for the rate-distortion subregion is illustrated in Fig. 4(a). The shaded area at the bottom plane is the set of source rates that allow the reconstruction of the underlying source at distortion $\mathbf{d}$, while the shaded area at the top plane is the set of source rates that allow reconstruction at distortion $\mathbf{d}'$. For example, two source coding schemes (A and B) achieve two distortion levels ($\mathbf{d}'$ and $\mathbf{d}$) at source rates ($\mathbf{s}'$ and $\mathbf{s}$), respectively. Using scheme A for a fixed portion of the time and scheme B for the remaining time achieves all points between A and B. The dashed line between the new point $\begin{pmatrix}\mathbf{s}'\\ \mathbf{d}'\end{pmatrix}_A^{(t)}$ and $\mathcal{R}_{\rm sub}^{(t-1)}$ represents the convexifying process by time sharing. A similar process applies to power-capacity subregion as shown in Fig. 4(b).

C. Dual Decomposition Algorithm with Column-Generation

We now present our main algorithm: a dual decomposition algorithm for joint source coding, routing, and power allocation problem. This dual algorithm, when combined with a column-generation method, allows an efficient solution for the joint optimization problem (1).

Algorithm 4: Dual Decomposition Algorithm with Column Generation

1) Initialize $\lambda^{(0)}$ and subregions $\mathcal{C}_{\rm sub}^{(0)}$, $\mathcal{R}_{\rm sub}^{(0)}$.
2) At time t, set $\lambda = \lambda^{(t)}$, $\mu^T = \lambda^T A$.
   2.1) Compute

$$\begin{aligned}&\text{Source Coding Algorithm 2} \;\rightarrow\; \mathbf{s}^G, \mathbf{d}^G\\ &\text{Power Control Algorithm 3} \;\rightarrow\; \mathbf{c}^G, \mathbf{p}^G\\ &\left\{\min_{\mathbf{s},\mathbf{d}} \alpha^T\mathbf{d} + \lambda^T\mathbf{s} \;\middle|\; \begin{pmatrix}\mathbf{s}\\ \mathbf{d}\end{pmatrix} \in \mathcal{R}_{\rm sub}^{(t)}\right\} \;\rightarrow\; \mathbf{s}_{\rm sub}^*, \mathbf{d}_{\rm sub}^*\\ &\left\{\max_{\mathbf{c},\mathbf{p}} \mu^T\mathbf{c} - \beta^T\mathbf{p} \;\middle|\; \begin{pmatrix}\mathbf{c}\\ \mathbf{p}\end{pmatrix} \in \mathcal{C}_{\rm sub}^{(t)}\right\} \;\rightarrow\; \mathbf{c}_{\rm sub}^*, \mathbf{p}_{\rm sub}^*\end{aligned}$$

   2.2) Update the source rate and the rate-distortion subregion:

$$\begin{aligned}&\text{If} \quad \alpha^T\mathbf{d}^G + \lambda^T\mathbf{s}^G < \alpha^T\mathbf{d}_{\rm sub}^* + \lambda^T\mathbf{s}_{\rm sub}^*\\ &\quad \mathbf{s}^* = \mathbf{s}^G,\ \mathcal{R}_{\rm sub}^{(t+1)} = \mathrm{Co}\left\{\mathcal{R}_{\rm sub}^{(t)} \bigcup \begin{pmatrix}\mathbf{s}^G\\ \mathbf{d}^G\end{pmatrix}\right\}\\ &\text{If} \quad \alpha^T\mathbf{d}^G + \lambda^T\mathbf{s}^G \geq \alpha^T\mathbf{d}_{\rm sub}^* + \lambda^T\mathbf{s}_{\rm sub}^*\\ &\quad \mathbf{s}^* = \mathbf{s}_{\rm sub}^*,\ \mathcal{R}_{\rm sub}^{(t+1)} = \mathcal{R}_{\rm sub}^{(t)}\end{aligned}$$

   2.3) Update the link capacity and the power-capacity subregion:

$$\begin{aligned}&\text{If} \quad \mu^T\mathbf{c}^G - \beta^T\mathbf{p}^G > \mu^T\mathbf{c}_{\rm sub}^* - \beta^T\mathbf{p}_{\rm sub}^*\\ &\quad \mathbf{c}^* = \mathbf{c}^G,\ \mathcal{C}_{\rm sub}^{(t+1)} = \mathrm{Co}\left\{\mathcal{C}_{\rm sub}^{(t)} \bigcup \begin{pmatrix}\mathbf{c}^G\\ \mathbf{p}^G\end{pmatrix}\right\}\\ &\text{If} \quad \mu^T\mathbf{c}^G - \beta^T\mathbf{p}^G \leq \mu^T\mathbf{c}_{\rm sub}^* - \beta^T\mathbf{p}_{\rm sub}^*\\ &\quad \mathbf{c}^* = \mathbf{c}_{\rm sub}^*,\ \mathcal{C}_{\rm sub}^{(t+1)} = \mathcal{C}_{\rm sub}^{(t)}\end{aligned}$$

3) In dual domain, update λ using the following rule:

$$\lambda^{(t+1)} = \left[\lambda^{(t)} + \epsilon^{(t)}(\mathbf{s}^* - A\mathbf{c}^*)\right]^+ \tag{20}$$

4) Set $t = t+1$. Return to Step 2 until subregions converge.
5) Given the converged subregions, solve the following linear programming problem to obtain a solution for problem (1):

$$\begin{aligned}\text{minimize} \quad & \alpha^T\mathbf{d} + \beta^T\mathbf{p} \qquad (21)\\ \text{subject to} \quad & \begin{pmatrix}\mathbf{s}\\ \mathbf{d}\end{pmatrix} = \sum_k \gamma_k \begin{pmatrix}\mathbf{s}\\ \mathbf{d}\end{pmatrix}_k^{\rm sub}\\ & \sum_k \gamma_k = 1,\ \gamma_k \geq 0\\ & \begin{pmatrix}\mathbf{c}\\ \mathbf{p}\end{pmatrix} = \sum_j \pi_j \begin{pmatrix}\mathbf{c}\\ \mathbf{p}\end{pmatrix}_j^{\rm sub}\\ & \sum_k \pi_j = 1,\ \pi_j \geq 0\\ & A\mathbf{c} \geq \mathbf{s}\end{aligned}$$

In the initialization step, the subregions are set to be empty sets. By convention, the source coding minimization subproblem returns a value of infinity when the rate-distortion subregion is an empty set. Similarly, the power allocation maximization subproblem returns a value of zero when the power-capacity subregion is an empty set.

In Step 2.1, source coding Algorithm 2 and power control Algorithm 3 are used to obtain new local optimal solutions $(\mathbf{s}^G, \mathbf{d}^G)$ and $(\mathbf{c}^G, \mathbf{p}^G)$ for the application-layer and physical-layer subproblems, respectively. Since the dual variables λ and μ are newly updated in each step, $(\mathbf{s}^G, \mathbf{d}^G)$ and $(\mathbf{c}^G, \mathbf{p}^G)$ can potentially be new extreme points of the current achievable rate-distortion and power-capacity subregions. The subregions $\mathcal{R}_{\text{sub}}$ and $\mathcal{C}_{\text{sub}}$ are updated when this happens. To test whether the new points are outside of the current subregions, Steps 2.2 and 2.3 compare the new points with the optimal points $(\mathbf{s}^*_{\text{sub}}, \mathbf{d}^*_{\text{sub}})$ and $(\mathbf{c}^*_{\text{sub}}, \mathbf{p}^*_{\text{sub}})$ inside the current $\mathcal{R}^{(t)}_{\text{sub}}$ and $\mathcal{C}^{(t)}_{\text{sub}}$ using the current λ and μ. Since the subregions are polytopes defined by their extreme points, the solutions for the subproblems can be obtained efficiently by linear programming. In particular, let $\begin{pmatrix} \mathbf{s} \\ \mathbf{d} \end{pmatrix}^{\text{sub}}_{k=1,\cdots,K}$ and $\begin{pmatrix} \mathbf{p} \\ \mathbf{d} \end{pmatrix}^{\text{sub}}_{j=1,\cdots,J}$ denote the extreme points in subregions. The rate-distortion region problem is

$$\begin{aligned} \text{minimize} \quad & \alpha^T \mathbf{d} + \lambda^T \mathbf{s} \\ \text{subject to} \quad & \begin{pmatrix} \mathbf{s} \\ \mathbf{d} \end{pmatrix} = \sum_k \gamma_k \begin{pmatrix} \mathbf{s} \\ \mathbf{d} \end{pmatrix}^{\text{sub}}_k \\ & \sum_k \gamma_k = 1, \ \gamma_k \geq 0 \end{aligned} \tag{22}$$

The power-capacity region problem is

$$\begin{aligned} \text{maximize} \quad & \mu^T \mathbf{c} - \beta^T \mathbf{p} \\ \text{subject to} \quad & \begin{pmatrix} \mathbf{c} \\ \mathbf{p} \end{pmatrix} = \sum_j \pi_j \begin{pmatrix} \mathbf{c} \\ \mathbf{p} \end{pmatrix}^{\text{sub}}_j \\ & \sum_j \pi_j = 1, \ \pi_j \geq 0 \end{aligned} \tag{23}$$

Note that the solutions to the above linear program problems may be nonunique. Nevertheless, the algorithm works as long as one valid solution is given.

In Step 3, the dual variables are updated according to the rate demand $\mathbf{s}^*$ and rate supply $A\mathbf{c}^*$. Step 4 increases the outer loop time index and checks for the convergence of the subregions. Since the subregions $\mathcal{R}_{\text{sub}}$ and $\mathcal{C}_{\text{sub}}$ are upper bounded by the full regions and they are monotonically non-contracting in the iterative process, the algorithm is guaranteed to converge. Once the subregions converge, Step 5 computes the final solution based on the extreme points of the subregions $\begin{pmatrix} \mathbf{s} \\ \mathbf{d} \end{pmatrix}^{\text{sub}}_k$ and $\begin{pmatrix} \mathbf{c} \\ \mathbf{p} \end{pmatrix}^{\text{sub}}_k$ using another linear programming step.

Note that the outer loop of the Algorithm 4 is indexed by t. The inner loops, i.e. the source coding Algorithm 2 and power control Algorithm 3 of Step 2.1 are indexed by τ.

The solution from Step 5 is always primal feasible, but it may not be globally optimal in general, because only local optimal points are generated in the iterative steps. For small networks, it is possible to bound the difference between the current solution and the true optimum using a duality-gap technique (combined with exhaustive search); see e.g., [7]. Our simulation experience on small networks suggests that the proposed algorithm is sometimes capable of reaching a solution close to the global optimum. However, because the underlying capacity region maximization problem is inherently nonconvex, it is in general difficult to guarantee that the solution would always be close to the optimum.

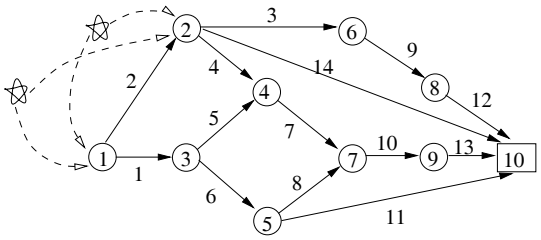


Fig. 5. Topology of a wireless sensor network.

The solution obtained from the proposed algorithm can either take the form of a single local optimal point of a subregion (i.e. a single column), or the time-sharing of multiple local optimal points. The time-sharing factor is automatically determined by the linear programming step (i.e. Step 5). If the obtained solution is a single local optimal point, the appropriate quantization and power allocation settings can be delivered directly to the sensors. To implement a time-sharing solution, a synchronization mechanism needs to be implemented. Alternatively, time-sharing of multiple physical-layer power allocation schemes can also be implemented in the frequency domain.

The column-generation method presented here is inspired by the work of [7]. However, the current approach also differs from that of [7] in the sense that Algorithm 4 does not require the power control subproblem (19) to be solved optimally, which is computationally difficult in general. Instead, we allow local optimal points, which can be computed efficiently using iterative methods, to be added in each step. Further, the current work also generalizes that of [7] by applying the column-generation method to the rate-distortion problem.

The column-generation method requires centralized control because of the need for time-sharing operations that convexify the subregions. This suggests that Algorithm 4 should be run at CEO, which usually has more computational resource and is able to centrally collect and estimate network parameters such as observation channel matrix and wireless channel coefficients.

V. Simulation Example

We now present simulation results on a wireless sensor network example to illustrate the main ideas of this paper. The network topology is shown in Fig. 5. The underlying physical source is modeled as a two-dimension Gaussian vector with an identity covariance matrix. For the sake of simplicity, we assume that two nodes closest to the source (e.g., sensors 1 and 2) are active in sensing the field, while the rest of the nodes act as relays.

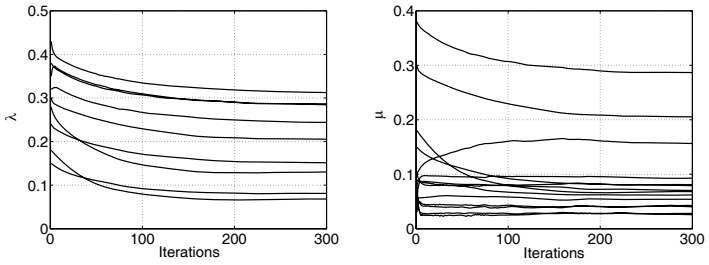


Fig. 6. Convergence process of the dual variables. Each curve in the figure on the left represents a dual variable λ_i for each node. Each curve in the figure on the right represents a dual variable μ_l for each link. The x-axis corresponds to the iteration index t in Algorithm 4.

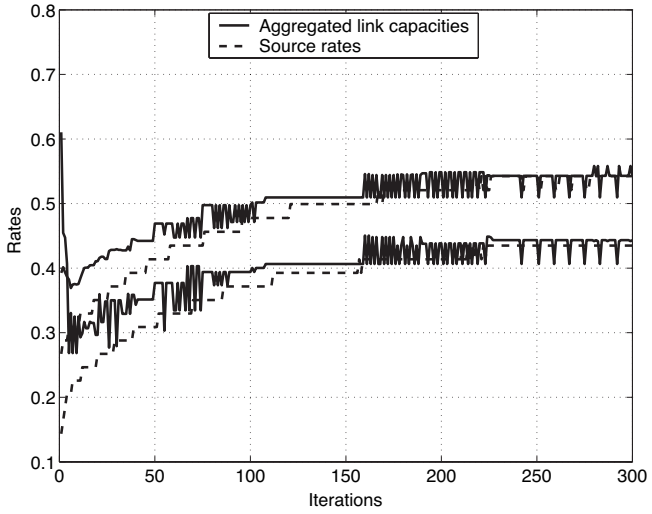


Fig. 7. Convergence of the source rates and the aggregated outgoing link capacities at the two sensing nodes. The source rates eventually match the aggregated link capacities. The link capacities oscillate between two modes of operation, implying that the optimal solution is a time-sharing point of the two modes. The x-axis corresponds to the iteration index t in Algorithm 4.

The wireless channel coefficients G_{lj} are modeled using a deterministic path loss model, i.e. $G_{lj} = K_{lj} d_{lj}^{-\rho}$, where d_{lj} is the distance between the transmitter of link j and the receiver of the link l, ρ is a deterministic path loss exponent, and K_{lj} is a normalization constant, which depends on the propagation characteristics of the environment. In this example, ρ is set to 4, and K_{lj}'s are generated from a log-normal distribution, modeling log-normal shadowing. In addition, the sensor nodes are assumed to be capable of transmitting in multiple frequency subbands in the physical layer. Independent transmission takes place in each subband; transmit power may be freely allocated among the subbands; the achievable link capacity is the sum of subband transmission rates. This enables the implementation of time-sharing solution in the frequency domain.

We use the proposed dual decomposition algorithm with column-generation to find the jointly optimal source coding, routing and power allocation for the network. Fig. 6 shows the convergence process of the dual variables λ and μ. Each curve in the left figure corresponds to a dual variable λ_i for each node; each curve in the right figure corresponds to μ_l for each link. As shown in the figure, all dual variables converge.

Fig. 7 plots the convergence behavior of the source rates at the two sensing nodes (dashed line) and the aggregated outgoing link rates from the sensing nodes (solid line). The convergence process illustrates the interaction between the aggregated application-layer demand and the aggregated physical-layer supply of rates. At the beginning, the application-layer demand for rates is low, while the physical-layer supply for link capacities is high. As the supply is greater than the demand, the node prices λ would decrease as shown in Fig. 6. Due to the change in prices, the physical layer would reduce its capacities, while the application layer would increase its source rates. This negotiation process is coordinated by the dual variables. Eventually, the source rates reach a steady state. However, the link capacities fluctuate among two different modes of operation. This is due to the fact that the power-capacity subregion is a polytope. Thus, a small change in the dual variables may cause the optimal primal variable to oscillate between neighboring vertices. This implies that time-sharing between the vertices is needed to achieve the optimum. As in Step 5 of Algorithm 4, the optimal time-sharing factors can be found by solving the linear programming problem (21). Further, it is important to realize that oscillation occurs only among a few points. Thus, the computational complexity of linear programming can be reduced by considering only these few points[4].

For the particular example presented here, the optimal solution consists of source rates $s_1 = 0.435$, $s_2 = 0.538$, $s_3 = \cdots = s_9 = 0$ and link capacities $c_1 = 0.436$, $c_2 = 0$, $c_3 = 0.309$, $c_4 = 0.229$, $c_5 = 0.126$, $c_6 = 0.317$, $c_7 = 0.360$, $c_8 = 0.286$, $c_9 = 0.309$, $c_{10} = 0.649$, $c_{11} = 0.035$, $c_{12} = 0.309$, $c_{13} = 0.650$, and $c_{14} = 0$. The corresponding distortion is $\alpha^T \mathbf{d} = 1.62$. It is interesting to note that the optimized capacities for link 2 and link 14 are both zero. The optimal capacity for link 14 is zero, because operating long distance links is not power efficient. Moreover, the fact that link 2 has a zero optimal capacity indicates that node 1 can find an alternative route to the CEO without going through node 2.

It is also interesting to observe that the solution of link capacities and source rates from Algorithm 4 satisfies flow conservation exactly, i.e. $A\mathbf{c} = \mathbf{s}$. Therefore, the solution is energy efficient in the sense that there are no slack link capacities.

The iteration number in Fig. 6 and Fig. 7 corresponds to the number of dual variable λ update (indexed by t). This is the number of outer-loop iterations in Algorithm 4. For each outer iteration step, the source coding Algorithm 2 and the power control Algorithm 3 are used to obtain local optimum solutions for the two subproblems. For the particular setup of Fig. 5, Algorithm 4 takes around 300 outer-loop iterations to converge, while the source coding algorithm takes an average of 2 inner-loop iterations and the power control algorithm takes an average of 8 inner-loop iterations to converge.

[4]Note that during the iteration process, columns that are in the interior of the rate-distortion and power-capacity subregions may be pruned to maintain an economic representation of the subregions.

VI. Conclusions

This paper proposes a general framework for the cross-layer optimization of wireless sensor networks. Using a Lagrangian dual approach, we show that the joint source coding, routing and power allocation problem in a wireless sensor network can be naturally decomposed into two subproblems: a power allocation subproblem at the physical layer and a source coding subproblem at the application layer, with a set of dual variables coordinating the two layers. This dual decomposition approach allows the overall problem to be solved efficiently in a modular fashion.

Our approach is facilitated by a number of technical developments. We introduce a simplified source coding model to characterize the rate-distortion region at the application layer; we devise iterative algorithms for achieving local optimal solutions to the source coding and power allocation problems; we incorporate a column-generation method to ensure the convexity of the rate-distortion and power-capacity regions and the efficient solution of the overall joint optimization problem. Our approach can be thought of as a generalization of the network utility maximization problem in which the overall utility is coupled across the nodes. This paper presents a natural framework to model and to utilize this coupling – a crucial task in sensor network design.

References

[1] F. P. Kelly, A. Maulloo, and D. Tan, "Rate control for communication networks: shadow prices, proportional fairness and stability," *J. Operations Research Society*, vol. 49, no. 3, pp. 237–252, Mar. 1998.

[2] S. H. Low and D. E. Lapsley, "Optimization flow control, I: basic algorithm and convergence," *IEEE/ACM Trans. Networking*, vol. 7, no. 6, pp. 861–875, Dec. 1999.

[3] L. Xiao, M. Johansson, and S. Boyd, "Simultaneous routing and resource allocation via dual decomposition," *IEEE Trans. Commun.*, vol. 52, no. 7, pp. 1136–1144, July 2004.

[4] M. Chiang, "Balancing transport and physical layers in wireless multihop networks: jointly optimal congestion control and power control," *IEEE J. Select. Areas Commun.*, vol. 23, no. 1, pp. 104–116, Jan. 2005.

[5] J. Wang, L. Li, S. H. Low, and J. C. Doyle, "Cross-layer optimization in TCP/IP networks," *IEEE/ACM Trans. Networking*, vol. 13, no. 3, pp. 582–568, June 2005.

[6] X. Lin and N. B. Shroff, "The impact of imperfect scheduling on cross-layer rate control in wireless netowrks," *IEEE/ACM Trans. Networking*, vol. 14, no. 2, pp. 302–315, Apr. 2006.

[7] M. Johansson and L. Xiao, "Cross-layer optimization of wireless networks using nonlinear column generation," *IEEE Trans. Wireless Commun.*, vol. 5, no. 2, pp. 435–445, Feb. 2006.

[8] G. Kim, A. Rajeswaran, and R. Negi, "Joint power adaptation, scheduling and routing framework for wireless ad-hoc networks," in *Proc. the 6th IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, New York, June 2005.

[9] J. Xiao, S. Cui, Z.-Q. Luo, and A. J. Goldsmith, "Power scheduling of universal decentralized estimation in sensor networks," *IEEE Trans. Signal Processing*, vol. 54, no. 2, pp. 413–422, Feb. 2006.

[10] J.-J. Xiao and Z.-Q. Luo, "Multiterminal source-channel communication under orthogonal multiple access," in *Proc. 43th Allerton Conference on Communication, Control and Computing*, Sept. 2005.

[11] A. Scaglione and S. D. Servetto, "On the interdependence of routing and data compression in multi-hop sensor neworks," in *Proc. the 8th ACM International Conference on Mobile Computing and Networking (MobiCom)*, Atlanta, GA, Sept. 2002.

[12] M. Gastpar, "Distributed source-channel coding for wireless sensor networks," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Montreal, Canada, May 2004.

[13] S. C. Draper and G. W. Wornell, "Side information aware coding strategies for sensor networks," *IEEE J. Select. Areas Commun.*, vol. 22, no. 6, pp. 966–976, June 2004.

[14] D. Bertsekas and R. G. Gallager, *Data Networks*. Prentice Hall, 1991.

[15] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[16] N. Z. Shor, *Minimization Methods for Non-Differentiable Functions*. Springer, 1985.

[17] S. Boyd, L. Xiao, and A. Mutapcic, "Subgradient methods," 2003. [Online] http://www.stanford.edu/class/ee392o/subgrad-method.pdf.

[18] W. R. Bennett, "Spectra of quantized signals," *Bell Syst. Tech. J.*, vol. 27, pp. 446–472, July 1948.

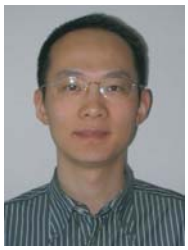
[19] B. Widrow, "Statistical analysis of amplitude quantized sampled data systems," *Trans. Amer. Inst. Elec. Eng. Pt. II: Applications and Industry*, vol. 79, pp. 555–568, Jan. 1960.

[20] T. Berger, "Multiterminal source coding," in *The Information Theory Approach to Communications*, G. Longo, Ed., vol. 229 of *CISM Courses and Lectures*, pp. 171–231. Springer-Verlag, 1978.

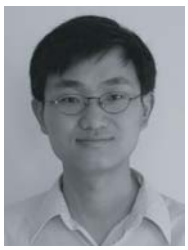
[21] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 471–480, July 1973.

[22] J. Yuan, *Optimization Techniques for Wireless Networks*, Ph.D. dissertation, University of Toronto, Department of Electrical and Computer Engineering, Sept. 2007.

[23] D. Bertsekas, *Nonlinear Programming*. Athena Scientific, 1999.



Jun Yuan (S'04) received the B.E. degree in Electrical Engineering and B.S. degree in Applied Mathematics from Shanghai Jiao Tong University, Shanghai, China in 2001. He received the M.S. degree in Electrical and Computer Engineering from Queen's University, Kingston, Ontario, Canada, in 2003, and the Ph.D. degree in Electrical and Computer Engineering from the University of Toronto, Toronto, Ontario, Canada in 2007. He is currently a MIMO-OFDM system designer with the Wireless Technology Labs, Nortel, Ottawa, Ontario, Canada. His research interests include wireless communications, optimization and cross-layer design.



Wei Yu (S'97-M'02) received the B.A.Sc. degree in Computer Engineering and Mathematics from the University of Waterloo, Waterloo, Ontario, Canada in 1997 and M.S. and Ph.D. degrees in Electrical Engineering from Stanford University, Stanford, CA, in 1998 and 2002, respectively. Since 2002, he has been an Assistant Professor with the Electrical and Computer Engineering Department at the University of Toronto, Toronto, Ontario, Canada, where he also holds a Canada Research Chair. His main research interests include multiuser information theory, optimization, wireless communications and broadband access networks.

Prof. Wei Yu was an Editor for *IEEE Transactions on Wireless Communications* from 2004 to 2007. He was a Guest Editor of *IEEE Journal on Selected Areas in Communications* for a special issue on "Nonlinear Optimization of Communications Systems" in 2006, and a Guest Editor of *EURASIP Journal on Applied Signal Processing* for a special issue on "Advanced Signal Processing for Digital Subscriber Lines" in 2005. He received the Early Researcher Award from Ontario in 2006, and the Early Career Teaching Award from the Faculty of Applied Science and Engineering, University of Toronto, in 2007.