Greedy Adaptive Linear Compression in Signal-Plus-Noise Models

Entao Liu, *Member, IEEE*, Edwin K. P. Chong, *Fellow, IEEE*, and Louis L. Scharf, *Life Fellow, IEEE*

***Abstract*— In this paper, we examine adaptive compression policies, when the sequence of vector-valued measurements to be compressed is noisy and the compressed variables are themselves noisy. The optimization criterion is information gain. In the case of sequential scalar compressions, the unit-norm compression vectors that greedily maximize per-stage information gain are eigenvectors of an *a priori* error covariance matrix, and the greedy policy selects them according to eigenvalues of a posterior covariance matrix. These eigenvalues depend on all previous compressions and are computed recursively. A water-filling solution is given for the optimum compression policy that maximizes net information gain, under a constraint on the average norm of compression vectors. We provide sufficient conditions under which the greedy policy for maximizing stepwise information gain actually is optimal in the sense of maximizing the net information gain. In the case of scalar compressions, our examples and simulation results illustrate that the greedy policy can be quite close to optimal when the noise sequences are white.**

***Index Terms*— Entropy, information gain, compressive sensing, compressed sensing, greedy policy, optimal policy.**

I. Introduction

A. Background

Consider a signal of interest $\mathbf{x}$, which is a random vector taking values in $\mathbb{R}^N$ with prior distribution $\mathcal{N}(\boldsymbol{\mu}, \mathbf{P}_0)$ (i.e., $\mathbf{x}$ is Gaussian distributed with mean $\boldsymbol{\mu}$ and $N \times N$ covariance matrix $\mathbf{P}_0$). Note that $\mathbf{P}_0$ may be singular, indicating a linear dependency among the components of $\mathbf{x}$. Whenever this happens, we can simply sample a subset of components $\mathbf{x}'$ whose covariance matrix is nonsingular and $\mathbf{x}$ is linearly determined by $\mathbf{x}'$. Hence, we naturally assume $\mathbf{P}_0$ is nonsingular throughout this paper. The signal $\mathbf{x}$ is carried over a noisy channel to a sensor, according to the model $\mathbf{z} := \mathbf{H}\mathbf{x} + \mathbf{n}$ where $\mathbf{H} \in \mathbb{R}^{K \times N}$ is a full-rank sensor matrix. For simplicity, in this paper we focus on the case where $K \geq N$, though analogous results are obtained when $K < N$. Our aim is to compress m realizations of $\mathbf{z}$, denoted $\mathbf{z}_k = \mathbf{H}\mathbf{x} + \mathbf{n}_k$, $k = 1, \ldots, m$, where m is specified upfront. Because the implementation of each compression has a noise penalty, the kth compressed measurement is

$$\mathbf{y}_k = \mathbf{A}_k(\mathbf{H}\mathbf{x} + \mathbf{n}_k) + \mathbf{w}_k \tag{1}$$

where the compression matrix $\mathbf{A}_k$ is $L \times K$. We will mainly consider scalar models where $L = 1$ throughout the paper, but we will also briefly consider a vector case in Section V-A. Consequently, the measurement $\mathbf{y}_k$ takes values in $\mathbb{R}^L$. Assume that the measurement noise $\mathbf{w}_k \in \mathbb{R}^L$ has distribution $\mathcal{N}(\mathbf{0}, \mathbf{R}_{ww})$ and sensor noise $\mathbf{n}_k \in \mathbb{R}^K$ has distribution $\mathcal{N}(\mathbf{0}, \mathbf{R}_{nn})$. The measurement and sensor noise sequences are independent over k and independent of each other. We can rewrite (1) as

$$\mathbf{y}_k = \mathbf{A}_k\mathbf{H}\mathbf{x} + (\mathbf{A}_k\mathbf{n}_k + \mathbf{w}_k) \tag{2}$$

and consider $\mathbf{A}_k\mathbf{n}_k + \mathbf{w}_k$ as the total noise with distribution $\mathcal{N}(\mathbf{0}, \mathbf{A}_k\mathbf{R}_{nn}\mathbf{A}_k^T + \mathbf{R}_{ww})$.

We consider the following adaptive (sequential) compression problem. For each $k = 1, \ldots, m$, we are allowed to choose the compression matrix $\mathbf{A}_k$ (possibly subject to some constraint). Moreover, our choice is allowed to depend on the entire history of measurements up to that point: $\mathcal{I}_{k-1} = \{\mathbf{y}_1, \ldots, \mathbf{y}_{k-1}\}$.

It is easy to check that the posterior distribution of $\mathbf{x}$ given $\mathcal{I}_k$ is still Gaussian. Let $\mathbf{x}_k$ denote the mean and $\mathbf{P}_k$ the covariance matrix. Concretely, $\mathbf{P}_k$ can be written recursively for $k = 1, \ldots, m$ as

$$\mathbf{P}_k = \mathbf{P}_{k-1} - \mathbf{P}_{k-1}\mathbf{B}_k^T(\mathbf{B}_k\mathbf{P}_{k-1}\mathbf{B}_k^T + \mathbf{N}_k)^{-1}\mathbf{B}_k\mathbf{P}_{k-1} \tag{3}$$

where $\mathbf{B}_k := \mathbf{A}_k\mathbf{H}$ and $\mathbf{N}_k := \mathbf{A}_k\mathbf{R}_{nn}\mathbf{A}_k^T + \mathbf{R}_{ww}$. Assume that all components of total noise $\mathbf{A}_k\mathbf{n}_k + \mathbf{w}_k$ are linearly independent, which implies that $\mathbf{N}_k$ is invertible. Moreover, recursively from the assumption of the invertibility of $\mathbf{P}_0$, we deduce that the $\mathbf{P}_{k-1}$ are nonsingular. Then by the Woodbury identity a simpler version of (3) is

$$\mathbf{P}_k = \left(\mathbf{P}_{k-1}^{-1} + \mathbf{B}_k^T\mathbf{N}_k^{-1}\mathbf{B}_k\right)^{-1}. \tag{4}$$

In order to introduce the optimization criterion based on information gain, let us recall the definition of the *entropy* of the posterior distribution of $\mathbf{x}$ given $\mathcal{I}_k$ [1]:

$$H_k = \frac{1}{2}\log\det(\mathbf{P}_k) + \frac{N}{2}\log(2\pi e), \tag{5}$$

Manuscript received July 12, 2012; revised January 10, 2014; accepted February 18, 2014. Date of publication February 25, 2014; date of current version March 13, 2014. This work was supported in part by DARPA under Contract N66001-11-C- 4023, in part by ONR under Contract N00014-08-1-110, in part by NSF under Grant CFF-1018472, and in part by AFOSR under Contract FA-9550-10-1-0241. This paper was presented at the 2012 Conference on Information Sciences and Systems and Asilomar 2012.

E. Liu is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: liuentao@gmail.com).

E. K. P. Chong is with the Department of Electrical and Computer Engineering and Department of Mathematics, Colorado State University, Fort Collins, CO 80523 USA (e-mail: edwin.chong@colostate.edu).

L. L. Scharf is with the Department of Mathematics and Statistics, Colorado State University, Fort Collins, CO 80523 USA (e-mail: scharf@engr.colostate.edu).

Communicated by D. Guo, Associate Editor for Shannon Theory.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2014.2308258

where the first term $\det(\mathbf{P}_k)$ is actually proportional to the volume of the error concentration ellipse for $\mathbf{x} - \mathbb{E}[\mathbf{x}|\mathcal{I}_k]$. We now focus on a common information-theoretic criterion for choosing the compression matrices: for the kth compression matrix, we pick $\mathbf{A}_k$ to maximize the *per-stage information gain*, defined as $H_{k-1} - H_k$. For reasons that will be made clear later, we refer to this strategy as a *greedy* policy. The term *policy* simply refers to a rule for picking $\mathbf{A}_k$ for each k based on $\mathcal{I}_{k-1}$.

Suppose that the overall goal is to maximize the *net information gain*, defined as $H_0 - H_m$. We ask the following questions: Does the greedy policy achieve this goal? If not, then what policy achieves it? How much better is such a policy than the greedy one? Are there cases where the greedy policy does achieve this goal? The objective of this paper is to provide some answer (occasionally partial) to these questions. In Section II, we analyze the greedy policy and compute its net information gain. In Section III, to find the net information gain of the optimal policy, we introduce a relaxed optimization problem, which can be solved as a water-filling problem. In Section IV, we derive two sufficient conditions under which the greedy policy is optimal. In Section V, we give examples and numerical simulations for which the greedy policy is not optimal, but close to optimal.

B. Relevance of the Problem and Relation of This Paper to Prior Work

There is by now a burgeoning literature on compressed sensing, but a relatively small subset of it is addressed to the problem of compressing noisy measurements for transmission over a noisy channel, based on statistical models for measurements and signals. In the following paragraphs we review a representative set of papers which do bear on this problem, so that our work may be placed in context with prior work.

Let us establish a distinction between compression in a *single experiment problem* and in a *multiple-experiment problem*. In a single experiment problem, references [2] and [3] address the problem of compression of $\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{n}$ into a sequence of scalar $u_k = \mathbf{a}_k^T \mathbf{z}$ for subsequent estimation of $\mathbf{x}$. References [4]–[10] address a variation on the problem of measuring $\mathbf{y}$ in the model $\mathbf{y} = \mathbf{C}\mathbf{z} + \mathbf{w}$, where $\mathbf{z} = \mathbf{H}\mathbf{x} + \mathbf{n}$ is a transformation and noisy sensing of the signal $\mathbf{x}$ and $\mathbf{C}$ is a channel matrix. The compression problem is to compress $\mathbf{z}$ as $\Phi\mathbf{z} = \Phi(\mathbf{H}\mathbf{x} + \mathbf{n})$, under a power constraint, in which case the lower dimensional measurement $\mathbf{y} = \mathbf{C}\Phi(\mathbf{H}\mathbf{x} + \mathbf{n}) + \mathbf{w}$ is to be processed for an estimator of $\mathbf{x}$. The question is how to design the compression matrix Φ so that performance is managed, according to some appropriate measure. Common measures are trace and determinant of the resulting error covariance matrix, which in the multivariate normal model are related to mutual information and differential information rate. In [4] the authors introduce Renyi information as a performance measure. The important point to be made is that in the single experiment problem there is *only one* realization of the experiment, which for $\mathbf{C} = \mathbf{I}$ produces the realization $\mathbf{y} = \Phi(\mathbf{H}\mathbf{x} + \mathbf{n}) + \mathbf{w}$. In references [5] and [6] there is no transmission of the compression $\Phi\mathbf{z}$ over a channel, and the resulting designs are called reduced-rank filterings. In references [7] and [8] the compression $\Phi(\mathbf{H}\mathbf{x} + \mathbf{n})$ is transmitted over a channel to be received as $\mathbf{y} = \mathbf{C}\Phi(\mathbf{H}\mathbf{x}+\mathbf{n})+\mathbf{w}$, but the sensor noise $\mathbf{n}$ is zero and the sensor matrix $\mathbf{H}$ is the identity. These designs are called precoder designs. Reference [4] fits this paradigm, but notably, the multivariate normal model of [5]–[10] is replaced by a Gaussian mixture model, and an algorithm for sequentially designing the compression is given. In references [9] and [10], the general model is considered, wherein the noisy sensor output $\mathbf{z} = \mathbf{H}\mathbf{x}+\mathbf{n}$ is compressed, to be received as the noisy measurement $\mathbf{y} = \mathbf{C}\Phi(\mathbf{H}\mathbf{x} + \mathbf{n}) + \mathbf{w}$.

In this paper we treat a different model, the *multiple-experiment model* $\mathbf{y}_k = \mathbf{z}_k+\mathbf{w}_k$, where $\mathbf{z}_k = \mathbf{H}\mathbf{x}+\mathbf{n}_k$, and seek to compress each realization $\mathbf{y}_k$, as $\mathbf{y}_k = \mathbf{A}_k(\mathbf{H}\mathbf{x} + \mathbf{n}_k) + \mathbf{w}_k$. In this model there is a *sequence* of compressed measurements $\mathbf{y}_k$, each of which brings a compressed measurement of the noisy version of the signal $\mathbf{x}$, observed in noise $\mathbf{w}_k$. So the signal is fixed at $\mathbf{x}$, but it is sensed in an uncorrelated sensor noise sequence $\mathbf{n}_k$. This sequence is compressed by the sequence of compressors $\mathbf{A}_k$ and observed in measurement noise as $\mathbf{y}_k = \mathbf{A}_k(\mathbf{H}\mathbf{x}+\mathbf{n}_k)+\mathbf{w}_k$, where the measurement noise sequence $\mathbf{w}_k$ is uncorrelated. The problem we address is the design of the *sequence* of compressors $\mathbf{A}_k$ so that the *sequence* of measurements $\mathbf{y}_k$ conserves "information" as efficiently as possible.

The relevance of our model is this: in radar, sonar, geophysics, chemical detection, and other applications, there is a state of nature $\mathbf{x}$ that is unchanging over some interval of time, during which it may be actively interrogated or passively sensed as $\mathbf{H}\mathbf{x} + \mathbf{n}_k$, compressed as $\mathbf{A}_k(\mathbf{H}\mathbf{x} + \mathbf{n}_k)$, and then transmitted or stored over a noisy channel as $\mathbf{y}_k = \mathbf{A}_k(\mathbf{H}\mathbf{x} + \mathbf{n}_k)+\mathbf{w}_k$. In the work of [4]–[10] there is one sensing $\mathbf{H}\mathbf{x}+\mathbf{n}$, one compression, and one transmission $\mathbf{y} = \mathbf{C}\Phi(\mathbf{H}\mathbf{x}+\mathbf{n})+\mathbf{w}$. Why is this distinction important? Because many problems do produce multiple experiments, and in such cases, we expect a sequence of very low-dimensional compressions $\mathbf{A}_k$ to produce a sequence of low-dimensional measurements from which the signal $\mathbf{x}$ may be estimated. That is, in contrast to the problem of [4]–[10], where a matrix-valued compression of one realization is expected to preserve much information, the multiplicity of measurements in our model is expected to compensate for scalar compression of each measurement. Or said another way: matrix-valued compression is achieved with a sequence of scalar-valued compressions, by virtue of the fact that the scalar-valued compressions may be applied to a multiplicity of measurements.

We have adopted an information-theoretic objective function, namely the net information gain. This choice is worth some discussion. Let $p(\mathbf{x})$ denote the prior probability distribution of multivariate normal $\mathbf{x}$. After m measurement steps, the posterior distribution becomes $p(\mathbf{x}|\mathcal{I}_m)$. The "information" provided by these measurements reduces uncertainty, and this reduction can be measured by the volume of the concentration ellipse. Minimizing the volume of the concentration ellipse is equivalent to minimizing $\det(\mathbf{P}_k)$ and maximizing the net information gain in the multivariate Gaussian case. An alternative is to minimize the mean-squared error, given by $\mathrm{tr}\,(\mathbf{P}_k)$. There is no general answer to the question of

how minimization of $\det(\mathbf{P}_k)$ corresponds to minimization of $\mathrm{tr}\,(\mathbf{P}_k)$. A good design for one may be bad for another, as one of them manages geometric mean of eigenvalues and the other arithmetic mean. But certainly determinant minimization traps the state into a set of smaller volume than does trace minimization. For more on this issue, see [11] and [12]. In particular, [12] has an introduction that makes the argument for using an information-theoretic objective function.

It is worth noting that we may dispense with the assumption of Gaussian distributed variables and argue that we are simply minimizing $\det(\mathbf{P}_k)$, which is proportional to the volume of the error concentration ellipse defined by $(\mathbf{x}-\hat{\mathbf{x}}_{k-1})^T\mathbf{P}_k^{-1}(\mathbf{x}-\hat{\mathbf{x}}_{k-1}) \leq 1$, where $\hat{\mathbf{x}}_{k-1} = \mathbb{E}[\mathbf{x}|\mathcal{I}_{k-1}]$. Notice that the greedy policy does not use the values of $\mathbf{y}_1, \ldots, \mathbf{y}_{k-1}$; its choice of $\mathbf{A}_k$ depends only on $\mathbf{P}_{k-1}$, $\mathbf{R}_{nn}$ and $\mathbf{R}_{ww}$. In fact, the formulas above show that information gain is a deterministic function of the model matrices (in our particular setup). This implies that in principle the optimal policy can be computed by deterministic dynamic programming, though in practice its complexity would make this computation practically intractable. In general, we would not expect the greedy policy to solve such a dynamic programming problem. However, as we will see in following sections, there are cases where it does.

II. Greedy Policy

A. Preliminaries

We now explore how the *greedy policy* performs for the adaptive measurement problem. Before proceeding, we first make some remarks on the information gain criterion:

- Information gain as defined in this paper also goes by the name *mutual information* between $\mathbf{x}$ and $\mathbf{y}_k$ given $\mathcal{I}_{k-1}$ in the case of per-stage information gain, and between $\mathbf{x}$ and $\mathcal{I}_m$ in the case of net information gain.
- The net information gain can be written as the cumulative sum of the per-stage information gains:

$$H_0 - H_m = \sum_{k=1}^{m}(H_{k-1} - H_k).$$

  This is why the greedy policy is so named: at each stage k, the greedy policy simply maximizes the immediate (short-term) contribution $H_{k-1} - H_k$ to the overall cumulative sum.
- Using the formulas (3) and (5) for H_k and $\mathbf{P}_k$, we can write

$$H_{k-1} - H_k = -\frac{1}{2}\log\det\big(\mathbf{I}_N - \mathbf{P}_{k-1}\mathbf{B}_k^T(\mathbf{B}_k\mathbf{P}_{k-1}\mathbf{B}_k^T + \mathbf{N}_k)^{-1}\mathbf{B}_k\big) \quad (6)$$

  where $\mathbf{I}_N$ is the $N \times N$ identity matrix. In other words, at the kth stage, the greedy policy minimizes (with respect to $\mathbf{A}_k$)

$$\log\det(\mathbf{I}_N - \mathbf{P}_{k-1}\mathbf{B}_k^T(\mathbf{B}_k\mathbf{P}_{k-1}\mathbf{B}_k^T + \mathbf{N}_k)^{-1}\mathbf{B}_k). \quad (7)$$

- Equivalently, by the other formula (4) for $\mathbf{P}_k$, the greedy policy maximizes

$$\log\det\big(\mathbf{P}_{k-1}^{-1} + \mathbf{B}_k^T\mathbf{N}_k^{-1}\mathbf{B}_k\big) \quad (8)$$

  at each stage. For the purpose of optimization, the log function in the objective functions above can be dropped, owing to its monotonicity.

B. Sequential Scalar Measurements

This subsection is devoted to the special case where $L = 1$ (i.e., each measurement is a scalar). Accordingly, we can write $\mathbf{A}_k = \mathbf{a}_k^T$, where $\mathbf{a}_k \in \mathbb{R}^K$. Moreover, we shall assume throughout that for all k, $\mathbf{R}_{ww} = \sigma_w^2$, and $\mathbf{R}_{nn} = \sigma_n^2\mathbf{I}_K$. Then the scalar measurement y_k is given by

$$y_k = \mathbf{a}_k^T(\mathbf{H}\mathbf{x} + \mathbf{n}_k) + w_k, \quad (9)$$

for $k = 1, \ldots, m$. The problem is to design the rows of the compression matrix $\Phi = [\mathbf{a}_1, \ldots, \mathbf{a}_m]^T$ sequentially, one at a time. In the special case $\mathbf{n}_k = \mathbf{0}$, the measurement model for the sequence of compressions may be written

$$\mathbf{y} = \Phi\mathbf{H}\mathbf{x} + \mathbf{w}, \quad (10)$$

where $\mathbf{y} \in \mathbb{R}^m$ is called the measurement vector, and $\mathbf{w}$ is a white Gaussian noise vector. In this context, the construction of a "good" compression matrix Φ to convey information about $\mathbf{x}$ is also a topic of interest. When $\mathbf{y} = \Phi\mathbf{x} + \mathbf{w}$, this is a problem of greedy adaptive noisy compressive sensing. Our solution is a more general solution than this for the more general problem (9). In this more general problem, the sequence of uncompressed measurements $\mathbf{z}_k = \mathbf{H}\mathbf{x} + \mathbf{n}_k$ is a noisy version of the filtered state $\mathbf{H}\mathbf{x}$, and compression by $\mathbf{a}_k$ introduces measurement noise w_k and colors the sensor noise $\mathbf{n}_k$.

The concept of sequential scalar measurements in a closed-loop fashion has been discussed in a number of recent papers; e.g., [4], [13]–[22]. The objective function for the optimization here can take a number of possible forms, besides the net information gain. For example, in [14], the objective is to maximize the posterior variance of the expected measurement.

If the $\mathbf{a}_k$ can only be chosen from a prescribed *finite* set, the optimal design of Φ is essentially a sensor selection problem (see [23], [24]), where the greedy policy has been shown to perform well. For example, in the problem of sensor selection with a submodular objective function subject to a uniform matroid constraint [25], the greedy policy is suboptimal with a provable bound on its performance, using bounds from optimization of submodular functions [26], [27].

Consider a constraint of the form $\|\mathbf{a}_k\| \leq 1$ for $k = 1, \ldots, m$ (where $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^K$), which is much more relaxed than a prescribed finite set. The constraint that Φ has unit-norm rows is a standard setting for compressive sensing [28]. In this special case $L = 1$, the expression in (7) simplifies to

$$\log\det\left(\mathbf{I}_N - \frac{\mathbf{P}_{k-1}\mathbf{H}^T\mathbf{a}_k\mathbf{a}_k^T\mathbf{H}}{\mathbf{a}_k^T\mathbf{H}\mathbf{P}_{k-1}\mathbf{H}^T\mathbf{a}_k + \sigma_n^2\|\mathbf{a}_k\|^2 + \sigma_w^2}\right), \quad (11)$$

which further reduces (see [29, Lemma 1.1]) to

$$\log\left(1 - \frac{\mathbf{a}_k^T\mathbf{H}\mathbf{P}_{k-1}\mathbf{H}^T\mathbf{a}_k}{\mathbf{a}_k^T\mathbf{H}\mathbf{P}_{k-1}\mathbf{H}^T\mathbf{a}_k + \sigma_n^2\|\mathbf{a}_k\|^2 + \sigma_w^2}\right). \quad (12)$$

Combining (6) and (12), the information gain at the kth step is

$$H_{k-1} - H_k = -\frac{1}{2}\log\left(1 - \frac{1}{1 + (\sigma_n^2\|\mathbf{a}_k\|^2 + \sigma_w^2)/\mathbf{a}_k^T\mathbf{H}\mathbf{P}_{k-1}\mathbf{H}^T\mathbf{a}_k}\right). \tag{13}$$

It is obvious that the greedy policy maximizes

$$\frac{\mathbf{a}_k^T\mathbf{H}\mathbf{P}_{k-1}\mathbf{H}^T\mathbf{a}_k}{\sigma_n^2\|\mathbf{a}_k\|^2 + \sigma_w^2} \tag{14}$$

to obtain the maximal information gain in the kth step.

Notice that the compression y_k may be written as

$$\begin{aligned} y_k &= \mathbf{a}_k^T\left(\mathbf{H}\hat{\mathbf{x}}_{k-1} + \mathbf{H}(\mathbf{x} - \hat{\mathbf{x}}_{k-1}) + \mathbf{n}_k\right) + w_k \\ &= \mathbf{a}_k^T\mathbf{H}\hat{\mathbf{x}}_{k-1} + \mathbf{a}_k^T\mathbf{H}(\mathbf{x} - \hat{\mathbf{x}}_{k-1}) + \mathbf{a}_k^T\mathbf{n}_k + w_k. \end{aligned} \tag{15}$$

Therefore (14) is simply the ratio of variance components: the numerator is $\mathbb{E}\left[\mathbf{a}_k^T\mathbf{H}(\mathbf{x}-\hat{\mathbf{x}}_{k-1})(\mathbf{x}-\hat{\mathbf{x}}_{k-1})^T\mathbf{H}^T\mathbf{a}_k\right]$, $\hat{\mathbf{x}}_{k-1} = \mathbb{E}[\mathbf{x}|\mathcal{I}_{k-1}]$, and the denominator is $\mathbb{E}(\mathbf{a}_k^T\mathbf{n}_k + w_k)^2$. So the goal for the greedy policy is to select $\mathbf{a}_k$ to maximize signal-to-noise ratio, where the signal is taken to be the part of the measurement y_k that is due to error $\mathbf{x} - \hat{\mathbf{x}}_{k-1}$ in the state estimate and noise is taken to be the sum of $\mathbf{a}_k^T\mathbf{n}_k$ and w_k. This is reasonable, as $\hat{\mathbf{x}}_{k-1}$ is now fixed by $\mathcal{I}_{k-1}$, and only variance components can be controlled by the measurement vector $\mathbf{a}_k$.

The greedy policy can be described succinctly in terms of certain eigenvectors, as follows. Denote the eigenvalues of $\mathbf{D}_k := \mathbf{H}\mathbf{P}_k\mathbf{H}^T$ by $\lambda_1^{(k)} \geq \lambda_2^{(k)} \geq \cdots \geq \lambda_N^{(k)} \geq \lambda_{N+1}^{(k)} = \cdots = \lambda_K^{(k)} = 0$. For simplicity, when $k = 0$ we may omit the superscript and write $\lambda_i := \lambda_i^{(0)}$ for $i = 1, \ldots, K$. Since $\mathbf{P}_0$ is a covariance matrix, which is symmetric, and $\mathbf{D}_0$ is also symmetric, there exist corresponding orthonormal eigenvectors $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_K\}$. Clearly,

$$\frac{\mathbf{a}_1^T\mathbf{D}_0\mathbf{a}_1}{\sigma_n^2\|\mathbf{a}_1\|^2 + \sigma_w^2} \leq \frac{\lambda_1\|\mathbf{a}_1\|^2}{\sigma_n^2\|\mathbf{a}_1\|^2 + \sigma_w^2} \leq \frac{\lambda_1}{\sigma_n^2 + \sigma_w^2}. \tag{16}$$

The equalities hold when $\mathbf{a}_1$ equals $\mathbf{v}_1$, which is the eigenvector of $\mathbf{D}_0$ corresponding to its largest eigenvalue λ_1; we take this to be what the greedy policy picks. If eigenvalues are repeated, we simply pick the eigenvector with smallest index i. After picking $\mathbf{a}_1 = \mathbf{v}_1$, by (3) we have

$$\mathbf{P}_1 = \mathbf{P}_0 - \frac{\mathbf{P}_0\mathbf{H}^T\mathbf{v}_1\mathbf{v}_1^T\mathbf{H}\mathbf{P}_0}{\sigma^2 + \lambda_1} \tag{17}$$

where $\sigma^2 := \sigma_n^2 + \sigma_w^2$. We can verify the following:

$$\begin{aligned} \mathbf{D}_1\mathbf{v}_i &= \left(\mathbf{H}\big(\mathbf{P}_0 - \frac{\mathbf{P}_0\mathbf{H}^T\mathbf{v}_1\mathbf{v}_1^T\mathbf{H}\mathbf{P}_0}{\sigma^2 + \lambda_1}\big)\mathbf{H}^T\right)\mathbf{v}_i \\ &= \left(\mathbf{D}_0 - \frac{\mathbf{D}_0\mathbf{v}_1\mathbf{v}_1^T\mathbf{D}_0}{\sigma^2 + \lambda_1}\right)\mathbf{v}_i \\ &= \left(\mathbf{D}_0 - \frac{\lambda_1^2\mathbf{v}_1\mathbf{v}_1^T}{\sigma^2 + \lambda_1}\right)\mathbf{v}_i \\ &= \lambda_i\mathbf{v}_i, \quad \text{for } i = 2\ldots, K \end{aligned} \tag{18}$$

and

$$\begin{aligned} \mathbf{D}_1\mathbf{v}_1 &= \left(\mathbf{D}_0 - \frac{\lambda_1^2\mathbf{v}_1\mathbf{v}_1^T}{\sigma^2 + \lambda_1}\right)\mathbf{v}_1 \\ &= \left(\frac{1}{\lambda_1} + \frac{1}{\sigma^2}\right)^{-1}\mathbf{v}_1. \end{aligned} \tag{19}$$

So $\mathbf{D}_1$ turns out to have the same collection of eigenvectors as $\mathbf{D}_0$, and the nonzero eigenvalues of $\mathbf{D}_1$ are $(1/\lambda_1 + 1/\sigma^2)^{-1}$, $\lambda_2, \ldots, \lambda_N$. By induction, we conclude that, when applying the greedy policy, all the $\mathbf{D}_k$s for $k = 0, \ldots, m$ have the same collection of eigenvectors and the greedy policy always picks the compressors $\mathbf{a}_k$, $k = 1, \ldots, m$, from the set of eigenvectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$. The implication is that this basis for the invariant subspace $\langle\mathbf{V}\rangle$ for the prior measurement covariance $\mathbf{D}_0$ may be used to define a prescribed finite set of compression vectors from which compressors are to be drawn. The greedy policy then amounts to selecting the compressor $\mathbf{a}_k$ to be the eigenvector of $\mathbf{D}_k$ with eigenvalue $\lambda_1^{(k)}$. In other words, the greedy policy simply re-sorts the eigenvectors of $\mathbf{D}_0$, step-by-step, and selects the one with maximum eigenvalue.

Consequently, after applying m iterations of the greedy policy, the net information gain is

$$\begin{aligned} H_0 - H_m &= \sum_{k=1}^{m}\max_{\|\mathbf{a}_k\|\leq 1}(H_{k-1} - H_k) \\ &= -\frac{1}{2}\sum_{k=1}^{m}\log\left(\frac{\sigma^2}{\lambda_1^{(k-1)} + \sigma^2}\right) \\ &= \frac{1}{2}\log\prod_{k=1}^{m}\left(1 + \frac{\lambda_1^{(k-1)}}{\sigma^2}\right) \end{aligned} \tag{20}$$

where $\lambda_1^{(k-1)}$, the largest eigenvalue of $\mathbf{D}_{k-1}$, is computed iteratively from the sequence $\mathbf{P}_0, \ldots, \mathbf{P}_{k-1}$.

One of the key results of our paper is that once the eigenvectors of $\mathbf{H}\mathbf{P}_0\mathbf{H}$ are computed, then the greedy policy simply selects its next compression vector from this set, revisiting eigenvectors according to the greedy policy of the paper. Now, for nonsingular $\mathbf{P}_0$, this set of eigenvectors will be a basis for $\mathbb{R}^n$, and therefore the optimal greedy policy for an alternative prior covariance, call it $\mathbf{Q}_0$, would use a linear combination of these eigenvectors for its greedy policy. So there is a question of mismatch between the eigenvectors used in the greedy policy for an assumed $\mathbf{P}_0$ and those that would be used for the actual $\mathbf{Q}_0$. But qualitatively we expect this mismatch to lead to a requirement for more measurements for a target value of $\det(\mathbf{Q}_n)$ than would be required in the case of no mismatch. This mismatch analysis is a very interesting problem of the type that is discussed on related work [30].

C. Example of the Greedy Policy

Suppose that the uncompressed measurements are $\mathbf{z}_k = \mathbf{x} + \mathbf{n}_k$, $k = 1, \ldots, m$, with $\mathbf{P}_0 = \lambda\mathbf{I}_N$, indicating no prior indication of shape for the error covariance matrix. In the model $y_k = \mathbf{a}_k^T\mathbf{z}_k + w_k = \mathbf{a}_k^T(\mathbf{x} + \mathbf{n}_k) + w_k$, $k = 1\ldots, m$, assume that w_k has distribution $\mathcal{N}(0, \sigma_w^2)$ and $\mathbf{n}_k$ has distribution $\mathcal{N}(\mathbf{0}, \sigma_n^2\mathbf{I}_N)$. The choice of orthonormal eigenvectors for

$\mathbf{D}_0 = \mathbf{P}_0$ is arbitrary, with $\mathbf{V} = \mathbf{E} = [\mathbf{e}_1, \ldots, \mathbf{e}_N]$ (the standard basis for $\mathbb{R}^N$) a particular choice that minimizes the complexity of compression. So compressed measurements will consist of the noisy measurements $y_k = \mathbf{e}_{(k)}^T \mathbf{z}_k + w_k$, where $\mathbf{e}_{(k)}^T$ is the k-th choice of compressor from $\mathbf{E}$.

After picking $\mathbf{a}_1 = \mathbf{e}_1$, the eigenvalues of $\mathbf{P}_1$ are $\lambda_1^{(1)} = \cdots = \lambda_{N-1}^{(1)} = \lambda$, $\lambda_N^{(1)} = (\frac{1}{\lambda} + \frac{1}{\sigma^2})^{-1}$. Analogously, after picking $\mathbf{a}_2 = \mathbf{e}_2$, the eigenvalues of $\mathbf{P}_2$ are $\lambda_1^{(2)} = \cdots = \lambda_{N-2}^{(2)} = \lambda$, $\lambda_{N-1}^{(2)} = \lambda_N^{(2)} = (\frac{1}{\lambda} + \frac{1}{\sigma^2})^{-1}$, and so on. If $m \leq N$, then after m iterations of the greedy policy the eigenvalues of $\mathbf{D}_m$ are $\lambda_1^{(m)} = \cdots = \lambda_{N-m}^{(m)} = \lambda$, $\lambda_{N-m+1}^{(m)} = \cdots = \lambda_N^{(m)} = (\frac{1}{\lambda} + \frac{1}{\sigma^2})^{-1}$. In the first m iterations, the per-stage information gain is $\frac{1}{2}\log(1 + \lambda/\sigma^2)$.

If $m > N$, after N iterations of the greedy policy, $\lambda_1^{(N)} = \cdots = \lambda_N^{(N)} = (\frac{1}{\lambda} + \frac{1}{\sigma^2})^{-1}$. We now simply encounter a similar situation as in the very beginning. We update $\lambda \leftarrow (\frac{1}{\lambda} + \frac{1}{\sigma^2})^{-1}$ and $m \leftarrow (m - N)$. The analysis above then applies again, leading to a round-robin selection of measurements.

III. Optimal Policy and Relaxed Optimal Policy

A. Optimal Policy

We now turn to the problem of maximizing the net information gain, subject to the unit-norm constraint:

$$\begin{aligned} &\text{maximize } \sum_{k=1}^{m}(H_{k-1} - H_k), \\ &\text{subject to } \|\mathbf{a}_k\| \leq 1, \quad k = 1, \ldots, m. \end{aligned} \tag{21}$$

The policy that maximizes (21) is called the *optimal policy*. The objective function can be written as

$$\begin{aligned} &\sum_{k=1}^{m}(H_{k-1} - H_k) \\ &\quad = -\frac{1}{2}\sum_{k=1}^{m}\log\frac{\det(\mathbf{P}_k)}{\det(\mathbf{P}_{k-1})} \\ &\quad = \frac{1}{2}\log\frac{\det(\mathbf{P}_0)}{\det(\mathbf{P}_m)} \\ &\quad = \frac{1}{2}\log\det(\mathbf{P}_0)\det\Big(\mathbf{P}_0^{-1} + \sum_{k=1}^{m}\frac{\mathbf{H}^T\mathbf{a}_k\mathbf{a}_k^T\mathbf{H}}{\|\mathbf{a}_k\|^2\sigma_n^2 + \sigma_w^2}\Big) \\ &\quad = \frac{1}{2}\log\det(\mathbf{I}_m + \mathbf{C}^T\mathbf{D}_0\mathbf{C}) \end{aligned} \tag{22}$$

where

$$\begin{aligned} \mathbf{C} &:= [\mathbf{c}_1, \ldots, \mathbf{c}_m] \\ &:= \left[\frac{\mathbf{a}_1}{\sqrt{\|\mathbf{a}_1\|^2\sigma_n^2 + \sigma_w^2}}, \ldots, \frac{\mathbf{a}_m}{\sqrt{\|\mathbf{a}_m\|^2\sigma_n^2 + \sigma_w^2}}\right]. \end{aligned} \tag{23}$$

Recall that the eigenvalue decomposition $\mathbf{D}_0 = \mathbf{V}\Lambda\mathbf{V}^T$, where $\Lambda = \mathrm{Diag}(\lambda_1, \lambda_2, \ldots, \lambda_K)$ and $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_K]$. (The notation $\mathrm{Diag}(\lambda_1, \lambda_2, \ldots, \lambda_K)$ means the diagonal matrix with diagonal entries $\lambda_1, \ldots, \lambda_K$.) Then, continuing from (22),

$$\begin{aligned} &\sum_{k=1}^{m}(H_{k-1} - H_k) \\ &\quad = \frac{1}{2}\log\det\Big(\mathbf{I}_m + \mathbf{C}^T\mathbf{V}\Lambda\mathbf{V}^T\mathbf{C}\Big) \\ &\quad = \frac{1}{2}\log\det\Big(\mathbf{I}_m + \mathbf{G}^T\Lambda\mathbf{G}\Big) \end{aligned} \tag{24}$$

where

$$\mathbf{G} := [\mathbf{g}_1, \ldots, \mathbf{g}_m] := \mathbf{V}^T\mathbf{C}. \tag{25}$$

Since $\mathbf{V}$ is nonsingular, the map $\mathbf{c}_k \mapsto \mathbf{g}_k = \mathbf{V}^T\mathbf{c}_k$ is one-to-one.

The constraint $\|\mathbf{a}_k\| \leq 1$ implies that $\|\mathbf{g}_k\|^2 = \|\mathbf{c}_k\|^2 \leq \sigma^{-2}$. Conversely, $\|\mathbf{g}_k\|^2 = \|\mathbf{c}_k\|^2 = \frac{\|\mathbf{a}_k\|^2}{\|\mathbf{a}_k\|^2\sigma_n^2 + \sigma_w^2} \leq \sigma^{-2}$ also implies $\|\mathbf{a}_k\| \leq 1$, as $\frac{\|\mathbf{a}_k\|^2}{\|\mathbf{a}_k\|^2\sigma_n^2 + \sigma_w^2}$ is a monotonically increasing function in terms of $\|\mathbf{a}_k\|^2$. So the constraint in (21) can be written as $(\mathbf{G}^T\mathbf{G})_{ii} \leq \sigma^{-2}$ for $i = 1, \ldots, m$.

B. Relaxed Optimal Policy

To help characterize the optimal policy (solution to (21)), we now consider an alternative optimization problem with the same objective function in (21) but a relaxed constraint:

$$\begin{aligned} &\text{maximize } \sum_{k=1}^{m}(H_{k-1} - H_k), \\ &\text{subject to } \frac{1}{m}\sum_{k=1}^{m}\|\mathbf{a}_k\| \leq 1, \end{aligned} \tag{26}$$

i.e., the rows of Φ have *average* unit norm. We will call the policy that maximizes (26) the *relaxed optimal policy*.

The average norm constraint in (26) is equivalent to $\mathrm{tr}\,\mathbf{G}^T\mathbf{G} = \sum_{k=1}^{m}\|\mathbf{g}_k\|^2 \leq \sigma^{-2}m$. With the scaling

$$\widetilde{\mathbf{G}} := \sigma m^{-1/2}\mathbf{G}, \tag{27}$$

the constraint $\mathrm{tr}\,\mathbf{G}^T\mathbf{G} \leq \sigma^{-2}m$ becomes $\mathrm{tr}\,\widetilde{\mathbf{G}}^T\widetilde{\mathbf{G}} \leq 1$. Hence, the relaxed optimization problem (26) is equivalent to

$$\begin{aligned} &\text{maximize } \frac{1}{2}\log\det(\mathbf{I}_m + \widetilde{\mathbf{G}}^T\widetilde{\Lambda}\widetilde{\mathbf{G}}), \\ &\text{subject to } \mathrm{tr}\,\widetilde{\mathbf{G}}^T\widetilde{\mathbf{G}} \leq 1 \end{aligned} \tag{28}$$

where $\widetilde{\Lambda} = \mathrm{Diag}(\Lambda_1, \ldots, \Lambda_N)$ and $\Lambda_i := \frac{m\lambda_i}{\sigma^2}$, for $i = 1, \ldots, N$.

To solve (28), we need the following known results from [31].

Lemma 1: Given any $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_q > 0$, there exists a unique integer r, with $1 \leq r \leq q$, such that for $1 \leq k \leq r$ we have

$$\frac{1}{\lambda_k} < \frac{1}{k}\Big(1 + \sum_{j=1}^{k}\frac{1}{\lambda_j}\Big), \tag{29}$$

while for indices k, if any, satisfying $r < k \leq q$ we have

$$\frac{1}{\lambda_k} \geq \frac{1}{k}\Big(1 + \sum_{j=1}^{k}\frac{1}{\lambda_j}\Big). \tag{30}$$

Lemma 2: For $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_q > 0$ and r as in Lemma 1, the sequence

$$M_k = \Big(\frac{1}{k} + \frac{1}{k}\sum_{j=1}^{k}\frac{1}{\lambda_j}\Big)^k \prod_{i=1}^{k}\lambda_i, \qquad k = 1, \ldots, q, \quad (31)$$

is strictly increasing.

By [31, Theorem 2], the optimal value of the relaxed maximization problem (28) is

$$\frac{1}{2}\log\Big(\Big(\frac{1}{r} + \frac{1}{r}\sum_{j=1}^{r}\frac{1}{\Lambda_j}\Big)^r \prod_{i=1}^{r}\Lambda_i\Big) = \frac{1}{2}\log\Big(\prod_{i=1}^{r}\Big(\frac{\Lambda_i}{r} + \frac{1}{r}\sum_{j=1}^{r}\frac{\Lambda_i}{\Lambda_j}\Big)\Big) \quad (32)$$

where r is defined by Lemma 1. Specifically, r is defined by the largest eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_q$ of $\mathbf{D}_0$, where in our case we set $q := \min(m, N)$.

Indeed the optimal value (32) may also be derived from the solution to the well-known water-filling problem (see [32] for details). It is known from [31] that the optimal value of the maximization problem

$$\text{maximize } \prod_{i=1}^{q}(1 + \Lambda_i p_i), \quad \text{subject to } \sum_{i=1}^{q} p_i \leq 1, \quad (33)$$

is

$$\prod_{i=1}^{r}\Big(\frac{\Lambda_i}{r} + \frac{1}{r}\sum_{j=1}^{r}\frac{\Lambda_i}{\Lambda_j}\Big). \quad (34)$$

This optimal value is only obtained when

$$p_i = \Big(\mu - \frac{1}{\Lambda_i}\Big)^+, \quad i = 1, 2, \ldots, q, \quad (35)$$

where

$$\mu := \frac{1}{r}\Big(1 + \sum_{i=1}^{r}\frac{1}{\Lambda_i}\Big) \quad (36)$$

is called the *water level*. By taking a close look at (35), we know that $p_1 \geq \ldots \geq p_r > 0$ and $p_{r+1} = \ldots = p_q = 0$. Fig. 1 illustrates the relation among Λ_i, p_i, and water level μ.

With the values of p_i defined in (35), we can determine the $\widetilde{\mathbf{G}}$ that solves the maximization problem (28). The optimal $\widetilde{\mathbf{G}}$ is obtained for, and only for, the following two cases. Let $\mathbf{G}_0$ be the $K \times m$ matrix with $(\mathbf{G}_0)_{ii} = \sqrt{p_i}$, $i = 1, \ldots, r$, and all other elements zero.

- Case 1. $\lambda_r > \lambda_{r+1}$ for $r = N$. Then $\widetilde{\mathbf{G}} = \mathbf{G}_0\mathbf{U}$ where $\mathbf{U}$ is any $m \times m$ orthonormal matrix.
- Case 2. $\lambda_i = \lambda_r$ for and only for $r - \alpha < i \leq r + \beta$ with $\alpha \geq 1$, $\beta \geq 1$. Then $\widetilde{\mathbf{G}} = \text{block-Diag}(\mathbf{I}_{r-\alpha}, \mathbf{U}_2, \mathbf{I}_{K-r-\beta})\mathbf{G}_0\mathbf{U}_1$ where $\mathbf{U}_1$ is any $m \times m$ orthonormal matrix and $\mathbf{U}_2$ any $(\alpha+\beta)\times(\alpha+\beta)$ orthonormal matrix. This case is only possible when $r = q = m < N$. (The notation block-Diag$(\mathbf{I}_{r-\alpha}, \mathbf{U}_2, \mathbf{I}_{K-r-\beta})$ denotes a block diagonal matrix with diagonal blocks $\mathbf{I}_{r-\alpha}$, $\mathbf{U}_2$, and $\mathbf{I}_{K-r-\beta}$.)

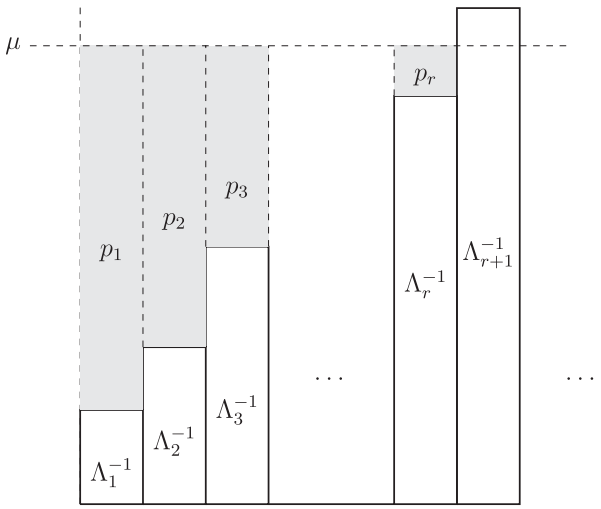

Fig. 1. Water-filling solution.

After obtaining $\widetilde{\mathbf{G}}$, we can extract the optimal solution $\Phi = [\mathbf{a}_1, \ldots, \mathbf{a}_m]^T$ for the relaxed constraint problem (26) by using (27), (25), and (23).

Our main motivation to relax the constraint to an *average* norm constraint is our knowledge of the relaxed optimal solution. Particularly, for the multivariate Gaussian signal $\mathbf{x}$ the maximal net information gain under the relaxed constraint is given by the water-filling solution. This helps us to identify cases where the greedy policy is in fact optimal, as discussed in the next section.

IV. When Greedy Is Optimal

In the preceding sections, we have discussed three types of policies: the greedy policy, the optimal policy, and the relaxed optimal policy. Denote by H_G, H_O, and H_R the net information gains associated with these three policies respectively. Clearly,

$$H_G \leq H_O \leq H_R. \quad (37)$$

In the rest of this section, we characterize H_G, H_O, and H_R. In general, we do not expect to have $H_G = H_O$; in other words, in general, greedy is not optimal. However, it is interesting to explore cases where greedy *is* optimal. In the rest of this section, we provide sufficient conditions for the greedy policy to be optimal.

Before proceeding, we make the following observation on the net information gain. In (28) denote $\widetilde{\Gamma} := \widetilde{\mathbf{G}}\widetilde{\mathbf{G}}^T$; then the determinant in the objective function becomes

$$\det(\mathbf{I}_m + \widetilde{\mathbf{G}}^T\widetilde{\Lambda}\widetilde{\mathbf{G}}) = \det(\mathbf{I}_K + \widetilde{\Lambda}\widetilde{\Gamma}). \quad (38)$$

Under the unit-norm constraint,

$$\widetilde{\Gamma} = \frac{\sigma^2}{m}\mathbf{G}\mathbf{G}^T = \frac{\sigma^2}{m}\Big(\sum_{i=1}^{m}\mathbf{g}_i\mathbf{g}_i^T\Big) = \frac{1}{m}\mathbf{V}^T\Big(\sum_{i=1}^{m}\mathbf{a}_i\mathbf{a}_i^T\Big)\mathbf{V}. \quad (39)$$

Remark 3: In the maximization problem (21), if the $\mathbf{a}_k$s were only picked from $\{\mathbf{v}_1, \ldots, \mathbf{v}_K\}$, by (39) $\widetilde{\Gamma} = \mathrm{Diag}(\gamma_1, \ldots, \gamma_K)$ where each γ_i is an integer multiple of $1/m$ and $\sum_{k=1}^{K} \gamma_k = 1$. This integer γ_i would be determined by the multiplicity of appearances of $\mathbf{v}_i$ among $\mathbf{a}_1, \ldots, \mathbf{a}_m$. Thus the net information gain would be

$$\frac{1}{2}\log\det(\mathbf{I}_K + \widetilde{\Lambda}\widetilde{\Gamma}) = \frac{1}{2}\log\prod_{i=1}^{K}(1+\Lambda_i\gamma_i) = \frac{1}{2}\log\prod_{i=1}^{N}(1+\Lambda_i\gamma_i) \quad (40)$$

where we use the fact that $\Lambda_{N+1} = \cdots = \Lambda_K = 0$. Clearly, to maximize the net information gain by selecting compressors from $\{\mathbf{v}_1, \ldots, \mathbf{v}_K\}$, we should never pick $\mathbf{a}_k$ from $\{\mathbf{v}_{N+1}, \ldots, \mathbf{v}_K\}$, because (40) is not a function of $\gamma_{N+1}, \ldots, \gamma_K$. In particular, the greedy policy picks $\mathbf{a}_k$ from $\{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$. After m iterations of the greedy policy, the net information gain can be computed by the right hand side of (40).

We now provide two sufficient conditions (in Theorems 4 and 5) under which $H_G = H_O$ holds for the sequential scalar measurements problem (9).

Theorem 4: Suppose that the optimal $\mathbf{a}_k$, $k = 1, \ldots, m$, are constrained to be picked from the prescribed set $S \subseteq \{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$, which is a subset of the orthonormal eigenvectors of $\mathbf{D}_0$. If $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\} \subseteq S$, then the greedy policy is optimal, i.e., $H_G = H_O$.

Proof: See Appendix A. ∎

Next, assume that we can pick $\mathbf{a}_k$ to be any arbitrary vector with unit norm. In this much more complicated situation, we establish when $H_G = H_O$ by directly showing that $H_G = H_R$, which implies that $H_G = H_O$ in light of (37).

Theorem 5: Assume that $\mathbf{a}_k$, $k = 1, \ldots, m$, can be selected to be any vector with $\|\mathbf{a}_k\| \leq 1$. If $1/\lambda_k - 1/\lambda_{k+1} = n_k/\sigma^2$, where n_k is some nonnegative integer, for $k = 1, \ldots, r-1$, and r divides $(m - \sum_{k=1}^{r-1} k n_k)$, then the greedy policy is optimal, i.e., $H_G = H_O$.

Proof: See Appendix B. ∎

The two theorems above furnish conditions under which greedy is optimal. However, these conditions are quite restrictive. Indeed, as pointed out earlier, in general the greedy policy is not optimal. The restrictiveness of the sufficient conditions above help to highlight this fact. In the next section, we provide examples of cases where greedy is *not* optimal.

V. When Greedy Is Not Optimal

A. An Example With Non-Scalar Measurements

In this subsection we give an example where the greedy policy is not optimal for the scenario $\mathbf{z} = \mathbf{x}$ and $\mathbf{y}_k = \mathbf{A}_k\mathbf{x} + \mathbf{w}_k$. Suppose that we are restricted to a set of only three choices for $\mathbf{A}_k$:

$$\mathcal{A} = \left\{\mathrm{Diag}(1,0), \mathrm{Diag}(0,1), \frac{1}{2}\mathrm{Diag}(1,1)\right\}.$$

In this case, $L = N = 2$. Moreover, set $m = 2$, $\mathbf{P}_0 = 16\mathbf{I}_2$, and $\mathbf{R}_{ww} = \mathbf{I}_2$.

Let us see what the greedy policy would do in this case. For $k = 1$, it would pick $\mathbf{A}_1$ to maximize

$$\det\left(\frac{1}{16}\mathbf{I}_2 + (\mathbf{A}_1)^2\right).$$

A quick calculation shows that for $\mathbf{A}_1 = \mathrm{Diag}(1,0)$ or $\mathrm{Diag}(0,1)$, we have

$$\det\left(\frac{1}{16}\mathbf{I}_2 + (\mathbf{A}_1)^2\right) = \frac{17}{256},$$

whereas for $\mathbf{A}_1 = \frac{1}{2}\mathrm{Diag}(1,1)$,

$$\det\left(\frac{1}{16}\mathbf{I}_2 + (\mathbf{A}_1)^2\right) = \frac{25}{256},$$

So the greedy policy picks $\mathbf{A}_1 = \frac{1}{2}\mathrm{Diag}(1,1)$, which leads to $\mathbf{P}_1 = \frac{16}{5}\mathbf{I}_2$.

For $k = 2$, we go through the same calculations: for $\mathbf{A}_2 = \mathrm{Diag}(1,0)$ or $\mathrm{Diag}(0,1)$, we have

$$\det\left(\frac{5}{16}\mathbf{I}_2 + (\mathbf{A}_2)^2\right) = \frac{105}{256}$$

whereas for $\mathbf{A}_2 = \frac{1}{2}\mathrm{Diag}(1,1)$,

$$\det\left(\frac{5}{16}\mathbf{I}_2 + (\mathbf{A}_2)^2\right) = \frac{81}{256}.$$

So, this time the greedy policy picks $\mathbf{A}_2 = \mathrm{Diag}(1,0)$ (or $\mathrm{Diag}(0,1)$), after which $\det(\mathbf{P}_2) = 256/105$.

Consider the alternative policy that picks $\mathbf{A}_1 = \mathrm{Diag}(1,0)$ and $\mathbf{A}_2 = \mathrm{Diag}(0,1)$. In this case,

$$\mathbf{P}_2^{-1} = \frac{1}{16}\mathbf{I}_2 + \mathrm{Diag}(1,0) + \mathrm{Diag}(0,1) = \frac{17}{16}\mathbf{I}_2 \quad (41)$$

and so $\det(\mathbf{P}_2) = 256/289$, which is clearly provides greater net information gain than the greedy policy. Call this alternative policy the *alternating policy* (because it alternates between $\mathrm{Diag}(1,0)$ and $\mathrm{Diag}(0,1)$).

In conclusion, for this example the greedy policy is not optimal with respect to the objective of maximizing the net information gain. How much worse is the objective function of the greedy policy relative to that of the optimal policy? On the face of it, this question seems related to the well-known fact that the net information gain is a submodular function. As mentioned before, in this case we would expect to be able to bound the suboptimality of the greedy policy compared to the optimal policy (though we do not explicitly do that here).

Nonetheless, it is worthwhile exploring this question a little further. Suppose that we set $\mathbf{P}_0 = \alpha^{-1}\mathbf{I}_2$ and let the third choice in $\mathcal{A}$ be $\alpha^{1/4}\mathbf{I}_2$, where $\alpha > 0$ is some small number. (Note that the numerical example above is a special case with $\alpha = 1/16$.) In this case, it is straightforward to check that the greedy policy picks $\mathbf{A}_1 = \alpha^{1/4}\mathbf{I}_2$ and $\mathbf{A}_2 = \mathrm{Diag}(1,0)$ (or $\mathrm{Diag}(0,1)$) if α is sufficiently small, resulting in

$$\det(\mathbf{P}_2) = \frac{1}{\sqrt{\alpha}(1+\sqrt{\alpha})(1+\sqrt{\alpha}+\alpha)},$$

which increases unboundedly as $\alpha \to 0$. However, the alternating policy results in

$$\det(\mathbf{P}_2) = \frac{1}{(1+\alpha)^2},$$

which converges to 1 as $\alpha \to 0$. Hence, letting α get arbitrarily small, the ratio of $\det(\mathbf{P}_2)$ for the greedy policy to that of the alternating policy can be made arbitrarily large. Insofar as we accept minimizing $\det(\mathbf{P}_2)$ to be an equivalent objective to maximizing the net information gain (which differs by the normalizing factor $\det(\mathbf{P}_0)$ and taking log), this means that *the greedy policy is arbitrarily worse than the alternating policy*.

What went wrong? The greedy policy was "fooled" into picking $\mathbf{A}_1 = \alpha^{1/4}\mathbf{I}_2$ at the first stage, because this choice maximizes the per-stage information gain in the first stage. But once it does that, it is stuck with its resulting covariance matrix $\mathbf{P}_1$. The alternating policy trades off the per-stage information gain in the first stage for the sake of better net information gain over two stages. The first measurement matrix $\mathrm{Diag}(1, 0)$ "sets up" the covariance matrix $\mathbf{P}_1$ so that the second measurement matrix $\mathrm{Diag}(0, 1)$ can take advantage of it to obtain a superior covariance matrix $\mathbf{P}_2$ after the second stage, embodying a form of "delayed gratification."

Interestingly, the argument above depends on the value of α being sufficiently small. For example, if $\alpha = 0.347809$, then the greedy policy has the same net information gain as the alternating policy, and is in fact optimal.

An interesting observation to be made here is that the submodularity of the net information gain as an objective function depends crucially on including the log function. In other words, although for the purpose of optimization we can dispense with the log function in the objective function in view of its monotonicity, bounding the suboptimality of the greedy policy with respect to the optimal policy turns on submodularity, which relies on the presence of the log function in the objective function. In particular, if we adopt the volume of the error concentration ellipse as an equivalent objective function, we can no longer bound the suboptimality of the greedy policy relative to the optimal policy—the greedy policy is provably *arbitrarily worse* in some scenarios, as our example above shows.

B. An Example With Scalar Measurements

Consider the channel model $\mathbf{z}_k = \mathbf{x} + \mathbf{n}_k$ and scalar measurements

$$y_k = \mathbf{a}_k^T \mathbf{z}_k + w_k. \tag{42}$$

Assume that

$$\mathbf{P}_0 = \begin{bmatrix} 3 & 2 \\ 2 & 3 \end{bmatrix},$$

$\mathbf{R}_{ww} = 0.5$, $\mathbf{R}_{nn} = 0.5\mathbf{I}_2$ and set $m = 2$. Our goal is to find $\|\mathbf{a}_1\|, \|\mathbf{a}_2\| \le 1$ such that $\mathbf{a}_1$, $\mathbf{a}_2$ maximize the net information gain:

$$H_0 - H_2 = \frac{1}{2}\log\det(\mathbf{P}_0)\det(\mathbf{P}_0^{-1} + \mathbf{a}_1\mathbf{a}_1^T + \mathbf{a}_2\mathbf{a}_2^T). \tag{43}$$

By simple computation, we know that the eigenvalues of $\mathbf{P}_0$ are $\lambda_1^{(0)} = 5$ and $\lambda_2^{(0)} = 1$. If we follow the greedy policy, the eigenvalues of $\mathbf{P}_1$ are $\lambda_1^{(1)} = 1$ and $\lambda_2^{(1)} = 5/6$. By (20), the net information gain for the greedy policy is

$$H_0 - H_2 = \frac{1}{2}\log(1+5)(1+1) = \frac{1}{2}\log(12).$$

We are now in a position to solve for the optimal solution. Let $\mathbf{a}_1 = [a_1, a_2]^T$. By (4), we have

$$\mathbf{P}_1 = \begin{bmatrix} \dfrac{5a_2^2+3}{3a_1^2+4a_1a_2+3a_2^2+1} & \dfrac{-(5a_1a_2-2)}{3a_1^2+4a_1a_2+3a_2^2+1} \\ \dfrac{-(5a_1a_2-2)}{3a_1^2+4a_1a_2+3a_2^2+1} & \dfrac{5a_1^2+3}{3a_1^2+4a_1a_2+3a_2^2+1} \end{bmatrix}.$$

We compute that

$$\lambda_1^{(1)} = \frac{(25a_1^4+50a_1^2a_2^2-80a_1a_2+25a_2^4+16)^{1/2}}{6a_1^2+8a_1a_2+6a_2^2+2} + \frac{5a_1^2+5a_2^2+6}{6a_1^2+8a_1a_2+6a_2^2+2}. \tag{44}$$

When we choose $\mathbf{a}_2$ in the second stage, we can simply maximize the information gain in that stage. In this special case when $m = 2$, the second stage is actually the last one. If $\mathbf{a}_1$ is given, maximizing the net information gain is identical to maximizing the information gain in the second stage. Therefore, the second step is essentially a greedy step. By (20),

$$H_1 - H_2 = -\frac{1}{2}\log\left(1 - \frac{1}{1+1/\lambda_1^{(1)}}\right) = \frac{1}{2}\log(1+\lambda_1^{(1)}). \tag{45}$$

By (13), we know

$$H_0 - H_1 = -\frac{1}{2}\log\det\left(\mathbf{I}_2 - \frac{\mathbf{P}_0\mathbf{a}_1\mathbf{a}_1^T}{\mathbf{a}_1^T\mathbf{P}_0\mathbf{a}_1+1}\right) = \frac{1}{2}\log(4+4a_1a_2). \tag{46}$$

Using $\|\mathbf{a}_1\| = 1$, we simplify (45) and (46) to obtain

$$H_0 - H_2 = \frac{1}{2}\log\Big(\frac{1}{2}\big((41-80a_1a_2)^{1/2} + 19 + 8a_1a_2\big)\Big). \tag{47}$$

This expression reaches its maximal value when $a_1a_2 = 1/5$. Consequently, the optimal net information gain is $\frac{1}{2}\log(12.8)$, when

$$\mathbf{a}_1 = \left[\left(\frac{-\sqrt{21}+5}{10}\right)^{1/2}, \left(\frac{\sqrt{21}+5}{10}\right)^{1/2}\right]^T$$

and

$$\mathbf{a}_2 = \left[\left(\frac{\sqrt{21}+5}{10}\right)^{1/2}, \left(\frac{-\sqrt{21}+5}{10}\right)^{1/2}\right]^T.$$

This indicates that the greedy policy is not optimal.

For illustration, we also consider the cases $m = 3, \ldots, 6$. Fig. 2 shows the information gains obtained by greedy policy and optimal policy for $m = 2, \ldots, 6$, where we can see the difference between the optimal policy and greedy policy is very close, when sensor and measurement noises are white.

In order to verify the claim that optimal policy and greedy policy are very close, for the same model as in (42) we

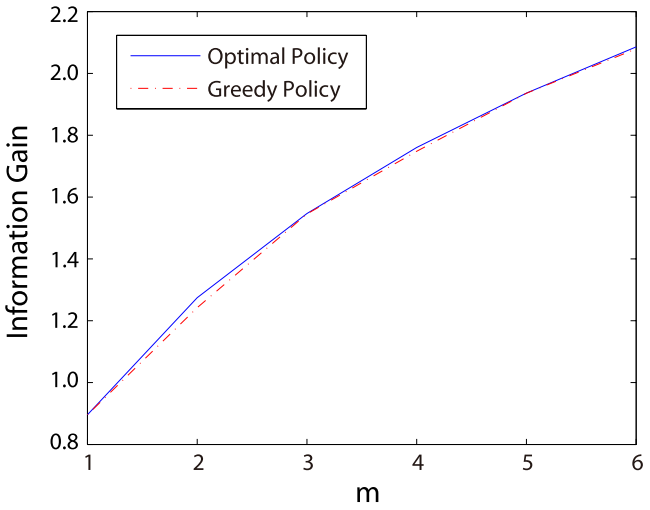

Fig. 2. Information gains using optimal policy and greedy policy for different m.

TABLE I
RELATIVE DIFFERENCE BETWEEN THE INFORMATION GAINS OBTAINED BY OPTIMAL POLICY AND GREEDY POLICY FOR DIFFERENT m

m	2	3	4	5	6
average of $(H_O - H_G)/H_O$	0.05%	1.88%	0.01%	0.61%	0.01%

generate 100 realizations of $\mathbf{P}_0$ whose eigenvalues are uniformly distributed on $[2, 4]$, $\mathbf{R}_{ww} = 0.5$, and $\mathbf{R}_{nn} = 0.5\mathbf{I}_2$. Then compute the information gains using optimal policy and greedy policy. Table I shows the average relative difference between optimal polices and greedy polices, i.e. the average of $(H_O - H_G)/H_O$.

VI. CONCLUSION

We have investigated adaptive compression policies for compressing a sequence of transformed and noisy measurements of a Gaussian signal into a sequence of scalar compression variables. The optimization criterion is to maximize the net information gain (mutual information) between the signal and its measurements. Under a relaxed constraint on the average norm of compression vectors, the optimal policy is a water-filling policy. The optimal *greedy* policy, which maximizes information gain at each compression under a unit-norm constraint, draws its vector-valued compression vectors from the eigenvectors of the prior error covariance matrix for the signal to be estimated, with the draws determined by the eigenvalues of its posterior error covariance matrix. These eigenvalues are computed and sorted recursively. Unsurprisingly, the greedy policy, which maximizes the per-stage information gain, is not optimal in general. However, we provide sufficient conditions under which the greedy policy is actually optimal. In the case of scalar compressions, our examples and simulation results illustrate that the greedy policy can be quite close to optimal when the noise sequences are white.

APPENDIX A
PROOF OF THEOREM 4

If $\mathbf{a}_k$, $k = 1, \ldots, m$, can only be picked from $\{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$, then by (40) the net information gain is $\frac{1}{2}\log\prod_{i=1}^{N}(1+\Lambda_i\gamma_i)$. We can simply manage γ_i in each channel to maximize the net information gain. Rewrite

$$\prod_{k=1}^{N}(1+\Lambda_k\gamma_k) = \Big(\prod_{k=1}^{N}\Lambda_k\Big)\prod_{k=1}^{N}\Big(\frac{1}{\Lambda_k}+\gamma_k\Big). \tag{48}$$

As we claimed before, $\sum_{i=1}^{N}\gamma_k = 1$ where γ_k, $k = 1, \ldots, N$, is an integer multiple of $1/m$. Inspired by the water-filling algorithm, we can consider $(\gamma_1, \ldots, \gamma_N)$ as an allocation of m blocks (each with size $1/m$) into N channels. In contrast to water-filling, we refer to this problem as *block-filling* (or, to be more evocative, *ice-cube-filling*). The original heights of these channels are $1/\Lambda_1 \le \ldots \le 1/\Lambda_N$. Finally, the net information gain is determined by the product $\prod_{k=1}^{N}(\frac{1}{\Lambda_k}+\gamma_k)$ of the final heights. The optimal solution can be extracted from an optimal allocation that maximizes (48).

Because $\Lambda_1 \ge \ldots \ge \Lambda_N$, to maximize $\prod_{k=1}^{N}(1+\Lambda_k\gamma_k)$ we should allocate nonzero values of γ_k in the first $q = \min(m, N)$ channels. Accordingly, there exists an optimal solution $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)$ such that

$$\prod_{k=1}^{N}(1+\Lambda_k\alpha_k) = \prod_{k=1}^{q}(1+\Lambda_k\alpha_k). \tag{49}$$

Assume that we pick $\mathbf{a}_k$, $k = 1, \ldots, m$, using the greedy policy. By (18) and (19), we see that the kth iteration of the greedy algorithm only changes $\Lambda_1^{(k-1)}$ into $(\frac{1}{\Lambda_1^{(k-1)}}+\frac{1}{m})^{-1}$, which is equivalent to changing $\frac{1}{\Lambda_1^{(k-1)}}$ into $\frac{1}{\Lambda_1^{(k-1)}}+\frac{1}{m}$. Consider this greedy policy in the viewpoint of block-filling. The greedy policy fills blocks to the lowest channel one by one. If there are more than one channel having the same lowest height, it adds to the channel with the smallest index. Likewise, since the original heights of the channels are $1/\Lambda_1 \le \ldots \le 1/\Lambda_N$, the greedy policy only fills blocks to the first q channels, i.e., greedy solution $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_N)$ also satisfies

$$\prod_{k=1}^{N}(1+\Lambda_k\eta_k) = \prod_{k=1}^{q}(1+\Lambda_k\eta_k). \tag{50}$$

We now provide a necessary condition for both optimal and greedy solutions.

Lemma 6: Assume that an allocation $\boldsymbol{\alpha}$ is determined by either an optimal solution or a greedy solution. If α_k is nonzero, then $\alpha_k + \frac{1}{\Lambda_k}$ is bounded in the interval $(\mu - \frac{1}{m}, \mu + \frac{1}{m})$. Moreover, it suffices for the optimal and greedy solutions to pick from the set $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$.

Proof: First, assume that $\boldsymbol{\alpha}$ is given by an optimal solution. Recall that $\alpha_k + \frac{1}{\Lambda_k}$ is the final height of the kth channel. By examining the total volumes of water and blocks, we deduce the following. If $\alpha_i > 0$ and $\alpha_i + \frac{1}{\Lambda_i} > \mu$ for some $1 \le i \le q$, where μ is the water level defined in (36), then there exists some channel $1 \le j \le r$ such that $\alpha_j + \frac{1}{\Lambda_j} < \mu$. For the purpose of proof by contradiction, let us assume that $\alpha_i + \frac{1}{\Lambda_i} \ge \mu + \frac{1}{m}$. We move the top block of the ith channel to the jth channel to get another allocation $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)$. Clearly, $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ have the same entries except the ith and jth components. The argument in this paragraph is illustrated in Fig. 3.

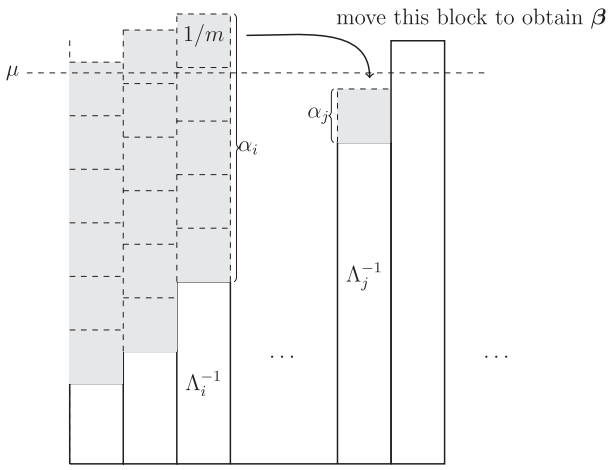

Fig. 3. Obtain allocation $\boldsymbol{\beta}$ from $\boldsymbol{\alpha}$.

For simplicity, denote $\delta_k := \alpha_k + \Lambda_k^{-1} - \mu$ for $k = 1, \ldots, m$. So

$$\begin{aligned}
&\frac{\prod_{k=1}^{m}(1+\Lambda_k\beta_k)}{\prod_{k=1}^{m}(1+\Lambda_k\alpha_k)} \\
&= \frac{(1+\Lambda_i(\mu+\delta_i-\Lambda_i^{-1}-m^{-1}))}{(1+\Lambda_i(\mu+\delta_i-\Lambda_i^{-1}))} \\
&\quad \frac{(1+\Lambda_j(\mu+\delta_j-\Lambda_j^{-1}+m^{-1}))}{(1+\Lambda_j(\mu+\delta_j-\Lambda_j^{-1}))} \\
&= \frac{(\mu+\delta_i-m^{-1})(\mu+\delta_j+m^{-1})}{(\mu+\delta_i)(\mu+\delta_j)} \\
&= \frac{(\mu+\delta_i)(\mu+\delta_j)+m^{-1}(\delta_i-\delta_j)-m^{-2}}{(\mu+\delta_i)(\mu+\delta_j)} > 1, \quad (51)
\end{aligned}$$

because $\delta_i - \delta_j > \frac{1}{m}$. Thus $\boldsymbol{\beta}$ gives a better allocation, which contradicts the optimality of $\boldsymbol{\alpha}$. By a similar argument, we obtain that for any optimal solution $\boldsymbol{\alpha}$, there also does not exist i such that $\alpha_i > 0$ and $\alpha_i + \frac{1}{\Lambda_i} \leq \mu - \frac{1}{m}$. In conclusion, the final height $\alpha_i + \frac{1}{\Lambda_i}$, $i = 1, \ldots, r$, in each channel in the optimal solution is bounded in the interval $(\mu - \frac{1}{m}, \mu + \frac{1}{m})$. Additionally, in both cases when $r = q$ and $r < q$, $\alpha_{r+1} = \cdots = \alpha_N = 0$. This means that it suffices for the optimal solution to pick from the set $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$.

Next, we assume that $\boldsymbol{\alpha}$ is determined by a greedy solution. If $\alpha_i > 0$ and $\alpha_i + \frac{1}{\Lambda_i} > \mu$, for some $1 \leq i \leq q$, then there exists a channel with index $1 \leq j \leq r$ such that $\eta_j + \frac{1}{\Lambda_j} < \mu$. For the purpose of proof by contradiction, let us assume that $\alpha_i + \frac{1}{\Lambda_i} \geq \mu + \frac{1}{m}$. This implies that when the greedy algorithm fills the top block to the ith channel, it does not add that block to the jth channel with a lower height. This contradicts how the the greedy policy actually behaves. By a similar argument, there does not exist some channel i such that $\alpha_i > 0$ and $\alpha_i + \frac{1}{\Lambda_i} \leq \mu - \frac{1}{m}$. In conclusion, the final height $\alpha_i + \frac{1}{\Lambda_i}$, $i = 1, \ldots, r$, in each channel in the greedy solution is bounded in the interval $(\mu - \frac{1}{m}, \mu + \frac{1}{m})$. Moreover, $\alpha_{r+1} = \cdots = \alpha_N = 0$. This means that it suffices for the greedy solution to pick from the set $\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$. ■

We now proceed to the equivalence between the optimal solution and the greedy solution. To show this equivalence, let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)$ be an arbitrary allocation of m blocks satisfying the necessary condition in Lemma 6. Next, we will show how to modify $\boldsymbol{\theta}$ to obtain an optimal allocation. After that, we will also show how to modify $\boldsymbol{\theta}$ to obtain an allocation that is generated by the greedy policy. It will then be evident that these two resulting allocations have the same information gain.

To obtain an optimal allocation from $\boldsymbol{\theta}$, we first remove the top block from each channel whose height is above μ to get an auxiliary allocation $\boldsymbol{\theta}' = (\theta_1', \ldots, \theta_r')$. Assume that the total number of removed blocks is m'. This auxiliary $\boldsymbol{\theta}'$ is unique, because each θ_k' is simply the maximal number of blocks can be filled in the kth channel to obtain a height not above the water level: this number is uniquely determined by Λ_k, μ, and m. We now show how to re-allocate the removed m' blocks, so that, together with $\boldsymbol{\theta}'$, we have an optimal allocation of all m blocks.

Note that by Lemma 6, to obtain an optimal solution we cannot allocate more than one block to any channel, because that would make the height of that channel above $\mu + \frac{1}{m}$. We claim that the optimal allocation simply re-allocates the m' removed blocks to the lowest m' channels in $\boldsymbol{\theta}'$. We can show this by contradiction. Assume that the optimal allocation adds one block to the ith channel instead of a lower jth channel in $\boldsymbol{\theta}'$. This means that $\theta_i' > \theta_j'$, $\theta_i = \theta_i' + 1/m$, and $\theta_j = \theta_j'$. By an argument similar to (51), if we move the top block in the ith channel to the jth channel, we would obtain a better allocation (which gives a larger net information gain). This contradiction verifies our claim.

Next, we concentrate on the allocation provided by the greedy policy. First, we recall that at each step of the greedy algorithm it never fills a block to some higher channel instead of a lower one. So after the greedy algorithm fills one block to some channel, its height cannot differ from a lower channel by more than $1/m$. If we apply the greedy policy for picking $\mathbf{a}_k$, $k = 1, \ldots, (m - m')$, then we obtain the same allocation as $\boldsymbol{\theta}'$. This is because any other allocation of $(m - m')$ blocks would result in a channel, after its top block filled, with a height deviating by more than $1/m$ from some other channel. This allocation contradicts the behavior of the greedy policy. Continuing with $\boldsymbol{\theta}'$, the greedy policy simply allocates the remaining m' blocks to the lowest m' channels one by one. So the greedy policy gives the same final heights as the optimal allocation. The only possible difference is the order of these heights. Therefore, the greedy solution is equivalent to the optimal solution in the sense of giving the same net information gain, i.e., $H_G = H_O$. This completes the proof of Theorem 4.

Appendix B
Proof of Theorem 5

We have studied the performance of the greedy policy in the viewpoint of block-filling in the proof of Theorem 4. For the purpose of simplicity, we rewrite $1/\lambda_k - 1/\lambda_{k+1} = n_k/\sigma^2$ as

$$\frac{1}{\Lambda_k} - \frac{1}{\Lambda_{k+1}} = \frac{n_k}{m} \quad (52)$$

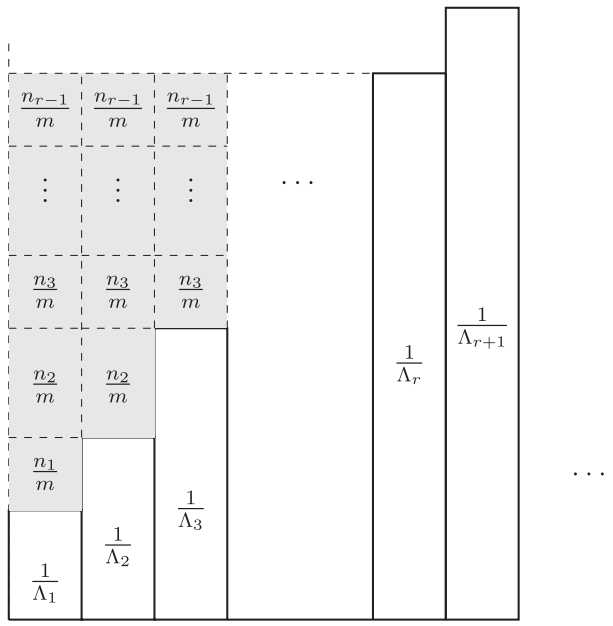

Fig. 4. Heights of channels after $\hat{m}$ iterations of the greedy policy.

where $\Lambda_i = \frac{m\lambda_i}{\sigma^2}$. After $\hat{m} := \sum_{k=1}^{r-1} kn_k$ iterations of the greedy policy, the heights in the first r channels give a flat top, which is illustrated in Fig. 4.

There are $m - \hat{m}$ blocks remaining after $\hat{m}$ iterations. If r divides $m - \hat{m}$, the final heights of the first r channels still give a flat top coinciding with μ in each channel. Therefore $H_G = H_R$. From (37), we conclude that $H_G = H_O$.

REFERENCES

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley, 1991, ch. 9, pp. 224–238.

[2] J. S. Goldstein, I. S. Reed, and L. L. Scharf, "A multistage representation of the wiener filters based on orthogonal projections," *IEEE Trans. Inf. Theory*, vol. 44, no. 7, pp. 2943–2959, Nov. 1998.

[3] L. L. Scharf, E. K. P. Chong, M. D. Zoltowski, J. S. Goldstein, and I. S. Reed, "Subspace expansion and the equivalence of conjugate direction and multistage wiener filters," *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 5013–5019, Oct. 2008.

[4] W. R. Carson, M. Chen, M. R. D. Rodrigues, R. Calderbank, and L. Carin, "Communications-inspired projection design with application to compressive sensing," *SIAM J. Imag. Sci.*, vol. 5, no. 4, pp. 1185–1212, 2012.

[5] L. Scharf, *Statistical Signal Processing*. Reading, MA, USA: Addison-Wesley, 1991, pp. 330–333.

[6] Y. Hua, M. Nikpour, and P. Stoica, "Optimal reduced-rank estimation and filtering," *IEEE Trans. Signal Process.*, vol. 49, no. 3, pp. 457–469, Mar. 2001.

[7] A. Scaglione, P. Stoica, S. Barbarossa, G. B. Giannakis, and H. Sampath, "Optimal designs for space-time linear precoders and decoders," *IEEE Trans. Signal Process.*, vol. 50, no. 5, pp. 1051–1064, May 2002.

[8] F. Pérez-Cruz, M. R. Rodrigues, and S. Verdú, "MIMO Gaussian channels with arbitrary inputs: Optimal precoding and power allocation," *IEEE Trans. Inf. Theory*, vol. 56, no. 3, pp. 1070–1084, Mar. 2010.

[9] I. D. Schizas, G. B. Giannakis, and Z. Luo, "Distributed estimation using reduced-dimensionality sensor observations," *IEEE Trans. Signal Process.*, vol. 55, no. 8, pp. 4284–4299, Aug. 2007.

[10] Y. Wang, H. Wang, and L. L. Scharf, "Optimum compression of a noisy measurement for transmission over a noisy channel," *IEEE Trans. Signal Process.*, vol. 62, no. 5, pp. 1279–1289, Mar. 2014.

[11] J. Haupt and R. Nowak, "Adaptive sensing for sparse recovery," in *Compressed Sensing: Theory and Applications*, Y. C. Eldar and G. Kutyniok, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2012, ch. 5, pp. 269–304.

[12] A. O. Hero, "Information theoretic approaches to sensor management," in *Foundations and Applications of Sensor Management*, A. O. Hero, D. A. Castanon, D. Cochran, and K. Kastella, Eds. New York, NY, USA: Springer-Verlag, 2007, ch. 3, pp. 33–57.

[13] M. Elad, "Optimized projections for compressed sensing," *IEEE Trans. Signal Process.*, vol. 55, no. 12, pp. 5695–5702, Dec. 2007.

[14] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2346–2356, Jun. 2008.

[15] S. Ji, D. Dunson, and L. Carin, "Multitask compressive sensing," *IEEE Trans. Signal Process.*, vol. 57, no. 1, pp. 92–106, Jan. 2009.

[16] R. Castro, J. Haupt, R. Nowak, and G. Raz, "Finding needles in noisy haystacks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Las Vegas, NV, USA, Apr. 2008, pp. 5133–5136.

[17] J. Haupt, R. Castro, and R. Nowak, "Distilled sensing: Adaptive sampling for sparse detection and estimation," *IEEE Trans. Inf. Theory*, vol. 57, no. 9, pp. 6222–6235, Sep. 2011.

[18] J. Haupt, R. Castro, and R. Nowak, "Improved bounds for sparse recovery from adaptive measurements," in *Proc. IEEE ISIT*, Austin, TX, USA, Jun. 2010, pp. pp. 1563–1567.

[19] E. Liu and E. K. P. Chong, "On greedy adaptive measurements," in *Proc. 46th Annu. CISS*, Princeton, NJ, USA, Mar. 2012, pp. 1–6.

[20] E. Liu, E. K. P. Chong, and L. L. Scharf, "On greedy adaptive measurements," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, USA, Nov. 2012, pp. 1229–1232.

[21] A. Ashok, J. L. Huang, and M. A. Neifeld, "Information-optimal adaptive compressive imaging," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, USA, 2011, pp. 1255–1259.

[22] J. Ke, A. Ashok, and M. A. Neifeld, "Object reconstruction from adaptive compressive measurements in feature-specific imaging," *Appl. Opt.*, vol. 49, no. 34, pp. H27–H39, 2010.

[23] S. Joshi and S. Boyd, "Sensor selection via convex optimization," *IEEE Trans. Signal Process.*, vol. 57, no. 2, pp. 451–462, Feb. 2009.

[24] H. Rowaihy, S. Eswaran, M. Johnson, D. Verma, A. Bar-Noy, T. Brown, *et al.*, "A survey of sensor selection schemes in wireless sensor networks," *Proc. SPIE*, vol. 6562, pp. 65621A-1–65621A-13, Apr. 2007.

[25] Z. Zhang, Z. Wang, E. K. P. Chong, A. Pezeshki, and W. Moran, "Near optimality of greedy strategies for string submodular functions with forward and backward curvature constraints," in *Proc. 52nd IEEE Conf. Decision Control*, Florence, Italy, Dec. 2013, pp. 5156–5161.

[26] G. L. Nemhauser and L. A. Wolsey, "Best algorithms for approximating the maximum of a submodular set function," *Math. Oper. Res.*, vol. 3, no. 3, pp. 177–188, 1978.

[27] G. Calinescu, C. Chekuri, M. Pál, and J. Vondrák, "Maximizing a monotone submodular function subject to a matroid constraint," *SIAM J. Comput.*, vol. 40, no. 6, pp. 1740–1766, 2011.

[28] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[29] J. Ding and A. Zhou, "Eigenvalues of rank-one updated matrices with some applications," *Appl. Math. Lett.*, vol. 20, no. 12, pp. 1223–1226, 2007.

[30] Y. Chi, L. Scharf, A. Pezeshki, and A. R. Calderbank, "Sensitivity to basis mismatch in compressed sensing," *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 2182–2195, May 2011.

[31] H. S. Witsenhausen, "A determinant maximization problem occurring in the theory of data communication," *SIAM J. Appl. Math.*, vol. 29, no. 3, pp. 515–522, 1975.

[32] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY, USA: Wiley, 1968.

Entao Liu (S'10–M'11) received B.S. degree in Applied Mathematics from Shandong University, Jinan, China in 2005 and Ph.D. degree also in Applied Mathematics from the University of South Carolina, Columbia, SC in 2011.

Currently, he is a postdoc in the School of Electrical and Computer Engineering, Georgia Institute of Technology. His research interests include approximation theory, signal processing, inverse problem, compressed sensing, etc.

Edwin K. P. Chong (F'04) received the B.E. degree with First Class Honors from the University of Adelaide, South Australia, in 1987; and the M.A. and Ph.D. degrees in 1989 and 1991, respectively, both from Princeton University, where he held an IBM Fellowship. He joined the School of Electrical and Computer Engineering at Purdue University in 1991, where he was named a University Faculty Scholar in 1999. Since August 2001, he has been a Professor of Electrical and Computer Engineering and Professor of Mathematics at Colorado State University. His current research interests span the areas of stochastic modeling and control, optimization methods, and communication and sensor networks. He coauthored the best-selling book, *An Introduction to Optimization* (4th Edition, Wiley-Interscience, 2013). He received the NSF CAREER Award in 1995 and the ASEE Frederick Emmons Terman Award in 1998. He was a co-recipient of the 2004 Best Paper Award for a paper in the journal *Computer Networks*. In 2010, he received the IEEE Control Systems Society Distinguished Member Award.

Prof. Chong was the founding chairman of the IEEE Control Systems Society Technical Committee on Discrete Event Systems, and served as an IEEE Control Systems Society Distinguished Lecturer. He is currently a Senior Editor of the IEEE TRANSACTIONS ON AUTOMATIC CONTROL, and has also served on the editorial boards of *Computer Networks* and the *Journal of Control Science and Engineering*. He was a member of the IEEE Control Systems Society Board of Governors and is currently Vice President for Financial Activities. He has also served on the organizing committees of several international conferences. He has been on the program committees for the IEEE Conference on Decision and Control, the American Control Conference, the IEEE International Symposium on Intelligent Control, IEEE Symposium on Computers and Communications, and the IEEE Global Telecommunications Conference. He has also served in the executive committees for the IEEE Conference on Decision and Control, the American Control Conference, the IEEE Annual Computer Communications Workshop, the International Conference on Industrial Electronics, Technology & Automation, and the IEEE International Conference on Communications. He was the Conference (General) Chair for the Conference on Modeling and Design ofWireless Networks, part of SPIE ITCom 2001. He was the General Chair for the 2011 Joint 50th IEEE Conference on Decision and Control and European Control Conference.

Louis L. Scharf (S'67–M'69-SM'77–F'86–LF'07) received the Ph.D. degree from the University of Washington, Seattle.

From 1971 to 1982, he served as Professor of Electrical Engineering and Statistics with Colorado State University (CSU), Ft. Collins. From 1982 to 1985, he was Professor and Chairman of Electrical and Computer Engineering, University of Rhode Island, Kingston. From 1985 to 2000, he was Professor of Electrical and Computer Engineering, University of Colorado, Boulder. In January 2001, he rejoined CSU as Professor of Electrical and Computer Engineering and Statistics. He is currently Research Professor of Mathematics at CSU. Prof. Scharf has held several visiting positions here and abroad, including the Ecole Superieure dElectricite, Gif-sur-Yvette, France; Ecole Nationale Superieure des Telecommunications, Paris, France; EURECOM, Nice, France; the University of La Plata, La Plata, Argentina; Duke University, Durham, NC; the University of Wisconsin, Madison; and the University of Tromso, Tromso, Norway. His interests are in statistical signal processing, as it applies to radar, sonar, and wireless communication.

Prof. Scharf was Technical Program Chair for the 1980 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Denver, CO; Tutorials Chair for ICASSP 2001, Salt Lake City, UT; and Technical Program Chair for the Asilomar Conference on Signals, Systems, and Computers 2002. He is past-Chair of the Fellow Committee for the IEEE Signal Processing Society. He has received numerous awards for his research contributions to statistical signal processing, including a College Research Award, an IEEE Distinguished Lectureship, an IEEE Third Millennium Medal, and the Technical Achievement and Society Awards from the IEEE Signal Processing Society.