# 1 Adaptive Sensing for Sparse Recovery

Jarvis Haupt[1] and Robert Nowak[2]
[1]University of Minnesota; [2]University of Wisconsin–Madison

## 1.1 Introduction

High-dimensional inference problems cannot be accurately solved without enormous amounts of data or prior assumptions about the nature of the object to be inferred. Great progress has been made in recent years by exploiting intrinsic low-dimensional structure in high-dimensional objects. *Sparsity* is perhaps the simplest model for taking advantage of reduced dimensionality. It is based on the assumption that the object of interest can be represented as a linear combination of a small number of elementary functions. The specific functions needed in the representation are assumed to belong to a larger collection or dictionary of functions, but are otherwise unknown. The *sparse recovery* problem is to determine which functions are needed in the representation based on measurements of the object. This general problem can usually be cast as a problem of identifying a vector $x \in \mathbb{R}^n$ from measurements. The vector is assumed to have $k \ll n$ non-zero elements, however the locations of the non-zero elements are unknown.

Most of the existing theory and methods for the sparse recovery problem are based on non-adaptive measurements. In this chapter we investigate the advantages of sequential samplingschemes that adapt to $x$ using information gathered throughout the sampling process. The distinction between adaptive and non-adaptive measurement can be made more precise, as follows. Information is obtained from samples or measurements of the form $y_1(x), y_2(x) \ldots$, where $y_t$ are functionals from a space $\mathcal{Y}$ representing all possible measurement forms and $y_t(x)$ are the values the functionals take for $x$. We distinguish between two types of information:

**Non-Adaptive Information**: $y_1, y_2, \cdots \in \mathcal{Y}$ are chosen non-adaptively (deterministically or randomly) and independently of $x$.

**Adaptive Information**: $y_1, y_2, \cdots \in \mathcal{Y}$ are selected sequentially, and the choice of $y_{t+1}$ may depend on the previously gathered information, $y_1(x), \ldots, y_t(x)$.

In this chapter we will see that adaptive information can be significantly more powerful when the measurements are contaminated with additive noise. In particular, we will discuss a variety of adaptive measurement procedures that

gradually focus on the subspace, or sparse support set, where $x$ lives, allowing for increasingly precise measurements to be obtained. We explore adaptive schemes in the context of two common scenarios, which are described in some detail below.

### 1.1.1 Denoising

The classic denoising problem deals with the following. Suppose we observe $x$ in noise according to the *non-adaptive* measurement model

$$y = x + e, \tag{1.1}$$

where $e \in \mathbb{R}^n$ represents a vector of additive Gaussian white noise; i.e., $e_j \overset{i.i.d.}{\sim} \mathcal{N}(0,1),\ j = 1, \ldots, n,$ where i.i.d. stands for *independent and identically distributed* and $\mathcal{N}(0,1)$ denotes the standard Gaussian distribution. It is sufficient to consider unit variance noises in this model, since other values can be accounted for by an appropriate scaling of the entries of $x$.

Let $x$ be deterministic and sparse, but otherwise unknown. The goal of the denoising problem we consider here is to determine the locations of the non-zero elements in $x$ from the measurement $y$. Because the noises are assumed to be i.i.d., the usual strategy is to simply threshold the components of $y$ at a certain level $\tau$, and declare those that exceed the threshold as detections. This is challenging for the following simple reason. Consider the probability $\Pr(\max_j e_j > \tau)$ for some $\tau > 0$. Using a simple bound on the Gaussian tail and the union bound, we have

$$\Pr(\max_j e_j > \tau) \leq \frac{n}{2} \exp\left(-\frac{\tau^2}{2}\right) \;=\; \exp\left(-\frac{\tau^2}{2} + \log n - \log 2\right) . \tag{1.2}$$

This shows that if $\tau > \sqrt{2 \log n}$, then the probability of false detections can be controlled. In fact, in the high-dimensional limit [1]

$$\Pr\left(\lim_{n\to\infty} \frac{\max_{j=1,\ldots,n} e_j}{\sqrt{2\log n}} = 1\right) = 1 \tag{1.3}$$

and therefore, for large $n$, we see that false detections cannot be avoided with $\tau < \sqrt{2 \log n}$. These basic facts imply that this classic denoising problem cannot be reliably solved unless the non-zero components of $x$ exceed $\sqrt{2 \log n}$ in magnitude. This dependence on the problem size $n$ can be viewed as a statistical "curse of dimensionality."

The classic model is based on non-adaptive measurements. Suppose instead that the measurements could be performed sequentially as follows. Assume that each measurement $y_j$ results from integration over time or averaging of repeated independent observations. The classic non-adaptive model allocates an equal portion of the full *measurement budget* to each component of $x$. In the sequential adaptive model, the budget can be distributed in a more flexible and adaptive manner. For example, a sequential sensing method could first measure all of

the components using a third of the total budget, corresponding to observations of each component plus an additive noise distributed as $\mathcal{N}(0, 3)$. The measurements are very noisy, but may be sufficiently informative to reliably rule out the presence of non-zero components at a large fraction of the locations. After ruling out many locations, the remaining two thirds of the measurement budget can be directed at the locations still in question. Now, because there are fewer locations to consider, the variance associated with the subsequent measurements can be even smaller than in the classic model. An illustrative example of this process is depicted in Figure 1.1. We will see in later sections that such sequential measurement models can effectively mitigate the curse of dimensionality in high-dimensional sparse inference problems. This permits the recovery of signals having nonzero components whose magnitudes grow much more slowly than $\sqrt{2 \log n}$.

### 1.1.2 Inverse Problems

The classic inverse problem deals with the following observation model. Suppose we observe $x$ in noise according to the *non-adaptive* measurement model

$$y = Ax + e, \tag{1.4}$$

where $A \in \mathbb{R}^{m \times n}$ is a known *measurement matrix*, $e \in \mathbb{R}^n$ again represents a vector of independent Gaussian white noise realizations, and $x$ is assumed to be deterministic and sparse, but otherwise unknown. We will usually assume that the columns of $A$ have unit norm. This normalization is used so that the SNR is not a function of $m$, the number of rows. Note that in the denoising problem we have $A = I_{n \times n}$, the identity operator, which also has unit norm columns.

The goal of the inverse problem is to recover $x$ from $y$. A natural approach to this problem is to find a solution to the constrained optimization

$$\min_x \|y - Ax\|_2^2 \ , \quad \text{subject to } \|x\|_0 \leq k \ , \tag{1.5}$$

where, as stated in Chapter 1, $\|x\|_0$ is the $\ell_0$ (pseudo-)norm which counts the number of non-zero components in $x$. It is common to refer to an $\ell_0$ constraint as a *sparsity constraint*. Note that in the special case where $A = I_{n \times n}$ the solution of the optimization (1.5) corresponds to hard-thresholding of $y$ at the level of the magnitude of the minimum of the $k$ largest (in magnitude) components of $y$. Therefore the $\ell_0$-constrained optimization (1.5) coincides with the denoising problem described above.

For the general inverse problem, $A$ is not proportional to the identity matrix and it may even be non-invertible. Nevertheless, the optimization above can still have a unique solution due to the sparsity constraint. Unfortunately, in this case the optimization (1.5) is combinatorial in nature, generally requiring a brute-force search over all $\binom{n}{k}$ sparsity patterns. A common alternative is to instead

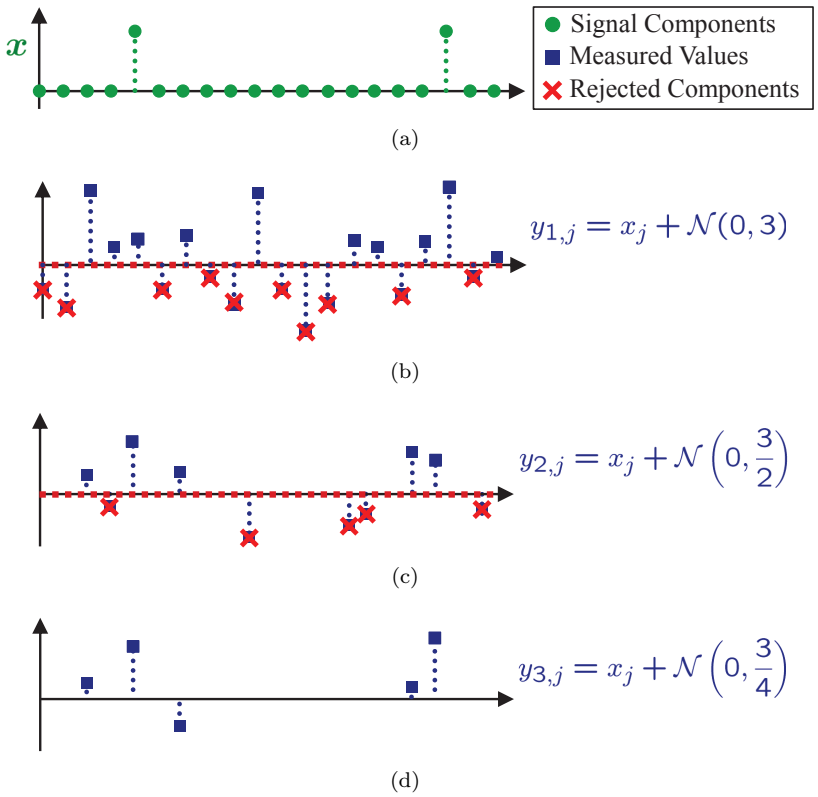


**Figure 1.1** Qualitative illustration of a sequential sensing process. A total of 3 observation steps are utilized, and the measurement budget is allocated uniformly over the steps. The original signal is depicted in panel (a). In the first observation step, shown in panel (b), all components are observed and a simple test identifies two subsets—one corresponding to locations to be measured next, and another set of locations to subsequently ignore. In the second observation step (panel (b)), each observation has twice the precision as the measurements in the previous step, since the same portion of the measurement budget is being used to measure half as many locations. Another refinement step leads to the final set of observations depicted in panel (d). Note that a single-step observation process would yield measurements with variance 1, while the adaptive procedure results in measurements with lower variance at the locations of interest.

solve a convex relaxation of the form

$$\min_x \|y - Ax\|_2^2 \ , \quad \text{subject to } \|x\|_1 \leq \tau \ , \tag{1.6}$$

for some $\tau > 0$. This $\ell_1$-constrained optimization is relatively easy to solve using convex optimization techniques. It is well known that the solutions of the optimization (1.6) are sparse, and the smaller $\tau$, the sparser the solution.

If the columns of $A$ are not too correlated with one another and $\tau$ is chosen appropriately, then the solution to this optimization is close to the solution of the $\ell_0$-constrained optimization. In fact in the absence of noise, perfect recovery of

the sparse vector $x$ is possible. For example, compressed sensing methods often employ an $A$ comprised of realizations of i.i.d. symmetric random variables. If $m$ (the number of rows) is just slightly larger than $k$, then every subset of $k$ columns from such an $A$ will be close to orthogonal [2, 3, 4]. This condition suffices to guarantee that any sparse signal with $k$ or fewer non-zero components can be recovered from $\{y, A\}$ – see, for example, [5].

When noise is present in the measurements, reliably determining the locations of the non-zero components in $x$ requires that these components are significantly large relative to the noise level. For example, if the columns of $A$ are scaled to have unit norm, recent work [6] suggests that the optimization in (1.6) will succeed (with high probability) only if the magnitudes of the non-zero components exceed a fixed constant times $\sqrt{\log n}$. In this chapter we will see that this fundamental limitation can again be overcome by sequentially designing the rows of $A$ so that they tend to focus on the relevant components as information is gathered.

### 1.1.3 A Bayesian Perspective

The denoising and inverse problems each have a simple Bayesian interpretation which is a convenient perspective for the development of more general approaches. Recall the $\ell_0$-constrained optimization in (1.5). The Lagrangian formulation of this optimization is

$$\min_x \left\{ \|y - Ax\|_2^2 + \lambda \|x\|_0 \right\} \ , \tag{1.7}$$

where $\lambda > 0$ is the Lagrange multiplier. The optimization can be viewed as a Bayesian procedure, where the term $\|y - Ax\|_2^2$ is the negative Gaussian log-likelihood of $x$, and $\lambda\|x\|_0$ is the negative log of a prior distribution on the support of $x$. That is, the mass allocated to $x$ with $k$ non-zero elements is uniformly distributed over the $\binom{n}{k}$ possible sparsity patterns. Minimizing the sum of these two quantities is equivalent to solving for the *Maximum a Posteriori* (MAP) estimate.

The $\ell_1$-constrained optimization also has a Bayesian interpretation. The Lagrangian form of that optimization is

$$\min_x \left\{ \|y - Ax\|_2^2 + \lambda \|x\|_1 \right\} \ . \tag{1.8}$$

In this case the prior is proportional to $\exp(-\lambda\|x\|_1)$, which models components of $x$ independently with a heavy-tailed (double-exponential, or Laplace) distribution. Both the $\ell_0$ and $\ell_1$ priors, in a sense, reflect a belief that the $x$ we are seeking is sparse (or approximately so) but otherwise unstructured in the sense that all patterns of sparsity are equally probable a priori.

### 1.1.4 Structured Sparsity

The Bayesian perspective also provides a natural framework for more structured models. By modifying the prior (and hence the penalizing term in the optimization), it is possible to encourage solutions having more structured patterns of sparsity. A very general information-theoretic approach to this sort of problem was provided in [7], and we adopt that approach in the following examples. Priors can be constructed by assigning a binary code to each possible $x$. The prior probability of any given $x$ is proportional to $\exp(-\lambda K(x))$, where $\lambda > 0$ is a constant and $K(x)$ is the bit-length of the code assigned to $x$. If $A$ is an $m \times n$ matrix with entries drawn independently from a symmetric binary-valued distribution, then the expected mean square error of the estimate

$$\widehat{x} \;=\; \arg\min_{x \in \mathcal{X}} \left\{ \|y - Ax\|_2^2 + \lambda K(x) \right\} \tag{1.9}$$

selected by optimizing over a set $\mathcal{X}$ of candidates (which, for example, could be a discretized subset of the set of all vectors in $\mathbb{R}^n$ with $\ell_2$ norm bounded by some specified value), satisfies

$$\mathbb{E}\|\widehat{x} - x^*\|_2^2/n \;\leq\; C \min_{x \in \mathcal{X}} \left\{ \|x - x^*\|_2^2/n + c\, K(x)/m \right\} \;. \tag{1.10}$$

Here, $x^*$ is the vector that generated $y$ and $C, c > 0$ are constants depending on the choice of $\lambda$. The notation $\mathbb{E}$ denotes expectation, which here is taken with respect to the distribution on $A$ and the additive noise in the observation model (1.4). The $\|x\|_0$ prior/penalty is recovered as a special case in which $\log n$ bits are allocated to encode the location and value of each non-zero element of $x$ (so that $K(x)$ is proportional to $\|x\|_0 \log n$). Then the error satisfies the bound

$$\mathbb{E}\|\widehat{x} - x^*\|_2^2/n \;\leq\; C' \|x^*\|_0 \log n/m \;, \tag{1.11}$$

for some constant $C' > 0$.

The Bayesian perspective also allows for more structured models. To illustrate, consider a simple sparse binary signal $x^*$ (i.e., all non-zero components take the value 1). If we make no assumptions on the sparsity pattern, then the location of each non-zero component can be encoded using $\log n$ bits, resulting in a bound of the same form as (1.11). Suppose instead that the sparsity pattern of $x$ can be represented by a binary tree whose vertices correspond the elements of $x$. This is a common model for the typical sparsity patterns of wavelet coefficients, for example see [8]. The tree-structured restriction means that a node can be non-zero if and only if its "parent" node is also non-zero. Thus, each possible sparsity pattern corresponds to a particular branch of the full binary tree. There exist simple prefix codes for binary trees, and the codelength for a tree with $k$ vertices is at most $2k + 1$ (see, for example, [9]). In other words, we require just over 2 bits per component, rather than $\log n$. Applying the general error bound (1.10) we obtain

$$\mathbb{E}\|\widehat{x} - x^*\|_2^2/n \;\leq\; C'' \|x^*\|_0/m \;, \tag{1.12}$$

for some constant $C'' > 0$ which, modulo constants, is a factor of $\log n$ better than the bound under the unstructured assumption. Thus, we see that the Bayesian perspective provides a formalism for handling a wider variety of modeling assumptions and deriving performance bounds. Several authors have explored various other approaches to exploiting structure in the patterns of sparsity [10, 11].

Another possibility offered by the Bayesian perspective is to customize the sensing matrix in order to exploit more informative prior information (other than simple unstructured sparsity) that may be known about $x$. This has been formulated as a Bayesian experimental design problem [12, 13]. Roughly speaking, the idea is to identify a good prior distribution for $x$ and then optimize the choice of the sensing matrix $A$ in order to maximize the expected information of the measurement. In the next section we discuss how this idea can be taken a step further, to sequential Bayesian experimental designs that automatically adapt the sensing to the underlying signal in an on-line fashion.

## 1.2 Bayesian Adaptive Sensing

The Bayesian perspective provides a natural framework for sequential adaptive sensing, wherein information gleaned from previous measurements is used to automatically adjust and focus the sensing. In principle the idea is very simple. Let $\mathcal{Q}_1$ denote a probability measure over all $m \times n$ matrices having unit Frobenius norm in expectation. This normalization generalizes the column normalization discussed earlier. It still implies that the SNR is independent of $m$, but it also allows for the possibility of distributing the measurement budget more flexibly throughout the columns. This will be crucial for adaptive sensing procedures. For example, in many applications the sensing matrices have entries drawn i.i.d. from a symmetric distribution (see Chapter 5 for a detailed discussion of random matrices). Adaptive sensing procedures, including those discussed in later sections of this chapter, are often also constructed from entries drawn from symmetric, but not identical, distributions. By adaptively adjusting the variance of the distributions used to generate the entries, these sensing matrices can place more or less emphasis on certain components of the signal.

Now consider how we might exploit adaptivity in sparse recovery. Suppose that we begin with a prior probability distribution $p(x)$ for $x$. Initially collect a set of measurements $y \in \mathbb{R}^m$ according to the sensing model $y = Ax + w$ with $A \sim \mathcal{Q}_1$, where $\mathcal{Q}_1$ is a prior probability distribution on $m \times n$ sensing matrices. For example, $\mathcal{Q}_1$ could correspond to drawing the entries of $A$ independently from a common symmetric distribution. A *posterior* distribution for $x$ can be calculated by combining these data with a prior probability model for $x$, using Bayes' rule. Let $p(x|y)$ denote this posterior distribution. It then becomes natural to ask, which sensing actions will provide the most new information about $x$? In other words, we are interested in designing $\mathcal{Q}_2$ so that the next measurement

using a sensing matrix $A \sim \mathcal{Q}_2$ maximizes our gain in information about $x$. For example, if certain locations are less likely (or even completely ruled-out) given the observed data $y$, then $\mathcal{Q}_2$ should be designed to place little (or zero) probability mass on the corresponding columns of the sensing matrix. Our goal will be to develop strategies that utilize information from previous measurements to effectively "focus" the sensing energy of subsequent measurements into subspaces corresponding to the true signal of interest (and away from locations of less interest). An example depicting the notion of focused sensing is shown in Figure 1.2.

More generally, the goal of the next sensing action should be to reduce the uncertainty about $x$ as much as possible. There is a large literature dealing with this problem, usually under the topic of "sequential experiments." The classical Bayesian perspective is nicely summarized in the work of DeGroot [14]. He credits Lindley [15] with first proposing the use of Shannon entropy as a measure of uncertainty to be optimized in the sequential design of experiments. Using the notion of Shannon entropy, the "information-gain" of an experiment can be quantified by the change that the new data produces in the entropy associated with the unknown parameter(s). The optimal design of a number of sequential experiments can be defined recursively and viewed as a dynamic programming problem. Unfortunately, the optimization is intractable in all but the most simple situations. The usual approach, instead, operates in a greedy fashion, maximizing the information-gain at each step in a sequence of experiments. This can be suboptimal, but often is computationally feasible and effective.

An adaptive sensing procedure of this sort can be devised as follows. Let $p(x)$ denote the probability distribution of $x$ after the $(t)$-th measurement step. Imagine that in the $(t+1)$-th step we measure $y = Ax + e$, where $A \sim \mathcal{Q}$ and $\mathcal{Q}$ is a distribution we can design as we like. Let $p(x|y)$ denote the posterior distribution according to Bayes' rule. The "information" provided by this measurement is quantified by the Kullback-Leibler (KL) divergence of $p(x)$ from $p(x|y)$ which is given by

$$\mathbb{E}_X \left[ \log \frac{p(x|y)}{p(x)} \right], \tag{1.13}$$

where the expectation is with respect to the distribution of a random variable $X \sim p(x|y)$. Notice that this expression is a function of $y$, which is undetermined until the measurement is made. Thus, it is natural to consider the expectation of the KL divergence with respect to the distribution of $y$, which depends on the prior $p(x)$, the distribution of the noise, and most importantly, on the choice of $\mathcal{Q}$. Let $p(y)$ denote the distribution of the random measurement obtained using the observation matrix $A \sim \mathcal{Q}$. The *expected* information gain from a measurement based on $A$ is defined to be

$$\mathbb{E}_{Y_\mathcal{Q}} \mathbb{E}_X \left[ \log \frac{p(x|y)}{p(x)} \right], \tag{1.14}$$

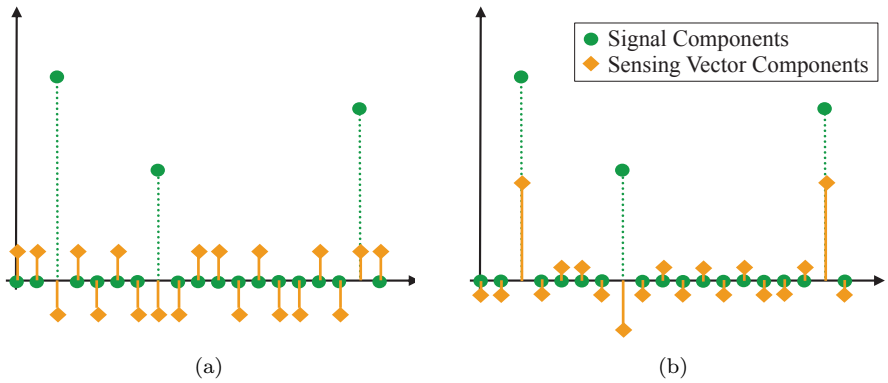


**Figure 1.2** Traditional vs. focused sensing. Panel (a) depicts a sensing vector that may be used in a traditional non-adaptive measurement approach. The components of the sensing vector have uniform amplitudes, implying that an equal amount of "sensing energy" is being allocated to all locations regardless of the signal being measured. Panel (b) depicts a focused sensing vector where most of the sensing energy is focused on a small subset of the components corresponding to the relevant entries of the signal.

where the outer expectation is with respect to the distribution of a random variable $Y_{\mathcal{Q}} \sim p(y)$. This suggests choosing a distribution for the sensing matrix for the next measurement to maximize the expected information gain, that is

$$\mathcal{Q}_{t+1} \;=\; \arg\max_{\mathcal{Q}} \mathbb{E}_{Y_{\mathcal{Q}}}\mathbb{E}_X \left[\log \frac{p(x|y)}{p(x)}\right] \;, \tag{1.15}$$

where the optimization is over a space of possible distributions on $m \times n$ matrices.

One useful interpretation of this selection criterion follows by observing that maximizing the expected information gain is equivalent to minimizing the conditional entropy of the posterior distribution [15]. Indeed, simplifying the above expression we obtain

$$\begin{aligned}
\mathcal{Q}_{t+1} &= \arg\max_{\mathcal{Q}} \mathbb{E}_{Y_{\mathcal{Q}}}\mathbb{E}_X \left[\log \frac{p(x|y)}{p(x)}\right] \\
&= \arg\min_{\mathcal{Q}} -\mathbb{E}_{Y_{\mathcal{Q}}}\mathbb{E}_X \log p(x|y) + \mathbb{E}_{Y_{\mathcal{Q}}}\mathbb{E}_X \log p(x) \\
&= \arg\min_{\mathcal{Q}} H(X|Y_{\mathcal{Q}}) - H(X) \\
&= \arg\min_{\mathcal{Q}} H(X|Y_{\mathcal{Q}}),
\end{aligned} \tag{1.16}$$

where $H(X)$ denotes the Shannon entropy and $H(X|Y_{\mathcal{Q}})$ the entropy of $X$ conditional on $Y_{\mathcal{Q}}$. Another intuitive interpretation of the information gain criterion follows from the fact that

$$\mathbb{E}_{Y_{\mathcal{Q}}}\mathbb{E}_X \left[\log \frac{p(x|y)}{p(x)}\right] = \mathbb{E}_{X,Y_{\mathcal{Q}}} \left[\log \frac{p(x,y)}{p(x)p(y)}\right] \tag{1.17}$$

where the right-hand side is just the mutual information between the random variables $X$ and $Y_{\mathcal{Q}}$. Thus, the information gain criterion equivalently suggests that the next measurements should be constructed in a way that maximizes the mutual information between $X$ and $Y_{\mathcal{Q}}$.

Now, given this selection of $\mathcal{Q}_{t+1}$, we may draw $A \sim \mathcal{Q}_{t+1}$, collect the next measurement $y = Ax + e$, and use Bayes' rule to obtain the new posterior. The rationale is that at each step we are choosing a sensing matrix that maximizes the expected information gain, or equivalently minimizes the expected entropy of the new posterior distribution. Ideally, this adaptive and sequential approach to sensing will tend to focus on $x$ so that sensing energy is allocated to the correct subspace, increasing the SNR of the measurements relative to non-adaptive sensing. The performance could be evaluated, for example, by comparing the result of several adaptive steps to that obtained using a single non-adaptively chosen $A$.

The approach outlined above suffers from a few inherent limitations. First, while maximizing the expected information gain is a sensible criterion for focusing, the exposition makes no guarantees about the performance of such methods. That is, one cannot immediately conclude that this procedure will lead to an improvement in performance. Second, and perhaps more importantly in practice, selecting the sensing matrix that maximizes the expected information gain can be computationally prohibitive. In the next few sections, we discuss several efforts where approximations or clever choices of the prior are employed to alleviate the computational burden of these procedures.

### 1.2.1 Bayesian Inference Using a Simple Generative Model

To illustrate the principles behind the implementation of Bayesian sequential experimental design, we begin with a discussion of the approach proposed in [16]. Their work employed a simple signal model in which the signal vector $x \in \mathbb{R}^n$ was assumed to consist of only a single nonzero entry. Despite the potential model misspecification, this simplification enables the derivation of closed-form expressions for model parameter update rules. It also leads to a simple and intuitive methodology for the shaping of projection vectors in the sequential sampling process.

#### 1.2.1.1 Single Component Generative Model

We begin by constructing a generative model for this class of signals. This model will allow us to define the problem parameters of interest, and to perform inference on them. First, we define $L$ to be a random variable whose range is the set of indices of the signal, $j = \{1, 2, \ldots, n\}$. The entries of the probability mass function of $L$, denoted by $q_j = \Pr(L = j)$, encapsulate our belief regarding which index corresponds to the true location of the single nonzero component. The amplitude of the single nonzero signal component is a function of its location $L$, and is denoted by $\alpha$. Further, conditional on the outcome $L = j$, we model

the amplitude of the nonzero component as a Gaussian random variable with location-dependent mean and variance, $\mu_j$ and $\nu_j$, respectively. That is, the distribution of $\alpha$ given $L = j$ is given by

$$p(\alpha|L = j) \sim \mathcal{N}(\mu_j, \nu_j). \tag{1.18}$$

Thus, our prior on the signal $x$ is given by $p(\alpha, L)$, and is described by the hyperparameters $\{q_j, \mu_j, \nu_j\}_{j=1}^n$.

We will perform inference on the hyperparameters, updating our knowledge of them using scalar observations collected according to the standard observation model,

$$y_t = A_t x + e_t, \tag{1.19}$$

where $A_t$ is a $1 \times n$ vector and the noises $\{e_t\}$ are assumed to be i.i.d. $\mathcal{N}(0, \sigma^2)$ for some known $\sigma > 0$. We initialize the hyperparameters of the prior to $q_j(0) = 1/n$, $\mu_j(0) = 0$, and $\nu_j(0) = \sigma_0^2$ for some specified $\sigma_0$, for all $j = 1, 2, \ldots, n$. Now, at time step $t \geq 1$, the posterior distribution for the unknown parameters at a particular location $j$ can be written as

$$p(\alpha, L = j|y_t, A_t) = p(\alpha|y_t, A_t, L = j) \cdot q_j(t-1). \tag{1.20}$$

Employing Bayes' rule, we can rewrite the first term on the right-hand side to obtain

$$p(\alpha|y_t, A_t, L = j) \propto p(y_t|A_t, \alpha, L = j) \cdot p(\alpha|L = j), \tag{1.21}$$

and thus the posterior distribution for the unknown parameters satisfies

$$p(\alpha, L = j|y_t, A_t) \propto p(y_t|A_t, \alpha, L = j) \cdot p(\alpha|L = j) \cdot q_j(t-1). \tag{1.22}$$

The proportionality notation has been used to suppress the explicit specification of the normalizing factor. Notice that, by construction, the likelihood function $p(y_t|A_t, \alpha, L = j)$ is conjugate to the prior $p(\alpha|L = j)$, since each is Gaussian. Substituting in the corresponding density functions, and following some straightforward algebraic manipulation, we obtain the following update rules for the hyperparameters:

$$\mu_j(t) = \frac{A_{t,j}\nu_j(t-1)y_t + \mu_j(t-1)\sigma^2}{A_{t,j}^2\nu_j(t-1) + \sigma^2}\ , \tag{1.23}$$

$$\nu_j(t) = \frac{\nu_j(t-1)\sigma^2}{A_{t,j}^2\nu_j(t-1) + \sigma^2}\ , \tag{1.24}$$

$$q_j(t) \propto \frac{q_j(t-1)}{\sqrt{A_{t,j}^2\nu_j(t-1) + \sigma^2}} \exp\left(-\frac{1}{2}\frac{(y_t - A_{t,j}\mu_j(t-1))^2}{A_{t,j}^2\nu_j(t-1) + \sigma^2}\right). \tag{1.25}$$

#### 1.2.1.2 Measurement Adaptation

Now, as mentioned above, our goal here is twofold. One one hand, we want to *estimate* the parameters corresponding to the location and amplitude of the

unknown signal component. On the other hand, we want to devise a strategy for *focusing* subsequent measurements onto the features of interest to boost the performance of our inference methods. This can be accomplished by employing the information gain criterion, that is, selecting our next measurement to be the most informative measurement that can be made given our current state of knowledge of the unknown quantities. This knowledge is encapsulated by our current estimates of the problem parameters.

We adopt the criterion in (1.16), as follows. Suppose that the next measurement vector $A_{t+1}$ is drawn from some distribution $\mathcal{Q}$ over $1 \times n$ vectors. Let $Y_{\mathcal{Q}}$ denote the random measurement obtained using this choice of $A_{t+1}$. Our goal is to select $\mathcal{Q}$ to minimize the conditional entropy of a random variable $X$ distributed according to our generative model with parameters that reflect information obtained up to time $t$, given $Y_{\mathcal{Q}}$. In other words, the information gain criterion suggests that we choose the distribution from which the next sensing vector will be drawn according to (1.16).

To facilitate the optimization, we will consider a simple construction for the space from which $\mathcal{Q}$ is to be chosen. Namely, we will assume that the next projection vector is given by the element-wise product between a random sign vector $\xi \in \{-1, 1\}^n$ and a non-negative weight vector $\psi \in \mathbb{R}^n$, so that $A_{t+1,j} = \xi_j \psi_j$. Further, we assume that the entries of the sign vector are equally likely and independent. In other words, we will assume that the overall observation process is as depicted in Figure 1.3, and our goal will be to determine the weight vector $\psi$.

Let us focus on the objective term $H(X|Y_{\mathcal{Q}})$. First, note that conditional on the outcome $y_{t+1}$, $x$ is distributed according to a Gaussian mixture with hyperparameters $\{q_j(t), \mu_j(t), \nu_j(t)\}_{j=1}^n$. There is no closed-form expression for the entropy of a Gaussian mixture. Instead, using the fact that the conditional differential entropy is a lower bound for the differential entropy [17], and conditioning on the selection of the mixture component, the upper-bound

$$\begin{aligned} H(X|Y_{\mathcal{Q}} = y_t) &\leq H(q) + \frac{1}{2}\sum_{j=1}^{n} q_j(t) \log\left(2\pi e\left(\frac{\nu_j(t)\sigma^2}{(\xi_j\psi_j)^2\nu_j(t) + \sigma^2}\right)\right) \\ &= H(q) + \frac{1}{2}\sum_{j=1}^{n} q_j(t) \log\left(2\pi e\left(\frac{\nu_j(t)\sigma^2}{\psi_j^2\nu_j(t) + \sigma^2}\right)\right) \end{aligned}$$

can be obtained following some straightforward algebraic manipulations. Now, notice that this bound is independent of the actual outcome of the sign vector at location $j$ because only the square of that term appears in the expression. Similarly, the actual observed value $y_{t+1}$ does not appear at all. Thus, we can easily take the expectation with respect to the distribution of the observed values to obtain

$$H(X|Y_{\mathcal{Q}}) \leq H(q) + \frac{1}{2}\sum_{j=1}^{n} q_j(t) \log\left(2\pi e\left(\frac{\nu_j(t)\sigma^2}{\psi_j^2\nu_j(t) + \sigma^2}\right)\right) \tag{1.26}$$

Now, we consider choosing the weights $\psi_j$ for $j = 1, 2, \ldots, n$ by optimizing over this upper bound:

$$\begin{aligned}\psi &= \arg \min_{z \in \mathbb{R}^n : \|z\|_2 = 1} H(p) + \frac{1}{2} \sum_{j=1}^{n} q_j(t) \log \left( 2\pi e \left( \frac{\nu_j(t)\sigma^2}{z_j^2 \nu_j(t) + \sigma^2} \right) \right) \\ &= \arg \max_{z \in \mathbb{R}^n : \|z\|_2 = 1} \sum_{j=1}^{n} q_j(t) \log \left( z_j^2 \nu_j(t) + \sigma^2 \right). \end{aligned} \tag{1.27}$$

This optimization can be solved by a simple application of Lagrange multipliers, but it is perhaps more illustrative to consider one further simplification that is appropriate in low-noise settings. In particular, let us assume that $\sigma^2 \approx 0$, then the optimization becomes

$$\psi = \arg \max_{z \in \mathbb{R}^n : \|z\|_2 = 1} \sum_{j=1}^{n} q_j(t) \log \left( z_j^2 \right). \tag{1.28}$$

Now, it is easy to show that the objective in this simplified formulation is maximized by selecting $z_j = \sqrt{q_j(t)}$.

The focusing criterion obtained here is generally consistent with our intuition for this problem. It suggests that the amount of "sensing energy" that should be allocated to a given location $j$ be proportional to our current belief that the nonzero signal component is indeed at location $j$. Initially, when we assume that the location of the nonzero component is uniformly distributed among the set of indices, this criterion instructs us to allocate our sensing energy uniformly, as is the case in traditional "non-adaptive" CS methods. On the other hand, as we become more confident in our belief that we have identified a set of promising locations at which the the nonzero component could be present, the criterion suggests that we focus our energy on those locations to reduce the measurement uncertainty (ie, to obtain the highest SNR measurements possible).

The procedure outlined here can be extended, in a straightforward way, to settings where the unknown vector $x$ has multiple nonzero entries. The basic idea is to identify the nonzero entries of the signal one-at-a-time, using a sequence of iterations of the proposed procedure. For each iteration, the procedure is executed as described above until one entry of the posterior distribution for the location parameter exceeds a specified threshold $\tau \in (0, 1)$. That is, the current iteration of the sequential sensing procedure terminates when the posterior likelihood of a true nonzero component at any of the locations becomes large, which corresponds to the event that $q_j(t) > \tau$ for any $j \in \{1, 2, \ldots, n\}$, for a specified $\tau$ that we choose to be close to 1. At that point, we conclude that a nonzero signal component is present at the corresponding location. The sequential sensing procedure is then restarted and the parameters $\{q_j, \mu_j, \nu_j\}_{j=1}^n$ are reinitialized, except that the initial values of $\{q_j\}_{j=1}^n$ are set to zero at locations identified as signal components in previous iterations of the procedure, and uniformly distributed over the remaining locations. The resulting multi-step procedure is akin to an "onion peeling" process.

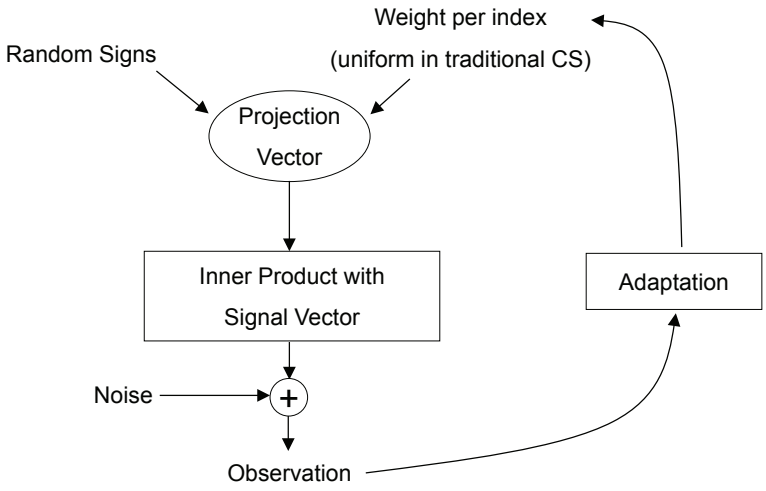


**Figure 1.3** Block diagram of the adaptive focusing procedure. Previous observations are utilized to "shape" the weights associated with each location of the random vectors which will be used in the sensing process.

### 1.2.2 Bayesian Inference Using Multi-Component Models

The simple single-component model for the unknown signal $x$ described above is but one of many possible generative models that might be employed in a Bayesian treatment of the sparse inference problem. Another, perhaps more natural, option is to employ a more sophisticated model that explicitly allows for the signal to have multiple nonzero components.

#### 1.2.2.1 Multi-component Generative Model

As discussed above, a widely used sparsity promoting prior is the Laplace distribution,

$$p(x|\lambda) = \left(\frac{\lambda}{2}\right)^n \cdot \exp\left(-\lambda \sum_{j=1}^{n} |x_j|\right). \tag{1.29}$$

From an analytical perspective in Bayesian inference, however, this particular choice of prior on $x$ can lead to difficulties. In particular, under a Gaussian noise assumption, the resulting likelihood function for the observations (which is conditionally Gaussian given $x$ and the projection vectors) is not conjugate to the Laplace prior, and so closed-form update rules cannot be easily obtained.

Instead, here we discuss the method that was examined in [18], which utilizes a *hierarchical* prior on the signal $x$, similar to a construction proposed in the context of sparse Bayesian learning in [19]. As before, we begin by constructing a generative model for the signal $x$. To each $x_j$, $j = 1, 2, \ldots, n$, we associate a parameter $\rho_j > 0$. The joint distribution of the entries of $x$, conditioned on the parameter vector $\rho = (\rho_1, \rho_2, \ldots, \rho_n)$, is given in the form of a product distribu-

tion,

$$p(x|\rho) = \prod_{j=1}^{n} p(x_j|\rho_j), \tag{1.30}$$

and we let $p(x_j|\rho_j) \sim \mathcal{N}(0, \rho_j^{-1})$. Thus, we may interpret the $\rho_j$ as precision or "inverse variance" parameters. In addition, we impose a prior on the entries of $\rho$, as follows. For global parameters $\alpha, \beta > 0$, we set

$$p(\rho|\alpha, \beta) = \prod_{j=1}^{n} p(\rho_j|\alpha, \beta), \tag{1.31}$$

where $p(\rho_j|\alpha, \beta) \sim \text{Gamma}(\alpha, \beta)$ is distributed according to a Gamma distribution with parameters $\alpha$ and $\beta$. That is,

$$p(\rho_j|\alpha, \beta) = \frac{\rho_j^{\alpha-1}\beta^{\alpha}\exp(-\beta\rho_j)}{\Gamma(\alpha)}, \tag{1.32}$$

where

$$\Gamma(\alpha) = \int_0^{\infty} z^{\alpha-1}\exp(-z)dz \tag{1.33}$$

is the Gamma function. As in the previous section we will assume that the noise is Gaussian-distributed, but here we also impose a Gamma prior on the distribution of the noise precision, resulting in a hierarchical prior similar to that utilized for the signal vector. Formally, we model our observations using the standard matrix-vector formulation,

$$y = Ax + e, \tag{1.34}$$

where $y \in \mathbb{R}^m$ and $A \in \mathbb{R}^{m\times n}$, and we let $p(e|\rho_0) \sim \mathcal{N}(0, \rho_0 I_{m\times m})$, and $p(\rho_0|\gamma, \delta) \sim \text{Gamma}(\gamma, \delta)$. A graphical summary of the generative signal and observation models is depicted in Figure 1.4.

Now, the hierarchical model was chosen primarily to facilitate analysis, since the Gaussian prior on the signal components is conjugate to the Gaussian (conditional) likelihood of the observations. Generally speaking, a Gaussian prior itself will not promote sparsity; however, incorporating the effect of the Gamma hyperprior lends some additional insight into the situation here. By marginalizing over the parameters $\rho$, we can obtain an expression for the overall prior distribution of the signal components in terms of the parameters $\alpha$ and $\beta$,

$$p(x|\alpha, \beta) = \prod_{j=1}^{n} \int_0^{\infty} p(x_j|\rho_j) \cdot p(\rho_j|\alpha, \beta)d\rho_j. \tag{1.35}$$

The integral(s) can be evaluated directly, giving

$$p(x_j|\alpha,\beta) = \int_0^\infty p(x_j|\rho_j)\cdot p(\rho_j|\alpha,\beta)d\rho_j$$
$$= \frac{\beta^\alpha \Gamma(\alpha+1/2)}{(2\pi)^{1/2}\Gamma(\alpha)}\left(\beta+\frac{x_j^2}{2}\right)^{-(\alpha+1/2)}. \tag{1.36}$$

In other words, the net effect of the prescribed hierarchical prior on the signal coefficients is that of imposing a Student-t prior distribution on each signal component. The upshot is that, for certain choices of the parameters $\alpha$ and $\beta$, the product distribution can be strongly-peaked about zero, similar (in spirit) to the Laplace distribution—see [19] for further discussion.

Given the hyperparameters $\rho$ and $\rho_0$, as well as the observation vector $y$ and corresponding measurement matrix $A$, the posterior for $x$ is conditionally a multivariate Gaussian distribution with mean $\mu$ and covariance matrix $\Sigma$. Letting $R = \mathrm{diag}(\rho)$, and assuming that the matrix $\left(\rho_0 A^T A + R\right)$ is full-rank, we have

$$\Sigma = \left(\rho_0 A^T A + R\right)^{-1}, \tag{1.37}$$

and

$$\mu = \rho_0 \Sigma A^T y. \tag{1.38}$$

The goal of the inference procedure, then, is to estimate the hyperparameters $\rho$ and $\rho_0$ from the observed data $y$. From Bayes' rule, we have that

$$p(\rho,\rho_0|y) \propto p(y|\rho,\rho_0)p(\rho)p(\rho_0). \tag{1.39}$$

Now, following the derivation in [19], we consider improper priors obtained by setting the parameters $\alpha,\beta,\gamma$, and $\delta$ all to zero, and rather than seeking a fully-specified posterior for the hyperparameters we instead obtain point estimates via a maximum likelihood procedure. In particular, the maximum likelihood estimates of $\rho$ and $\rho_0$ are obtained by maximizing

$$p(y|\rho,\rho_0) = (2\pi)^{-m/2}\left|\frac{1}{\rho_0}I_{m\times m} + AR^{-1}A^T\right|^{-1/2}$$
$$\exp\left\{-\frac{1}{2}y^T\left(\frac{1}{\rho_0}I_{m\times m} + AR^{-1}A^T\right)^{-1}y\right\}. \tag{1.40}$$

This yields the the following update rules:

$$\rho_j^{\text{new}} = \frac{1-\rho_j\Sigma_{j,j}}{\mu_j^2}, \tag{1.41}$$

$$\rho_0^{\text{new}} = \frac{m-\sum_{j=1}^n(1-\rho_j\Sigma_{j,j})}{\|y-A\mu\|_2^2}. \tag{1.42}$$

Overall, the inference procedure alternates between solving for $\rho_0$ and $\rho$ as functions of $\mu$ and $\Sigma$ using (1.41) and (1.42), and solving for $\mu$ and $\Sigma$ as functions of $\rho_0$ and $\rho$ using (1.37) and (1.38).

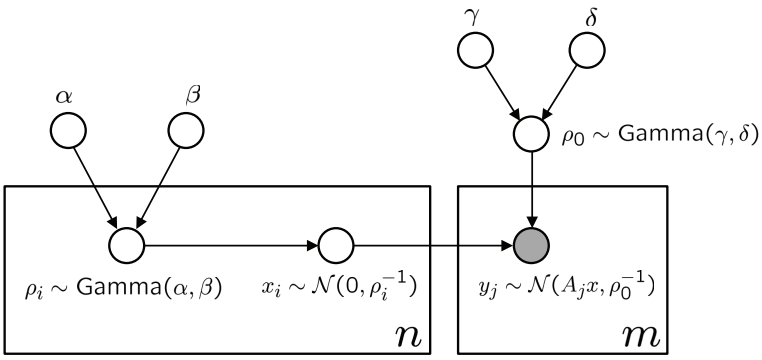


**Figure 1.4** Graphical model associated with the multi-component Bayesian CS model.

#### 1.2.2.2 Measurement Adaptation

As in the previous section, we may devise a sequential sensing procedure by first formulating a criterion under which the next projection vector can be chosen to be the most informative. Let us denote the distribution of $x$ given the first $t$ measurements by $p(x)$. Suppose that the $(t+1)$-th measurement is obtained by projecting onto a vector $A_{t+1} \sim \mathcal{Q}$, and let $p(x|y)$ denote the posterior. Now, the criterion for selecting the distribution $\mathcal{Q}_{t+1}$ from which the next measurement vector should be drawn is given by (1.16). As in the previous example, we will simplify the criterion by first restricting the space of distributions over which the objective is to be optimized. In this case, we will consider a space of degenerate distributions. We assume that each $\mathcal{Q}$ corresponds to a distribution that takes a deterministic value $Q \in \mathbb{R}^n$ with probability one, where $\|Q\|_2 = 1$. The goal of the optimization, then, is to determine the "direction" vector $Q$.

Recall that by construction, given the hyperparameters $\rho_0$ and $\rho$ the signal $x$ is multivariate Gaussian with mean and variance $\mu$ and $\Sigma$ as given in (1.38) and (1.37), respectively. The hierarchical prior(s) imposed on the hyperparameters $\rho_0$ and $\rho$ make it difficult to evaluate $H(X|Y_Q)$ directly. Instead, we simplify the problem further by assuming that $x$ is unconditionally Gaussian (ie, $\rho_0$ and $\rho$ are deterministic and known). In this case the objective function of the information gain criterion can be evaluated directly, and the criterion for selecting the next measurement vector becomes

$$A_{t+1} = \arg \min_{Q \in \mathbb{R}^n, \|Q\|_2 = 1} -\frac{1}{2} \log \left(1 + \rho_0 Q \Sigma Q^T\right), \tag{1.43}$$

where $\Sigma$ and $\rho_0$ reflect the knowledge of the parameters up to time $t$. From this it is immediately obvious that $A_{t+1}$ should be in the direction of the eigenvector corresponding to the largest eigenvalue of the covariance matrix $\Sigma$.

As with the simple single-component signal model case described in the previous section, the focusing rule obtained here also lends itself to some intuitive explanations. Recall that at a given step of the sequential sensing procedure, $\Sigma$ encapsulates our knowledge of both our level of uncertainty about which entries

of the unknown signal are relevant as well as our current level of uncertainty about the actual component value. In particular, note that under the zero-mean Gaussian prior assumption on the signal amplitudes large values of the diagonal entries of $R$ can be understood to imply the existence of a true nonzero signal component at the corresponding location. Thus, the focusing criterion described above suggests that we focus our sensing energy onto locations at which we are both fairly certain that a signal component is present (as quantified by large entries of the diagonal matrix $R$), and fairly uncertain about its actual value because of the measurement noise (as quantified by the $\rho_0 A^T A$ term in (1.37)). Further, the relative contribution of each is determined by the level of the additive noise or, more precisely, our current estimate of it.

### 1.2.3 Quantifying Performance

The adaptive procedures discussed in the previous sections can indeed provide realizable performance improvements relative to non-adaptive CS methods. It has been shown, via simulation, that these adaptive sensing procedures can outperform traditional CS in noisy settings. For example, adaptive methods can provide a reduction in mean-square reconstruction error, relative to non-adaptive CS, in situations where each utilizes the same total number of observations. Similarly, it has been shown that in some settings adaptive methods can achieve the same error performance as non-adaptive methods using a smaller number of measurements. We refer the reader to [16, 18], as well as [20, 21] for extensive empirical results and more detailed performance comparisons of these procedures.

A complete analysis of these adaptive sensing procedures would ideally also include an analytical performance evaluation. Unfortunately, it appears to be very difficult to devise quantitative error bounds, like those known for non-adaptive sensing, for Bayesian sequential methods. Because each sensing matrix depends on the data collected in the previous steps, the overall process is riddled with complicated dependencies that prevent the use of the usual approaches to obtain error bounds based, for example, on concentration of measure and other tools.

In the next section, we present a recently-developed alternative to Bayesian sequential design called *distilled sensing* (DS). In essence, the DS framework encapsulates the spirit of sequential Bayesian methods, but uses a much simpler strategy for exploiting the information obtained from one sensing step to the next. The result is a powerful, computationally efficient procedure that is also amenable to analysis, allowing us to quantify the dramatic performance improvements that can be achieved through adaptivity.

## 1.3 Quasi-Bayesian Adaptive Sensing

In the previous section, the Bayesian approach to adaptive sensing was discussed, and several examples were reviewed to show how this approach might be implemented in practice. The salient aspect of these techniques, in essence, was the use of information from prior measurements to guide the acquisition of subsequent measurements in an effort to obtain samples that are most informative. This results in sensing actions that focus sensing resources toward locations that are more likely to contain signal components, and away from locations that likely do not. While this notion is intuitively pleasing, its implementation introduces statistical dependencies that make an analytical treatment of the performance of such methods quite difficult.

In this section we discuss a recently-developed adaptive sensing procedure called *distilled sensing* (DS) [22] which is motivated by Bayesian adaptive sensing techniques, but also has the added benefit of being amenable to theoretical performance analysis. The DS procedure is quite simple, consisting of a number of iterations, each of which is comprised of an observation stage followed by a refinement stage. In each observation stage, measurements are obtained at a set of locations which could potentially correspond to nonzero components. In the corresponding refinement stage, the set of locations at which observations were collected in the measurement stage is partitioned into two disjoint sets—one corresponding to locations at which additional measurements are to be obtained in the next iteration, and a second corresponding to locations to subsequently ignore. This type of adaptive procedure was the basis for the example in Figure 1.1. The refinement strategy utilized in DS is a sort of "poor-man's Bayesian" methodology intended to approximate the focusing behavior achieved by methods that employ the information gain criterion. The upshot here is that this simple refinement is still quite effective at focusing sensing resources toward locations of interest. In this section we examine the performance guarantees that can be attained using the DS procedure.

For the purposes of comparison, we begin with a brief discussion of the performance limits for non-adaptive sampling procedures, expanding on the discussion of the denoising problem in Section 1.1.1. We then present and discuss the DS procedure in some detail, and we provide theoretical guarantees on its performance which quantify the gains that can be achieved via adaptivity. In the last subsection we discuss extensions of DS to underdetermined compressed sensing observation models, and we provide some preliminary results on that front.

### 1.3.1 Denoising using Non-Adaptive Measurements

Consider the general problem of recovering a sparse vector $x \in \mathbb{R}^n$ from its samples. Let us assume that the observations of $x$ are described by the simple model

$$y = x + e, \tag{1.44}$$

where $e \in \mathbb{R}^n$ represents a vector of additive perturbations, or "noise." The signal $x$ is assumed to be sparse, and for the purposes of analysis in this section we will assume that all the non-zero components of $x$ take the same value $\mu > 0$. Even with this restriction on the form of $x$, we will see that non-adaptive sensing methods cannot reliably recover signals unless the amplitude $\mu$ is considerably larger than the noise level. Recall that the support of $x$, denoted by $\mathcal{S} = \mathcal{S}(x) = \text{supp}(x)$, is defined to be the set of all indices at which the vector $x$ has a nonzero component. The sparsity level $\|x\|_0$ is simply the cardinality of this set, $\|x\|_0 = |\mathcal{S}|$. To quantify the effect of the additive noise, we will suppose that the entries of $e$ are i.i.d. $\mathcal{N}(0,1)$. Our goal will be to perform *support recovery* (also called model selection), or to obtain an accurate estimate of the support set of $x$, using the noisy data $y$. We denote our support estimate by $\hat{\mathcal{S}} = \hat{\mathcal{S}}(y)$.

Any estimation procedure based on noisy data is, of course, subject to error. To assess the quality of a given support estimate $\hat{\mathcal{S}}$, we define two metrics to quantify the two different types of errors that can occur in this setting. The first type of error corresponds to the case where we declare that nonzero signal components are present at some locations where they are not, and we refer to such mistakes as *false discoveries*. We quantify the number of these errors using the false discovery proportion (FDP), defined here as

$$\text{FDP}(\hat{\mathcal{S}}) := \frac{|\hat{\mathcal{S}} \backslash \mathcal{S}|}{|\hat{\mathcal{S}}|}, \tag{1.45}$$

where the notation $\hat{\mathcal{S}} \backslash \mathcal{S}$ denotes the set difference. In words, the FDP of $\hat{\mathcal{S}}$ is the ratio of the number of components falsely declared as non-zero to the total number of components declared non-zero. The second type of error occurs when we decide that a particular location does not contain a true nonzero signal component when it actually does. We refer to these errors as *non-discoveries*, and we quantify them using the non-discovery proportion (NDP), defined as

$$\text{NDP}(\hat{\mathcal{S}}) := \frac{|S \backslash \hat{\mathcal{S}}|}{|S|}. \tag{1.46}$$

In words, the NDP of $\widehat{\mathcal{S}}$ is the ratio of the number of non-zero components missed to the number of actual non-zero components. For our purposes, we will consider a testing procedure to be effective if its errors in these two metrics are suitably small.

In contrast to the Bayesian treatments discussed above, here we will assume that $x$ is fixed, but it is otherwise unknown. Recall that by assumption the nonzero components of $x$ are assumed to be nonnegative. In this case it is natural to focus on a specific type of estimator for $\mathcal{S}$, which is obtained by applying a simple, coordinate-wise, one-sided thresholding test to the outcome of each of the observations. In particular, the support estimate we will consider here is

$$\hat{\mathcal{S}} = \hat{\mathcal{S}}(y, \tau) = \{j : y_j > \tau\}, \tag{1.47}$$

where $\tau > 0$ is a specified threshold.

To quantify the error performance of this estimator, we examine the behavior of the resulting FDP and NDP for a sequence of estimation problems indexed by the dimension parameter $n$. Namely, for each value of $n$, we consider the estimation procedure applied to a signal $x \in \mathbb{R}^n$ having $k = k(n)$ nonzero entries of amplitude $\mu = \mu(n)$, observed according to (1.44). Analyzing the procedure for increasing values of $n$ is a common approach to quantify performance in high-dimensional settings, as a function of the corresponding problem parameters. To that end we consider letting $n$ tend to infinity to identify a critical value of the signal amplitude $\mu$ below which the estimation procedure fails, and above which it succeeds. The result is stated here as a theorem [23, 22].

**Theorem 1.1.** *Assume $x$ has $n^{1-\beta}$ non-zero components of amplitude $\mu = \sqrt{2r \log n}$ for some $\beta \in (0,1)$ and $r > 0$. If $r > \beta$, there exists a coordinate-wise thresholding procedure with corresponding threshold value $\tau(n)$ that yields an estimator $\hat{\mathcal{S}}$ for which*

$$\mathrm{FDP}(\hat{\mathcal{S}}) \xrightarrow{P} 0\,, \quad \mathrm{NDP}(\hat{\mathcal{S}}) \xrightarrow{P} 0\,, \tag{1.48}$$

*as $n \to \infty$, where $\xrightarrow{P}$ denotes convergence in probability. Moreover, if $r < \beta$, then there does not exist a coordinate-wise thresholding procedure that can guarantee that both the FDP and NDP tend to 0 as $n \to \infty$.*

This result can be easily extended to settings where the nonzero entries of $x$ are both positive and negative, and may also have unequal amplitudes. In those cases, an analogous support estimation procedure can be devised which applies the threshold test to the magnitudes of the observations. Thus, Theorem 1.1 can be understood as a formalization of the general statement made in Section. 1.1.1 regarding the denoising problem. There it was argued, based on simple Gaussian tail bounds, that the condition $\mu \approx \sqrt{2 \log n}$ was required in order to reliably identify the locations of the relevant signal components from noisy entry-wise measurements. The above result was obtained using a more sophisticated analysis, though the behavior with respect to the problem dimension $n$ is the same. In addition, and perhaps more interestingly, Theorem 1.1 also establishes a *converse* result—that reliable recovery from non-adaptive measurements is *impossible* unless $\mu$ increases in proportion to $\sqrt{\log n}$ as $n$ gets large. This result gives us a baseline with which to compare the performance of adaptive sensing, which is discussed in the following section.

### 1.3.2 Distilled Sensing

We begin our discussion of the distilled sensing procedure by introducing a slight generalization of the sampling model (1.44). This will facilitate explanation of the procedure and allow for direct comparison with non-adaptive methods. Suppose that we are able to collect measurements of the components of $x$ in a sequence

of $T$ observation steps, according to the model

$$y_{t,j} = x_j + \rho_{t,j}^{-1/2} e_{t,j}, \ j = 1, 2, \ldots, n, \quad t = 1, 2, \ldots, T, \tag{1.49}$$

where $e_{t,j}$ are i.i.d. $\mathcal{N}(0,1)$ noises, $t$ indexes the observation step, and the $\rho_{t,j}$ are non-negative "precision" parameters that can be chosen to modify the noise variance associated with a given observation. In other words, the variance of additive noise associated with observation $y_{t,j}$ is $\rho_{t,j}^{-1}$, so larger values of $\rho_{t,j}$ correspond to more precise observations. Here, we adopt the convention that setting $\rho_{t,j} = 0$ for some pair $(t, j)$ means that component $j$ is not observed at step $t$.

This multi-step observation model has natural practical realizations. For example, suppose that observations are obtained by measuring at each location one or more times and averaging the measurements. Then $\sum_{t,j} \rho_{t,j}$ expresses a constraint on the total number of measurements that can be made. This measurement budgetcan be distributed uniformly over the locations (as in non-adaptive sensing), or non-uniformly and adaptively. Alternatively, suppose that each observation is based on a sensing mechanism that integrates over time to reduce noise. The quantity $\sum_{t,j} \rho_{t,j}$, in this case, corresponds to a constraint on the total observation time. In any case, the model encapsulates an inherent flexibility in the sampling process, in which sensing resources may be preferentially allocated to locations of interest. Note that, by dividing through by $\rho_{t,j} > 0$, we arrive at an equivalent observation model, $\tilde{y}_{t,j} = \rho_{t,j}^{1/2} x_j + e_{t,j}$, which fits the general linear observation model utilized in the previous sections. Our analysis would proceed similarly in either case; we choose to proceed here using the model as stated in (1.49) because of its natural interpretation.

To fix the parameters of the problem, and to facilitate comparison with non-adaptive methods, we will impose a constraint on the overall measurement budget. In particular, we assume that $\sum_{t=1}^{T} \sum_{j=1}^{n} \rho_{t,j} \leq B(n)$. In the case $T = 1$ and $\rho_{1,j} = 1$ for $j = 1, 2, \ldots, n$, which corresponds to the choice $B(n) = n$, the model (1.49) reduces to the canonical non-adaptive observation model (1.44). For our purposes here we will adopt the same measurement budget constraint, $B(n) = n$.

With this framework in place, we now turn to the description of the DS procedure. To begin, we initialize by selecting the number of observation steps $T$ that are to be performed. The total measurement budget $B(n)$ is then divided among the $T$ steps so that a portion $B_t$ is allocated to the $t$-th step, for $t = 1, 2, \ldots, T$, and $\sum_{t=1}^{T} B_t \leq B(n)$. The set of indices to be measured in the first step is initialized to be the set of all indices, $\mathcal{I}_1 = \{1, 2, \ldots, n\}$. Now, the portion of the measurement budget $B_1$ designated for the first step is allocated uniformly over the indices to be measured, resulting in the precision allocation $\rho_{1,j} = B_1/|\mathcal{I}_1|$ for $j \in \mathcal{I}_1$. Noisy observations are collected, with the given precision, for each entry $j \in \mathcal{I}_1$. The set of observations to be measured in the next step, $\mathcal{I}_2$, is obtained by applying a simple threshold test to each of the observed values. Specifically,

**Algorithm 1.1** (Distilled sensing)**.**

**Input:**

Number of observation steps $T$

Resource allocation sequence $\{B_t\}_{t=1}^T$ satisfying $\sum_{t=1}^T B_t \leq B(n)$

**Initialize:**

Initial index set $\mathcal{I}_1 = \{1, 2, \ldots, n\}$

**Distillation:**

For $t = 1$ to $T$

Allocate resources: $\rho_{t,j} = \left\{ \begin{array}{cc} B_t/|\mathcal{I}_t| & j \in \mathcal{I}_t \\ 0 & j \notin \mathcal{I}_t \end{array} \right\}$

Observe: $y_{t,j} = x_j + \rho_{t,j}^{-1/2} e_{t,j}, j \in \mathcal{I}_t$

Refine: $\mathcal{I}_{t+1} = \{j \in \mathcal{I}_t : y_{t,j} > 0\}$

End for

**Output:**

Final index set $\mathcal{I}_T$

Distilled observations $y_T = \{y_{T,j} : j \in \mathcal{I}_T\}$

we identify the locations to be measured in the next step as those corresponding to observations that are strictly greater than zero, giving $\mathcal{I}_2 = \{j \in \mathcal{I}_1 : y_j > 0\}$. This procedure is repeated for each of the $T$ measurement steps, where (as stated above) the convention $\rho_{t,j} = 0$ implies that the signal component at location $j$ is not observed in measurement step $t$. The output of the procedure consists of the final set of locations measured, $\mathcal{I}_T$, and the observations collected at those locations $y_{T,j}$, $j \in \mathcal{I}_T$. The entire process is summarized as Algorithm 1.1.

A few aspects of the DS procedure are worth further explanation. First, we comment on the apparent simplicity of the refinement step, which identifies the set of locations to be measured in the subsequent observation step. This simple criterion encapsulates the notion that, given that the nonzero signal components are assumed to have positive amplitude, we expect that their corresponding noisy observation should be nonnegative as well. Interpreting this from a Bayesian perspective, the hard-thresholding selection operation encapsulates the idea that the probability of $y_{t,j} > 0$ given $x_j = \mu$ and $\rho_{t,j} > 0$ is approximately equal to one. In reality, using a standard bound on the tail of the Gaussian distribution, we have that

$$\Pr(y_{t,j} > 0 | \rho_{t,j} > 0, x_j = \mu) \geq 1 - \exp\left(-\frac{\rho_{t,j}\mu^2}{2}\right), \tag{1.50}$$

suggesting that the quality of this approximation may be very good, depending on the particular values of the signal amplitude $\mu$ and the precision parameter for the given observation, $\rho_{t,j}$.

Second, as in the simple testing problem described in Section 1.3.1, the DS procedure can also be extended in a straight-forward way to account for signals with both positive and negative entries. One possible approach would be to further divide the measurement budget allocation for each step $B_t$ in half, and then perform the whole DS procedure twice. For the first pass, the procedure is performed as stated in Algorithm 1.1 with the goal of identifying positive signal components. For the second pass, replacing the refinement criterion by $\mathcal{I}_{t+1} = \{i \in \mathcal{I}_t : y_{t,j} < 0\}$ would enable the procedure to identify the locations corresponding to negative signal components.

#### 1.3.2.1 Analysis of Distilled Sensing

The simple adaptive behavior of DS, relative to a fully-Bayesian treatment of the problem, renders the procedure amenable to analysis. As in Section 1.3.1, our objects of interest here will be sparse vectors $x \in \mathbb{R}^n$ having $n^{1-\beta}$ nonzero entries, where $\beta \in (0,1)$ is a fixed (and typically unknown) parameter. Recall that our goal is to obtain an estimate $\hat{\mathcal{S}}$ of the signal support $\mathcal{S}$, for which the errors as quantified by the False Discovery Proportion (1.45) and Non-Discovery Proportion (1.46) are simultaneously controlled. The following result shows that the DS procedure results in significant improvements over the comparable non-adaptive testing procedure using the same measurement budget [22]. This is achieved by carefully calibrating the problem parameters, i.e., the number of observation steps $T$ and the measurement budget allocation $\{B_t\}_{t=1}^T$.

**Theorem 1.2.** *Assume $x$ has $n^{1-\beta}$ non-zero components, where $\beta \in (0,1)$ is fixed, and that each nonzero entry has amplitude exceeding $\mu(n)$. Sample $x$ using the distilled sensing procedure with*

- $T = T(n) = \max\{\lceil \log_2 \log n \rceil, 0\} + 2$ *measurement steps,*
- *measurement budget allocation $\{B_t\}_{t=1}^T$ satisfying $\sum_{t=1}^T B_t \leq n$, and for which*
  - $B_{t+1}/B_t \geq \delta > 1/2$, *and*
  - $B_1 = c_1 n$ *and* $B_T = c_T\, n$ *for some* $c_1, c_T \in (0,1)$

*If $\mu(n) \to \infty$ as a function of $n$, then the support set estimator constructed using the output of the DS algorithm*

$$\hat{\mathcal{S}}_{\mathrm{DS}} := \{j \in \mathcal{I}_T : y_{T,j} > \sqrt{2/c_T}\} \tag{1.51}$$

*satisfies*

$$\mathrm{FDP}(\hat{\mathcal{S}}_{\mathrm{DS}}) \xrightarrow{P} 0\,, \quad \mathrm{NDP}(\hat{\mathcal{S}}_{\mathrm{DS}}) \xrightarrow{P} 0, \tag{1.52}$$

*as $n \to \infty$.*

This result can be compared directly to the result of Theorem 1.1, where it was shown that the errors associated with the estimator obtained from non-adaptive observations would converge to zero in probability only in the case $\mu > \sqrt{2\beta \log n}$. In contrast, the result of Theorem 1.2 states that the same performance metrics can be met for an estimator obtained from adaptive samples, under the much weaker constraint $\mu(n) \to \infty$. This includes signals whose nonzero components have amplitude on the order of $\mu \sim \sqrt{\log \log n}$, or $\mu \sim \sqrt{\log \log \log \cdots \log n}$, in fact, the result holds if $\mu(n)$ is *any arbitrarily slowly growing function of* $n$. If we interpret the ratio between the squared amplitude of the nonzero signal components and the noise variance as the SNR, the result in Theorem 1.2 establishes that adaptivity can provide an improvement in *effective* SNR of up to a factor of $\log n$ over comparable non-adaptive methods. This improvement can be very significant in high-dimensional testing problems where $n$ can be in the hundreds or thousands, or more.

Interpreted another way, the result of Theorem 1.2 suggests that adaptivity can dramatically mitigate the "curse of dimensionality," in the sense that the error performance for DS exhibits much less dependence on the ambient signal dimension than does the error performance for non-adaptive procedures. This effect is demonstrated in finite-sample regimes by the simulation results in Figure 1.5. Each panel of the figure depicts a scatter plot of the FDP and NDP values resulting from 1000 trials of both the adaptive DS procedure, and the non-adaptive procedure whose performance was quantified in Theorem 1.1. Each trial used a different (randomly-selected) threshold value to form the support estimate. Panels (a)-(d) correspond to four different values of $n$: $n = 2^{10}$, $2^{13}$, $2^{16}$, and $2^{19}$, respectively. In all cases, the signals being estimated have 128 nonzero entries of amplitude $\mu$, and the SNR is fixed by the selection $\mu^2 = 8$. For each value of $n$, the measurement budget allocation parameters $B_t$ were chosen so that $B_{t+1} = 0.75 B_t$ for $t = 1, \ldots, T-2$, $B_1 = B_T$, and $\sum_{t=1}^{T} B_t = n$. Comparing the results across panels, we see that the error performance of the non-adaptive procedure degrades significantly as a function of the ambient dimension, while the error performance of DS is largely unchanged across 9 orders of magnitude. This demonstrates the effectiveness of DS for acquiring high-precision observations primarily at the signal locations of interest.

The analysis of the DS procedure relies inherently upon two key ideas pertaining to the action of the refinement step(s) at each iteration. First, for any iteration of the procedure, observations collected at locations where no signal component is present will be independent samples of a zero-mean Gaussian noise process. Despite the fact that the variance of the measured noise will depend on the allocation of sensing resources, the symmetry of the Gaussian distribution ensures that the value obtained for each such observation will be (independently) positive with probability 1/2. This notion can be made formal by a straightforward application of Hoeffding's inequality.

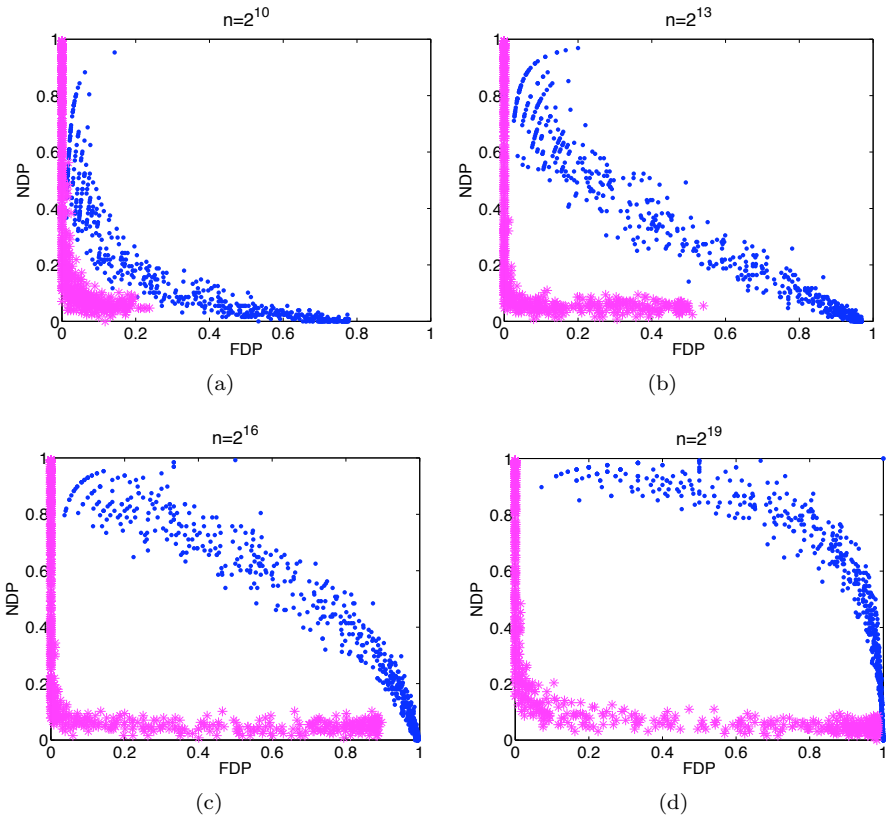


**Figure 1.5** The curse of dimensionality and the virtue of adaptivity. Each panel depicts a scatter plot of FDP and NDP values resulting for non-adaptive sensing ($\bullet$) and the adaptive DS procedure ($*$). Not only does DS outperform the non-adaptive method, it exhibits much less dependence on the ambient dimension.

**Lemma 1.1.** *Let* $\{y_j\}_{j=1}^m \overset{iid}{\sim} \mathcal{N}(0,\sigma^2)$. *For any* $0 < \epsilon < 1/2$, *the number of* $y_j$ *exceeding zero satisfies*

$$\left(\frac{1}{2}-\epsilon\right)m \leq \left|\left\{j \in \{1,2,\ldots,m\} : y_j > 0\right\}\right| \leq \left(\frac{1}{2}+\epsilon\right)m, \tag{1.53}$$

*with probability at least* $1 - 2\exp\left(-2m\epsilon^2\right)$.

In other words, each refinement step will eliminate about half of the (remaining) locations at which no signal component is present with high probability.

The second key idea is that the simple refinement step will not incorrectly eliminate too many of the locations corresponding to nonzero signal components from future consideration. A formal statement of this result, which is fundamentally a statement about the tails of the Binomial distribution , is given in the following lemma [22]. The proof is repeated here for completeness.

**Lemma 1.2.** *Let* $\{y_j\}_{j=1}^m \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ *with* $\sigma > 0$ *and* $\mu > 2\sigma$. *Let*

$$\delta = \frac{\sigma}{\mu\sqrt{2\pi}}, \tag{1.54}$$

*and note that* $\delta < 0.2$, *by assumption. Then,*

$$(1-\delta)\, m \leq \left| \left\{ j \in \{1, 2, \ldots, m\} : y_j > 0 \right\} \right| \leq m, \tag{1.55}$$

*with probability at least*

$$1 - \exp\left(-\frac{\mu m}{4\sigma\sqrt{2\pi}}\right). \tag{1.56}$$

*Proof.* Let $q = \Pr(y_j > 0)$. Using a standard bound on the tail of the Gaussian distribution, we have

$$1 - q \;\leq\; \frac{\sigma}{\mu\sqrt{2\pi}} \exp\left(-\frac{\mu^2}{2\sigma^2}\right). \tag{1.57}$$

Next, we employ the Binomial tail bound from [24]: for any $0 < b < \mathbb{E}[\sum_{j=1}^m \mathbf{1}_{\{y_j>0\}}] = mq$,

$$\Pr\left(\sum_{j=1}^m \mathbf{1}_{\{y_j>0\}} \leq b\right) \leq \left(\frac{m-mq}{m-b}\right)^{m-b} \left(\frac{mq}{b}\right)^b. \tag{1.58}$$

Note that $\delta > 1 - q$ (or equivalently, $1 - \delta < q$), so we can apply the Binomial tail bound to the sum $\sum_{j=1}^m \mathbf{1}_{\{y_j>0\}}$ with $b = (1-\delta)m$ to obtain

$$\Pr\left(\sum_{j=1}^m \mathbf{1}_{\{y_j>0\}} \leq (1-\delta)m\right) \leq \left(\frac{1-q}{\delta}\right)^{\delta m} \left(\frac{q}{1-\delta}\right)^{(1-\delta)m} \tag{1.59}$$

$$\leq \exp\left(-\frac{\mu^2\delta m}{2\sigma^2}\right) \left(\frac{1}{1-\delta}\right)^{(1-\delta)m}. \tag{1.60}$$

Now, to establish the stated result, it suffices to show that

$$\exp\left(-\frac{\mu^2\delta m}{2\sigma^2}\right) \left(\frac{1}{1-\delta}\right)^{(1-\delta)m} \leq \exp\left(-\frac{\mu m}{4\sigma\sqrt{2\pi}}\right). \tag{1.61}$$

Taking logarithms and dividing through by $\delta m$, the condition to establish becomes

$$-\frac{\mu^2}{2\sigma^2} + \left(\frac{1-\delta}{\delta}\right) \log\left(\frac{1}{1-\delta}\right) \leq -\frac{\mu}{4\delta\sigma\sqrt{2\pi}}$$

$$= \; -\frac{\mu^2}{4\sigma^2}, \tag{1.62}$$

where the last equality follows from the definition of $\delta$. The bound holds provided $\mu \geq 2\sigma$, since $0 < \delta < 1$ and

$$\left(\frac{1-\delta}{\delta}\right) \log\left(\frac{1}{1-\delta}\right) \leq 1 \tag{1.63}$$

for $\delta \in (0,1)$. □

Overall, the analysis of the DS procedure entails the repeated application of these two lemmas across iterations of the procedure. Note that the result in Lemma 1.1 is independent of the noise power, while the parameter $\delta$ in Lemma 1.2 is a function of both the signal amplitude and the observation noise variance. The latter is a function of how the sensing resources are allocated to each iteration and how many locations are being measured in that step. In other words, statistical dependencies are present across iterations with this procedure, as in the case of the Bayesian methods described above. However, unlike in the Bayesian methods, here the dependencies can be tolerated in a straight-forward manner by conditioning on the output of the previous iterations of the procedure.

Rather than presenting the full details of the proof here, we instead provide a short sketch of the general idea. To clarify the exposition, we will find it useful to fix some additional notation. First, we let $S_t = |\mathcal{S} \cap \mathcal{I}_t|$ be the number of locations corresponding to nonzero signal components that are to be observed in step $t$. Similarly, let $N_t = |\mathcal{S}^c \cap \mathcal{I}_t| = |\mathcal{I}_t| - S_t$ denote the number of remaining locations that are to be measured in the $(t)$-th iteration. Let $\sigma_1 = \sqrt{|\mathcal{I}_1|/B_1}$ denote the standard deviation of the observation noise in the first iteration, and let $\delta_1$ be the corresponding quantity from Lemma 1.2 described in terms of the quantity $\sigma_1$. Notice that since the quantity $|\mathcal{I}_1|$ is fixed and known, the quantities $\sigma_1$ and $\delta_1$ are deterministic.

Employing Lemmas 1.1 and 1.2, we determine that the result of the refinement step in the first iteration is that for any $0 < \epsilon < 1/2$, the bounds $(1-\delta_1)S_1 \leq S_2 \leq S_1$ and $(1/2-\epsilon)\,N_1 \leq N_2 \leq (1/2+\epsilon)\,N_1$ hold simultaneously, except in an event of probability no greater than

$$2\exp\left(-2N_1\epsilon^2\right) + \exp\left(-\frac{\mu S_1}{4\sigma_1\sqrt{2\pi}}\right). \tag{1.64}$$

To evaluate the outcome of the second iteration, we condition on the event that the bounds on $S_2$ and $N_2$ stated above hold. In this case, we can obtain bounds on the quantity $\mathcal{I}_2 = S_2 + N_2$, which in turn imply an upper bound on the variance of the observation noise in the second iteration. Let $\sigma_2$ denote such a bound, and $\delta_2$ its corresponding quantity from Lemma 1.2. Following the second iteration step, we have that the bounds $(1-\delta_1)(1-\delta_2)S_1 \leq S_3 \leq S_1$ and $(1/2-\epsilon)^2\,N_1 \leq N_3 \leq (1/2+\epsilon)^2\,N_1$ hold simultaneously, except in an event

of probability no greater than

$$2\exp\left(-2N_1\epsilon^2\right) + \exp\left(-\frac{\mu S_1}{4\sigma_1\sqrt{2\pi}}\right) + \tag{1.65}$$

$$2\exp\left(-2(1-\epsilon)N_1\epsilon^2\right) + \exp\left(-\frac{\mu(1-\delta_1)S_1}{4\sigma_2\sqrt{2\pi}}\right). \tag{1.66}$$

The analysis proceeds in this fashion, by iterated applications of Lemmas 1.1 and 1.2 conditioned on the outcome of all previous refinement steps. The end result is a statement quantifying the probability that the bounds $\prod_{t=1}^{T-1}(1-\delta_t)S_1 \leq S_T \leq S_1$ and $(1/2-\epsilon)^{T-1} N_1 \leq N_s \leq (1/2+\epsilon)^{T-1} N_1$ hold simultaneously following the refinement step in the $(T-1)$-st iteration, prior to the $(T)$-th observation step. It follows that the final testing problem is equivalent in structure to a general testing problem of the form considered in Section 1.1.1, but with a different *effective* observation noise variance. The final portion of the proof of Theorem 1.2 entails a careful balancing between the design of the resource allocation strategy, the number of observation steps $T$, and the specification of the parameter $\epsilon$. The goal is to ensure that as $n \to \infty$ the stated bounds on $S_T$ and $N_T$ are valid with probability tending to one, the fraction of signal components missed throughout the refinement process tends to zero, and the effective variance of the observation noise for the final set of observations is small enough to enable the successful testing of signals with very weak features. The full details can be found in [22].

### 1.3.3 Distillation in Compressed Sensing

While the results above demonstrate that adaptivity in sampling can provide a tremendous improvement in effective measurement SNR in certain sparse recovery problems, the benefits of adaptivity are somewhat less clear with respect to the other problem parameters. In particular, the comparison outlined above was made on the basis that each procedure was afforded the same measurement budget, as quantified by a global quantity having a natural interpretation in the context of a total sample budget or a total time constraint. Another basis for comparison would be the total number of measurements collected with each procedure. In the non-adaptive method in Section 1.3.1, a total of $n$ measurements were collected (one per signal component). In contrast, the number of measurements obtained via the DS procedure is necessarily larger, since each component is directly measured at least once, and some components may be measured up to a total of $T$ times—once for each iteration of the procedure. Strictly speaking, the total number of measurements collected during the DS procedure is a random quantity which depends implicitly on the outcome of the refinements at each step, which in turn are functions of the noisy measurements. However, our high-level intuition regarding the behavior of the procedure allows us to make some illustrative approximations. Recall that each refinement step eliminates (on average) about half of the locations at which no signal component is present. Fur-

ther, under the sparsity level assumed in our analysis, the signals being observed are vanishingly sparse—that is, the fraction of locations of $x$ corresponding to non-zero components tends to zero as $n \to \infty$. Thus, for large $n$, the number of measurements collected in the $t$-th step of the DS procedure is *approximately* given by $n \cdot 2^{-(t-1)}$, which implies (upon summing over $t$) that the DS procedure requires on the order of $2n$ total measurements.

By this analysis, the SNR benefits of adaptivity come at the expense of a (modest) relative increase in the number of measurements collected. Motivated by this comparison, it is natural to ask whether the distilled sensing approach might also be extended to the so-called underdetermined observation settings, such as those found in standard compressed sensing (CS) problems. In addition, and perhaps more importantly, can an analysis framework similar to that employed for DS be used to obtain performance guarantees for adaptive CS procedures? We will address these questions here, beginning with a discussion of how the DS procedure might be applied in CS settings.

At a high level, the primary implementation differences relative to the original DS procedure result from the change in observation model. Recall that, for $\rho_{t,j} > 0$, the observation model (1.49) from the previous section could alternatively be written as

$$y_{t,j} = \rho_{t,j}^{1/2} x_j + e_{t,j}, \;\; j = 1, 2, \ldots, n, \quad t = 1, 2, \ldots, T, \tag{1.67}$$

subject to a global constraint on $\sum_{t,j} \rho_{t,j}$. Under this alternative formulation, the overall sampling process can be effectively described using the matrix-vector formulation $y = Ax + e$ where $A$ is a matrix whose entries are either zero (at times and locations where no measurements were obtained) or equal to some particular $\rho_{t,j}^{1/2}$. The first point we address relates to the specification of the sampling or measurement budget. In this setting, we can interpret our budget of measurement resources in terms of the matrix $A$, in a natural way. Recall that in our original formulation, the constraint was imposed on the quantity $\sum_{t,j} \rho_{t,j}$. Under the matrix-vector formulation, this translates directly to a constraint on the sum of the squared entries of $A$. Thus, we can generalize the measurement budget constraint to the current setting by imposing a condition on the Frobenius norm of $A$. To account for the possibly random nature of the sensing matrix (as in traditional CS applications), we impose the constraint in expectation:

$$\mathbb{E}\left[\|A\|_F^2\right] = \mathbb{E}\left[\sum_{t,j} A_{t,j}^2\right] \leq B(n). \tag{1.68}$$

Note that, since the random matrices utilized in standard CS settings typically are constructed to have unit-norm columns, they satisfy this constraint when $B(n) = n$.

The second point results from the fact that each observation step will now comprise a number of noisy projection samples of $x$. This gives rise to another set of algorithmic parameters to specify how many measurements are obtained

in each step, and these will inherently depend on the sparsity of the signal being acquired. In general, we will denote by $m_t$ the number of rows in the measurement matrix utilized in step $t$.

The final point to address in this setting pertains to the refinement step. In the original DS formulation, because the measurement process obtained direct samples of the signal components plus independent Gaussian noises, the simple one-sided threshold test was a natural choice. Here the problem is slightly more complicated. Fundamentally the goal is the same—to process the current observations in order to accurately determine promising locations to measure in subsequent steps. However in the current setting, the decisions must be made using (on average) much less than one measurement per location. In this context, each refinement decision can itself be thought of as a coarse-grained model selection task.

We will discuss one instance of this procedure, which we call *Compressive Distilled Sensing* (CDS), corresponding to particular choices of the algorithm parameters and refinement strategy. Namely, for each step, indexed by $t = 1, 2, \ldots, T$, we will obtain measurements using an $m_t \times n$ sampling matrix $A_t$ constructed as follows. For $u = 1, 2, \ldots, m_t$ and $v \in \mathcal{I}_t$, the $(u, v)$-th entry of $A_t$ is drawn independently from the distribution $\mathcal{N}(0, \tau_t/m_t)$ where $\tau_t = B_t/|\mathcal{I}_t|$. The entries of $A_t$ are zero otherwise. Notice that this choice automatically guarantees that the overall measurement budget constraint $\mathbb{E}\left[\|A\|_F^2\right] \leq B(n)$ is satisfied. The refinement at each step is performed by coordinate-wise thresholding of the crude estimate $\widehat{x}_t = A_t^T y_t$. Specifically, the set $\mathcal{I}_{t+1}$ of locations to subsequently consider is obtained as the subset of $\mathcal{I}_t$ corresponding to locations at which $\widehat{x}_t$ is positive. This approach is outlined in Algorithm 1.2.

The final support estimate is obtained by applying the Least Absolute Shrinkage and Selection Operator (LASSO) to the distilled observations. Namely, for some $\lambda > 0$, we obtain the estimate

$$\widetilde{x} = \arg\min_{z \in \mathbb{R}^n} \|y_T - A_T z\|_2^2 + \lambda \|z\|_1, \tag{1.69}$$

and from this, the support estimate $\hat{\mathcal{S}}_{\mathrm{DS}} = \{j \in \mathcal{I}_T : \widetilde{x}_j > 0\}$ is constructed. The following theorem describes the error performance of this support estimator obtained using the CDS adaptive compressive sampling procedure. The result follows from iterated application of Lemmas 1 and 2 in [25], which are analogous to Lemmas 1.1 and 1.2 here, as well as the results in [26] which describe the model selection performance of the LASSO.

**Theorem 1.3.** *Let $x \in \mathbb{R}^n$ be a vector having at most $k(n) = n^{1-\beta}$ nonzero entries for some fixed $\beta \in (0, 1)$, and suppose that every nonzero entry of $x$ has the same value $\mu = \mu(n) > 0$. Sample $x$ using the compressive distilled sensing procedure described above with*

- *$T = T(n) = \max\{\lceil \log_2 \log n \rceil, 0\} + 2$ measurement steps,*
- *measurement budget allocation $\{B_t\}_{t=1}^T$ satisfying $\sum_{t=1}^T B_t \leq n$, and for which*

**Algorithm 1.2** (Compressive distilled sensing)**.**

**Input:**

Number of observation steps $T$
Measurement allocation sequence $\{m_t\}_{t=1}^T$
Resource allocation sequence $\{B_t\}_{t=1}^T$ satisfying $\sum_{t=1}^T B_t \leq B(n)$

**Initialize:**

Initial index set: $\mathcal{I}_1 = \{1, 2, \ldots, n\}$

**Distillation:**

For $t = 1$ to $T$
  Construct $m_t \times n$ measurement matrix:
    $A_t(u, v) \sim \mathcal{N}\left(0, \frac{B_t}{m_t|\mathcal{I}_t|}\right),\ u = 1, 2, \ldots, m_t,\ v \in \mathcal{I}_t$
    $A_t(u, v) = 0,\ u = 1, 2, \ldots, m_t,\ v \in \mathcal{I}_t^c$
  Observe: $y_t = A_t x + e_t$
  Compute: $\widehat{x}_t = A_t^T y_t$
  Refine: $\mathcal{I}_{t+1} = \{i \in \mathcal{I}_t : \widehat{x}_t > 0\}$
End for

**Output:**

Index sets $\{\mathcal{I}_t\}_{t=1}^T$
Distilled observations $\{y_t, A_t\}_{t=1}^T$

– $B_{t+1}/B_t \geq \delta > 1/2$, *and*
– $B_1 = c_1 n$ *and* $B_T = c_T\, n$ *for some* $c_1, c_T \in (0, 1)$.

*There exist constants* $c, c', c'' > 0$ *and* $\lambda = O(1)$ *such that if* $\mu \geq c\sqrt{\log\log\log n}$ *and the number of measurements collected satisfies* $m_t \geq c' \cdot k \cdot \log\log\log n$ *for* $t = 1, \ldots, T-1$ *and* $m_T \geq c'' \cdot k \cdot \log n$*, then the support estimate* $\hat{\mathcal{S}}_{\mathrm{DS}}$ *obtained as described above satisfies*

$$\mathrm{FDP}(\hat{\mathcal{S}}_{\mathrm{DS}}) \xrightarrow{P} 0\,, \quad \mathrm{NDP}(\hat{\mathcal{S}}_{\mathrm{DS}}) \xrightarrow{P} 0, \tag{1.70}$$

*as* $n \to \infty$.

A few comments are in order regarding results of Theorems 1.2 and Theorem 1.3. First, while Theorem 1.2 guaranteed recovery provided only that $\mu(n)$ be a growing function of $n$, the result in Theorem 1.3 is slightly more restrictive, requiring that $\mu(n)$ grow like $\sqrt{\log\log\log n}$. Even so, this still represents a dramatic improvement relative to the non-adaptive testing case in Section 1.3.1. Second, we note that Theorem 1.3 actually *requires* that the signal components

have *the same amplitudes* (or, more precisely, that their amplitudes be within a constant multiple of each other), whereas the result in Theorem 1.2 placed no restrictions on the values of the signal amplitudes relative to each other. In essence these two points arise from the choice of refinement procedure. Here, the threshold tests are no longer statistically independent as they were in the original DS formulation, and the methods employed to tolerate this dependence give rise to these subtle differences.

The effectiveness of CDS can also be observed in finite sample regimes. Here, we examine (by experiment) the performance of CDS relative to a non-adaptive compressed sensing that utilizes a random measurement matrix with i.i.d. zero-mean Gaussian entries. For both cases, the support estimators we consider are constructed as the positive components of the LASSO estimate that is obtained using the corresponding adaptive or non-adaptive measurements. Our application of the CDS recovery procedure differs slightly from the conditions of Theorem 1.3, in that we apply the LASSO to *all* of the adaptively collected measurements.

The results of the comparison are depicted in Figure 1.6. Each panel of the figure shows a scatter plot of the FDP and NDP values resulting from 1000 trials of both the CDS procedure and the non-adaptive sensing approach, each using a different randomly selected LASSO regularization parameter. For each trial, the unknown signals $x \in \mathbb{R}^n$ were constructed to have 128 nonzero entries of uniform (positive) amplitude $\mu$, and the SNR is fixed by the selection $\mu^2 = 12$. Panels (a)-(d) correspond to $n = 2^{13}$, $2^{14}$, $2^{15}$, and $2^{16}$ respectively, and the number of measurements in all cases was $m = 2^{12}$.

The measurement budget allocation parameters for CDS, $B_t$, were chosen so that $B_{t+1} = 0.75B_t$ for $j = 1, \ldots, T-2$, $B_1 = B_T$, and $\sum_{t=1}^{T} B_t = n$, where $n$ is the ambient signal dimension in each case. Measurement allocation parameters $m_t$ were chosen so that $\lfloor m/3 \rfloor = 1365$ measurements were utilized for the last step of the procedure, and the remaining $\lfloor 2m/3 \rfloor$ measurements were equally allocated to the first $T-1$ observation steps. Comparing the results across all panels of Figure 1.6, we see that CDS exhibits much less dependence on the ambient dimension than the non-adaptive procedure. As with the examples for DS above, we see that CDS is an effective approach to mitigate the "curse of dimensionality" here as well. In particular, the performance of CDS shows much less dependence on the ambient dimension $n$ than does the nonadaptive procedure.

In conclusion, we note that the result of Theorem 1.3 has successfully addressed our initial question, at least in part. We have shown that in some special settings, the CDS procedure can achieve similar performance to the DS procedure but using many fewer total measurements. In particular, the total number of measurements required to obtain the result in Theorem 1.3 is $m = O(k \cdot \log\log\log n \cdot \log\log n + k \log n) = O(k \log n)$, while the result of Theorem 1.2 required $O(n)$ total measurements. The discussion in this section demonstrates that it is possible to obtain the benefits of both adaptive sampling

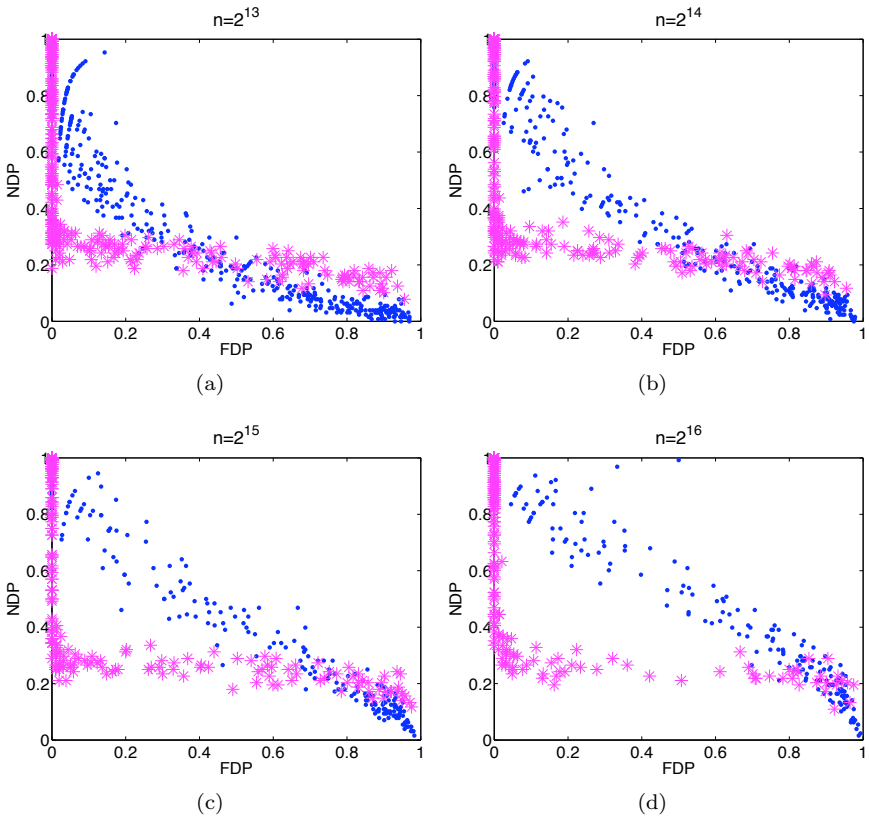


**Figure 1.6** Adaptivity in compressed sensing. Each panel depicts a scatter plot of FDP and NDP values resulting for non-adaptive CS (•) and the adaptive CDS procedure (∗).

*and* compressed sensing. This is a significant step toward a full understanding of the benefits of adaptivity in CS.

## 1.4 Related Work and Suggestions for Further Reading

Adaptive sensing methods for high-dimensional inference problems are becoming increasingly common in many modern applications of interest, primarily due to the continuing tremendous growth of our acquisition, storage, and computational abilities. For instance, multiple testing and denoising procedures are an integral component of many modern bioinformatics applications (see [27] and the references therein), and sequential acquisition techniques similar in spirit to those discussed here are becoming quite popular in this domain. In particular, two stage testing approaches in gene association and expression studies were examined in [28, 29, 30]. Those works described procedures where a large number of genes is initially tested to identify a promising subset, which is then

examined more closely in a second stage. Extensions to multi-stage approaches were discussed in [31]. Two-stage sequential sampling techniques have also been examined recently in the signal processing literature. In [32], two-stage target detection procedures were examined, and a follow-on work examined a Bayesian approach for incorporating prior information into such two-step detection procedures [33].

The problem of target detection and localization from sequential compressive measurements was recently examined in [34]. That work examined a multi-step binary bisection procedure to identify signal components from noisy projection measurements, and provided bounds for its sample complexity. Similar adaptive compressive sensing techniques based on binary bisection were examined in [35]. In [36], an adaptive compressive sampling method for acquiring wavelet-sparse signals was proposed. Leveraging the inherent tree structure often present in the wavelet decompositions of natural images, that work discussed a procedure where the sensing action is guided by the presence (or absence) of significant features at a given scale to determine which coefficients to acquire at finer scales.

Finally, we note that sequential experimental design continues to be popular in other fields as well, such as in computer vision and machine learning. We refer the reader to the survey article [37] as well as [38, 39] and the references therein for further information on active vision and active learning.

# References


[1] Leadbetter MR, Lindgren G, Rootzen H. Extremes and Related Properties of Random Sequences and Processes. Springer-Verlag; 1983.

[2] Donoho D. Compressed sensing. IEEE Transactions on Information Theory. 2006 April;52(4):1289–1306.

[3] Candes E, Tao T. Near optimal signal recovery from random projections: Universal encoding strategies? IEEE Transactions on Information Theory. 2006 December;52(12):5406–5425.

[4] Baraniuk R, Davenport M, DeVore R, Wakin M. A simple proof of the restricted isometry property for random matrices. Constructive Approximation. 2008 December;28(3):253–263.

[5] Candes E. The restricted isometry property and its implications for compressed sensing. C R Math Acad Sci Paris. 2008;346:589–592.

[6] Candes E, Plan Y. Near-ideal model selection by $\ell_1$ minimization. Annals of Statistics. 2009;37(5A):2145–2177.

[7] Haupt J, Nowak R. Signal reconstruction from noisy random projections. IEEE Transactions on Information Theory. 2006;52(9):4036–4048.

[8] Crouse M, Nowak R, Baraniuk R. Wavelet -Based Statistical Signal Processing using Hidden Markov Models. IEEE TransSignal Processing. 1998;46(4):886–902.

[9] Jacobson G. Space-efficient static trees and graphs. In: Proc. 30th Annual Symp. on Foundations of Comp. Sci.; 1989. p. 549–554.

[10] Huang J, Zhang T, Metaxas D. Learning with structured sparsity. In: ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning. New York, NY, USA: ACM; 2009. p. 417–424.

[11] Baraniuk R, Cevher V, Duarte M, Hedge C. Model-based Compressive Sensing. IEEE Trans Info Th. 2010;56(4):1982–2001.

[12] Seeger M. Bayesian Inference and Optimal Design in the Sparse Linear Model. Journal of Machine Learning Research. 2008;9:759–813.

[13] Seeger M, Nickisch H, Pohmann R, Schoelkopf B. Optimization of k-Space Trajectories for Compressed Sensing by Bayesian Experimental Design. Magnetic Resonance in Medicine. 2009;63:116–126.

[14] DeGroot M. Uncertainty, information, and sequential experiments. Annals of Mathematical Statistics. 1962;33(2):404–419.

[15] Lindley D. On the measure of the information provided by an experiment. Annals of Mathematical Statistics. 1956;27(4):986–1005.

[16] Castro R, Haupt J, Nowak R, Raz G. Finding needles in noisy haystacks. In: Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing. Las Vegas, NV;

2008. p. 5133–5136.

[17] Cover T, Thomas J. Elements of Information Theory. 2nd ed. Wiley; 2006.

[18] Ji S, Xue Y, Carin L. Bayesian Compressive Sensing. IEEE Transactions on Signal Processing. 2008 June;56(6):2346–2356.

[19] Tipping M. Sparse Bayesian learning and the relevance vector machine. Journal of Machine Learning Research. 2001 June;1:211–244.

[20] Seeger M. Bayesian Inference and Optimal Design for the Sparse Linear Model. Journal of Machine Learning Research. 2008 April;9:759–813.

[21] Seeger M, Nickisch H. Compressed Sensing and Bayesian Experimental Design. In: Proc. Intl. Conf. on Machine Learning; 2008. .

[22] Haupt J, Castro R, Nowak R. Distilled Sensing: Adaptive sampling for sparse detection and estimation. submitted manuscript. 2010 January;Preprint available online at `http://www.ece.umn.edu/~jdhaupt/publications/sub10_ds.pdf`.

[23] Donoho D, Jin J. Higher criticsm for detecting sparse heterogenous mixtures. Annals of Statistics. 2004;32(3):962–994.

[24] Chernoff H. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. Annals of Statistics. 1952;23(4):493–507.

[25] Haupt J, Baraniuk R, Castro R, Nowak R. Compressive distilled sensing: Sparse recovery using adaptivity in compressive measurements. In: Proc. 43rd Asilomar Conf. on Signals, Systems, and Computers. Pacific Grove, CA; 2009. p. 1551–1555.

[26] Wainwright M. Sharp thresholds for high-dimensional and noisy sparsity pattern recovery using $\ell_1$ constrained quadratic programming (Lasso). IEEE Transactions in Information Theory. 2009 May;55(5):2183–2202.

[27] Dudoit S, van der Laan M. Multiple testing procedures with applications to genomics. Springer Series in Statistics. Springer; 2008.

[28] Muller HH, Pahl R, Schafer H. Including Sampling and Phenotyping Costs into the Optimization of Two Stage Designs for Genomewide Association Studies. Genetic Epidemiology. 2007;31(8):844–852.

[29] Zehetmayer S, Bauer P, Posch M. Two-stage Designs for Experiments with large number of hypotheses. Bioinformatics. 2005;21(19):3771–3777.

[30] Satagopan J, Elston R. Optimal two-stage Genotyping in Population-based association studies. Genetic Epidemiology. 2003;25(2):149–157.

[31] Zehetmayer S, Bauer P, Posch M. Optimized Multi-Stage Designs Controlling the False Discovery or the Family-wise Error Rate. Statistics in Medicine. 2008;27(21):4145–4160.

[32] Bashan E, Raich R, Hero A. Optimal two-stage search for sparse targets using convex criteria. IEEE Transactions on Signal Processing. 2008 November;56(11):5389–5402.

[33] Newstadt G, Bashan E, Hero A. Adaptive search for sparse targets with informative priors. In: Proc. International Conference on Acoustics, Speech, and Signal Processing. Dallas, TX; 2010. p. 3542–3545.

[34] Iwen M, Tewfik A. Adaptive group testing strategies for target detection and localization in noisy environments. submitted. 2010 June;.

[35] Aldroubi A, Wang H, Zarringhalam K. Sequential adaptive compressed sampling via Huffman codes. preprint. 2009;`http://arxiv.org/abs/0810.4916v2`.

[36] Deutsch S, Averbuch A, Dekel S. Adaptive compressed image sensing based on wavelet modeling and direct sampling. In: Proc. 8th International Conference on Sampling Theory and Applications. Marseille, France; 2009. .

[37] Various authors. Promising Directions in Active Vision. International Journal of Computer Vision. 1991;11(2):109–126.

[38] Cohn D. Neural Network Exploration Using Optimal Experiment Design. In: Advances in Neural Information Processing Systems (NIPS); 1994. p. 679–686.

[39] Cohn D, Ghahramani Z, Jordan M. Active learning with statistical models. Journal of Artificial Intelligence Research. 1996;4:129–145.

# Index