CHAPTER 5

General theory and further applications

Residual statics estimation is only one of many inverse problems that arise in the analysis of geophysical data. Geophysical inverse problems are usually solved by minimizing a function defined over the space of model parameters—the same approach used for estimating statics. Questions of uniqueness aside, these optimization problems can be broadly classified into two types: those that contain only one minimum, and those that contain more than one. The type with one minimum is called linear inversion because the solution may be obtained by solving a set of linear equations. When there are many minima no such set of linear equations exists unless a good initial guess can be made. In the absence of a good initial guess, the problem becomes *nonlinear* inversion. In general, nonlinear inverse problems pose the same computational difficulties that occur in residual statics estimation: a global minimum must be located among many local minima.

I present in this chapter a general theory for the use of simulated annealing to solve nonlinear inverse problems. My goal is to define a class of problems for which simulated annealing can be a practical tool for nonlinear inversion. After developing the theory, I return to the problem of residual statics estimation to show how it conforms to this general class of problems. I then propose three further applications of simulated annealing in reflection seismology: frequency-dependent statics estimation, deconvolution, and velocity estimation. In concluding the chapter, I briefly examine two potentially useful concepts from statistical mechanics.

5.1 NONLINEAR INVERSION AND OPTIMIZATION

Consider a physical system (for example, the Earth) that is characterized by a finite set of unknown model parameters $\mathbf{X} = \{X_1, \ldots, X_M\}$ in the M-dimensional parameter space. An experiment performed in this physical system produces a finite set of data $\mathbf{D} = \{D_1, \ldots, D_R\}$ in the R-dimensional data space. $\mathbf{X}$ and $\mathbf{D}$ are (random) variables that assume specific values $\mathbf{x} = \{x_1, \ldots, x_M\}$ or $\mathbf{d} = \{d_1, \ldots, d_R\}$.

Let a set of R (nonlinear) functions G_i be denoted by $\mathbf{G}$. $\mathbf{G}$ is a function of the model parameters $\mathbf{x}$ and describes the theoretical outcome of the experiment. The observed data $\mathbf{d}$ is contaminated by additive noise and is represented as

$$\mathbf{d} = \mathbf{G}(\mathbf{x}) + \mathbf{n} \quad , \tag{5.1}$$

where $\mathbf{n} = \{n_1, \ldots, n_R\}$ is a realization of the random noise $\mathbf{N}$, which is assumed to be independent, identically distributed, and independent of $\mathbf{X}$. It is assumed that only discrete values are allowed when $\mathbf{X}$, $\mathbf{D}$, or $\mathbf{N}$ is the argument of a probability distribution.

Solving equation (5.1) for an $\mathbf{x}$ that represents the entire underlying set of model parameters is the most ambitious of geophysical inverse problems. In a more realistic approach, $\mathbf{x}$ represents a vector of parameters for a smaller problem, in which many physical quantities are previously defined. In such an approach, for example, $\mathbf{x}$ represents shot and receiver statics, and it is assumed that stacking velocities are known.

Inversion of (5.1) is generally performed by optimization. One seeks the $\mathbf{x}$ that solves the minimization problem

$$\min_{\mathbf{x}} f\ [\ \mathbf{d},\ \mathbf{G}(\mathbf{x})\ ] \quad . \tag{5.2}$$

In least-squares formulations, f is typically the squared difference between $\mathbf{G}(\mathbf{x})$ and $\mathbf{d}$. Using $\mathbf{G}(\mathbf{x})$ implies that forward modeling is done; alternatively one can minimize a function that depends on $\mathbf{G}^{-1}$, the (approximate) inverse of $\mathbf{G}$. In this case the optimization problem is

$$\min_{\mathbf{x}} f\ [\ \mathbf{G}^{-1}(\mathbf{d};\mathbf{x})\ ] \quad . \tag{5.3}$$

$\mathbf{G}^{-1}$ operates on $\mathbf{d}$, but the precise form of $\mathbf{G}^{-1}$ may depend on $\mathbf{x}$.

Whether approach (5.2) or (5.3) is chosen, optimization remains a basic problem. In many geophysical inverse problems the function f, whatever its form, is rife with local minima. One might naively suggest that an exhaustive evaluation of all possible solutions be made. As we have seen with residual statics estimation, however, this suggestion is usually not practicable. Specifically, if M parameters can each assume one of N values, then there are N^M possible solutions, usually far too many for an exhaustive search to be performed.

Despite this assumed complexity, equations (5.2) and (5.3) do not necessarily pose intractable optimization problems. The key to their solution, in a general sense, is prior information. For example, conventional approaches to nonlinear inversion usually

incorporate an initial guess $\mathbf{x}^0$ for $\mathbf{x}$. The remaining perturbation $\Delta\mathbf{x} = \mathbf{x} - \mathbf{x}^0$ is then assumed to approximately satisfy the linear relation

$$\mathbf{G}(\mathbf{x}) \approx \mathbf{G}(\mathbf{x}^0) + \mathbf{F}\Delta\mathbf{x} \ ,$$

where $\mathbf{F}$ is a matrix of partial derivatives $\partial G_i / \partial x_j$ evaluated at $\mathbf{x}^0$. Then, letting $\mathbf{d}^0 = \mathbf{G}(\mathbf{x}^0) + \mathbf{n}$ and $\Delta\mathbf{d} = \mathbf{d} - \mathbf{d}^0$, one solves for the $\Delta\mathbf{x}$ that satisfies

$$\mathbf{F}\Delta\mathbf{x} = \Delta\mathbf{d} \ .$$

This yields a possible solution $\mathbf{x}^1 = \mathbf{x}^0 + \Delta\mathbf{x}$, which may or may not be satisfactory for minimizing (5.2) or (5.3). If it is not, the procedure is iterated by replacing $\mathbf{x}^0$ with $\mathbf{x}^1$, etc. Iterative techniques of this general form are widely used—reviews are contained in Parker (1977), Aki and Richards (1980), and Lines and Treitel (1984). The basic shortcoming of these techniques, however, is their reliance upon a good initial guess, without which they might fail severely.

What can be done, then, if there is no basis for an initial guess? Prior information, in the form of Bayesian inference, may still light the way. If a prior probability distribution $P(\mathbf{X}{=}\mathbf{x})$ can be formulated, low probabilities can be assigned to much of the parameter space; that part of the parameter space is thereby effectively eliminated from consideration. The prior distribution represents relative weights assigned to all possible $\mathbf{x}$ before any data are collected. After the data are observed, standard statistical analysis is used to combine the prior distribution and the observed data to obtain the posterior probability distribution $P(\mathbf{X}{=}\mathbf{x} \mid \mathbf{D}{=}\mathbf{d})$ via Bayes' theorem (Box and Tiao, 1973):

$$P\ (\mathbf{X} = \mathbf{x}\ \mid\ \mathbf{D} = \mathbf{d}\ ) \ = \ \frac{P\ (\mathbf{D} = \mathbf{d}\ \mid\ \mathbf{X} = \mathbf{x}\ )\ P\ (\mathbf{X} = \mathbf{x}\ )}{P\ (\mathbf{D} = \mathbf{d}\ )} \ . \qquad (5.4)$$

With this formulation, one method of solution is to find the $\mathbf{x}$ that maximizes the posterior probability. This step is often called maximum a posteriori (MAP) estimation, and presents yet another optimization problem.

One interesting generalized approach to the incorporation of prior information in nonlinear inversion is described by Tarantola and Valette (1982). Although their formalism is appealing, the generalized use of prior information alone does not necessarily make a problem computationally tractable. Successful nonlinear geophysical inversion often requires more—specifically, a method that employs available knowledge to obtain a *computable* solution. Simulated annealing is one such method.

In the next section, I define a class of problems for which the prior distribution is the Gibbs distribution of statistical mechanics. Then, following Geman and Geman

(1984), I show in §5.3 how a Bayesian formulation leads to an expression for the posterior distribution that is also a Gibbs distribution. The application of simulated annealing then arises naturally from the desire to find the value of the model parameters that maximizes the posterior probability distribution.

5.2 GIBBS-MARKOV MODELS

Large-scale inverse problems can often be subdivided naturally to a smaller, computationally more convenient scale. The surface-consistent analysis of seismic data, for example, can be divided into related sub-units, the size of which depends on the length of the seismic cable (typically much less than the length of a seismic survey). The analysis that follows attempts to unify estimates of parameters made in sub-units with estimates that are consistent with the entire dataset. The statistical model I use can be described in either of two ways that have been shown to be formally equivalent. One model, a *Markov random field*, is a microscopic viewpoint derived from probability theory. The other model, the *Gibbs distribution* of statistical mechanics, is a macroscopic description of the problem. The goal is to divide a large problem into smaller, more manageable parts while maintaining the relationships between the smaller parts as precisely as possible. Ultimately, this goal leads to the application of simulated annealing for the solution of a large-scale nonlinear inverse problem. The Gibbs-Markov model forms the foundation of this objective.

Markov random fields (Kindermann and Snell, 1980) describe the microscopic structure common to the particular class of problems I address. A Markov random field is the spatial counterpart of the one-dimensional Markov chain. Recall that the simplest form of a Markov chain is a stationary sequence in which the conditional probability of an event at time t depends only on the value of the sequence at time $t-1$ [see equation (3.11)]. A straightforward generalization leads to the consideration of a two-dimensionally indexed set of random variables $\mathbf{X} = \{X_{ij}\}$. The X_{ij} define a Markov random field if the value of each X_{ij} depends only on a *neighborhood* A_{ij} of (i, j). A_{ij} might contain only the nearest neighbors of x_{ij}:

$$A_{ij} = \{ (i+1, j), (i-1, j), (i, j+1), (i, j-1) \} \quad . \tag{5.5}$$

Higher dimensionality and more complex neighborhoods are possible. In general, neighborhoods contain only those model parameters that most immediately influence the values a given parameter may assume. A two-dimensional Markov random field with neighborhood A_{ij} is stated as

$$P[X_{ij}=x_{ij} \mid X_{kl}=x_{kl}, (k,l)\neq(i,j)] = P[X_{ij}=x_{ij} \mid X_{kl}=x_{kl}, (k,l)\in A_{ij}] \quad . \tag{5.6a}$$

It is also required that all possible parameter vectors have positive probability:

$$P(\mathbf{X} = \mathbf{x}) > 0 \text{ for all } \mathbf{x} . \tag{5.6b}$$

Figure 5.1 shows an example of a Markov random field on a two-dimensional lattice.

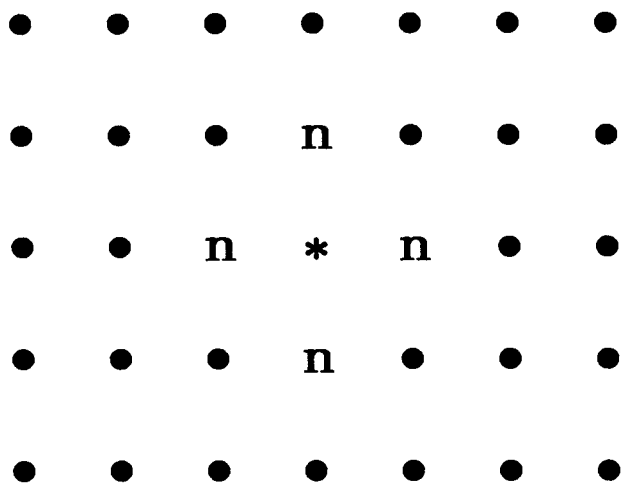

FIG. 5.1. A schematic representation of a Markov random field with nearest-neighbor interactions on a two-dimensional lattice. The probability distribution of the parameter located at position * depends only on its nearest neighbors, the n's; thus $P(\ * \mid \text{all else}) = P(\ * \mid \text{neighbors})$.

The notion of a Markov random field describes only the general form of local, conditional probabilities. While useful in concept, these local conditional probabilities provide information only on the microscopic interactions of the parameters, not on the behavior of the complete set of parameters taken as a whole. Information at the macroscopic level should be in the form of a joint probability distribution $P(\mathbf{X}=\mathbf{x})$. This form is fortunately available, because all Markov random fields exhibit a Gibbs probability distribution, and all Gibbs distributions define a Markov random field (Geman and Geman, 1984; Kindermann and Snell, 1980; Moussouris, 1974). The Gibbs distribution was introduced in §3.3.1. Recall that $\mathbf{X}$ has a Gibbs distribution if

$$P(\mathbf{X} = \mathbf{x}) = \frac{1}{Z} e^{\frac{-E(x)}{T}} . \tag{5.7}$$

In general, the energy $E(\mathbf{x})$ can be expressed as the sum of *local potentials* $V_C(\mathbf{x})$ such that

$$E(\mathbf{x}) = -\sum_C V_C(\mathbf{x}) \quad . \tag{5.8}$$

Each V_C depends only on a subset C of $\mathbf{X}$; each member of C is a neighbor of the other member. The subsets C are called *cliques*. For the two-dimensional model with neighborhoods A_{ij} given by equation (5.5), each clique contains one pair of parameters.

The Gibbs-Markov equivalence lends physical significance to the purely probabilistic concept of a Markov random field. If a problem can be divided into sub-units that satisfy equations (5.6a) and (5.6b), then physical insight gleaned from the statistical mechanics expressed by equations (5.7) and (5.8) is applicable to that problem. The key to intuitively connecting the physics with the statistics lies in the energy function (5.8). Here it is evident why Gibbs distributions and Markov random fields are closely related: the same neighborhoods limit the spatial range of both the conditional probabilities (5.6a) and the x_{ij} needed for the evaluation of the local potentials V_C. To prove formally that a Gibbs distribution defines a Markov random field, define $\mathbf{X} = \{X_{ij}\}$ to be a random variable with a Gibbs distribution. Then

$$P[X_{ij}=x_{ij} \mid X_{kl}=x_{kl}, (k,l)\neq(i,j)] = \frac{\exp\{-E(\mathbf{x})/T\}}{\sum_{\mathbf{x}^{ij}}\exp\{-E(\mathbf{x}^{ij})/T\}} \quad , \tag{5.9}$$

where $\mathbf{x}^{ij} = \mathbf{x}$ everywhere except possibly at x_{ij}. Substituting equation (5.8) into equation (5.9),

$$P[X_{ij}=x_{ij} \mid X_{kl}=x_{kl}, (k,l)\neq(i,j)] = \frac{\exp\{\sum_C V_C(\mathbf{x})/T\}}{\sum_{\mathbf{x}^{ij}}\exp\{\sum_C V_C(\mathbf{x}^{ij})/T\}} \quad . \tag{5.10}$$

Because $\mathbf{x}^{ij}$ can differ from $\mathbf{x}$ only at x_{ij}, $V_C(\mathbf{x}) = V_C(\mathbf{x}^{ij})$ for any clique C that does not contain x_{ij}. Thus all terms corresponding to cliques that do not contain x_{ij} cancel from the numerator and denominator in equation (5.10). The remaining terms then include only x_{ij} and its neighbors x_{kl}, $(k,l)\in A_{ij}$. Thus the Markov probabilities of equation (5.6a) hold, and it is therefore shown that a Gibbs distribution defines a Markov random field.

The utility of the Gibbs representation will be evident shortly. Note now, however, not only that a large problem has been subdivided into smaller parts (Markov probabilities), but also that the joint Gibbs distribution describes the presumed interactions of these individual parts by supplying the prior probability of any given

parameter *vector*. Moreover, given the prior, the posterior distribution can now be derived, and the statistical statement of the problem can be completed.

5.3 BAYESIAN FORMULATION

Assuming that $\mathbf{X}$ is a Markov random field, one may write the joint prior probability distribution for the model parameters as

$$P(\mathbf{X} = \mathbf{x}) = \frac{1}{Z} e^{\frac{-E(\mathbf{x})}{T}} . \tag{5.11}$$

Following Geman and Geman (1984), I now show that the posterior probability $P(\mathbf{X} = \mathbf{x} \mid \mathbf{D} = \mathbf{d})$ is also a Gibbs distribution.

Starting with Bayes' theorem [equation (5.4)], substitute the Gibbs prior for $P(\mathbf{X}=\mathbf{x})$ and take $P(\mathbf{D} = \mathbf{d})$ to be constant. We then obtain

$$P(\mathbf{X} = \mathbf{x} \mid \mathbf{D} = \mathbf{d}) = \frac{1}{Z} P(\mathbf{D} = \mathbf{d} \mid \mathbf{X} = \mathbf{x}) e^{\frac{-E(\mathbf{x})}{T}} \tag{5.12}$$

where Z is now a new constant. The noise $\mathbf{N} = \{N_1, \ldots, N_R\}$ is assumed to be independent, identically distributed, and independent of $\mathbf{X}$. For analytic convenience the probability distribution of the noise is assumed to be zero-mean with the form

$$P(\mathbf{N} = \mathbf{n}) = c^{-1} e^{-\frac{1}{2}\left(\frac{||\mathbf{n}||_p}{\sigma}\right)^p} \tag{5.13}$$

where c and σ are constants and $||\bullet||_p$ is the L^p norm such that $(||\mathbf{n}||_p)^p = R^{-1}\sum_i^r n_i^p$. If $p = 2$ the noise is Gaussian, and if $p = 1$ the noise is exponential.

Now solve for the posterior. Rewrite equation (5.12) as

$$\begin{aligned} P(\mathbf{X} = \mathbf{x} \mid \mathbf{D} = \mathbf{d}) &= \frac{1}{Z} P[\mathbf{D} = \mathbf{G}(\mathbf{x}) + \mathbf{n} \mid \mathbf{X} = \mathbf{x}] e^{\frac{-E(\mathbf{x})}{T}} \\ &= \frac{1}{Z} P[\mathbf{N} = \mathbf{d} - \mathbf{G}(\mathbf{x}) \mid \mathbf{X} = \mathbf{x}] e^{\frac{-E(\mathbf{x})}{T}} . \end{aligned}$$

Because $\mathbf{N}$ is independent of $\mathbf{X}$,

$$P(\mathbf{X} = \mathbf{x} \mid \mathbf{D} = \mathbf{d}) = \frac{1}{Z} P[\mathbf{N} = \mathbf{d} - \mathbf{G}(\mathbf{x})] e^{\frac{-E(\mathbf{x})}{T}}$$

and by substituting from equation (5.13), we obtain

$$P(\mathbf{X} = \mathbf{x} \mid \mathbf{D} = \mathbf{d}) = \frac{1}{Z} e^{\frac{-E(\mathbf{x})}{T} - \frac{1}{2}\left(\frac{||\mathbf{n}||_p}{\sigma}\right)^p}$$

where Z is again a new normalizing constant. By writing

$$E'(\mathbf{x},\mathbf{d}) = E(\mathbf{x}) + \frac{T}{2}\left(\frac{||\mathbf{d} - \mathbf{G}(\mathbf{x})||_p}{\sigma}\right)^p$$

one sees that the posterior distribution is also Gibbs with energy function $E'(\mathbf{x},\mathbf{d})$:

$$P(\mathbf{X} = \mathbf{x} \mid \mathbf{D} = \mathbf{d}) = \frac{1}{Z} e^{\frac{-E'(\mathbf{x},\mathbf{d})}{T}} . \tag{5.14}$$

Note that the form of the Gibbs posterior is unaffected by the value of p; thus the noise need not be Gaussian.

Geman and Geman (1984) derive some additional results showing that the posterior neighborhood structure is slightly modified to include "second-order" neighbors (i.e., neighbors of neighbors). For computational purposes, however, the prior and posterior neighborhood structures can be taken to be equal.

The model parameters that fit the data best, from the viewpoint of Bayesian inference, are determined by maximizing the posterior probability (5.14). This is *maximum a posteriori* (MAP) estimation. Maximizing this posterior probability by conventional gradient techniques is often not possible for nonlinear problems, because many local extrema may exist. However, because simulated annealing creates random solutions $\mathbf{x}$ drawn from a Gibbs distribution, the method is, in theory, drawing random solutions from the posterior distribution (5.14). The posterior distribution may be multimodal in a multidimensional space, but if T is chosen carefully the random solutions have a high probability of being near the global maximum of that distribution. In fact, under conditions of strict equilibrium, the probability of the maximum being attained is given by $P(\mathbf{X}=\mathbf{x}_{\max} \mid \mathbf{D}=\mathbf{d})$. As T decreases, this probability becomes greater.

5.4 RESIDUAL STATICS ESTIMATION: A NEW PERSPECTIVE

The analysis in the previous sections yields the following conclusions. If a problem conforms to a Gibbs-Markov model, then the prior probability distribution of the model parameters can be a Gibbs distribution. Furthermore, if the noise is independent of the model parameters, then the posterior probability distribution of the model parameters is also a Gibbs distribution, but now the energy is explicitly a function of the model parameters *and* the data. In the previous chapters, no distinction was made between prior and posterior energy functions; the energy function implicitly included

the influence of the data [see, for example, equation (3.3)]. The statics problem was modeled with a Gibbs distribution, and simulated annealing was employed to obtain the solution with the greatest stack power. Using the perspective of this chapter, one can now see that this method is equivalent to obtaining the most probable solution, given the Gibbs prior and the observed data.

The structure of the statics problem is closely related to that of a Markov random field, though the statics problem is not in the form of a simple two-dimensional lattice. See Figure 5.2, and then refer back to Figure 5.1. Because the seismic cable is shorter than the length of a seismic survey, an individual shot or receiver static affects the stack of only those CMP gathers to which that shot or receiver contributes seismograms. The relevant midpoints physically span a cablelength. Thus, as Figure 5.2 shows, an individual stack-power calculation depends only on the value of the shot and receiver statics located within a cablelength of the shot (or receiver) static of interest. These shot and receiver statics are the "nearest neighbors" in the sense used in a Gibbs-Markov model.

Local potentials also have a well-defined role. Using the notation of Chapter 3, define the power in the stack of a single CMP gather y by

$$V_{C_y}(\mathbf{s},\mathbf{r}) \;=\; \sum_t \left[\sum_h d_{yh}\left(t + s_{i(y,h)} + r_{j(y,h)} \right) \right]^2 \;.$$

The objective function for statics estimation [equation (3.3)] can now be rewritten as

$$E(\mathbf{s},\mathbf{r}) \;=\; -\sum_y V_{C_y}(\mathbf{s},\mathbf{r}) \;. \tag{5.15}$$

This form of the objective function is the practical implementation of the general statement given by equation (5.8). In equation (5.15), each V_{C_y} depends only on a subset C_y of $\mathbf{s}$ and $\mathbf{r}$. Each member of C_y is a neighbor of each other member; the members of C_y are simply the shot and receiver statics associated with CMP gather y. Thus CMP gathers assume the role of the cliques described previously.

A nearest-neighbor model was implicitly used in the description of the Metropolis algorithm given in §3.3.2. For example, equation (3.5), which describes the contribution to stack power from shot static s_i, can be rewritten as

$$\phi_{s_i} \;=\; \sum_{y \in Y_{s_i}} V_{C_y}(\mathbf{s},\mathbf{r}) \;. \tag{5.16}$$

The range of the summation is limited by the subset Y_{s_i} of all midpoints y. The shot and receiver statics associated with these midpoints are the nearest neighbors of s_i.

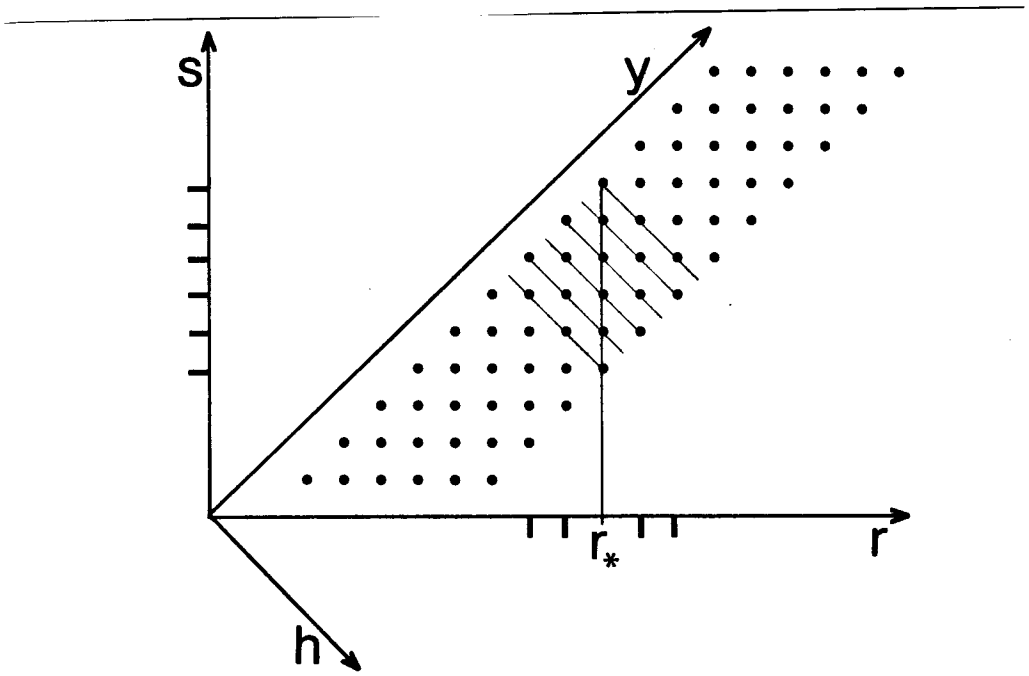

FIG. 5.2. Schematic representation of a seismic survey conducted with a 6-trace cable. The shot, midpoint, receiver, and offset axes are denoted by s, y, r, and h, respectively. Each dot represents a single seismic trace (the time axis may be taken to go into the page). Each trace is uniquely specified by any 2 of the 4 coordinate axes. In this picture, all traces recorded at receiver location r_* are connected by a vertical line. Calculation of the stack power ϕ_{r_*} given by equation (3.6) requires summing over the h-axis for all midpoint gathers containing a trace recorded at location r_*; this sum over h is depicted by the diagonal lines perpendicular to the y-axis. This calculation of stack power depends only on the shot statics and receiver statics located on the s- and r-axes at the positions marked by a dash. These are the "nearest neighbors" of r_* in the sense used in a Gibbs-Markov model.

The computations in the one-step heat-bath algorithm are limited by the same nearest-neighbor model. Equation (3.7), the transformation of a crosscorrelation function to a probability distribution, is essentially the same as equation (5.9). Simply equate the crosscorrelation function $q_m(\tau_p)$ to the energy $E(\mathbf{x})$. The entire objective function need not be calculated, however; only that part of it which is affected by the mth parameter needs to be computed. This local computation is the surface-consistently averaged crosscorrelation function given by equation (3.8). Equation (3.8) is thus the same as equation (5.16) with normalized crosscorrelation substituted for stack power.

5.5 FUTURE APPLICATIONS

Several other problems in reflection seismology can be placed in the framework of a Gibbs-Markov model. For the model to be successfully applied, it appears that the computations required for the evaluation of the local potentials V_C associated with an individual parameter must not require too much effort. Specifically, the computations of V_C required for a single update of one parameter must be considerably easier than the computations required for the evaluation of the entire objective function $E(\mathbf{x})$. If this is not the case, $E(\mathbf{x})$ must itself be easily computable.

In this section, I mention three possible future applications: frequency-dependent statics estimation, deconvolution, and velocity estimation. I survey each in the form of a brief proposal; considerable further research is required in each case.

5.4.1 Frequency-dependent statics estimation

In theory, the residual statics algorithm can be extended to the estimation of frequency-dependent statics. The statics model discussed thus far uses simple linear phase shifts. In a frequency-dependent model, however, phase shifts can be a more general function of frequency.

The problem of frequency-dependent statics is difficult to solve with traveltime models similar to equation (3.2) because phase shifts greater than $|\pi|$ are computationally ambiguous (Sword, 1983). Some form of phase unwrapping (Tribolet, 1979) is usually thought to be necessary. In principle, adaptation of the present statics algorithm is straightforward and requires only the application of two elementary theorems from Fourier transform theory (Bracewell, 1978). The Rayleigh-Parseval theorem states that power in the time domain equals power in the frequency domain. So for a function $f(t)$ and its Fourier transform $F(\omega)$,

$$\sum_t |\, f(t)\,|^2 = \sum_\omega |\, F(\omega)\,|^2 \ .$$

In addition, the shift theorem states that the Fourier analog of a time shift is multiplication by a complex exponential:

$$f(t-\tau) \supset e^{i\omega\tau} F(\omega) \ .$$

Then by letting the Fourier transform of $d_{yh}(t)$ be denoted by $D_{yh}(\omega)$, one can include frequency dependence in equation (3.5) by writing

$$\phi_{s_i}[\mathbf{s}(\omega), \mathbf{r}(\omega)] = \sum_{y \in Y_{s_i}} \sum_\omega \left| \sum_h e^{i\omega[\, s_{i(y,h)}(\omega) + r_{j(y,h)}(\omega)\,]} D_{yh}(\omega) \right|^2 . \qquad (5.17)$$

Similar changes can be made to equation (3.6). Note that the s_i and r_j are now functions of ω.

As I have noted previously, solutions to statics estimation problems are inherently nonunique. This problem of nonuniqueness is worse in the frequency-dependent case if each ω-component is treated independently of the others. A physical model of frequency-dependent phase shifts should therefore require that the phase shifts be locally correlated with each other. This condition may be incorporated into equation (5.17) by smoothing $\mathbf{s}(\omega)$ and $\mathbf{r}(\omega)$ over ω. When these smoothed functions are represented by $\overline{\mathbf{s}}(\omega)$ and $\overline{\mathbf{r}}(\omega)$, the energy function for frequency-dependent statics is

$$E\,[\,\overline{\mathbf{s}}(\omega),\,\overline{\mathbf{r}}(\omega)\,]\;=\;-\sum_i \phi_{s_i}\,[\,\overline{\mathbf{s}}(\omega),\,\overline{\mathbf{r}}(\omega)\,]-\sum_j \phi_{r_j}\,[\,\overline{\mathbf{s}}(\omega),\,\overline{\mathbf{r}}(\omega)\,]\quad .$$

5.4.2 Deconvolution

Frequency-dependent statics estimation is closely related to surface-consistent deconvolution. Although the application of frequency-dependent statics would not change the power spectra of the data, it would equalize phase surface-consistently, and thus solve a considerable part of the deconvolution problem.

One of the fundamental problems in seismic deconvolution is the definition of an objective function. Objective functions have commonly been in the form of a sum-squared error resulting from, for example, the linear least-squares problem posed by prediction-error filtering (Claerbout, 1976). Deconvolution need not be a linear problem, however. For example, the method of minimum entropy deconvolution (Wiggins, 1978; Donoho, 1981) employs an objective function that cannot be minimized by linear least-squares methods; iterative descent from an initial guess is usually employed to estimate the coefficients of the deconvolution filter. Wiggins (1985) claims that suboptimal local minima are one of the drawbacks of the minimum entropy approach. My own tests (not shown) confirm this observation, but these tests also show that the problem of local minima is not nearly as severe in minimum entropy deconvolution as it is in statics estimation.

The prospect of applying simulated annealing to deconvolution creates new possibilities for the construction of suitable objective functions. Any function that adequately describes the objective of deconvolution is admissible; there should be no concern with local minima or even differentiability. When simulated annealing is applied to deconvolution, the model parameters would be the filter coefficients and the "neighborhoods" would encompass the entire filter.

5.4.3 Velocity estimation

Velocity estimation can also be viewed as an extension of the residual statics algorithm. Statics are essentially the components of a one-dimensional velocity function. In velocity estimation, a two-dimensional grid would be parameterized by velocity, and one would seek the velocity distribution yielding the maximum stack power. Toldi (1985) presents a method of stack-power maximization in which perturbations to a velocity model are linearly related to estimates of stacking velocities. Fowler (1985) discusses an extension of this technique in which perturbations to a similar velocity model are linearly related to estimates of time-migration velocities. In both of these methods, the results are dependent on an initial guess of the velocity model. Maximizing stack power by simulated annealing could be a valuable approach if the initial guess were difficult to obtain or if the dependence of the final solution on the initial guess were strong. The application of simulated annealing to velocity estimation might require too much computational effort, however. Construction of the appropriate form of the functions V_C would probably entail far more work than the simple shifts and sums needed for residual statics estimation. For more details, see Rothman (1985).

5.5 TWO USEFUL CONCEPTS FROM STATISTICAL MECHANICS

Many concepts from statistical mechanics may be useful in studying the application of simulated annealing to nonlinear inversion. I discuss two of these concepts below.

5.5.1 Critical temperature

The notion of a critical temperature, previously encountered in Chapter 3 and 4, is perhaps the single most important issue in the application of simulated annealing to statics estimation, and probably nonlinear inversion in general. As defined in physics, a critical temperature is the temperature at which a liquid changes to a solid, or the temperature at which a ferromagnetic substance acquires permanent magnetization. These examples of the spontaneous ordering of matter are called *phase transitions* and have been the object of extensive study [see, for example, Stanley (1971)]. In Monte Carlo statics estimation, the critical temperature T_c may be broadly defined to be the largest value of T that leads to significant correlations between shot and receiver statics. More generally, T_c is the temperature at which significant correlations between parameters extend well beyond the nearest neighbors. Convergence is possible only below T_c. As I discussed in Chapter 4, the critical temperature is presently estimated

empirically. An analytic approximation or an efficient empirical method of estimating T_c remains an open research problem.

5.5.2 Ergodic average

The second useful concept from statistical mechanics is the *ergodic average*. In the original formulation of Metropolis et al. (1953), the Monte Carlo algorithm was used to estimate the ergodic averages

$$< f(\mathbf{x}) > = \sum_{\mathbf{x}} f(\mathbf{x}) P(\mathbf{X} = \mathbf{x}) = \frac{\sum_{\mathbf{x}} f(\mathbf{x}) \exp\{-E(\mathbf{x})/T_1\}}{\sum_{\mathbf{x}} \exp\{-E(\mathbf{x})/T_1\}} , \quad (5.18)$$

for a given $T = T_1$ and some function $f(\mathbf{x})$. These averages are valid only if the system has reached equilibrium. Equilibrium is achieved quickly at high T_1 and, as demonstrated in Chapter 4, relatively slowly when $T_1 \leq T_c$. If one is reasonably certain that equilibrium has been attained, the generation of the ergodic averages (5.18) can be useful for estimating means, variances, covariances, and other statistical quantities. Thus one can obtain not only a simple answer (the maximum a posteriori solution) but also estimates of resolution and accuracy. Importantly, the posterior probability distribution itself can be estimated by constructing a histogram of the output of each iteration at constant temperature [and, formally speaking, by substituting $P(\mathbf{X}{=}\mathbf{x} \mid \mathbf{D}{=}\mathbf{d})$ for $P(\mathbf{X}{=}\mathbf{x})$ in equation (5.18)]. The posterior probability distribution contains the most fundamental information that can be provided by a solution to an inverse problem (Tarantola and Valette, 1982).