12

Analytic continuation

The presence of dynamical information is a feature distinguishing a finite-temperature quantum Monte Carlo simulation from a classical one. We now discuss numerical methods for extracting this information that use techniques and concepts borrowed from an area of probability theory called Bayesian statistical inference. The use of these techniques and concepts provided a solution to the very difficult problem of analytically continuing imaginary-time Green's functions, estimated by a quantum Monte Carlo simulation, to the real-time axis. Baym and Mermin (1961) proved that a unique mapping between these functions exists. However, executing this mapping numerically, with a simulation's incomplete and noisy data, transforms the problem into one without a unique solution and thus into a problem of finding a "best" solution according to some reasonable criterion. Instead of executing the analytic continuation between imaginary- and real-time Green's functions, thereby obtaining real-time dynamics, we instead estimate the experimentally relevant spectral density function these Green's functions share. We present three "best" solutions and emphasize that making the simulation data consistent with the assumptions of the numerical approach is a key step toward finding any of these best solutions.

12.1 Preliminary comments

The title of this chapter, "Analytic Continuation," is unusual in the sense that it describes the task we wish to accomplish instead of the method we use to accomplish it. If we used the name of the method, the title would be something like "Bayesian Statistical Inference Using an Entropic Prior." A shorter title would be "The Maximum Entropy Method." We hope by the end of the chapter the reader will agree that using the short title is perhaps too glib and the longer one has meaningful content.

https://doi.org/10.1017/CBO9780511902581.013 Published online by Cambridge University Press

The task we want to accomplish is solving the integral equation

$$G(\tau) = \int_{-\infty}^{\infty} d\omega\, K(\tau,\omega) A(\omega), \quad K(\tau,\omega) = \frac{e^{-\tau\omega}}{1 \pm e^{-\beta\omega}} \tag{12.1}$$

for $A(\omega)$ using quantum Monte Carlo estimates of $G(\tau)$. We review the origin of this equation in the next section (Section 12.2). For now, it suffices to say that $G(\tau)$ is some dynamical (imaginary-time) many-body correlation function, and the equation is a standard result from many-body theory relating this correlation function to $A(\omega)$, which is called the *spectral density*. We also discuss in the next section that solving this equation is a surrogate for analytically continuing the dynamical correlations from $\tau \to it$.

Solving this linear integral equation seems simple, as discretizing τ and ω reduces it to a linear system of equations

$$G_i = \sum_j K_{ij} A_j.$$

As the number of knowns does not necessarily equal the number of unknowns, a least-squares solution for the A_i has to be found. Such a calculation is easily implemented. The problem is that this approach to the analytic continuation problem almost always fails to produce an acceptable solution.

Similar tasks have for decades been known to be major computational challenges. The exponential character of the kernel $K(\tau,\omega)$ makes solving the analytic continuation problem akin to parameterizing a two-nuclei radioactive decay problem

$$n(t) = a_1 e^{-\lambda_1 t} + a_2 e^{-\lambda_2 t},$$

where $n(t)$ is the number of decays as a function of time t. As discussed passionately by Acton (1970), this is an ill-posed problem as small changes in even very precise, closely spaced measurements produce large changes in the fitted values of a_1, a_2, λ_1, and λ_2. On the other hand, if we know λ_1 and λ_2 (the decay rates of the nuclei), good estimates of a_1 and a_2 (the relative proportions of the nuclei) are easily obtained.

A common way to get reasonable solutions to the radioactive decay and similar problems is to regularize the least-squares solution. In mathematics and statistics, particularly for machine learning and inverse problems, regularization refers to adding additional information to the solution to help solve an ill-posed problem or to avoid overfitting a least-squares problem. In the present case, we are dealing with both types of problems, and we take "regularization" to mean adding one or more constraints to the fit of the data. If we add one constraint, we maximize

$$\chi^2(A) = \lambda R(A) - \frac{1}{2} \sum_i \left(G_i - \sum_j K_{ij} A_j \right)^2,$$

where λ is a Lagrange multiplier and $R(A)$ is some function of our unknown.

https://doi.org/10.1017/CBO9780511902581.013 Published online by Cambridge University Press

One generally tries to use functions R that are consistent with known properties of the expected solution. Doing this forces some of the prior knowledge into the fit. For example, we might know that $A(\omega)$ is smooth, and thus we might want to constrain the solution so that its first derivative with respect to ω is continuous. As we discuss in the next section (Section 12.2), in the analytic continuation problem, we know at least that $A(\omega)$ is nonnegative, and its integral over ω is finite (that is, we know the integral satisfies a physical sum rule). For a physical solution, we need to take into account both of these pieces of prior information. The magnitude of λ controls how strongly this prior information is imposed on the solution. Fixing the numerical value of λ such that it does not undesirably bias the solution is often a problem in itself.

We choose to solve (12.1) and to use the prior information we know about $A(\omega)$ from the point of view of Bayesian statistical inference. We discuss this statistical approach in Section 12.3. As we will see, to use this approach, we need to recast our problem into the language of probability theory. In the present case, this recasting has one possible method of solution, maximizing an unnormalized probability function that looks like

$$\exp\left(\lambda R(A) - \frac{1}{2}\sum_i \left(G_i - \sum_j K_{ij}A_j\right)^2\right),$$

which of course is equivalent to maximizing the argument of the exponential. Hence, at this level of analysis, the Bayesian approach is equivalent to a regularized least-squares method of solution. Finding the maximum (the mode) of a probability function, that is, finding a regularized least-squares solution, is unfortunately often inadequate. Ultimately, we present a method for estimating a mean solution relative to this probability.

A component in the Bayesian inference approach is choosing something called a *prior probability*, which replaces the regularizer. It represents our belief about the solution prior to (before) the data. We use an entropic prior: The nonnegativity and boundedness of the spectral density, our minimal prior information, allows us to pretend it is a probability density, and consequently it has an associated information theory entropy (Section 2.7). Additionally, the character this quasi-regularizer imposes on the solution is something well documented. For instance, in the absence of data, varying any one A_i does not require any specific one or more of the remaining A_i to change, apart from the changes required to maintain the sum rule (Section 12.3). In effect, this prior probability induces structure in the solution only if that structure is enforced by the data or by other prior knowledge. We can even go beyond the standard regularization and develop a method to determine the

Lagrange multiplier (Section 12.4). The Bayesian-based analysis also unveils that the kernel of the integral "reveals" to the data only a handful of effective (latent) parameters. These parameters are not the A_i, but rather the A_i as functions of these effective parameters (Section 12.5).

We begin by discussing general properties of dynamical correlation functions and the kernel that connects them with the spectral density and the spectral densities with sum rules. We then discuss the basic principles of Bayesian statistical inference. We have a nonlinear optimization problem, but because we need to determine only a handful of latent variables, we can use a method that finds the solution to this optimization problem with negligible computational cost. Appendix O details this method.

12.2 Dynamical correlation functions

The fluctuations of a system in thermal equilibrium are characterized by time-correlation functions of the type $\langle C(t)B(0)\rangle$, where B and C are operators. Linear response theory (Negele and Orland, 1988) tells us that the dynamical response of the system to these operators is described by a retarded Green's function (with $\hbar = 1$)

$$iG_R(t > 0) = \left\langle \left[C(t), B(0)\right]_{\pm} \right\rangle, \tag{12.2}$$

where the angular brackets denote thermal averaging and the operators $C(t)$ and $B(0)$ are in the Heisenberg representation. The sign on the commutator is determined by whether the operators B and C (in the Schrödinger representation) satisfy Fermionic $(+)$ or Bosonic $(-)$ commutation relations.

The spectral density $A(\omega)$ associated with this Green's function satisfies (Negele and Orland, 1988)

$$G_R(\omega + i\eta) = \int_{-\infty}^{\infty} d\omega' \, \frac{A(\omega')}{\omega - \omega' + i\eta}, \quad 0 < \eta \ll 1, \tag{12.3}$$

where the frequency Fourier transform of $G_R(t)$ is defined by

$$G_R(\omega) = \frac{1}{2\pi} \int_0^{\infty} dt \, e^{i\omega t} G_R(t). \tag{12.4}$$

Knowing $A(\omega)$ thus yields the real-time and frequency-dependent retarded Green's functions: Substituting $A(\omega)$ into (12.3) gives $G_R(\omega)$, and then taking the inverse Fourier transform yields

$$G_R(t) = \int_{\infty}^{\infty} d\omega \, e^{-i\omega t} G_R(\omega).$$

https://doi.org/10.1017/CBO9780511902581.013 Published online by Cambridge University Press

Finite-temperature Monte Carlo simulations compute the imaginary-time Green's functions[1]

$$G(\tau) = \langle C(\tau)B(0)\rangle, \quad 0 \le \tau < \beta. \tag{12.5}$$

Because this type of Green's function is antiperiodic (Fermions) or periodic (Bosons) in imaginary time, that is, $G(\tau) = \mp G(\tau + \beta)$, where β is the inverse temperature, the Fourier transform of this correlation function is

$$\hat{G}(i\omega_n) = \int_0^\beta d\tau\, e^{i\omega_n \tau} G(\tau), \tag{12.6}$$

where ω_n is a Matsubara frequency equal to $(2n+1)\pi/\beta$ for Fermion and $2n\pi/\beta$ for Boson operators.

Knowing $A(\omega)$ yields the imaginary-time and the Matsubara frequency-dependent Green's functions: Substituting $A(\omega)$ into

$$\hat{G}(i\omega_n) = \int_{-\infty}^{\infty} d\omega' \, \frac{A(\omega')}{i\omega_n - \omega'} \tag{12.7}$$

gives $\hat{G}(i\omega_n)$. Then, the inverse transform

$$G(\tau) = \frac{1}{\beta} \sum_n e^{-i\omega_n \tau} \hat{G}(i\omega_n)$$

yields $G(\tau)$. We see that $\hat{G}(i\omega_n)$ in (12.7) and $G_R(\omega + i\eta)$ in (12.3) are the analytic continuations of each other: $i\omega_n \leftrightarrow \omega + i\eta$. This continuation connects the real- and imaginary-time Green's functions.

If (12.7) is substituted into the inverse transform, it yields

$$G(\tau) = \int_{-\infty}^{\infty} d\omega \, \frac{e^{-\tau\omega}}{1 \pm e^{-\beta\omega}} A(\omega). \tag{12.8}$$

This is the basic equation of this chapter. Given a Monte Carlo estimate of $G(\tau)$ from a finite-temperature simulation, we wish to solve this equation for $A(\omega)$. With $A(\omega)$, we then know $G_R(t)$ (Bonča and Gubernatis, 1993b).

Most often we do not go all the way to $G_R(t)$ but instead stop with $A(\omega)$. We stop with $A(\omega)$ because experimentally interesting properties of the system, such as the optical conductivity, NMR relaxation time, dynamic spin structure factors, and the like are related to various spectral densities. These functions give a picture of how interactions change the nature of the eigenstates of the Hamiltonian. Often, they give direct information about the existence of quasi-particles and collective modes

[1] Here, we use the phase convention of (7.22).

https://doi.org/10.1017/CBO9780511902581.013 Published online by Cambridge University Press

(e.g., Kawashima et al., 1996). A tool to extract this type of information from a simulation clearly enhances the value of the simulation.

Solving (12.1) for $A(\omega)$, given $G(\tau)$, is extremely difficult. The difficulty is that the kernel of the integral equation,

$$K(\tau,\omega) = \frac{e^{-\tau\omega}}{1 \pm e^{-\beta\omega}}, \tag{12.9}$$

becomes exponentially small at large positive and negative frequencies. In the forward problem, "Given A, what is G?," this behavior suppresses the sensitivity of $G(\tau)$ to the large-$|\omega|$ features of $A(\omega)$. In the inverse problem, "Given G, what is A?," the large-$|\omega|$ features of $A(\omega)$ depend on subtle features of $G(\tau)$, which are compromised by noise and incompleteness. In short, with incomplete and noisy data we are trying to extract features of A to which the data are insensitive. Many different spectral densities fit the data equally well.

Typically, a Hamiltonian has several interesting spectral densities. For illustration, we now discuss two such densities for the single-impurity (spin-degenerate) Anderson model

$$H = \sum_{k\sigma} \epsilon_k n_{k\sigma} + \sum_{k\sigma} \left(V_k c^\dagger_{k\sigma} d_\sigma + V^*_k d^\dagger_\sigma c_{k\sigma} \right) + \varepsilon_d \sum_\sigma d^\dagger_\sigma d_\sigma + U n_{d\uparrow} n_{d\downarrow},$$

where ϵ_k is the conduction band energy of an electron in momentum state k, V_k is the strength of the hybridization of the impurity orbital and the band, and U is the strength of the Coulomb repulsion between two electrons if both occupy the impurity orbital.

Of particular interest are spectral densities associated with the impurity orbital. If $G^\sigma(\tau) = \langle \mathcal{T} d^\dagger_\sigma(\tau) d_\sigma(0) \rangle$ for an electron with spin σ and $A_\sigma(\omega)$ is the associated spectral density, it is convenient to define $G(\tau) = \frac{1}{2}\sum_\sigma G^\sigma(\tau)$ and $A(\omega) = \frac{1}{2}\sum_\sigma A_\sigma(\omega)$. Then, because $d^\dagger_\sigma$ and d_σ satisfy Fermionic anticommutation relations,

$$G(\tau) = \int_{-\infty}^{\infty} d\omega \, \frac{e^{-\tau\omega}}{1 + e^{-\beta\omega}} A(\omega).$$

We can also show that $A(\omega) \geq 0$ and obeys the sum rule (Negele and Orland, 1988)

$$\int_{-\infty}^{\infty} d\omega \, A(\omega) = 1.$$

The symmetric Anderson impurity model describes a situation where the energy bands are symmetric around $E_{\text{Fermi}} = 0$ and $\varepsilon_d = -\frac{1}{2}U$. For this case, $A(\omega) = A(-\omega)$, and our integral equation becomes ($0 \leq \tau < \beta$)

$$G(\tau) = \frac{1}{2} \int_0^{\infty} d\omega \, \frac{e^{-\tau\omega} + e^{-(\beta-\tau)\omega}}{1 + e^{-\beta\omega}} A(\omega),$$

https://doi.org/10.1017/CBO9780511902581.013 Published online by Cambridge University Press

where the new kernel is

$$K(\tau,\omega) = \frac{1}{2}\frac{e^{-\tau\omega} + e^{-(\beta-\tau)\omega}}{1 + e^{-\beta\omega}}$$

and the sum rule becomes

$$\int_0^\infty d\omega\, A(\omega) = \frac{1}{2}.$$

The new kernel is an even function of ω.

Also of interest is the two-particle Green's function ($0 \le \tau < \beta$)

$$\chi(\tau) = \left\langle d_\uparrow^\dagger(\tau) d_\downarrow(\tau) d_\uparrow^\dagger(0) d_\downarrow(0) \right\rangle = \frac{1}{\pi}\int_{-\infty}^{\infty} d\omega\, \frac{e^{-\tau\omega}}{1 - e^{-\beta\omega}} \mathrm{Im}\chi(\omega).$$

Because $d_\uparrow^\dagger d_\downarrow$ commutes with itself, the kernel is Bosonic. For this Green's function, the spectral density $\mathrm{sign}(\omega)\mathrm{Im}\chi(\omega) \ge 0$ is the transverse magnetic susceptibility, which satisfies the sum rule

$$\frac{1}{\pi}\int_{-\infty}^{\infty} d\omega\, \frac{\mathrm{Im}\chi(\omega)}{\omega} = \chi(T).$$

$\chi(T)$ is the magnetic susceptibility at temperature T. Another quantity of interest is

$$\frac{1}{T_1 T} = K \lim_{\omega\to 0} \frac{\mathrm{Im}\chi(\omega)}{\omega},$$

where K is some constant that depends on the details of the coupling between the impurity nucleus and the d-electron spin. T_1 is the nuclear magnetic relaxation time. Because $\mathrm{Im}\chi(\omega)$ is an odd function of frequency, working with the antisymmetrized kernel

$$K(\tau,\omega) = \frac{1}{2}\frac{e^{-\tau\omega} - e^{-(\beta-\tau)\omega}}{1 - e^{-\beta\omega}}$$

restricts the problem to just the domain of positive frequencies.

As we soon discuss, our procedure for selecting a "best" solution of the integral equation (12.1) is information theory based. Increasing the amount of embodied information generally increases the quality of the solution. A symmetry is precise information. Symmetrizing the kernel as was done in the previous examples is important when using the numerical methods we eventually describe.

12.3 Bayesian statistical inference

Central to a Bayesian method is the use of probability theory and, concomitantly, the use of Bayes's theorem. We discussed this theorem in Section 2.1. To review,

https://doi.org/10.1017/CBO9780511902581.013 Published online by Cambridge University Press

if we have two sets of events, $X = (X_1, X_2, \ldots, X_m)$ and $Y = (Y_1, Y_2, \ldots, Y_n)$, to which we have assigned probabilities, then Bayes's theorem says (Section 2.1)

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}.$$

We use this theorem in the following manner: A given Hamiltonian fixes $A(\omega)$. The $G(\tau)$ data produced by the simulation, which we now denote as $\bar{G}(\tau)$, are thus conditioned on $A(\omega)$. However, we want $A(\omega)$ conditioned on the data $\bar{G}(\tau)$. Accordingly,

$$P(A|\bar{G}) = \frac{P(\bar{G}|A)P(A)}{P(\bar{G})}. \tag{12.10}$$

From this Bayesian perspective, the probability of the spectral density given the data, that is, $P(A|\bar{G})$, is the solution to this problem. As with any other probability function, we reduce its vast amount of information to fewer characteristic metrics, such as modes, means, variances, and the like. Obtaining the probability of the spectral density is constructive, because faced with an infinite number of possible solutions, we have quantitatively assigned degrees of belief about our options. We can now investigate them and decide how to reasonably designate something as the "best" solution. For example, if the probability $P(A|\bar{G})$ has a single sharp peak in the space of functions $A(\omega)$, then we can reasonably take this most probable $A(\omega)$, the mode of $P(A|\bar{G})$, as our solution. In fact, as we increase the amount of data, we a priori expect this situation to occur. If we have a single peak, but it is skewed and broad, then selecting an average spectral density, such as $\int \mathcal{D}A\, A\, P(A|\bar{G})$, would seem reasonable. On the other hand, a multiply-peaked $P(A|\bar{G})$ would require a situation-specific analysis.

The various probabilities appearing in Bayes's theorem have names. The probability of A, that is, $P(A)$, is called the *prior probability*. It represents the probability of A prior (logically, not temporally) to the data. It is the task of the data, and other information we may add to the problem, to pull A away from this prior knowledge. The probability of the data given the spectral density, that is, $P(\bar{G}|A)$, is called the *likelihood function*, and the probability of the spectral density given the data, $P(A|\bar{G})$, is called the *posterior probability*. Finally, the probability of the data, $P(\bar{G})$, is called the *evidence*. The evidence normalizes the posterior probability. To show this, let us functionally integrate

$$P(A|\bar{G})P(\bar{G}) = P(\bar{G}|A)P(A)$$

over A on both sides of the equation. Because $\int \mathcal{D}A\, P(A|\bar{G}) = 1$,

$$P(\bar{G}) = \int \mathcal{D}A\, P(\bar{G}|A)P(A).$$

https://doi.org/10.1017/CBO9780511902581.013 Published online by Cambridge University Press

Comparing this result with (12.10) establishes the evidence as the normalization of the posterior probability when we construct the posterior probability from the product of the likelihood function and prior probability, which is what we do. Thus, the evidence, as well as the posterior probability, depends on the likelihood function and prior probability.

We now begin defining our choices for the prior probability and the likelihood function. In defining the prior probability, we appeal to the *theory of most probable distributions* and the *principle of maximum entropy*, so we first discuss these two important concepts.

12.3.1 Principle of maximum entropy

Schrödinger championed the theory of most probable distributions as a simple and unified approach to generate the distribution functions at the foundation of statistical mechanics (Schrödinger, 1952). To illustrate his point, he considered finding the distribution of an energy E over an ensemble of N identical independent systems, each in one of many possible energy states ε_i. With n_i defined as the number of systems in state i, the number of possible states having the set of occupation numbers $(n_1, n_2, \ldots, n_i, \ldots)$ is

$$\Omega(n_1, n_2, \ldots, n_i, \ldots) = \frac{N!}{n_1! n_2! n_3! \cdots n_i! \cdots}.$$

Given that the energy of the ensemble is E, the numbers n_i must satisfy the constraints $\sum_i n_i = N$ and $\sum_i \varepsilon_i n_i = E$. Seeking the maximum of

$$\ln \Omega - \lambda \sum_i n_i - \mu \sum_i \varepsilon_i n_i,$$

where λ and μ are Lagrange multipliers, he assumed that N and the n_i are large, and then after using Stirling's formula

$$\ln n! \approx n \ln n - n, \tag{12.11}$$

he showed that the probability p_i to be in state i is

$$p_i = \frac{n_i}{N} = \frac{e^{-\mu\varepsilon_i}}{\sum_i e^{-\mu\varepsilon_i}}.$$

Thermodynamic consistency requires that $\mu = 1/kT$. In the large N and large n_i limits, the microcanonical entropy $S = k \ln \Omega$ becomes

$$S = -kN \sum_i p_i \ln p_i. \tag{12.12}$$

Hence, the entropy per system is $S/N = -k \sum_i p_i \ln p_i$.

https://doi.org/10.1017/CBO9780511902581.013 Published online by Cambridge University Press

The information theory approach to assigning probability densities, the principle of maximum entropy, maximizes a constrained entropy. It is similar to Schrödinger's use of the theory of most probable distributions. It, however, does not appeal to counting states and to the law of large numbers to define an entropy, but rather appeals to a small set of axioms (Shore and Johnson, 1980). The functional

$$S = -\int dx\, p(x) \ln\left(\frac{p(x)}{m(x)}\right), \tag{12.13}$$

is shown to satisfy them, up to an overall constant. These axioms are that the entropy functional should

- be unique
- have coordinate independence
- have system independence
- have subset independence.

They define the character of the entropy. Coordinate independence means invariance under a change of variables. System independence says it should not matter whether one accounts for independent information about independent systems in terms of separate distributions or in terms of a joint distribution. Finally, subset independence says it should not matter whether one treats an independent subset of systems in terms of a conditional density or in terms of the full system density.

For *discrete probabilities*, the entropy axioms of information theory say that up to an overall positive constant, the entropy expression is

$$S = -\sum_i p_i \log\left(\frac{p_i}{m_i}\right). \tag{12.14}$$

Seemingly this expression is the natural discretization of (12.13). In fact, a fifth axiom is required (Shore and Johnson, 1983). The entropy functional should

- be logically consistent,

that is, in the absence of additional information, maximizing the entropy for a discrete probability (12.14) must yield $p_i = m_i$.

Analogous results exist for a continuous distribution. In the entropy functional (12.13), $p(x)$ is a probability distribution and $m(x)$ is a measure necessary for invariance under a change of variables.[2] If we were to maximize (12.13), subject to the constraint that $\int dx\, p(x) = 1$, we would start with

[2] The invariance is easily seen by making the standard change of variables to the integration and recalling from Section 2.1 how probability distributions transform under a change of variables. The various Jacobians of the change of variables cancel.

$$Q = \lambda_0 \left[\int dx\, p(x) - 1 \right] - \int dx\, p(x) \ln \left(\frac{p(x)}{m(x)} \right).$$

Then, requiring the variations δQ in Q due to arbitrary variations δp in p to be zero,

$$\delta Q = - \int dx\, \delta p(x) \left[\ln \left(\frac{p(x)}{m(x)} \right) + 1 - \lambda_0 \right] = 0,$$

implies $p(x) = m(x) e^{-(1-\lambda_0)}$. Integrating both sides of the equation over x, using the normalization condition on $p(x)$, and assuming one for $m(x)$, for convenience, shows that $\exp[-(1 - \lambda_0)] = 1$, that is, $\lambda_0 = 1$. Therefore, $p(x) = m(x)$. Thus, we can view $m(x)$ as representing our prior knowledge of $p(x)$, that is, prior to the use of the data. In Bayesian analysis, this Lebesgue measure is sometimes called the *default model*. The last axiom says that in the absence of additional information we must recover our prior information.

Insight about m_i (and $m(x)$) comes from revisiting Schrödinger's problem, but now restricting the total number of energies to be M and assigning a probability m_i to each. Now, for a set of measures $(m_1, m_2, \ldots, m_M)$, which for convenience we assume are normalized,

$$1 = (m_1 + m_2 + \cdots + m_M)^N = \sum_{\substack{m_1, m_2, \ldots, m_M \\ m_1 + m_2 + \cdots + m_M = 1}} \frac{N!}{n_1! n_2! \cdots n_M!} m_1^{n_1} m_2^{n_2} \cdots m_M^{n_M}. \tag{12.15}$$

Accordingly, the probability of a given set $(n_1, n_2, \ldots, n_M)$ of occupation numbers, conditioned on $(m_1, m_2, \ldots, m_M)$, is

$$P(n_1, n_2, \ldots, n_M | m_1, m_2, \ldots, m_M) = \frac{N!}{n_1! n_2! \cdots n_M!} m_1^{n_1} m_2^{n_2} \cdots m_M^{n_M}. \tag{12.16}$$

For N and n_i large,

$$\ln P = -N \sum_i p_i \ln \left(\frac{p_i}{m_i} \right) = S, \tag{12.17}$$

where $p_i = n_i/N$. We note that p_i must be zero wherever m_i is zero.

The *principle of maximum entropy* says that to assign probabilities on the basis of partial information, we maximize the entropy, constrained by whatever information we know about the probability. What is produced is the least informative probability consistent with the constraints.

An iconic use of this principle is predicting the joint probability of a set of variables $x = (x_1, x_2, \ldots, x_N)$, each ranging from $-\infty$ to ∞, whose means and covariance matrix (Section 3.24) are

https://doi.org/10.1017/CBO9780511902581.013 Published online by Cambridge University Press

$$\langle x_k \rangle = \int dx\, x_k\, p(x),$$

$$C_{ij} = \langle (x_i - \langle x_i \rangle)\left(x_j - \langle x_j \rangle\right) \rangle.$$

Straightforward analysis yields

$$p(x|\{\langle x \rangle\}, C) = e^{-\frac{1}{2}\delta x^T C^{-1} \delta x} / \sqrt{\det(2\pi C)},$$

where the vector of deviations δx is $(x_1 - \langle x_1 \rangle, \ldots, x_N - \langle x_N \rangle)^T$. We express the result as a conditional probability, as logically the derivation is conditional on knowing the means and covariance matrix beforehand. We comment that if the variables did not range from $-\infty$ to ∞, we would need to compute the normalization constant by performing the integrations numerically. While the predicted probabilities would have the functional form of a Gaussian, they would rather be functions proportional to Gaussians.

In general, the principle of maximum entropy is useful for suggesting the functional forms of probabilities. In the next subsection, we adopt this multivariate Gaussian as a likelihood function, but we do not invoke the principle of maximum entropy. Entropy is part of our choice for a prior probability. The solutions to our integral equations involve competitions between maximizing the log-likelihood and maximizing the entropy.

12.3.2 The likelihood function and prior probability

We now discuss our choices for the likelihood function and prior probability. For the likelihood function, we start very generally with

$$P(\bar{G}|A) = e^{-\mathcal{L}(\bar{G},A)} / Z_{\mathcal{L}},$$

where $\mathcal{L}(\bar{G}, A)$ is some positive function and $Z_{\mathcal{L}}$ is the normalization constant. We now need to be more specific about the functional form of $\mathcal{L}(\bar{G}, A)$.

We want our choice of $\mathcal{L}$ to be compatible with the data the simulation generates. If $\bar{G}_i^{(j)}$ is the value of G_i for the j-th configuration of the simulation, then the simulation eventually gives an estimate of the mean

$$\bar{G}_i = \frac{1}{M} \sum_{j=1}^{M} \bar{G}_i^{(j)} \tag{12.18}$$

and the covariance matrix (3.24)

$$C_{ik} = \frac{1}{M(M-1)} \sum_{j=1}^{M} \left(\bar{G}_i^{(j)} - \bar{G}_i\right)\left(\bar{G}_k^{(j)} - \bar{G}_k\right). \tag{12.19}$$

https://doi.org/10.1017/CBO9780511902581.013 Published online by Cambridge University Press

With the data represented by the mean and covariance, it is natural and convenient to take

$$\mathcal{L}(\bar{G}, A) = \tfrac{1}{2}\chi^2(\bar{G}, A), \tag{12.20}$$

with

$$\chi^2(\bar{G}, A) = \sum_{i,j=1}^{L} \left(\bar{G}_i - G_i\right) \left[C^{-1}\right]_{ij} \left(\bar{G}_i - G_i\right), \tag{12.21}$$

where G_i is the exact value of $G(\tau)$ at τ_i for a given A. With this choice of $\mathcal{L}$, the normalization constant $Z_{\mathcal{L}}$ is $(2\pi)^{N/2}\sqrt{\det C}$.

This choice is the same as the one made for simple and regularized least-squares problems. The discussion there implicitly assumed that some Gaussian process has generated the data and hence defines the mean and the covariance. In our case, a Monte Carlo simulation generates the data from which we compute the mean and covariance. If we computed them with a sufficiently large amount of statistically independent information, then by the central limit theorem the means are distributed by a Gaussian whose width is defined by the covariance matrix.

Equation (12.17) hints at our choice of the prior probability: It says that in the absence of other information, the probability of a given set of occupation numbers is proportional to the exponential of the entropy. Instead of appealing to combinatorial arguments, we can appeal to the information theory arguments for the functional form of the entropy and use the same exponential form as a prior probability. In fact, in statistical inference, when assigning probabilities and needing to maintain positivity, the most common choice for the prior probability is the entropic prior

$$P(A) = e^{\alpha S(A)}/Z_S(\alpha), \tag{12.22}$$

where α sets the scale of the entropy's contribution and $Z_S(\alpha)$ is the normalization constant. In general, α is unknown a priori.

There are philosophical reasons for choosing an entropic prior. Our main reason, however, is practical: It is a convenient choice for maintaining the positivity and the normalization of the solution. Another reason is the choice that has value added: We constrain our solutions with the known characteristics prescribed by the axioms that establish the form of the entropy. In effect, because of the additive nature of the entropy, this prior probability induces structure in the solution only if that structure is enforced by the data or by other prior knowledge. Finally, we like the fact that the default model is an explicit mechanism for putting our prior information into the solution. For most physics problems, various approximations or exact results exist that we can use as default model.

These choices of the likelihood and prior enable us to state the joint distribution

$$P(A, \bar{G}) = P(\bar{G}|A)P(A) = \frac{e^{\alpha S - \mathcal{L}}}{Z_{\mathcal{L}} Z_S(\alpha)}. \tag{12.23}$$

The two remaining probabilities in Bayes's theorem are the posterior probability and the evidence. Our choices for the likelihood function and prior probability set both. Our task now is discussing how these choices affect our choice of a "best" solution to the integral equation.

12.3.3 The "best" solutions

We discuss three approaches to finding a solution. With our choices of the likelihood function (12.20) and prior probability (12.22), we can write the posterior probability in terms of the evidence as

$$P(A|\bar{G}) = \frac{P(A, \bar{G})}{P(\bar{G})} = \frac{e^{Q(A)}}{Z_{\mathcal{L}} Z_S(\alpha)}, \tag{12.24}$$

with

$$Q(A) = \alpha S(A) - \tfrac{1}{2}\chi^2(A). \tag{12.25}$$

With this posterior probability we now state our first best solution: We ignore the evidence, which does not depend on A, and take as our solution the most probable spectral density A given the data $\bar{G}$. We find this by maximizing $Q(A)$, which, in turn, maximizes the posterior probability. We call this the *constrained fit*, as it is simply a regularized least-squares fit with the regularizing function being the entropy. This solution depends on α. We choose α so that $\chi^2 = N$, where N is the number of $\bar{G}_i$. This way of choosing α is basically ad hoc and usually tends to under-fit. The resulting procedure is commonly called the *historic maximum entropy method*.

At this level of analysis, we clearly see a duality between fitting and inference. We can regard the constrained fit as minimizing χ^2, which corresponds to choosing the A_i parameters to fit the data as closely as possible, subject to the entropy constraints. Or we can regard this solution as maximizing the entropy, which puts the least amount of information into the solution, subject to the constraints of the data.

An alternative approach to choosing α is to use Bayesian analysis to guide the choice. We call this second approach to solving the integral equation the *Bayesian constrained fit*. More commonly, it is called the *classic maximum entropy method*.

In starting the development of this alternative approach, we first note that our prior probability is conditional on α and we write (12.23) as

$$P(A, \bar{G}|\alpha) = P(\bar{G}|A)P(A|\alpha) = \frac{e^{Q}}{Z_{\mathcal{L}} Z_S(\alpha)}. \tag{12.26}$$

https://doi.org/10.1017/CBO9780511902581.013 Published online by Cambridge University Press

Note that the likelihood function in (12.26) is determined by the physical dynamics and hence is not conditional on α, which was introduced to scale the contribution of the entropy to the entropic prior. We then define a new joint probability

$$P(A, \bar{G}, \alpha) = P(A, \bar{G}|\alpha)P(\alpha) = P(\alpha)\frac{e^{Q}}{Z_{\mathcal{L}}Z_{S}(\alpha)} \tag{12.27}$$

and a new posterior probability

$$P(A, \alpha|\bar{G}) = \frac{P(\alpha)}{P(\bar{G})}\frac{e^{Q}}{Z_{\mathcal{L}}Z_{S}(\alpha)}.$$

Introduced is a new probability, $P(\alpha)$, the probability of α. This number is usually chosen to be a constant or what is called Jeffery's prior, $P(\alpha) \propto 1/\alpha$ (Sivia and Skilling, 2006). Jeffery argued that the probability assigned to a scale-setting parameter should be done so that $p(x)dx = p(cx)d(cx)$. Since $d(cx) = cdx$, this requirement reduces to $p(x) = cp(cx)$, which is satisfied only by $p(x) \propto 1/x$. Normalization requires an accompanying restriction on the range of x. In practice, a good solution is reasonably insensitive to choosing $P(\alpha)$ as either a constant or $1/\alpha$. For a fixed α, we note that the maximum of the new posterior probability still occurs at the maximum of Q.

To move toward finding the "best" value of α, we write

$$P(A|\bar{G}) = \int d\alpha\, P(A, \alpha|\bar{G}) = \int d\alpha\, P(A|\alpha, \bar{G})P(\alpha|\bar{G}).$$

If the number of data points is large and well connected to the inference problem, it seems reasonable to expect that the many data restrict the possible values of the single parameter α significantly, making $P(\alpha|\bar{G})$ sharply peaked at some value $\alpha = \hat{\alpha}$; that is, $P(\alpha|\bar{G}) \approx \delta(\alpha - \hat{\alpha})$. The posterior probability becomes

$$P(A|\bar{G}) \approx P(A|\bar{G}, \hat{\alpha}).$$

For our second solution of the integral equation, we find $\hat{\alpha}$, the value of α at which $P(\alpha|\bar{G})$ peaks, and *then* for this value of α, we find the A that maximizes Q as the solution. These two steps specify the Bayesian constrained fit. We note that this solution is not equivalent to solving

$$\frac{\partial P(A, \alpha|\bar{G})}{\partial A} = 0, \quad \frac{\partial P(A, \alpha|\bar{G})}{\partial \alpha} = 0.$$

These equations fix A and α simultaneously.

In Section 12.4, we discuss how to estimate $P(\alpha|\bar{G})$ and find $\hat{\alpha}$. Here we derive the form of this probability. We start by marginalizing A from the joint probability (12.27):

https://doi.org/10.1017/CBO9780511902581.013 Published online by Cambridge University Press

$$P(\alpha, \bar{G}) = \int \mathcal{D}A\, P(A, \alpha, \bar{G}).$$

Then, using $P(\alpha, \bar{G}) = P(\alpha|\bar{G})P(\bar{G})$ and (12.27), we find that

$$P(\alpha|\bar{G}) = \frac{P(\alpha)}{P(\bar{G})} \int \mathcal{D}A\, \frac{e^{Q}}{Z_{\mathcal{L}} Z_S}. \tag{12.28}$$

The parameter α is called a nuisance parameter, and a nuisance parameter generally is best handled by integrating it out of the problem (marginalizing it) and working with $P(A|\bar{G})$ instead of $P(A, \alpha|\bar{G})$. In fact, our third solution to the problem does precisely this. We call this approach the *average spectrum method*. In the past, it was called *Bryan's method* (Bryan, 1990) to acknowledge its source. The switch in name emphasizes the nature of the solution being an average instead of being a mode. This nature has often been overlooked.

For this third method, we first find $\hat{A}_\alpha$ from

$$\left.\frac{\partial Q}{\partial A}\right|_{A=\hat{A}_\alpha} = 0,$$

that is, we obtain the constrained fit as a function of α, and then we choose as the solution the average defined by

$$\langle A \rangle = \int d\alpha\, \hat{A}_\alpha P(\alpha|\bar{G}),$$

where $P(\alpha|\bar{G})$ is given by (12.28). We need this solution because $P(\alpha|\bar{G})$ is sometimes not sharply peaked but broadly peaked and skewed.

All three methods require finding the A that maximizes $Q(A)$ for a fixed value of α, and two require knowing something about $P(\alpha|\bar{G})$. In the next section, we discuss the nature of the maximum of $Q(A)$ in detail and present an approximation for $P(\alpha|\bar{G})$. Before moving to that section, we first will say a few words about the evidence.

There are two kinds of evidence present in our methods of solution, $P(\bar{G})$ and $P(\bar{G}|\alpha)$. We now derive formal expressions for both. First, using

$$P(\bar{G}|\alpha) = \frac{P(\alpha|\bar{G})P(\bar{G})}{P(\alpha)}$$

and (12.28), we immediately find that

$$P(\bar{G}|\alpha) = \int \mathcal{D}A\, \frac{e^{Q}}{Z_{\mathcal{L}} Z_S}. \tag{12.29}$$

https://doi.org/10.1017/CBO9780511902581.013 Published online by Cambridge University Press

Next, integrating both sides of (12.28) over α yields

$$P(\bar{G}) = \int d\alpha\, P(\alpha) \int \mathcal{D}A\, \frac{e^Q}{Z_{\mathcal{L}} Z_S}.$$

From these two results we see that the conditional evidence $P(\bar{G}|\alpha)$ is central to both the evidence $P(\bar{G})$ and to $P(\alpha|\bar{G})$. Fortunately, the evidence $P(\bar{G})$ for our posterior probability $P(A|\bar{G})$ is the expectation value of the conditional evidence $P(\bar{G}|\alpha)$ with respect to the probability $P(\alpha)$. Germane to this point are a few remarks about how to know when the evidence is significant.

If we accept as our solution a quantity that depends on the mode of the posterior probability, the evidence $P(\bar{G})$, being its normalization factor, plays no role in this solution as it has no explicit dependence on the fitting parameters α and the values of A. This situation is the case for the constrained and the Bayesian constrained fits. The evidence does, however, play an essential role if our solution is an average over A and α. The conditional evidence plays a role in the Bayesian constrained fit and average spectrum method.

12.4 Analysis details and the Ockham factor

Finding the maximum of Q (12.25) as a function of A for a fixed α, which is central to all three of our approaches, is simple in principle because both the likelihood function and the entropic prior are concave functions of the A_i and hence a unique maximum exists. In practice, finding the maximum can be difficult because it may not be sharp. The concavity of the likelihood function is familiar: The covariance matrix is positive definite, so we can always transform the data into the coordinate system where this matrix is diagonal. Then, in the likelihood function, the curvature tensor is diagonal with positive matrix elements equal to the positive eigenvalues of the covariance matrix. The concavity of the entropic contribution to Q is also easily established. By direct calculation,

$$\frac{\partial^2 S}{\partial A_i \partial A_j} = -\frac{\delta_{ij}}{A_i} = -\frac{\delta_{ij}}{\sqrt{A_i A_j}}.$$

Hence, the curvature of the entropy equals $\sqrt{A_i A_j}\delta_{ij}$.

The nature of Q in the vicinity of this maximum determines how easy it is to find it. To describe the maximum, it is convenient to transform the deviations from it, that is, the $\delta\hat{A}_i = A_i - \hat{A}_i$, to the space of new variables X_i in which the entropy curvature is flat. The new coordinate system satisfies

$$\frac{\partial A_j}{\partial X_i} = \sqrt{A_j}\delta_{ij}. \tag{12.30}$$

https://doi.org/10.1017/CBO9780511902581.013 Published online by Cambridge University Press

With this change of variables,

$$Q(A,\alpha) \approx Q(\hat{A}_\alpha) - \tfrac{1}{2}\sum_{ij} \delta X_i \Gamma_{ij} \delta X_j,$$

where Γ is a positive-definite matrix $\Gamma_{ij} = \alpha\delta_{ij} + \Lambda_{ij}$ with

$$\Lambda_{ij} = \left[\sqrt{A_i}\frac{\partial^2 \mathcal{L}}{\partial A_i \partial A_j}\sqrt{A_j}\right]_{A=\hat{A}_\alpha}$$

and

$$\frac{\partial^2 \mathcal{L}}{\partial A_i \partial A_j} = \left[K^T \cdot C^{-1} \cdot K\right]_{ij} = \sum_{kl} K_{ki}\left[C^{-1}\right]_{kl} K_{lj}.$$

Here, K_{ij} is the time-frequency discretization of the kernel of the integral equation. In the new coordinate system, Γ controls the curvature of Q around the maximum. If the eigenvalues of Γ are small, the curvature is flat. A flat curvature complicates finding the maximum and leads to considerable uncertainty in the result. The covariance of the X_i is a measure of the uncertainty,

$$\left\langle \delta X_i \delta X_j \right\rangle = \int \mathcal{D}X\, \delta X_i \delta X_j P\left(A|\bar{G}\right) \approx \left[\Gamma^{-1}\right]_{ij}.$$

In a coordinate system where Γ is diagonal, $\left\langle \delta X_i \delta X_j \right\rangle \to \left\langle \delta X_i^2 \right\rangle \delta_{ij} = \delta_{ij}/\gamma_i$, where the γ_i are the eigenvalues of Γ. Similarly, we find that

$$\left\langle \delta A_i \delta A_j \right\rangle \approx \left[\sqrt{\hat{A}}\,\Gamma^{-1}\sqrt{\hat{A}}\right]_{ij} = \sqrt{\hat{A}_i}\left[\Gamma^{-1}\right]_{ij}\sqrt{\hat{A}_j}. \tag{12.31}$$

Thus, a small eigenvalue of Γ leads to a large variance in the result. We also note that taking $\sqrt{A_i}$ as the metric for the problem (12.30) yields $\mathcal{D}A = \prod_i dA_i/\sqrt{A_i}$, and with the change in variables, $\mathcal{D}A \to \mathcal{D}X = \prod_i dX_i$.

We now make explicit the normalization factors needed in the analysis. Physically, as $A(\omega)$ ranges between 0 and 1, the values of the Green's function for the different imaginary times range between another set of bounds. For example, if $G(\tau)$ describes Fermions of a particular spin, they range between 0 and 1. Hence, the normalization integrals are something that at first glance we need to do numerically. However, if the error in the data is sufficiently small to make the exponential of the likelihood function sharply peaked, which is to say the exponential dies off over the range of integration, we can extend the limits of integration over all space and then evaluate the integral analytically: If

$$Z_{\mathcal{L}} = \int \mathcal{D}\bar{G}\, e^{-\frac{1}{2}\chi^2} = \int \prod_i d\bar{G}_i e^{-\frac{1}{2}\chi^2}$$

https://doi.org/10.1017/CBO9780511902581.013 Published online by Cambridge University Press

with

$$\chi^2 = \sum_{i,j=1}^{L} \left(\bar{G}_i - G_i\right) \left[C^{-1}\right]_{ij} \left(\bar{G}_j - G_j\right),$$

then when the limits of the integrations are extended over all space, we find

$$Z_{\mathcal{L}} = (2\pi)^{L/2} \sqrt{\det C}.$$

Our likelihood function is thus the normalized Gaussian we assumed.

This redundant analysis illustrates an important point about the central limit theorem: While the range of random variables for a given distribution, such as the uniform distribution over $[0, 1]$, is bounded, the allowed range of the average of a large number of these variables is over $-\infty$ to ∞. As the number in the sum increases, the probability of the average lying outside the original range becomes vanishingly small.

For the normalization factor of the prior probability, that is, $Z_S(\alpha)$, the situation is different. We need to derive an approximate expression and do so by a similar analysis. We start with

$$Z_S(\alpha) = \int \mathcal{D}A \, e^{\alpha S} = \prod_i \int \frac{dA_i}{\sqrt{A_i}} \, e^{\alpha S_i},$$

where $S_i = A_i \ln(A_i/m_i)$. Next we approximate the exponential by a Gaussian centered at $A_i = m_i$, extend the range of the integration to be over all space, and find

$$\int \frac{dA_i}{\sqrt{A_i}} e^{\alpha S_i} \approx \int dX e^{-\alpha X^2/2} = \sqrt{2\pi/\alpha}.$$

Consequently,

$$Z_S(\alpha) \approx \left(\frac{2\pi}{\alpha}\right)^{N/2} = \frac{(2\pi)^{N/2}}{\sqrt{\det \alpha I}}, \tag{12.32}$$

where I is an $N \times N$ identity matrix. It is definitely true that in the vicinity of its maximum, the entropy is well approximated by a quadratic function. The width around that maximum is controlled by the size of $\alpha > 0$, and the prior is most sharply peaked when α is large.

We now derive an approximate expression for the conditional evidence

$$P(\bar{G}|\alpha) = \int \mathcal{D}A \, \frac{e^{Q}}{Z_{\mathcal{L}} Z_S(\alpha)}.$$

https://doi.org/10.1017/CBO9780511902581.013 Published online by Cambridge University Press

With it, the expressions we need for $P(\alpha|\bar{G})$ and $P(\bar{G})$ follow readily. We simply approximate the integrand by a Gaussian form centered at the maximum of Q:

$$P(\bar{G}|\alpha) \approx \frac{e^{Q(\hat{A}_\alpha)}}{Z_{\mathcal{L}} Z_S(\alpha)} \int \mathcal{D}X\, e^{-\frac{1}{2}\delta X^T \cdot (\alpha I + \Lambda)\cdot \delta X} = \frac{e^{Q(\hat{A}_\alpha)}}{Z_{\mathcal{L}} Z_S(\alpha)} \frac{(2\pi)^{N/2}}{\sqrt{\det[\alpha I + \Lambda(\hat{A}_\alpha)]}}. \tag{12.33}$$

If the peak around Q is sufficiently sharp, this approximation should be sufficiently good.

For the Bayesian constrained fit, we need to maximize $P(\alpha|\bar{G})$ with respect to α. This function is simply $P(\alpha)P(\bar{G}|\alpha)/P(\bar{G})$. As $P(\bar{G})$ is independent of α, we find the maximum of $P(\alpha|\bar{G})$ from the maximum of $P(\alpha)P(\bar{G}|\alpha)$. In general, the common choices of $P(\alpha)$ are featureless. What we effectively need is the maximum of $P(\bar{G}|\alpha)$ as a function of α. With (12.33), the condition $\partial \ln P(\bar{G}|\alpha)/\partial\alpha = 0$ leads to

$$-2\hat{\alpha}S(\hat{A}_{\hat{\alpha}}) = \mathrm{Tr}\left[\Lambda(\alpha I + \Lambda)^{-1}\right] - \mathrm{Tr}\left[\frac{d\ln\Lambda}{d\ln\alpha}\Lambda(\alpha I + \Lambda)^{-1}\right].$$

The logarithmic derivative is expected to be small. In any case, we drop it from this expression. The defining equation for the Bayesian constrained fit becomes

$$N_{\mathrm{good}} \equiv -2\hat{\alpha}S(\hat{A}_{\hat{\alpha}}) = \mathrm{Tr}\left[\Lambda(\alpha I + \Lambda)^{-1}\right] = \sum_i \frac{\lambda_i}{\alpha + \lambda_i}, \tag{12.34}$$

where the λ_i are the eigenvalues of Λ. The quantity $-2\hat{\alpha}S(\hat{A}_{\hat{\alpha}})$, called N_{good}, is a measure of the shift of the solution away from the default model (if the solution were the default model, the entropy would be zero). When a λ_i is much greater than α, it contributes a value of unity to N_{good}. When it is much smaller, it contributes zero. Thus, N_{good} measures the amount of good information in the solution and is a convenient indicator of the amount of structure in the solution.

When N_{good} is large, one expects $P(\alpha|\bar{G})$ to be sharply peaked. Unfortunately, the extremely ill-posed nature of the analytic continuation fitting problem is typically characterized by $N_{\mathrm{good}} \ll N$ with N_{good} often being between 5 and 10. When this figure of merit is small, we can expect to parameterize the locations and widths of only a few peaks.

While the Bayesian constrained fit often gives an acceptable result, the average spectrum method is generally more consistent with the information in the data. Once again, this solution is

$$\langle A \rangle = \int \mathcal{D}A\, d\alpha\, A(\alpha) P(A, \alpha|\bar{G}) \approx \int d\alpha\, \hat{A}_\alpha P(\alpha|\bar{G}), \tag{12.35}$$

where the integral over α is done numerically. The average spectrum solution reproduces the Bayesian constrained fit if $P(\alpha|\bar{G})$ is sharply peaked and returns the model in the absence of data. To obtain the average spectrum result, we need a normalized $P(\alpha|\bar{G})$. What is done is to evaluate $P(\alpha)P(\bar{G}|\alpha)$, using (12.33) for $P(\bar{G}|\alpha)$, for a number of discrete values of α and then compute the area under this curve. The area is the normalization factor. Dividing the values of $P(\alpha)P(\bar{G}|\alpha)$ by this factor gives $P(\alpha|\bar{G})$.

We concluded the last subsection with a few remarks about when the evidence is significant. We conclude this subsection with a few remarks about the significance of the evidence. We do so by first discussing the relationship between the conditional evidence and what is called the Ockham factor. With (12.32), the conditional evidence

$$P(\bar{G}|\alpha) = \frac{e^{Q(\hat{A}_\alpha)}}{Z_{\mathcal{L}}} \frac{(2\pi)^{N/2}}{Z_S(\alpha)\sqrt{\det[\alpha I + \Lambda]}} \approx \underbrace{e^{\alpha S(\hat{A})} \frac{e^{-\mathcal{L}(\hat{A}_\alpha)}}{Z_{\mathcal{L}}}}_{\text{best fit}} \underbrace{\sqrt{\frac{\det[\alpha I]}{\det[\alpha I + \Lambda]}}}_{\text{Ockham factor}} \tag{12.36}$$

is approximately the product of two factors. One comes from the mode of the posterior probability, which represents the constrained least-squares fitting, and the other modifies the fit. From this perspective the maximum for a given α results from a competition between fitting the values of the spectral density to good data, which is the tendency for small α, and because of the Ockham factor, defaulting to the model, which is the tendency for large α. When the data are closely fitted, the mode represents a solution with many parameters, while the Ockham factor favors fewer parameters. As we showed, the evidence is the expectation value of the conditional evidence, and as the common choices of $P(\alpha)$ are featureless, if not flat, the full evidence carries with it the Ockham character. The evidence helps establish a balance between the accuracy of the fit to the data and the number of parameters being fitted.

In Fig. 12.1, we depict the Ockham factor schematically. Before we have data, our knowledge of the solution is expressed by the prior probability. It admits a possible solution over some volume ΔA in the space of parameters and hence roughly equals $1/\Delta A$. With data, the likelihood function restricts the solution to some smaller volume δA centered around the mode. In general, if our prior information admits solutions over a wide range of parameter space, only a small portion of the prior contributes to the evidence. Typically, as we increase the number of parameters, we make δA smaller while making ΔA larger. The Ockham factor, $\delta A/\Delta A$, expresses a penalty for using too many parameters for the sake of getting a good fit. In short, the evidence contains an Ockham factor that favors a simpler physical model than the least-squares solution in the spirit of Ockham's

https://doi.org/10.1017/CBO9780511902581.013 Published online by Cambridge University Press

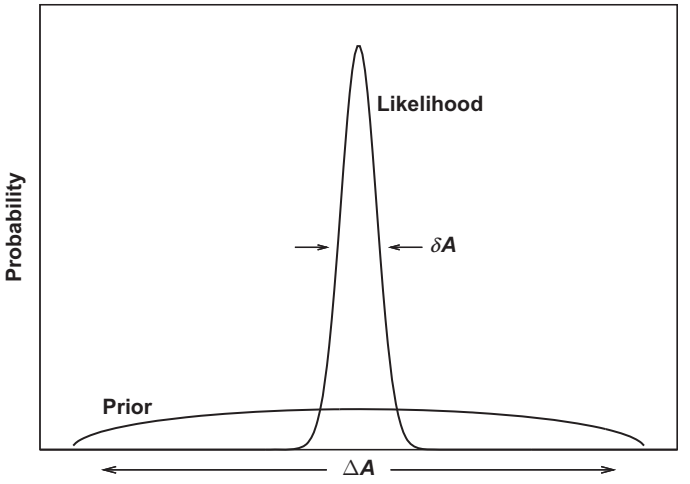

Figure 12.1 Schematic representation of the Ockham factor. This factor penalizes results for "wasting" volume in parameter space. The evidence $P(\bar{G}) = \int P(\bar{G}|A)P(A)\mathcal{D}A \approx P(\bar{G}|\hat{A})\delta A/\Delta A =$ maximum likelihood $\times$ Ockham factor.

centuries-old principle that one should always opt for an explanation in terms of the smallest possible number of causes, factors, or variables. The evidence helps avoid the "with enough parameters you can fit anything" syndrome.

12.5 Practical considerations

We now address several practical considerations associated with the use of our Bayesian entropic method to generate a solution to the integral equation (12.1). Each of the three basic steps to the method requires special considerations. These steps are:

1. Verifying that the data are consistent with the assumptions of the likelihood function
2. Solving the integral equation
3. Assessing the acceptance of the solution.

Jarrell and Gubernatis (1996) presented a detailed case study of these steps for the infinite-dimensional periodic Anderson model. A slightly updated version is given by Jarrell et al. (2008). Here, we highlight the main points of these studies.

1. Verifying that the data are consistent with the assumptions of the likelihood function has two parts. The first is establishing that the data used to calculate the covariance matrix are statistically independent. Here, we need to remove the correlations in the data that exist from one Monte Carlo step to another. The second

addresses removing the correlations that exist between the data at a given Monte Carlo step. Here, we remove these correlations by diagonalizing the covariance matrix.

The data consist of two types of measurements: the mean values (12.18) and the covariance matrix (12.19). As with other Monte Carlo estimates of means, the estimates of the $\bar{G}_i$ are unaffected by correlations existing between successive Monte Carlo steps. On the other hand, as with other Monte Carlo estimates of uncertainty, the estimates of the elements of the covariance matrix are affected. To generate the assumed statistically independent measurements needed to estimate the covariance properly, we use the method of blocked means, focusing on the *diagonal elements* of the covariance matrix. As explained in Section 3.4 of Chapter 3, we break up the data stream into larger and larger blocks until the block averages are statistically independent.

One difficulty in "verifying the data" lies with a number of estimates of the $\bar{G}_i$ being inclined to have a nonzero skewness and kurtosis. For example, the fluctuations in the Green's function $G(\tau)$ associated with the spectral density of the single-impurity Anderson model are bounded above by 1 when τ is close to 0 and bounded below by zero when τ is close to $\beta/2$ and β is large. Clearly, reducing the fluctuations in the block averages of correlations for these Green's function elements is more challenging than for others. If we recall the discussion in the previous subsection about the central limit theorem, we can begin to understand why. The central limit theorem locates the Gaussian at the mean value of the random variable. If that variable is bounded and its mean is close to the boundary, more reduction of the variance is needed before the distribution about the mean has a strongly Gaussian shape. In the present case, increasing the block sizes eventually promotes most of the *diagonal elements* of the covariance matrix to exhibit features of statistical independence and a Gaussian distribution.

Focusing on the diagonal elements of C to establish statistical independence is a practical approach more so than an insightful statistical one. It, however, is not the same as throwing away the off-diagonal elements of C. Throwing away removes the correlations among measurements at different τ values but leads to poor estimates of the errors of independent information as measured by C.

After the block size becomes reasonably established, increasing the number of blocks promotes the proper calculation of the positive-definite covariance matrix. Here, we are interested in diagonalizing this matrix so we can transform χ^2 into the standard estimate of the error,

$$\chi^2 = \sum_{i,j=1}^{L} (\bar{G}_i - G_i)\left[C^{-1}\right]_{ij}(\bar{G}_j - G_j) \to \sum_{i=1}^{L} (\bar{G}'_i - G'_i)^2/\sigma_i^2, \tag{12.37}$$

https://doi.org/10.1017/CBO9780511902581.013 Published online by Cambridge University Press

for uncorrelated measurements. If C is diagonalized by the similarity transformation $C = S\Sigma S^T$, then $G' = SG$ and $\bar{G}' = S\bar{G}'$. If the number of statistically independent blocks is insufficient, the eigenvalues of C, that is, the σ_i^2, when indexed from high to low, fall precipitously at some value of the index. The small eigenvalues are less accurate than the large ones. To prevent this break, it seems necessary that $N_{\text{blocks}} > 2L$. Producing a reasonably large number of sufficiently large blocks defines the numerical task of the quantum Monte Carlo method.

2. With the data qualified, we turn to the second step of our method and now seek a solution of the integral equation. We first need to discuss its discretization. Although our problem (12.1) is one of assigning a quasi-probability density $A(\omega)$, when we discretize the problem, we convert it to one of assigning probabilities. At times, a nonuniform discretization is advantageous to focus computational effort around peaks and not in smoothly varying featureless regions of frequency. When discretizing, assigning a mean value to the density over an interval is akin to assigning each interval a different probability. We need to do this in such a way that if we change, for example, from a uniform grid to a logarithmic one, we leave our entropy form invariant. (Recall the second axiom.)

After discretization, we have a system of linear equations to solve for the values of $A(\omega)$ at a set $(\omega_1, \omega_2, \ldots, \omega_N)$ of discrete values of ω using the values of $\bar{G}(\tau)$ at a set $(\tau_1, \tau_2, \ldots, \tau_L)$ of discrete values of τ. The system of equations is

$$\bar{G}_i = \sum_{j=1}^{N} K_{ij} A_j, \quad i = 1, \ldots, M, \tag{12.38}$$

where $\bar{G}_i = \bar{G}(\tau_i)$, $K_{ij} = K(\tau_i, \omega_j)$, and $A_i = A(\omega_i)\Delta\omega_i$. Similarly, we define $m_i = m(\omega_i)\Delta\omega_i$. With the latter two definitions, A_i represents the *probability* of being in the interval $(\omega_i, \omega_i + \Delta\omega_i)$, and the problem properly shifts from finding a probability density to assigning probabilities.

Each of the three solutions discussed in Section 12.3.3 requires maximizing

$$\begin{aligned} Q(A) &= Q(A_1, A_2, \ldots, A_N) \\ &= \alpha S(A_1, A_2, \ldots, A_N) - \frac{1}{2} \sum_{i=1}^{L} \left(\frac{\bar{G}_i - \sum_{j=1}^{N} K_{ij} A_j}{\sigma_i} \right)^2 \end{aligned} \tag{12.39}$$

for a given value of α. (For notational convenience, we dropped the primes on the $\bar{G}_i$.) The core strategy for each of these solutions is illustrated in Fig. 12.2. The contours with solid lines are values of the χ^2 misfit function, and those with dotted lines are isoentropic values. For large values of α, the solution is dominated by the default model, with a likely poor fit to the data, while for small values it is dominated by the least-squares fit, with likely a tight fit. Starting with a large value

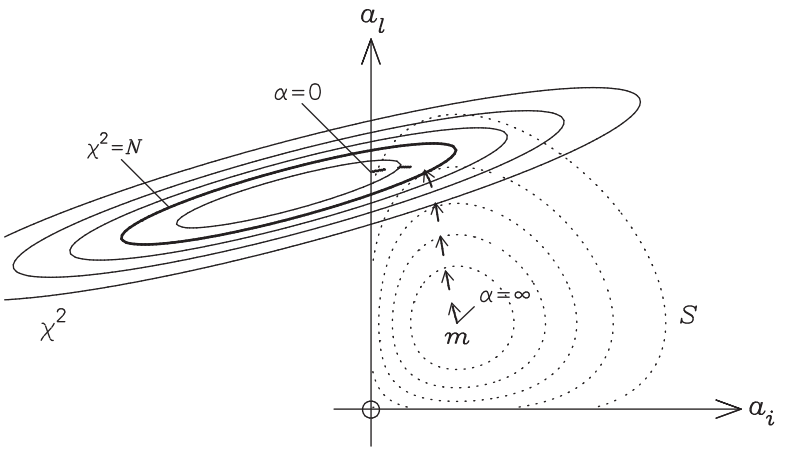

Figure 12.2 Schematic illustration of a solution trajectory. It starts with a default model m and takes small steps in α in such a way to reduce χ^2 while keeping the entropy S as large as possible (from Sivia and Skilling (2006)).

of α, we seek to take small steps reducing the misfit to the data and increasing the entropy (that is, moving it away from the default model). The concavity of both the misfit function and entropy means the solution is unique. We do the maximization by using a Levenberg-Marquardt algorithm (Golub and Loan, 1989; Press et al., 2007) after recasting Q to a form that reduces the complexity of the problem. We give the details of this algorithm in Appendix O. Here we summarize the main points of the overall procedure.

On the one hand, finding the maximum of Q is simple because it is unique, but on the other hand, it is touchy because the maximum is broad. This broadness is a reflection of $N_{\text{good}} \ll N$. The maximization method, however, is not directly applied to the original parameters A_i but rather to a smaller number of new parameters u_i that emerge after the problem is transformed into what is called the *dominant subspace*.

Normally, there may be a thousand or so values of $\bar{G}_i$ and a few times that for A_i. If we were to do a singular value composition (Golub and Loan, 1989; Press et al., 2007) of the transpose of the kernel, that is, $K^T = U \cdot D \cdot V^T$, where U is an $L \times N$ orthogonal matrix, D is an $N \times N$ diagonal matrix, and V is an $N \times N$ orthogonal matrix, then we would find that most of the diagonal elements of D, the singular values d_i^2 ordered from largest to smallest, lose all or almost all numerical accuracy. If s is the number of surviving accurate singular values, then the standard procedure is to use the factorization $K^T = \bar{U} \cdot \bar{D} \cdot \bar{V}^T$ where $\bar{U}$ is an $L \times s$ matrix obtained by retaining only the first s columns of the original U matrix, $\bar{D}$ is an $s \times s$ diagonal matrix, obtained by retaining only the first s columns and rows of the original D,

https://doi.org/10.1017/CBO9780511902581.013 Published online by Cambridge University Press

and $\bar{V}$ is an $N \times s$ matrix, obtained by retaining only the first s columns of the original V. The value of s is typically smaller than 10.

With this decomposition,

$$\bar{G} = K \cdot A \rightarrow (\bar{U}^T \cdot \bar{G}) = \bar{D} \cdot (\bar{V}^T \cdot A). \tag{12.40}$$

While this expression does not represent the procedure, it does illustrate that the severe ill-conditioned nature of the kernel reduces the effective dimension of the search space, the space of the latent parameters for the global maximum, to s or smaller. This is the dominant subspace. It also illustrates that the number of "good" parameters in this space is also s, that is, the number of elements of $(V^T \cdot A)$. The Levenberg-Marquardt method, a variant of Newton's method, is applied in this much smaller space whose dimension is principally set by the dimension of the dominant subspace of the kernel. Following Bryan (1990), we adopt a spectral function parameterized as

$$A_i = m_i \exp\left(\sum_{j=1}^{s} \bar{U}_{ij} u_j\right) \tag{12.41}$$

and Q is maximized with respect to the s parameters u_i. Because of the small dimension of the dominant subspace, for a given value of α, the maximization executes rapidly.

We can use this numerical procedure for finding the maximum of Q for a given α to obtain the constrained fit (historic maximum entropy solution) by applying a bisection technique to $\frac{1}{2}\chi(\alpha) - N$. We simply start with acceptable values of α small and large and with its value at the midpoint, and successively halve the intervals, until we locate the one containing $\frac{1}{2}\chi^2 - N = 0$. Convergence is quick. The Bayesian constrained fit (classic maximum entropy) requires the maximum of Q at the value of α that maximizes $P(\alpha|\bar{G})$ (12.33). Coupling a line search (Press et al., 2007) for this maximum with the procedure for maximizing Q is straightforward. The average spectrum method requires the numerical computation of an integral. Simpson's rule works well.

For the purpose of discussion, in Algorithm 41 we give a simpler description of the computational procedures. It is simpler in the sense that it does not compute explicitly any of the three solutions but rather computes the information needed to compute them. The information generated is the $A(\alpha)$ that maximize $Q(\alpha)$ for a given value of α, the misfit statistic $\chi^2(\alpha)$, and the probability $P(\alpha|\bar{G})$ associated with $A(\alpha)$. A graph of $\chi^2(\alpha)$ versus α enables a simple estimate of the value of α for which $\chi^2 \approx N$. Similarly, a graph of $P(\alpha|\bar{G})$ versus α enables an estimate of the $\hat{\alpha}$ that maximizes the conditional probability. Then, the constrained Bayesian solution $A(\hat{\alpha})$ is read from this table. With the tables, estimating the integral (12.35) is also

Algorithm 41 Core analytic continuation.

Input: Input $\bar{G}_i$, C_{ij}, K_{ij}, A_i, m_i, and ω_i. Specify $P(\alpha)$.
Diagonalize C and transform $\bar{G}$ and K^T to this basis ;
Perform a singular value decomposition on K and determine the size of the dominant subspace. Transform $\bar{G}$ to this basis (12.40) ;
Solve (12.41) for a set of starting parameters u_i ;
Choose a large starting α, a decrement $\Delta\alpha$, and number N of α values ;
for $i = 1$ to N **do**
  Find the u_i that maximizes $Q(\alpha)$ (Appendix O) ;
  From (12.41) compute $A_i(\alpha)$;
  Compute $\chi^2(\alpha)$;
  Compute $P(\alpha|\bar{G})$;
  $\alpha \leftarrow \alpha - \Delta\alpha$;
end for
return $A_i(\alpha)$, $\chi^2(\alpha)$, and $P(\alpha|\bar{G})$ as a function of α.

straightforward. However, this simpler approach to the solution is no substitute for the procedures mentioned in the previous paragraph.

In the algorithm, we see that first steps are inputting information generated in the "qualifying the data" step of the analysis and transforming it for use in the minimization procedure. We first define a grid of τ and ω values to discretize the integral equation (12.38), taking care of possible symmetries in the kernel. This same grid discretizes $A(\omega)$ and $m(\omega)$ as probabilities over intervals on this grid. We also need to choose the prior probability $P(\alpha)$, usually the flat or Jeffery's prior.

The key part of qualifying the data is accurately estimating the covariance matrix C. The first step in the maximization algorithm is diagonalizing C, and then transforming the Green's function data and the kernel to this space. Next, the transpose of the transformed kernel is factorized by a singular value decomposition, and the dimension s of the dominant subspace is determined. The matrix $\bar{U}$ from this factorization (used in (12.41)), with an initial guess for A and the choice of the default model, yields a starting point u for the algorithm that maximizes $Q(\alpha)$ for a sequence of decreasing values of α. Appendix O describes a specialized form of the Levenberg-Marquardt method that does this. It is similar to the Newton method's use for optimizing trial wave functions in the variational Monte Carlo method (Section 9.3) but is more involved as it exploits the fact that the ill-posed nature of the problem makes the effective parameter space much smaller than the number of parameters being fitted. Once a u is found that maximizes Q, it is a simple matter to find the corresponding A and to compute $\chi^2(\alpha)$ and $P(\alpha|\bar{G})$. The value of u is

https://doi.org/10.1017/CBO9780511902581.013 Published online by Cambridge University Press

used as the starting point for the solution for a smaller value of α. The algorithm returns the saved values of α, $A_i(\alpha)$, $\chi^2(\alpha)$, and $P(\alpha|\bar{G})$.

Computing all three solutions generates a comparative basis for confidence in the average spectrum result. There are several other things that we can do. One is studying how the solutions change if less or more statistically independent measurements are used. Another is reducing the number of τ values used. The imaginary-time Green's functions are very smooth, U-shaped curves. This behavior means successive τ values are correlated. The natural tendency is to improve the solution by increasing the number of τ values. However, because of the correlations, increasing the number of τ values only slowly adds new information to the solution.

3. With the solution stable relative to the quality of the numerical input, the third and final step of our procedure is to assess the acceptance of the solution with respect to its sensitivity to the default model and to estimate its errors. Does the result have a feature not present in the default model or does it lack such a feature? How does the situation change if more data or less data are used? If the result and the model have the same features, how does the result change if more or less data are used? Good solutions exhibit relative independence to the choice of the model. How does the result change if we change the model? For example, if the model were a Gaussian centered at a location where a single peak in the spectrum may be expected, we can ask how much the result would vary if we were to change the peak position and width of the Gaussian. A flat default model is the least commitment we can make about our prior knowledge. If we switch from a more physically motivated model to a flat one, what are the differences? Do we believe they are significant? While the default model can serve to represent our state of prior knowledge about the solution, one that is too informative (for example, the exact solution) can be counterproductive. In the absence of data the solution defaults to the model. Data pull the solution from the model toward one of many possible results consistent with the data. If the data do not support details of the model (or the exact solution), the result might prove difficult to accept.

We can perform a limited form of error estimation. We need to forgo the concept of an "error bar." We cannot estimate the error associated with individual values of A_i, because we do not know the amount of correlations between points,[3] but rather we can estimate the error in functions $f(\omega)A(\omega)$ integrated over a frequency interval. We recall the definitions of the posterior probability (12.24) and the covariance of the spectral density (12.31). With this covariance, we compute the error associated with the measurement

$$F = \int d\omega f(\omega)A(\omega)$$

[3] The computed $A(\omega)$ is smooth and hence correlations exist among neighboring values of ω.

https://doi.org/10.1017/CBO9780511902581.013 Published online by Cambridge University Press

of some function of frequency $f(\omega)$ via

$$\left\langle F^2 \right\rangle = \iint d\omega d\omega' f(\omega) f(\omega') \left\langle \delta A(\omega) \delta A(\omega') \right\rangle. \tag{12.42}$$

This follows from the quadratic approximation, $P\left(A|\bar{G}\right) \propto e^{-\frac{1}{2}\delta A^T \cdot \nabla\nabla Q \cdot \delta A}$. We can readily approximate the expectation $\langle \delta A \delta A \rangle$ from the inverse of the Hessian (Appendix O, (O.3)) after the Levenberg-Marquardt maximization converges. If the function $f(\omega)$ is unity, then we can estimate the errors associated with regions of $A(\omega)$. Doing so is useful for establishing confidence in peaks and shoulders in the solution.

12.6 Comments

Admittedly, our discussion has been abstract. As we noted, several reviews give a detailed case study of the methods (Jarrell and Gubernatis, 1996; Jarrell et al., 2008). Our intent was to provide more details of the basis of the methods and to highlight the assumptions and approximations.

In our presentation we opted to describe the three Bayesian-based solutions by names that differ from their original presentations. Often, the phrase "the maximum entropy method" is used for all three, obscuring their differences and leaving vague which one is actually being used. The historical development of these three solutions is a progressive sequence refining what to do about the parameter α. From one point of view α is a Lagrange multiplier, and fixing such multipliers is a challenge in almost all constrained least-squares problems. Our use of the term "average spectrum method" is similar in spirit but different in detail from the average spectrum method proposed by several authors, for example, Syljuåsen (2008), whose solution is

$$\langle A \rangle = \frac{\int \mathcal{D}A \, A P(A|\bar{G})}{\int \mathcal{D}A \, P(A|\bar{G})}.$$

Typically, the averages in these proposals are computed by a Monte Carlo evaluation of the integrals, and different strategies, sometimes ad hoc ones, are used to fix α. A number of these issues have been reviewed and discussed by Fuchs et al. (2010). *In many respects, the broader issue is better estimating the evidence.* We approximated it. Doing this estimate via Monte Carlo (or other means) is an active research topic in Bayesian analysis (Friel and Wyse, 2012). The method of nested sampling has become widely used within certain communities (Sivia and Skilling, 2006, chapter 9).

Still other variations and alternatives to the three solutions described can be found in the literature. Limited space here prevents their review and assessment. In general, these publications address the perceived complexity of the procedure,

its tendency to produce smooth results, and its difficulty with having multiple peak structures and peaks at high frequencies. Clearly, the described procedures are not absolute, but care is needed in distinguishing between them and alternatives targeted for more specific problems or burdened with shortcuts and misconceptions. Definitely, the solution procedures are not black boxes; that is, piping the quantum Monte Carlo output into any of them will not automatically produce an acceptable result. In some sense, the three procedures are hypotheses, testable by refining the input (the quantum Monte Carlo data and prior information), approximations, and assumptions. Hopefully, the just completed discussion imparts the understanding and insight to evaluate the past and promote future advances.

The presented analysis makes Gaussian approximations to multiple integrands, rendering exact expressions for the resulting integrals. The main purpose was to develop insight into what makes the solution difficult. A consequence was the ability to develop an algorithm that executes in a trivial amount of computer time. Proposed Monte Carlo approaches to analytical continuation are devoid of these approximations but require computer time that is a significant fraction of the time needed to generate the quantum Monte Carlo data. If the Gaussian approximations become invalid, then the numerical algorithm can yield unreliable results, and Monte Carlo solutions are the only hope. We note that in another field, for a quite different problem, the breakdown of these approximations has been reported (Skilling, 1998).

The Bayesian approach and the principle of maximum entropy have been used in a variety of quantum Monte Carlo contexts for purposes other than finding a spectral density. We note the use of this approach for extracting thermodynamics (Huscroft et al., 2000), ground state gaps (Cafferel and Ceperley, 1992), and excited states (Blume et al., 1997, 1998). Analytic continuation in the presence of a sign problem is discussed in Jarrell et al. (2008).

Suggested reading

M. Jarrell and J. E. Gubernatis, "Bayesian inference and the analytic continuation of imaginary-time quantum Monte Carlo data," *Phys. Rept.* **269**, 133 (1996).

D. S. Sivia and J. Skilling, *Data Analysis: A Bayesian Tutorial* (Oxford University Press, 2006).

Exercises

12.1 What distribution function results from maximizing the entropy subject to the constraints of normalization and knowledge of the mean? Recall that a Gaussian results if the constraints also include the variance.

12.2 Assign a probability p_i to each face of a die by maximizing the entropy $S = -\sum_{i=1}^{6} p_i \log p_i$ subject to the constraints $\sum_{i=1}^{6} p_i = 1$ and $\sum_{i=1}^{6} i p_i = 3.5$. The latter is the average face value if each of the six faces is equally likely.

1. How does the solution change if $3.5 \to 3.5 \pm 0.5$?
2. Will any prior knowledge of the mean give roughly 1/6 for all p_i?

12.3 Construct a discrete probability $p = (p_1, p_2, \ldots, p_N)$ that has two prominent peaks, and set the default model $m = (m_1, \ldots, m_N)$ equal to it. Keeping m fixed, permute the indices of p and study how values of $S = -\sum_i p_i \ln(p_i/m_i)$ change as the peak positions in the permuted p move relative to those in m. Note that $S' = -\sum_i p_i \ln p_i$ is invariant to the permutations and that S takes a minimum when $p = m$.

12.4 Consider the inverse problem

$$\begin{pmatrix} 11 \\ 2 \end{pmatrix} = \begin{pmatrix} 10 & 1 \\ 1 & 0.1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix},$$

where x_1 and x_2 are known to be positive, but not normalized, and the value of x_1 has a standard deviation of 0.01 and x_2 has one of 1. Using a flat default model, compare the exact values of x_1 and x_2 with the ones that maximize $Q = \alpha S - \frac{1}{2}\chi^2$ as a function of α obtained using a quadratic approximation to S. The quadratic approximation reduces the maximization problem to a linear one. Using a quadratic prior probability is called *Tikhonov regularization*.

12.5 Repeat the above problem using the exact expression for S.

12.6 Show that the alternative expression for the entropy,

$$S(A) = \sum_i \left[A_i - m_i - A_i \log\left(\frac{A_i}{m_i}\right) \right], \tag{12.43}$$

allows the normalization conditions on A_i and m_i to be relaxed.

12.7 Show that the entropy of a multivariate Gaussian is $\frac{1}{2}\ln[(2\pi e)^N \det C]$.

12.8 A third of all kangaroos have blue eyes and a quarter of all kangaroos are left-handed. On the basis of this information alone, what proportion are both blue-eyed and left-handed (Gull and Skilling, 1984)? This information alone will not allow a unique answer. Select an answer based on constraining the problem with (1) $-\sum_i p_i$ and (2) $-\sum_i p_i^2$, subject to the constraints of the available information. We have no prior knowledge that the handedness and eye color are correlated, so it is reasonable to expect that for a given eye color, left- and right-handed kangaroos are equally likely. Which functional form does not inject correlations that are not part of the prior information? Now constrain with (1) $\sum_i \log p_i$ and (2) $\sum_i \sqrt{p_i}$. How do the correlations change?

https://doi.org/10.1017/CBO9780511902581.013 Published online by Cambridge University Press