Exercises

N.6 **A fair split? Number partitioning.**[1] [2] (Computer science, Mathematics, Statistics) ③

A group of N kids want to split up into two teams that are evenly matched. If the skill of each player is measured by an integer, can the kids be split into two groups such that the sum of the skills in each group is the same?

This is the *number partitioning problem* (NPP), a classic and surprisingly difficult problem in computer science. To be specific, it is **NP**–complete—a category of problems for which no known algorithm can guarantee a resolution in a reasonable time (bounded by a polynomial in their size). If the skill a_j of each kid j is in the range $1 \leq a_J \leq 2^M$, the 'size' of the NPP is defined as NM. Even the best algorithms will, for the hardest instances, take computer time that grows faster than any polynomial in MN, getting exponentially large as the system grows.

In this exercise, we shall explore connections between this numerical problem and the statistical mechanics of disordered systems. Number partitioning has been termed 'the easiest hard problem'. It is genuinely hard numerically; unlike some other **NP**–complete problems, there are no good heuristics for solving NPP (i.e., that work much better than a random search). On the other hand, the random NPP problem (the ensembles of all possible combinations of skills a_j) has many interesting features that can be understood with relatively straightforward arguments and analogies. Parts of the exercise are to be done on the computer; hints can be found on the computer exercises portion of the book Web site [8].

We start with the brute-force numerical approach to solving the problem.

(a) *Write a function* `ExhaustivePartition(S)` *that inputs a list S of N integers, exhaustively searches through the 2^N possible partitions into two subsets, and returns the minimum cost (difference in the sums). Test your routine on the four sets [5] $S_1 =$ $[10, 13, 23, 6, 20]$, $S_2 = [6, 4, 9, 14, 12, 3, 15, 15]$, $S_3 = [93, 58, 141, 209, 179, 48, 225, 228]$, and $S_4 = [2474, 1129, 1388, 3752, 821, 2082, 201, 739]$.* Hint: S_1 has a balanced partition, and S_4 has a minumum cost of 48. You may wish to return the signs of the minimum-cost partition as part of the debugging process.

What properties emerge from studying ensembles of large partitioning problems? We find a *phase transition*. If the range of integers (M digits in base two) is large and there are relatively few numbers N to rearrange, it is unlikely that a perfect match can be found. (A random instance with $N = 2$ and $M = 10$ has a one chance in $2^{10} = 1024$ of a perfect match, because the second integer needs to be equal to the first.) If M is small and N is large it should be easy to find a match, because there are so many rearrangements possible and the sums are confined to a relatively small number of possible values. It turns out that it is the ratio $\kappa = M/N$ that is the key; for large random systems with $M/N > \kappa_c$ it becomes extremely unlikely that a perfect partition is possible, while if $M/N < \kappa_c$ a fair split is extremely likely.

(b) *Write a function* `MakeRandomPartitionProblem(N,M)` *that generates N integers randomly chosen from $\{1, \ldots, 2^M\}$, rejecting lists whose sum is odd (and hence cannot have perfect partitions). Write a function* `pPerf(N,M,trials)`*, which generates* `trials` *random lists and calls* `ExhaustivePartition` *on each, returning the fraction p_{perf} that can be partitioned evenly (zero cost). Plot p_{perf} versus $\kappa = M/N$, for $N = 3$, 5, 7 and 9, for all integers M with $0 < \kappa = M/N < 2$, using at least a hundred trials for each case. Does it appear that there is a phase transition for large systems where fair partitions go from probable to unlikely? What value of κ_c would you estimate as the critical point?*

Should we be calling this a phase transition? It emerges for large systems; only in the 'thermody-

[1] New exercise supplementing *Statistical Mechanics: Entropy, Order Parameters, and Complexity* by James P. Sethna, copyright Oxford University Press, 2007, page 7. A pdf of the text is available at pages.physics.cornell.edu/sethna/StatMech/ (select the picture of the text). Hyperlinks from this exercise into the text will work if the latter PDF is downloaded into the same directory/folder as this PDF.

[2] This exercise draws heavily from [5, chapter 7].

namic limit' where N gets large is the transition sharp. It separates two regions with qualitatively different behavior. The problem is much like a spin glass, with two kinds of random variables: the skill levels of each player a_j are fixed, 'quenched' random variables for a given random instance of the problem, and the assignment to teams can be viewed as spins $s_j = \pm 1$ that can be varied ('annealed' random variables)[3] to minimize the cost $C = |\sum_j a_j s_j|$.

(c) *Show that the square of the cost C^2 is of the same form as the Hamiltonian for a spin glass, $H = \sum_{i,j} J_{ij} s_i s_j$. What is J_{ij}?*

The putative phase transition in the optimization problem (part (b)) is precisely a zero-temperature phase transition for this spin-glass Hamiltonian, separating a phase with zero ground-state energy from one with non-zero energy in the thermodynamic limit.

We can understand both the value κ_c of the phase transition and the form of $p_{\rm perf}(N, M)$ by studying the distribution of possible 'signed' costs $E_{\mathbf{s}} = \sum_j a_j s_j$. These energies are distributed over a maximum total range of $E_{\max} - E_{\min} = 2\sum_{j=1}^{N} a_j \leq 2N\, 2^M$ (all players playing on the plus team, through all on the minus team). For the bulk of the possible team choices $\{s_j\}$, though, there will be some cancellation in this sum. The probability distribution $P(E)$ of these energies for a particular NPP problem $\{a_j\}$ is not simple, but the average probability distribution $\langle P(E) \rangle$ over the ensemble of NPP problems can be estimated using the central limit theorem. (Remember that the central limit theorem states that the sum of N random variables with mean zero and standard deviation σ converges rapidly to a normal (Gaussian) distribution of standard deviation $\sqrt{N}\sigma$.)

(d) *Estimate the mean and variance of a single term $s_j a_j$ in the sum, averaging over both the spin configurations s_j and the different NPP problem realizations $a_j \in [1, \ldots, 2^M]$, keeping only the most important term for large M.* (Hint: Approximate the sum as an integral, or use the explicit formula $\sum_1^K k^2 = K^3/3 + K^2/2 + K/6$ and keep only the most important term.) *Using the central limit theorem, what is the ensemble-averaged probability distribution $P(E)$ for a team with N players?* Hint: Here $P(E)$ is non-zero only for *even* integers E, so for large N $P(E) \approx (2/\sqrt{2\pi}\sigma) \exp(-E^2/2\sigma^2)$; the normalization is doubled.

Your answer to part (d) should tell you that the possible energies are mostly distributed among integers in a range of size $\sim 2^M$ around zero, up to a factor that goes as a power of N. The total number of states explored by a given system is 2^N. So, the expected number of zero-energy states should be large if $N \gg M$, and go to zero rapidly if $N \ll M$. Let us make this more precise.

(e) *Assuming that the energies for a specific system are randomly selected from the ensemble average $P(E)$, calculate the expected number of zero-energy states as a function of M and N for large N. What value of $\kappa = M/N$ should form the phase boundary separating likely from unlikely fair partitions? Does that agree well with your numerical estimate from part (b)?*

The assumption we made in part (e) ignores the correlations between the different energies due to the fact that they all share the same step sizes a_j in their random walks. Ignoring these correlations turns out to be a remarkably good approximation.[4] We can use the random-energy approximation to estimate $p_{\rm perf}$ that you plotted in part (b).

(f) *In the random-energy approximation, argue that $p_{\rm perf} = 1 - (1 - P(0))^{2^{N-1}}$. Approximating $(1 - A/L)^L \approx \exp(-A)$ for large L, show that*

$$p_{\rm perf}(\kappa, N) \approx 1 - \exp\left[-\sqrt{\frac{3}{2\pi N}} 2^{-N(\kappa - \kappa_c)} \right]. \quad (1)$$

Rather than plotting the theory curve through each of your simulations from part (b), we change variables to $x = N(\kappa - \kappa_c) + (1/2)\log_2 N$, where the

[3] *Quenched* random variables are fixed terms in the definition of the system, representing dirt or disorder that was frozen in as the system was formed (say, by quenching the hot liquid material into cold water, freezing it into a disordered configuration). *Annealed* random variables are the degrees of freedom that the system can vary to explore different configurations and minimize its energy or free energy.

[4] More precisely, we ignore correlations between the energies of different teams $\mathbf{s} = \{s_i\}$, *except* for swapping the two teams $\mathbf{s} \to -\mathbf{s}$. This leads to the $N - 1$ in the exponent of the exponent for $p_{\rm perf}$ in part (f). Notice that in this approximation, NPP is a form of the random energy model (REM, exercise N.5), except that we are interested in states of energy near $E = 0$, rather than minimum energy states.

theory curve

$$p_{\mathrm{perf}}^{\mathrm{scaling}}(x) = 1 - \exp\left[-\sqrt{\frac{3}{2\pi}}2^{-x}\right] \qquad (2)$$

is independent of N. If the theory is correct, your curves should converge to $p_{\mathrm{perf}}^{\mathrm{scaling}}(x)$ as N becomes large

(g) *Reusing your simulations from part (b), make a graph with your values of* $p_{\mathrm{perf}}(x, N)$ *versus* x *and* $p_{\mathrm{perf}}^{\mathrm{scaling}}(x)$. *Does the random-energy approximation explain the data well?*

Rigorous results show that this random-energy approximation gives the correct value of κ_c. The entropy of zero-cost states below κ_c, the probability distribution of minimum costs above κ_c (of the Weibull form, exercise N.4), and the probability distribution of the k lowest cost states are also correctly predicted by the random-energy approximation. It has also been shown that the correlations between the energies of different partitions vanish in the large (N, M) limit so long as the energies are not far into the tails of the distribution, perhaps explaining the successes of ignoring the correlations.

What does this random-energy approximation imply about the computational difficulty of NPP? If the energies of different spin configurations (arrangements of kids on teams) were completely random and independent, there would be no better way of finding zero-energy states (fair partitions) than an exhaustive search of all states. This perhaps explains why the best algorithms for NPP are not much better than the exhaustive search you implemented in part (a); even among **NP**–complete problems, NPP is unusually unyielding to clever methods.[5] It also lends credibility to the conjecture in the computer science community that $\mathbf{P} \neq \mathbf{NP}$–complete; any polynomial-time algorithm for NPP would have to ingeneously make use of the seemingly unimportant correlations between energy levels.

[5]The computational cost does peak near $\kappa = \kappa_c$. For small $\kappa \ll \kappa_c$ it's relatively easy to find a good solution, but this is mainly because there are so many solutions; even random search only needs to sample until it finds one of them. For $\kappa > \kappa_c$ showing that there is no fair partition becomes slightly easier as κ grows [5, fig 7.3].