Scientific Computing I: Algorithms

3 hours · 100 points

Answer all questions. This is a closed-book test; a calculator is allowed. Parts I and II need no justification. In Part III, show your reasoning: most of the credit is for the idea, not the arithmetic. Problems 13 and 14 use reduced units ($\varepsilon = \sigma = m = k_B = 1$).

(12 points) Part I: True/false questions

Mark each statement True (T) or False (F). Each question is worth 2 points.

1. ( ) A two-dimensional molecular-dynamics code uses a pair cutoff r_c. Replacing its all-pairs force loop by a cell list (square cells of side at least r_c; each atom searches its own cell and the eight neighbouring cells) changes the forces slightly, because some pairs within r_c are skipped.
2. ( ) When a Metropolis proposal is rejected, the unchanged configuration must be recorded again as the next sample. Averaging only over accepted configurations biases the estimates.
3. ( ) A program computes $|x|$ as `sqrt(x*x)`. An automatic-differentiation (AD) tool that uses the rule $d\sqrt{u}/du = 1/(2\sqrt{u})$ returns NaN (not a number) for the derivative at $x = 0$, in forward and in reverse mode.
4. ( ) A modern Hopfield network stores unit-norm patterns as the columns of X and updates its state by $x \leftarrow X\,\mathrm{softmax}(\beta X^T x)$. For every inverse temperature $\beta > 0$, one update returns the stored pattern with the largest inner product with x.
5. ( ) If loopy belief propagation (BP) converges, with its message residual down to 10^{-10}, its marginals are exact.
6. ( ) In branch and bound for a minimization, run to completion, the branching rule changes the size of the search tree but not the optimal objective value that is returned.

(18 points) Part II: Fill in the blanks

Each question is worth 3 points.

7. We estimate π as 4 times the fraction of random darts in the unit square that land inside the quarter circle. With $N = 10^6$ darts the standard error is 1.64×10^{-3}. A standard error of 10^{-4} needs about $N =$ _____ darts.
8. A diffusion model noises data by $x_t = \sqrt{\overline{\alpha}_t}x_0 + \sqrt{1-\overline{\alpha}_t}\varepsilon$ with $\varepsilon \sim \mathcal{N}(0,1)$. The data are $x_0 = \pm 1$ with equal probability. At $\overline{\alpha}_t = 0.64$, the ideal denoiser $\mathbb{E}[x_0 \mid x_t]$ returns _____ at $x_t = 0$ and _____ at $x_t = 0.45$.
9. Perplexity is e^L, with L the cross-entropy loss in nats per token. A model that gives equal probability to each of $V = 512$ tokens has perplexity _____. A model that assigns probability $1/2$ to the true next token at half of the positions and $1/8$ at the other half has perplexity _____.
10. In self-attention, a query q and a key k are independent, each with $d = 64$ independent entries of mean 0 and variance 1. Their dot product has standard deviation _____. Without rescaling, at large d the softmax output approaches a _____ vector and its gradients vanish.
11. Training a dense transformer with N non-embedding parameters on D tokens, at short context and counting a multiply–add as two operations, costs about $C \approx$ _____ ND floating-point operations.
12. The best rank-1 approximation M_1 of $M = \begin{bmatrix} 3 & 1 & 0 \\ 1 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ has relative Frobenius error $\|M - M_1\|_F / \|M\|_F =$ _____.

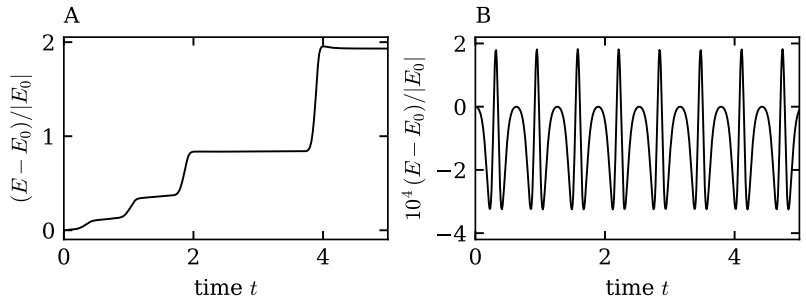
(70 points) Part III: Problems. Show all your work and justify your answers.

13. (14 points) Molecular dynamics and automatic differentiation. Two atoms interact through the Lennard-Jones potential $U(r) = 4(r^{-12} - r^{-6})$.

(a) (3 pts) Find the separation r_0 of the minimum and $U(r_0)$. Derive the force $F(r) = -dU/dr$. What does the sign of F mean?

(b) (5 pts) Write one step of velocity Verlet. How many new force evaluations does each step need? The figure shows the relative energy error of a dimer released from rest at $r = 1.2$, integrated with $\Delta t = 0.01$ by forward Euler ($x_{n+1} = x_n + \Delta t\, v_n$, $v_{n+1} = v_n + \Delta t\, a_n$) and by velocity Verlet. Which panel is which? Explain the jumps in panel A and the behaviour in panel B. Does halving Δt give Euler the long-time behaviour of Verlet? Why or why not?

(c) (6 pts) The energy is computed by the graph $a = r^{-6}$, $b = a^2$, $c = b - a$, $U = 4c$. (i) Propagate tangents in forward mode with $\dot r = 1$. (ii) Propagate adjoints in reverse mode with $\overline{U} = 1$. Which node collects two contributions? Which primal values must be stored? (iii) The energy of N atoms has $P = 3N$ inputs. Compare the cost of its gradient in the two modes. What does reverse mode pay instead? (iv) A central finite difference has truncation error $\sim h^2$ and rounding error $\sim 10^{-16}/h$. Estimate the order of magnitude of the best h.



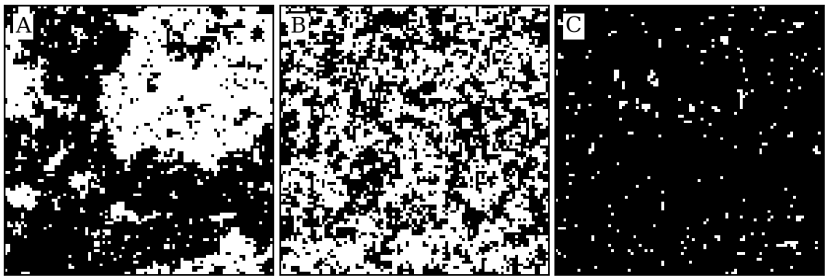
14. (14 points) Monte Carlo and honest error bars. The square-lattice Ising model has spins $s_i \in \{-1, +1\}$ on an $L \times L$ periodic lattice, energy $E = -\sum_{\langle ij \rangle} s_i s_j$ over nearest-neighbour pairs, Gibbs weights $P(s) = e^{-E/T}/Z$, and magnetization per spin $m = L^{-2} \sum_i s_i$.

(a) (4 pts) Write the Metropolis acceptance probability for a proposed flip of spin i. Show that $\Delta E = 2 s_i (s_1 + s_2 + s_3 + s_4)$, where s_i is the spin before the flip and $s_1, ..., s_4$ are its neighbours, and list the possible values. Compute the acceptance probabilities at $T = 2.3$. State detailed balance, and show that it makes P stationary. What else makes P the unique stationary distribution?

(b) (3 pts) The figure shows equilibrium snapshots (black and white are the two spin values) at $T = 1.9$, 2.269 and 3.2, in random order. Match them. For each, describe the size of the regions that fluctuate together (the correlation length). On a finite lattice $\langle m \rangle = 0$ at every temperature; why, and what do we measure instead?

(c) (4 pts) At $T = 2.3$ and $L = 64$, the integrated autocorrelation time of $|m|$ is $\tau_{\text{int}} = 1/2 + \sum_{t \geq 1} \rho(t) \approx 670$ sweeps, where $\rho(t)$ is the normalized autocorrelation at lag t. A run of $n = 10^5$ sweeps, one measurement per sweep, is analysed as if the samples were independent. By what factor is the naive error bar too small? How many effectively independent samples are there? Describe how binning reveals the honest error bar, and why this run is barely long enough.

(d) (3 pts) Near T_c the Metropolis autocorrelation time grows as $\tau \sim L^z$ with $z \approx 2.2$. Explain the origin of this critical slowing down. The Wolff algorithm grows a cluster from a random site, adds each aligned neighbour with probability $p = 1 - e^{-2/T}$, and flips the whole cluster with no acceptance test. Show that this obeys detailed balance. Why does it help near T_c?

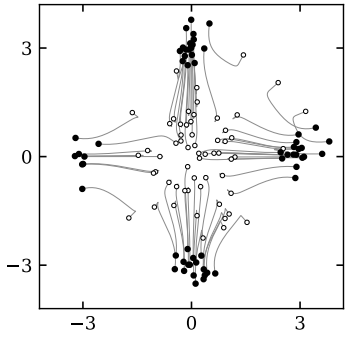


15. (14 points) Generative models: energy and transport.

(a) (5 pts) A restricted Boltzmann machine (RBM) has visible bits $v \in \{0,1\}^{n_v}$, hidden bits $h \in \{0,1\}^{n_h}$, energy $E(v,h) = -a^T v - b^T h - v^T W h$ and distribution $p_\theta(v,h) = e^{-E(v,h)}/Z$. Show that $p\big(h_j = 1 \mid v\big) = \sigma\Big(b_j + \sum_i v_i W_{ij}\Big)$, with $\sigma(u) = 1/(1+e^{-u})$, and that the hidden bits are independent given v. With the free energy $F(v) = -\ln \sum_h e^{-E(v,h)}$, the negative log-likelihood has gradient $\nabla \mathcal{L} = \mathbb{E}_{\text{data}}[\nabla F] - \mathbb{E}_{p_\theta}[\nabla F]$. Which term is intractable, and why? How does contrastive divergence (CD-k) replace it, and what is lost?

(b) (4 pts) A Hopfield network stores P patterns $\xi^\mu \in \{\pm 1\}^N$ with independent, uniformly random entries, using $J_{ij} = (1/N)\sum_\mu \xi_i^\mu \xi_j^\mu$ and $J_{ii} = 0$. Recall sets $s_i \leftarrow \text{sign}(h_i)$ with the local field $h_i = \sum_j J_{ij} s_j$. Set the network to pattern ξ^ν. Show that $\xi_i^\nu h_i$ equals about 1 plus a noise term of variance about $\alpha = P/N$. Bit i is unstable if $\xi_i^\nu h_i < 0$; estimate the unstable fraction using the standard normal cumulative distribution function Φ. Why does the network fail on binarized images of handwritten digits even at small loads?

(c) (5 pts) Flow matching trains $v_\theta(x,s)$ by minimizing the expected squared error $\mathbb{E}\|v_\theta(x_s,s) - u\|^2$ over the straight segments $x_s = (1-s)x_0 + sz$. Here x_0 is data at $s=0$, z is Gaussian noise at $s=1$ drawn independently of x_0, and $u = z - x_0$. What function does the trained network approximate? Write the continuity equation for the density p_s of x_s. The figure shows exact paths of the ordinary differential equation (ODE) $dx/ds = v(x,s)$, from noise at $s=1$ (open circles) to an equal-weight mixture of four Gaussian clusters centred at $(\pm 3, 0)$ and $(0, \pm 3)$ (filled). Every training segment is straight; why are the paths curved? What does one Euler step from $s=1$ directly to $s=0$ produce, and why?



16. (14 points) Tensor networks: inference and compression. A tensor network is a set of arrays with labelled indices; connected indices are summed and open ones form the output.

(a) (4 pts) Three $n \times n$ matrices are multiplied as $\sum_{jk} A_{ij} B_{jk} C_{kl}$, with $n = 10$. The cost of one pairwise contraction is the product of the sizes of all distinct indices that appear in either tensor (one operation per index combination). Find the total cost and the largest intermediate for the orders $(AB)C$ and $(AC)B$. Does the order change the result? Can a good order always keep the intermediates small?

(b) (4 pts) A Bayesian network has binary variables smoking S, lung cancer L and positive X-ray X, with arrows $S \to L \to X$. Each variable has states ordered (yes, no). Its tables are $p(S = \text{yes}) = 0.5$; $p(L = \text{yes} \mid S = \text{yes}) = 0.15$, $p(L = \text{yes} \mid S = \text{no}) = 0.05$; $p(X = \text{yes} \mid L = \text{yes}) = 0.9$, $p(X = \text{yes} \mid L = \text{no}) = 0.1$. Describe the tensor network for $p(L, X = \text{yes})$: which index is summed by contraction, and which vector is attached to the index of X? Compute $p(L = \text{yes} \mid X = \text{yes})$. Suppose the network also had bronchitis B with table $p(B \mid S)$, unobserved. Which vector goes on its index, and why does it not change the answer?

(c) (6 pts) An $m \times n$ matrix truncated to rank χ by the singular-value decomposition (SVD) is stored as two factors. (i) How many numbers does it need? For a 1000×1000 matrix, for which χ does truncation save storage? (ii) The boundary-MPS method contracts the Ising partition function of an open $L \times L$ square row by row. Why is exact row-by-row contraction exponential in L? How does a matrix product state (MPS) of bond dimension χ store the boundary instead, and with how many numbers? (iii) The specific-heat peaks on open $L \times L$ squares are at $T = 2.180$ ($L = 32$) and 2.224 ($L = 64$). Assuming $T_{\text{peak}}(L) = T_c + a/L$, estimate T_c. On periodic lattices the peaks approach T_c from above; why can the side differ?

17. (14 points) Hard problems: reductions and exact optimization. In (a) and (b), a bit $x \in \{0,1\}$ maps to a spin $s = 2x - 1$. A gadget is a small Ising energy $H = \sum_{i<j} K_{ij} s_i s_j + \sum_i h_i s_i$. Its lowest energy is its floor, and its floor states should be exactly the allowed rows of a truth table.

(a) (5 pts) Show that $H_{\mathrm{AND}} = s_a s_b - 2s_a s_z - 2s_b s_z - s_a - s_b + 2s_z$ has energy -3 on the four rows with $z = a \wedge b$. Compute the energy of the row $a = b = 1$, $z = 0$. Why can no gadget with bonds only (all $h_i = 0$) implement AND? Gadgets that share spins are added to build a multiplier; why can a violated gate not be compensated by another gate? A multiplier with two-bit inputs $p, q \in \{0,1,2,3\}$ has its output pinned to 7, which no such pair multiplies to. What is true of its ground state?

(b) (4 pts) At inverse temperature β, the partition function $Z = \sum_s \prod_\alpha e^{-\beta E_\alpha(s)}$, with one weight per bond or field term α, is a tensor contraction. Turn it into a contraction that returns $E_{\min} = -\lim_{\beta\to\infty}(\ln Z)/\beta$: what replaces each weight, each product and each sum? How do you also count the ground states? Deciding whether a spin glass with bonds between arbitrary pairs has a state with energy at most E_0 is in NP; explain why. Why does the same argument not let you verify a claimed number of ground states, and which complexity class contains this counting problem?

(c) (5 pts) Part (c) uses the integer-programming convention $H = \sum J_{ij} s_i s_j$, with no fields, so $J_{ij} > 0$ favours opposite spins. Set $x_i = (1 - s_i)/2$ and $z_{ij} = x_i$ XOR x_j, so $s_i s_j = 1 - 2z_{ij}$. Write four linear inequalities that force z_{ij} to equal the XOR when x is binary, and check them. A square of four spins (bonds 0–1, 1–2, 2–3, 3–0) has $J = (+1, +1, +1, -1)$ and ground-state energy -2. Show that the linear-programming (LP) relaxation, with $0 \le x_i, z_{ij} \le 1$, reaches -4 at $x_i = 1/2$ and cannot go lower. Give the integrality gap $E_{\mathrm{IP}} - E_{\mathrm{LP}}$. In branch and bound, when does a node stop being split? Most-fractional branching splits on the variable closest to $1/2$. Strong branching solves both child LPs for every candidate and picks the one that raises the bounds most. Why is most-fractional branching uninformative at the root of this square? Why can strong branching give a smaller tree yet run slower?