2024 Alibaba Global Mathematics Competition (Final Round)

Algebra & Number Theory

Question 1

Let $\mathrm{Mat}_2(\mathbb{Z})$ be the ring of 2×2 matrices with integral coefficients, and R the subring

$$\left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{Mat}_2(\mathbb{Z}) \middle| a \equiv d, c \equiv 0 \mod 2024 \right\}.$$

We consider $V := \mathbb{Q}^2$ as a left R-module via the natural left action of $\mathrm{Mat}_2(\mathbb{Z})$ on V. An R-lattice of V is a left R-submodule $L \subset V$ such that L is finitely generated as a $\mathbb{Z}$-module and $L \otimes_{\mathbb{Z}} \mathbb{Q} = V$. Two R-lattices are equivalent if they are isomorphic as left R-modules. Find the number (finite or infinite) of equivalence classes of R-lattices of V; justify your answer.

Question 2

We say that an ideal I of the polynomial ring $\mathbb{C}[x, y]$ is **scaling invariant**, if for every pair of elements $\lambda, \mu \in \mathbb{C}\backslash\{0\}$, the ideal generated by all elements of the form $f(\lambda x, \mu y)$ with $f(x, y) \in I$ recovers I itself. Find the number (finite or infinite) of scaling invariant ideals $I \subset \mathbb{C}[x, y]$ satisfying that $\dim_{\mathbb{C}} (\mathbb{C}[x, y]/I) = 6$; justify your answer.

Question 3

Let n and d be positive integers. Let $F_1, \ldots, F_m$ be homogeneous polynomials in $\mathbb{C}[X_0, \ldots, X_n]$ of degree at most d such that the set

$$V(F_1, \ldots, F_m) := \{[x_0 : \cdots : x_n] \in \mathbb{CP}^n | F_1(x_0, \ldots, x_n) = \cdots = F_m(x_0, \ldots, x_n) = 0\}$$

is finite, where $\mathbb{CP}^n$ denotes the n-dimensional complex projective space. Prove that the cardinality of $V(F_1, \ldots, F_m)$ is at most d^n.

Question 4

Let $p > 5$ be a prime number. Prove that the equation

$$\prod_{k=1}^{(p-1)/2} (X - 2\cos(2\pi k/p)Y) = p^2$$

has no solutions in $\mathbb{Z}^2$.

Question 5

Let H be the submonoid of $\mathrm{GL}_4(\mathbb{R})$ generated by matrices

$$\begin{pmatrix} 1 & a & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & b & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & c \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

for $a, b, c \geqslant 0$. Take elements $y_i, z_i \in H$ for $i = 1, 2, 3, \dots$ such that the sequence $(y_i z_i)_{i \geqslant 1}$ converges. Prove that there exists an infinite sub-sequence $(i_n)_{n \geqslant 1}$ of $(1, 2, 3, \dots)$, such that both sequences $(y_{i_n})_{n \geqslant 1}$ and $(z_{i_n})_{n \geqslant 1}$ converge.

Question 6

Let p be a prime number. Consider a semi-product group $G = L \rtimes H$ in which L is a cyclic p-group and H is a finite cyclic group, together with a finitely generated $\mathbb{F}_p[H]$-module M satisfying $\mathrm{Hom}_H(L, \mathrm{End}_{\mathbb{F}_p}(M)) = 0$. Here, an element $h \in H$ acts on $\varphi \in \mathrm{End}_{\mathbb{F}_p}(M)$ via the formula $(h\varphi)(m) = h\varphi(h^{-1}m)$ for $m \in M$. Suppose that we are given a decomposition

$$M = M_1 \oplus \cdots \oplus M_n$$

of $\mathbb{F}_p[H]$-modules satisfying $\mathrm{Hom}_H(M_i, M_j) = 0$ for $i \neq j$. Prove that for every positive integer d and every $(\mathbb{Z}/p^d\mathbb{Z})[G]$-module N satisfying $N \otimes_{\mathbb{Z}/p^d\mathbb{Z}} \mathbb{F}_p \simeq M$ as $\mathbb{F}_p[G]$-modules (where we regard H as a natural quotient of G), there is a unique decomposition

$$N = N_1 \oplus \cdots \oplus N_n$$

of $(\mathbb{Z}/p^d\mathbb{Z})[G]$-modules satisfying $N_i \otimes_{\mathbb{Z}/p^d\mathbb{Z}} \mathbb{F}_p \simeq M_i$ for $1 \leqslant i \leqslant n$.

Geometry & Topology

Question 1

Let (M, g) be a compact Riemannian manifold with nonnegative Ricci curvature. If for every positive δ, there exists a finite covering map $\pi : \hat{M} \to M$ such that the injectivity radius of $(\hat{M}, \pi^* g)$ is larger than δ, then (M, g) is flat.

Question 2

Let T^n be the n–dimensional torus, $f : T^n \to T^n$ be a continuous map, and $f_* : H_1(T^n; \mathbb{R}) \to H_1(T^n; \mathbb{R})$ be the induced map. Suppose that there exists a norm $|| \cdot ||$ on $H_1(T^n; \mathbb{R})$, such that for every nonzero $a \in H_1(T^n; \mathbb{Z})$, there exists a positive integer k with $||f_*^k(a)|| < ||a||$, where f^k is the kth iteration of f. Prove: f always has a fixed point.
(A *norm* on a vector space V over $\mathbb{R}$ is a map $|| \cdot || : V \to \mathbb{R}$, satisfying:
- $||v|| \geqslant 0$ for all $v \in V$, and the equality holds if and only if $v = 0$;
- $||\lambda v|| = |\lambda| \cdot ||v||$ for all $\lambda \in \mathbb{R}$ and $v \in V$;
- $||u + v|| \leqslant ||u|| + ||v||$ for all $u, v \in V$.)

Question 3

Consider a hypersurface in 3-dimensional complex projective space $\mathbb{C}P^3$ defined by the equation

$$z_0^n + z_1^n + z_2^n + z_3^n = 0,$$

where n is an integer, $n \geqslant 1$. Suppose this hypersurface carries a smooth circle action with only a prime number of isolated fixed points. Prove that $n = 1$ and construct such a circle action.

Question 4

Let N_g be the connected closed non-orientable surface of genus g. Let M be a connected closed oriented 3–manifold such that every smoothly embedded 2–sphere in M is the boundary of a 3–ball in M. Suppose that N_1 can be embedded into M smoothly. Prove that N_g can be embedded into M smoothly if and only if g is odd. (N_g is the g−fold connected sum of $\mathbb{R}P^2$ and g is called the genus of N_g.)

Question 5

Please construct a compact 4-manifold M with boundary a 3-torus having the following property: there exists an interior point P and four vector fields X_1, X_2, X_3, X_4 on $M \backslash \{P\}$ linearly independent everywhere such that the restrictions of X_1, X_2, X_3 on the boundary constitute the left invariant framing on the torus. For some metric g on M which restricts near the boundary to the product metric of $I \times S^1 \times S^1 \times S^1$, compute $\int_M (-\frac{\mathrm{Tr}(\Omega^2)}{8\pi^2})$, where Ω is the curvature form of the Levi-Civita connection of g.

Question 6

Let Σ be an embedded closed hypersurface in $\mathbb{R}^{n+1}$ ($n \geqslant 4$) with induced Riemannian metric g. Assume that for any $p \in \Sigma$ there is a local coordinate system $(x_1, \ldots, x_n)$ and a local smooth function u satisfying $g = e^u(\sum_i dx_i \otimes dx_i)$. Prove that at any point $p \in \Sigma$, there is a principal curvature with multiplicity at least $n-1$. Assume further that the set of non-umbilical point U is non-empty. Show that the distribution $\mathcal{D}$ on U generated by the eigenvectors associated to the multiplicity $n-1$ principal curvature is smooth and integrable.

Analysis & Differential Equations

Question 1

Given a positive constant ω, consider a nonzero tempered distribution $u \in \mathscr{S}'(\mathbb{R})$ satisfying the following equation (in the sense of distributions):

$$x\frac{d^2u}{dx^2} + (1-\omega^2 x)u = 0$$

(1) Prove that $u \in C(\mathbb{R}) \cap L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ (the L^p spaces are defined with respect to the Lebesgue measure on $\mathbb{R}$).

(2) Compute the value of

$$A := \frac{\left|\int_{\mathbb{R}} u(x)dx\right|^2}{\int_{\mathbb{R}} |u(x)|^2 dx}.$$

(3) For $\omega = \frac{1}{2}$, find the explicit formula of u.

Question 2

Let B be the set of all smooth, positive and periodic functions defined on the real line $\mathbb{R}$ with period 2π such that

$$f > 0, \quad \int_0^{2\pi} (f''(x))^2 dx \leqslant 1, \quad \forall f \in B.$$

For any given $k > 0$, let $S(k)$ be the set of $\alpha \in \mathbb{R}$ such that

$$\sup_{f \in B} \int_0^{2\pi} \frac{|f'(x)|^k}{(f(x))^\alpha} dx < \infty.$$

(1) Show that $S(4)$ is a closed interval and find the maximal of $S(4)$.

(2) Prove that there exists a constant C such that

$$|f'(x)| \leqslant C f^{\frac{1}{3}}(x), \quad \forall f \in B.$$

(3) Show that $S(2024)$ is a closed interval and find the maximal of $S(2024)$.

Question 3

For positive constants M and Q, define

$$f(r) = 1 - \frac{M}{r^2} + \frac{Q}{r^4} - r^2, \quad r > 0.$$

If f has three different positive roots $r_c > r_+ > r_- > 0$, show that $f'(r_+) + f'(r_-) < 0$.

Question 4

Consider the sequence

$$a_{n+1} = a_n + \frac{a_n^2}{n^2}, \quad 0 \leqslant a_1 < 1.$$

Show that the limit $\lim_{n\to\infty} a_n$ exists and is finite.

Question 5

Let Ω be a connected open set in $\mathbb{R}^d$ such that $\mathbb{R}^d \backslash \Omega$ contains an open cone $\mathcal{C}$.

Suppose that $u : \bar{\Omega} \to \mathbb{R}$ is a bounded continuous function, which is C^2 in Ω and satisfies

$$\begin{cases} \Delta u \geqslant 0 & \text{in } \Omega, \\ u \leqslant 0 & \text{on } \partial\Omega. \end{cases}$$

Show that

$$u \leqslant 0 \text{ in } \Omega.$$

Here the open cone $\mathcal{C}$ with vertex x_0 is the following open set in $\mathbb{R}^d$

$$\mathcal{C} = \{x : |x - x_0||v| \cos\theta < v \cdot (x - x_0)\}$$

for some $\theta \in (0, \frac{\pi}{2})$ and direction $v \in \mathbb{R}^d$, $v \neq 0$.

Question 6

Let $\mathcal{F}$ denote the set of all nondecreasing 1-Lipschitz functions $f : [0,1] \to [0,1]$, that is,

$$0 \leqslant f(x) - f(y) \leqslant x - y, \quad \forall 0 \leqslant y \leqslant x \leqslant 1, \quad \forall f \in \mathcal{F}.$$

(1) Prove that for any $\epsilon > 0$, there is is a positive constant τ, depending only on ϵ, such that for any $f \in \mathcal{F}$, there exists a subinterval $[a,b] \subset [0,1]$ with $b - a > \tau$ and $\sigma = \sigma(a,b) := \frac{f(b)-f(a)}{b-a}$,

$$f(a) + \sigma(x-a) + \epsilon(b-a) \geqslant f(x) \geqslant f(a) + \sigma(x-a) - \epsilon(b-a), \quad \forall x \in [a,b].$$

(2) Prove that for any $\epsilon > 0$, there exists a positive constant τ, depending only on ϵ, such that for any $f \in \mathcal{F}$, there is a decomposition of $[0,1]$ into consecutive intervals $\{[a_j, a_{j+1}]\}_{j=1}^J$ and $0 \leqslant \sigma_j < \sigma_{j+1} \leqslant 1$ such that $a_{j+1} - a_j \geqslant \tau$,

$$f(x) \geqslant f(a_j) + \sigma_j(x - a_j) - \epsilon(a_{j+1} - a_j), \forall x \in [a_j, a_{j+1}],$$

and

$$\sum_j \sigma_j(a_{j+1} - a_j) \geqslant f(1) - \epsilon.$$

Applied & Computational Mathematics

Question 1

For any $A, B \in \mathbb{C}^{n\times n}$, denote $\lambda(A)$ and $\lambda(B)$ the sets of eigenvalues of matrices A and B respectively. Suppose the Jordan decomposition of matrix A is given by $A = P^{-1}JP$, where J is the Jordan canonical form of A. For any $\mu \in \lambda(B)$, prove the following:

(1) $\|(\mu I - J)^{-1}\|_2^{-1} \leqslant \theta$ when $\mu \notin \lambda(A)$;

(2) $\min_{\lambda\in\lambda(A)} |\lambda - \mu| \leqslant 2(1+\theta)\ln(1+\theta^{1/m})$;

where $\theta = m\kappa_2(P)\|A-B\|_2$, $\kappa_2(P) = \|P\|_2\|P^{-1}\|_2$ represents the condition number of matrix P under the 2-norm, and m denotes the order of the largest Jordan block in J.

Question 2

Let $F(x;w)$ denote a deep neural network with scalar output, where x is the input and w denotes the weights. Assume F is continuously differentiable with respect to w and is overparameterized for training data $\{x_j, y_j\}_{j=1}^m$ in the sense that there exists $w^\star$ such that $F(x_j, w^\star) = y_j$ for all j. To study the local optimization dynamics of training neural networks at $w^\star$, we consider the linearized neural network $\widetilde{F}(x;w) = F(x;w^\star) + (w-w^\star)^\top \nabla F(x;w^\star)$, with training loss

$$\text{Loss}(w) := \frac{1}{2m}\sum_{j=1}^m (y_j - \widetilde{F}(x_j;w))^2.$$

Letting s denote the learning rate, the rule of gradient descent is $w_{i+1} = w_i - s\nabla\text{Loss}(w_i)$, while stochastic gradient descent is $w_{i+1} = w_i - s(\nabla\text{Loss}(w_i) + \epsilon_i)$, where ϵ_i is a noise term satisfying $\mathbb{E}\epsilon_i = 0$ and $\mathbb{E}\epsilon_i\epsilon_i^\top = M(w_i)/b$, with b being the mini-batch size. Assume that the state-dependent covariance M aligns with

$$\Sigma = \frac{1}{m}\sum_{j=1}^m \nabla F(x_j, w^\star)\nabla F(x_j, w^\star)^\top$$

in the sense that

$$\frac{\text{Tr}(M(w)\Sigma)}{2\text{Loss}(w)\|\Sigma\|_F^2} \geqslant \delta$$

for $\delta > 0$ and all w. Here $\|\cdot\|_F$ denotes the Frobenius norm.

(1) For gradient descent, prove that if the spectral norm of Σ satisfies

$$\|\Sigma\|_2 \leqslant \frac{2}{s},$$

then the dynamics is stable in the sense that $\text{Loss}(w_i)$ is bounded for all i. (Note that this yields a dimension-dependent bound $\|\Sigma\|_F \leqslant \frac{2\sqrt{d}}{s}$, where d is the dimension of w.)

(2) For stochastic gradient descent, if the dynamics is stable in the sense that $\mathbb{E}\text{Loss}(w_i)$ is bounded for all i, then the following dimension-independent bound must hold:

$$\|\Sigma\|_F \leqslant \frac{\sqrt{b/\delta}}{s}.$$

Question 3

Consider the following system

$$\partial_t f + v \cdot \nabla_x f + E \cdot \nabla_v f = 0, \quad \partial_t E = -\int_{\mathbb{R}^d} v f \mathrm{d}v\,. \tag{1}$$

Here $t \in \mathbb{R}^+$, $x \in \mathbb{R}^d$ and $v \in \mathbb{R}^d$ are independent variables, and f is an unknown function that depends on t, x and v. E only depends on t and x.

a) Show that the mass and energy

$$m(t) := \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} f \mathrm{d}v \mathrm{d}x\,, \qquad e(t) := \frac{1}{2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} |v|^2 f \mathrm{d}x \mathrm{d}v + \frac{1}{2}\int_{\mathbb{R}^d} |E|^2 \mathrm{d}x$$

are conserved over time.

b) Assume now $E(t,x)$ is given, then (1) reduces to

$$\partial_t f + v \cdot \nabla_x f + E \cdot \nabla_v f = 0. \tag{2}$$

We will use particle method to solve (2) with initial condition

$$f(t=0,x,v) = f_{\text{in}}(x,v).$$

Consider P particles denoted by $\{x_p(t), v_p(t)\}_{1\leqslant p\leqslant P}$, where $x_p(t)$ represents the location and $v_p(t)$ the velocity of the $p-$particle, each with a weight $\frac{1}{P}$. The evolution of $x_p(t)$ and $v_p(t)$ is governed by the following equations:

$$\frac{dx_p}{dt} = v_p, \quad \frac{dv_p}{dt} = E(x_p)\,.$$

Show that

$$f^P(t,x,v) := \frac{1}{P}\sum_{p=1}^{P} \delta(x - x_p(t))\delta(v - v_p(t))$$

is a weak solution to (2) associated with the initial condition $f_0^P = \frac{1}{P}\sum_{p=1}^{P}\delta(x - x_p(0))\delta(v - v_p(0))$ in the sense of distribution. Here δ is the Dirac-Delta function.

c) Going back to the original system (1). Consider the following particle method

$$\begin{aligned} x_p^{n+1} &= x_p^n + \Delta t v_p^n \\ v_p^{n+1} &= v_p^n - E(x_p^n)\Delta t \\ E_i^{n+1} &= E_i^n - J_i^n \Delta t\,. \end{aligned}$$

Here the superscript n refers to the time step, and $E(x_p^n)$ and J_i^n are defined respectively as follow

$$E(x_p^n) := \sum_i S(x_i - x_p^n) E_i^n \Delta x^d \,,$$

$$J_i^n := \frac{1}{P} \sum_{p=1}^{P} S(x_i - x_p^n) v_p^n \,.$$

In these expressions, x_i represents a uniform grid with grid size Δx, S is a spline function satisfying $\int_{\mathbb{R}^d} S(x) \mathrm{d}x = 1$. Does this method conserve energy? That is, do we have

$$\frac{1}{2P} \sum_p (v_p^n)^2 + \frac{1}{2} \sum_i (E_i^n)^2 \Delta x^d = \frac{1}{2P} \sum_p (v_p^{n+1})^2 + \frac{1}{2} \sum_i (E_i^{n+1})^2 \Delta x^d \,?$$

If yes, prove it. If no, construct a scheme that preserves the energy at the discrete level.

Question 4

Consider the following two optimization problems:

$$(\mathrm{A}): \begin{array}{ll} \min\limits_{\mathbf{x}} & f(\mathbf{x}), \\ \text{s. t.} & \mathbf{g}(\mathbf{x}) = 0, \\ & \mathbf{x}_i \in \mathcal{F}_i, \ i = 1, \ldots, n \end{array} \qquad \text{and} \qquad (\mathrm{B}): \begin{array}{ll} \min\limits_{\mathbf{x}} & f(\mathbf{x}) + \beta \mathbf{g}(\mathbf{x})^\top \mathbf{1}_p, \\ \text{s. t.} & \mathbf{x}_i \in \mathcal{F}_i, \ i = 1, \ldots, n, \end{array}$$

where $\mathbf{x} := [\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top]^\top \in \mathbb{R}^m$ (m, $n \in \mathbb{N}$) and $\mathbf{x}_i \in \mathbb{R}^{m_i}$ ($m_i \in \mathbb{N}$) for $i = 1, \ldots, n$ such that $\sum_{i=1}^n m_i = m$. The functions $f : \mathbb{R}^m \to \mathbb{R}$ and $\mathbf{g} : \mathbb{R}^m \to \mathbb{R}^p$ ($p \in \mathbb{N}$) are multi-affine, i.e., for any $i \in \{1, \ldots, n\}$, they are affine with respect to $\mathbf{x}_i$ after fixing the other $n-1$ blocks. Here, a function $\mathbf{h} : \mathbb{R}^q \to \mathbb{R}^r$ (q, $r \in \mathbb{N}$) is affine if

$$\mathbf{h}(a\mathbf{y}^{(1)} + (1-a)\mathbf{y}^{(2)}) = a\mathbf{h}(\mathbf{y}^{(1)}) + (1-a)\mathbf{h}(\mathbf{y}^{(2)})$$

holds for any $a \in \mathbb{R}$ and $\mathbf{y}^{(1)}$, $\mathbf{y}^{(2)} \in \mathbb{R}^q$. For any $i \in \{1, \ldots, n\}$, the set $\mathcal{F}_i \subseteq \mathbb{R}^{m_i}$ is a bounded polyhedron. The function $\mathbf{g}$ is nonnegative on $\bigtimes_{i=1}^n \mathcal{F}_i$, where "$\times$" refers to the Cartesian product of sets. In problem (B), the scalar β is a real number and $\mathbf{1}_p \in \mathbb{R}^p$ denotes the p-dimensional all-ones vector.

Please prove the following three statements.

1. Problem (B) has at least one optimal solution that is an extreme point of the feasible region (i.e., extreme-point optimal solution) for any $\beta \in \mathbb{R}$.
2. There exists a $\bar{\beta} \in \mathbb{R}$ such that any extreme-point optimal solution of problem (B) solves problem (A) whenever $\beta \geqslant \bar{\beta}$.
3. There exists a $\tilde{\beta} \in \mathbb{R}$ such that the optimal solution sets of both problems (A) and (B) coincide whenever $\beta \geqslant \tilde{\beta}$.

Question 5

Consider the following system of stochastic differential equations (SDEs)

$$\mathrm{d}x_i^t = -x_i^t \mathrm{d}t - \frac{1}{N}\sum_{j=1}^{N} \nabla W(x_i^t - x_j^t)\mathrm{d}t + \sqrt{\frac{2}{\beta}}\mathrm{d}B_t^i, \quad i = 1, 2, \cdots, N.$$

Here, each particle x_i^t belongs to the Euclidean space $\mathbb{R}^d$, and the parameter $\beta > 0$ represents the inverse temperature. Note that $(B_\cdot^i)$ represents N independent standard Brownian motions in $\mathbb{R}^d$. We assume further that $W \in C_c^\infty(\mathbb{R}^d)$, $W(0) = 0$, and $W(x) = W(-x)$ for all $x \in \mathbb{R}^d$.

1. Let the joint distribution of the N particles be represented by $\rho_N(t, \cdot)$. Please write down the Fokker-Planck equation solved by ρ_N explicitly.
2. Show that the Gibbs measure M_N, defined as

$$M_N(x_1, x_2, \cdots, x_N) = \frac{1}{Z_N}\exp\Big(-\beta\Big(\sum_{i=1}^{N}\frac{1}{2}|x_i|^2 + \frac{1}{N}\sum_{1\leqslant i<j\leqslant N} W(x_i - x_j)\Big)\Big),$$

   where the constant Z_N is chosen to make M_N a probability density, is the unique stationary solution of the Fokker-Planck equation for ρ_N.
3. Assume that a probability density ρ satisfies the following nonlinear equation

$$\rho = \frac{1}{Z}\exp\Big(-\beta\big(\frac{1}{2}|x|^2 + W * \rho(x)\big)\Big),$$

   where the normalizing constant Z is defined as

$$Z = Z(\rho) = \int_{\mathbb{R}^d} \exp\Big(-\beta\big(\frac{1}{2}|x|^2 + W * \rho(x)\big)\Big)\mathrm{d}x.$$

   Show that there exists a critical $\beta_c > 0$, such that when $\beta < \beta_c$, it holds that

$$\sup_{N\geqslant 2}\int_{(\mathbb{R}^d)^N} M_N \log\frac{M_N}{\rho^{\otimes N}}\mathrm{d}x_1\mathrm{d}x_2\cdots\mathrm{d}x_N < \infty,$$

   where $\rho^{\otimes N}(x_1, x_2, \cdots, x_N) = \rho(x_1)\rho(x_2)\cdots\rho(x_N)$.

   Hint: For the 3rd part, you can directly apply the following conclusion: there exists a universal constant $c_0 > 0$, such that when $\|\phi\|_{L^\infty} \leqslant c_0$, it holds that

$$\int_{(\mathbb{R}^d)^N} \rho^{\otimes N}\exp\Big(N\int_{\mathbb{R}^{2d}}\phi(x,y)\mathrm{d}(\mu_N(x) - \rho(x))\mathrm{d}(\mu_N(y) - \rho(y))\Big)\mathrm{d}x_1\mathrm{d}x_2\cdots\mathrm{d}x_N,$$

   is uniformly bounded with respect to N, where ρ is a probability measure, and $\mu_N = \frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}$ is the empirical measure associated to the point $(x_1, x_2, \cdots, x_N) \in (\mathbb{R}^d)^N$.

Question 6

Studying the scaling laws for large models is important for reducing the training cost: how does the final test loss scale with the number of training steps and model size? In this problem, we study the scaling laws in training linear models.

1. First, we focus on the setting of learning a one-dimensional linear model with gradient descent.

   - Let the data distribution $\mathcal{D}$ be a distribution over $\mathbb{R}^2$. Each data point is an input-output pair (x, y), where $x \sim \mathcal{N}(0, 1)$ and $y \sim \mathcal{N}(3x, 1)$.
   - We use gradient descent to learn the following linear model: $f_w(x) = w \cdot x$, where $w, x \in \mathbb{R}$. We initialize $w_0 = 0$ and perform multiple iterations. At each iteration, we sample $(x_t, y_t) \sim \mathcal{D}$ and then update w_t as $w_{t+1} \leftarrow w_t - \eta \nabla \ell_t(w_t)$, where $\ell_t(w) = \frac{1}{2}(f_w(x_t) - y_t)^2$ is the squared loss and $\eta > 0$ is the learning rate.

   If we run gradient descent with learning rate $\eta \in (0, \frac{1}{3}]$ for $T \geqslant 0$ steps, what is the expected test loss $\overline{\mathcal{L}}_{\eta,T} = \mathbb{E}_{w_T} \mathbb{E}_{(x,y)\sim D}[\frac{1}{2}(f_{w_T}(x) - y)^2]$?

2. In the setting of Part 1, consider the case where η is tuned optimally. Find a function $g(T)$ such that the following holds when $T \to +\infty$:

$$\left| \inf_{\eta \in (0, \frac{1}{3}]} \overline{\mathcal{L}}_{\eta,T} - g(T) \right| = O\left(\frac{(\log T)^2}{T^2} \right)$$

3. It has been usually observed that pretraining a large language model approximately follows the Chinchilla scaling law:

$$\overline{\mathcal{L}}_{N,T} \approx \frac{A}{N^\alpha} + \frac{B}{T^\beta} + C,$$

   where $\overline{\mathcal{L}}_{N,T}$ is the test loss of a model with N parameters trained after T steps, and A, B, α, β, C are constants. Now, we exemplify a setting of training a multidimensional linear model that also exhibits a similar scaling law.

   - Fix $a > 0, b \geqslant 1$. Every data point consists of an input $x_\bullet$ that is an infinite-dimensional vector (a sequence), and an output $y \in \mathbb{R}$. The data distribution $\mathcal{D}$ is defined as follows. First, sample k from a Zipf distribution $\Pr[k = i] \propto i^{-(a+1)}$ $(i \geqslant 1)$. Set $j := \lceil k^b \rceil$. Then, set the j-th coordinate x_j of $x_\bullet$ as a random sample from $\mathcal{N}(0, 1)$, and set all the other coordinates as 0. Finally, $y \sim \mathcal{N}(3x_j, 1)$. $\mathcal{D}$ is defined as the distribution of $(x_\bullet, y)$ generated in this way.
   - We study a linear model that focuses only on the first N input coordinates. Let $\phi_N(x_\bullet) = (x_1, \ldots, x_N)$. The linear model is parameterized by $\boldsymbol{w} \in \mathbb{R}^N$ and produces the output as $f_{\boldsymbol{w}}(x_\bullet) = \langle \boldsymbol{w}, \phi_N(x_\bullet) \rangle$.
   - We use gradient descent to learn such a linear model. We initialize $\boldsymbol{w}_0 = 0$ and perform multiple iterations. At each iteration, we sample $(x_{t,\bullet}, y_t) \sim \mathcal{D}$ and then update $\boldsymbol{w}_t$ as $\boldsymbol{w}_{t+1} \leftarrow \boldsymbol{w}_t - \eta \nabla \ell_t(\boldsymbol{w}_t)$, where $\ell_t(\boldsymbol{w}) = \frac{1}{2}(f_{\boldsymbol{w}}(x_{t,\bullet}) - y_t)^2$.

Let $\overline{\mathcal{L}}_{\eta,N,T} = \mathbb{E}_{\boldsymbol{w}_T}\mathbb{E}_{(\boldsymbol{x},y)\sim D}[\frac{1}{2}(f_{\boldsymbol{w}_T}(\boldsymbol{x}) - y)^2]$ be the expected test loss after running gradient descent on the linear model of N parameters with learning rate $\eta \in (0, \frac{1}{3}]$ for $T \geqslant 0$ steps. Find α, β, C such that $\exists \gamma > 0, \forall c > 0$, the following holds when $T = N^{c+o(1)}$ and N is large enough:

$$\epsilon(N,T) := \frac{\inf_{\eta\in(0,\frac{1}{3}]} \overline{\mathcal{L}}_{N,T} - C}{\frac{1}{N^\alpha} + \frac{1}{T^\beta}}, \qquad (\log N + \log T)^{-\gamma} \leqslant \epsilon(N,T) \leqslant (\log N + \log T)^{\gamma}.$$

That is, $\inf_{\eta\in(0,\frac{1}{3}]} \overline{\mathcal{L}}_{N,T} = \tilde{\Theta}(N^{-\alpha} + T^{-\beta}) + C$, where $\tilde{\Theta}$ ignores polylog factors of N and T.

Combinatorics & Probability

Question 1

Let m be a positive integer. Consider a Markov chain $X = (X_n)_{n \geqslant 0}$ on $\mathbb{Z}$ whose transition probability $p_{i,j} := \mathbb{P}[X_{n+1} = j | X_n = i]$ satisfies: (1). $p_{i,j} \neq 0$ if and only if $|j - i| = 1$; (2). $p_{i,i+1} = p_{j,j+1}$ when $j - i = m$. Let $Y_n = X_n \mod m$. Then $Y = (Y_n)_{n \geqslant 0}$ can be viewed as a Markov chain on $\{0, 1, \ldots, m-1\}$. Let $(\mu_i)_{0 \leqslant i < m}$ be the stationary distribution of Y. Let $A = \sum_{i=0}^{m-1} \mu_i p_{i,i+1}$ and $T = \inf\{n \geqslant 0 : X_n = m\}$. Prove that $(2A - 1)\mathbb{E}[T \mid X_0 = 0] = m$ if $A > \frac{1}{2}$.

Question 2

A directed graph G is called *simple* if it has no loops and there is at most one directed edge between any two vertices. Let u, v be two distinct vertices in $V(G)$. We write $u \to v$ for an edge directed from u to v, and we say that u is an in-neighbor of v, and v is an out-neighbor of u. The *distance* $d(u, v)$ from u to v is the length of the shortest directed path from u to v in G. For integers $j \geqslant 1$, let $N_j^+(u)$ denote the set of vertices $v \in V(G)$ satisfying $d(u, v) = j$.

Let G be a simple directed graph such that for any vertex $u \in V(G)$, the number of out-neighbors of u equals the number of in-neighbors of u. Assume that there are no three vertices u, v, w in G satisfying $u \to v, v \to w$ and $u \to w$. Prove that

$$\sum_{v \in V(G)} |N_2^+(v)| \geqslant \sum_{v \in V(G)} |N_1^+(v)|.$$

Question 3

Let Y be a random variable taking values in $(-1, 1)$. Write the binary representation of a real number $y \in (-1, 1)$ by

$$y = \sum_{k=1}^{\infty} a_k 2^{-k}, \text{ where } a_k \in \{-1, 1\} \text{ for each } k \in \mathbb{N}.$$

Here, the binary representation is unique by forbidding the existence of k_0 such that $a_n = 1$ for all $n \geqslant k_0$. Let $y_s = \sum_{k=1}^{s} a_k 2^{-k}$, i.e., the first s digits of y, for each $s \in \mathbb{N}$. Define the stochastic process $(Y_s)_{s \in \mathbb{N}}$ as above, i.e., gradually revealing the digits of Y. Note that $Y_s \to Y$ as $s \to \infty$ (point-wise). For a strictly increasing function $g : [-1, 1] \to \mathbb{R}$, define the stochastic process $(X_s)_{s \in \mathbb{N}} := (g(Y_s))_{s \in \mathbb{N}}$. Suppose that $(X_s)_{s \in \mathbb{N}}$ is a martingale with respect to its natural filtration. Prove that there exist $r < 0.9$ and $C > 0$ such that

$$\mathbb{E}[(X_s - g(Y))^2] \leqslant C r^s \text{ for all } s \in \mathbb{N}.$$

(The value 0.9 is not optimal.)

Hint: You may try to prove and use the following result: Suppose that $k \geqslant 3$ and the positive numbers $a_1, \ldots, a_k, b_1, \ldots, b_k$ satisfy

$$b_{s-2} \geqslant \min\{a_s, a_{s-1}\} \text{ for } s = 3, \ldots, k;$$
$$b_{s-1} \geqslant \min\{a_s, b_s\} \text{ for } s = 2, \ldots, k.$$

Then there exists $r < 0.9$ such that

$$\prod_{s=1}^{k} \frac{a_s}{a_s + b_s} \leqslant \sqrt{\frac{b_1}{b_k}}\, r^{k-2}.$$

Question 4

Suppose that C is a convex shape in $\mathbb{R}^2$ with area 1, and suppose that $\mathcal{S}$ is a (possibly infinite) set of convex shapes in $\mathbb{R}^2$. For every convex shape $D \in \mathcal{S}$, there exists a constant $k \in \mathbb{R}$ such that $D = kC := \{k\vec{x} \colon \vec{x} \in C\}$. We say that the convex shapes in $\mathcal{S}$ can be *packed* inside C using translations only if there exists a mapping $t \colon \mathcal{S} \to \mathbb{R}^2$ such that for every D in $\mathcal{S}$, the interior of D translated by the vector $t(D)$ is contained in C, and for any two distinct D and D' in $\mathcal{S}$, the interior of D translated by $t(D)$ does not overlap with the interior of D' translated by $t(D')$. Prove that if the total area of the convex shapes in $\mathcal{S}$ is at most 1/8, then they can be packed inside C using translations only.

Question 5

On day 0, a bond is worth 1 dollar. On day n, it is worth $S_n := \exp(X_1 + \cdots + X_n)$ dollars, where X_i's are i.i.d. random variables such that $P(X_i = 1) = P(X_i = -1) = 1/2$. Alice has some spare money that will not be needed until day $N \geqslant 1$. During this period she wishes to make an investment on this bond. As an investor with a special taste, she only wishes to invest at a "signal time", which is some day $K \in [1, \ldots, N-1]$, such that the value of bond is at the highest between day 0 and day K, but at the lowest between day K and day N.

Of course, even if there is such a day K, on that day she cannot say for sure this is really the signal time, but in hindsight, it is natural to wonder if there is really a such a day at all. Please prove that there exists universal constants $c, C > 0$ such that

$$c/\log N \leqslant P[\text{Such a day } K \text{ exists}] \leqslant C/\log N.$$

Hint: Define $p_n = P[S_i \geqslant 1,\ \forall i = 1, \ldots, n]$, and note that

$$p_n^2 \leqslant P[1 \leqslant S_i \leqslant S_n,\ \forall 1 \leqslant i \leqslant n] \leqslant p_{[n/2]}^2.$$

[You get partial credit too if you can give a proof of this hint!]

Question 6

A proper *q-colouring* of a graph G is an assignment of q available colours to the vertices, so that no edge is monochromatic. For a colouring σ and any $v \in V$, let $L_\sigma(v)$ be the set of colours that are available at v, namely the colours that do not show up in the neighbourhood of v under σ.

Show that there is a $d_0 \geqslant 1$ such that for any integer $d \geqslant d_0$, the following holds: for any triangle-free $G = (V, E)$ with maximum degree d and any $v \in V$,

$$\mathbb{E}_\sigma[|L_\sigma(v)|] \geqslant d/3,$$

where σ is a uniformly random proper d-colouring.