ML Notes 10.28

Group 6

November 4, 2025

Term Project II

LLM for mathematical discoveries

Reference paper

1. DeepMind AlphaEvolve:
2. Fun Search

Methodology summary

- The methods in both papers are similar: they utilize a Large Language Model (LLM) to generate code for problem-solving.
- The initial code may be very simple or based on a template.
- This code is executed to produce results, which are then used as feedback for the LLM.
- The LLM uses this feedback to perform reflection and optimization, generating new, improved code.

Kissing number

Definition

In an n-dimensional space, what is the maximum number of non-overlapping unit spheres that can simultaneously touch a central unit sphere of the same size?

Conditions

1. There is a central unit sphere (radius 1) centered at the origin.
2. Several other unit spheres (also radius 1) are placed around it.
3. All surrounding spheres must touch (kiss) the central sphere.
4. No two surrounding spheres can overlap (they can, at most, touch).

Objective

To find the maximum number of surrounding spheres that can be placed under these conditions. This maximum number is the kissing number for that dimension.

Kissing Numbers in Different Dimensions

- **Two Dimensions (2D)**: The kissing number is **6**.
- **Three Dimensions (3D)**: The kissing number is **12**.
- **Higher Dimensions**: Known values include 24 in 4 dimensions, 240 in 8 dimensions, and 196,560 in 24 dimensions.

AlphaEvolve's Contribution

- The paper treated this as a **construction** problem.
- The AI's task was to generate a piece of code. This code's function was to generate the coordinate positions of the sphere centers, thus completing a construction scheme.
- The AI optimized the scheme through a loop of: generating code $\rightarrow$ receiving feedback on the result $\rightarrow$ modifying the code.
- In **13-dimensional** space, the previously known construction could place 592 spheres. AlphaEvolve, using its method, increased this number to **593**.

Practical Algorithmic Optimization Duality

Zero-Sum Matrix Game

A zero-sum matrix game models two adversaries whose interests are perfectly opposed. The payoff matrix M encodes gains for the row player (Alice) and equal losses for the column player (Bob). When Alice commits to a row and Bob commits to a column, the outcome is M_{ij}; strategic uncertainty arises because each player wants to guard against the other's best counter-move. By analyzing deterministic and randomized strategies, we seek conditions under which an equilibrium value of the game exists.

Pure Strategy (Deterministic)

We compare two play orders: Alice selects a row before Bob answers with a column (Row-Column), versus Bob choosing first (Column-Row).

$$\min_i \max_j M_{ij} \geq \max_j \min_i M_{ij}$$

The left-hand side represents Alice's guaranteed payoff when she commits to her best pure row before Bob reacts. The right-hand side is Bob's guarantee when he moves first. Because Alice suffers when Bob tailors his response, the inequality is generally strict; equality only holds when a deterministic saddle point exists.

Mixed Strategy (Randomized)

Allowing mixed strategies means each player randomizes over their pure actions. Alice samples rows according to a distribution p, while Bob samples columns using q. Neither player reveals the sampled action until play occurs, but we analyze expected payoffs to capture risk-neutral preferences.

Row-Column

1. Alice chooses a strategy, a probability distribution p over rows.
2. Bob, observing the strategy chosen by Alice (p), chooses a probability distribution q over columns.
3. Expected value: $\min_p \max_q p^T M q$

Column-Row

$$\max_{q} \min_{p} p^T M q$$

Min-max Theorem

$$\min_{p} \max_{q} p^T M q = \max_{q} \min_{p} p^T M q$$

Von Neumann's min-max theorem guarantees that optimal mixed strategies exist for both players and that both optimization orders yield the same game value. Each player can secure this value regardless of the opponent's randomized response.

Theorem

From Brouwer fixed-point theorem:

$$\exists (p^*, q^*), \text{ such that } \forall p, q, \ (p^*)^T M q \leq (p^*)^T M q^* \leq p^T M q^*$$

The fixed-point argument identifies an equilibrium pair (p^*, q^*) where neither player improves by deviating. Any alternative p lowers Alice's payoff, and any alternative q raises Bob's loss, capturing the saddle-point structure of zero-sum games.

Theorem (Sion's Min-max Theorem)

If $f(x, y)$ is convex in x and concave in y, then

$$\min_{x} \max_{y} f(x, y) = \max_{y} \min_{x} f(x, y)$$

and there exist x^*, y^* such that

$$\forall x, y, \quad f(x^*, y) \leq f(x^*, y^*) \leq f(x, y^*)$$

Sion's theorem extends the min-max equality beyond finite games to compact convex domains with continuous payoffs. It ensures the existence of saddle points $f(x^*, y^*)$ in broad optimization settings, laying groundwork for primal-dual algorithms.

Lagrange Duality

Lagrange duality reframes constrained optimization by coupling the objective with weighted constraint violations. Introducing multipliers converts the constrained problem into an unconstrained saddle-point problem whose dual objective supplies lower bounds on the primal optimum. Under convexity and suitable regularity, primal and dual optima coincide, providing powerful certificates of optimality and algorithmic insights.

(P)

$$\min_x f(x)$$

subject to

$$g_i(x) \leq 0, \quad i \in [m]$$
$$h_i(x) = 0, \quad i \in [n]$$

Assume f, g_i are convex and h_i are linear.
Solution: x^*

Step I (Proposition)

$$(P) \iff \min_x \max_{\lambda,\mu} \big(f(x) + \sum_i \mu_i h_i(x) + \sum_i \lambda_i g_i(x)\big)$$

subject to $\lambda_i \geq 0$.

Step II

Define

$$L(x, \lambda, \mu) := f(x) + \sum_i \mu_i h_i(x) + \sum_i \lambda_i g_i(x)$$

Then

$$\min_x \max_{\mu \in \mathbb{R}^n, \lambda \geq 0} L(x, \mu, \lambda) = \max_{\mu \in \mathbb{R}^n, \lambda \geq 0} \min_x L(x, \mu, \lambda)$$

Step III

$$\max_{\mu \in \mathbb{R}^n, \lambda \geq 0} \Big[\min_x L(x, \mu, \lambda)\Big]$$

Fix $\lambda \geq 0, \mu$, then $\frac{\partial L}{\partial x} = 0 \implies x^* = \phi(\lambda, \mu)$.

Step IV (D)

$$\max_{\mu\in\mathbb{R}^n,\lambda\geq 0} L(\phi(\lambda,\mu),\lambda,\mu)$$

Solution: λ^*, μ^*

KKT Condition

Theorem (for optimization problem)

Karush–Kuhn–Tucker (KKT) conditions characterize optimal solutions of convex programs by combining primal feasibility, dual feasibility, and stationarity of the Lagrangian. They extend Lagrange multipliers to inequality constraints and provide a checklist for verifying convergence of optimization algorithms.

The following conditions for (x^*, λ^*, μ^*) are necessary and sufficient:

1. **Primal feasible:** $h_i(x^*) = 0,\ g_i(x^*) \leq 0$
2. **Dual feasible:** $\lambda_i^* \geq 0$
3. **Stationary:**
$$\nabla_x L(x,\lambda,\mu)|_{x^*,\lambda^*,\mu^*} = 0$$
4. **Complementary Slackness:**
$$\lambda_i^* g_i(x^*) = 0, \quad \forall i$$

Proof of necessity

Assume x^* solves the primal and strong duality holds. Any dual feasible (λ, μ) satisfies $L(x^*, \lambda, \mu) \leq f(x^*)$. Taking (λ^*, μ^*) that attains the dual optimum forces $L(x^*, \lambda^*, \mu^*) = f(x^*)$, which implies $g_i(x^*) \leq 0$ and $h_i(x^*) = 0$ (primal feasibility), $\lambda_i^* \geq 0$ (dual feasibility), and $\lambda_i^* g_i(x^*) = 0$ (complementary slackness). Differentiating L with respect to x at the optimum yields $\nabla_x L(x^*, \lambda^*, \mu^*) = 0$ because any feasible perturbation of x^* cannot decrease f, proving stationarity.

Proof of sufficiency

Conversely, suppose (x^*, λ^*, μ^*) satisfies the four KKT conditions. For any feasible x,

$$f(x) \geq L(x, \lambda^*, \mu^*) \geq L(x^*, \lambda^*, \mu^*) = f(x^*).$$

The first inequality uses $\lambda_i^* \geq 0$ and $g_i(x) \leq 0$, together with $h_i(x) = 0$. The second inequality follows from stationarity and convexity, which ensure x^* minimizes $L(\cdot, \lambda^*, \mu^*)$. The final equality comes from complementary slackness, yielding $L(x^*, \lambda^*, \mu^*) = f(x^*)$. Therefore no feasible point can decrease the objective, so x^* is optimal and the dual value equals the primal value.

Classification

We study binary linear classification with separable data. Given labeled samples (x_i, y_i) where $y_i \in \{-1, +1\}$, the goal is to find a hyperplane (w, b) that separates the classes with the largest possible margin, yielding robust generalization.

$$\max_{w,b,t} t \quad \text{s.t.} \quad y_i(w^T x_i + b) \geq t,\ \|w\| = 1 \iff \min_{w,b} \frac{1}{2}\|w\|^2 \quad \text{s.t.} \quad y_i(w^T x_i + b) \geq 1,\ \forall i$$

The left program maximizes the geometric margin t under unit-norm weights. Rescaling shows it is equivalent to minimizing the squared norm of w while enforcing functional margins $y_i(w^T x_i + b)$ of at least 1.

(P)

$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad \text{s.t.} \quad y_i(w^T x_i + b) \geq 1, \quad i \in [n]$$

$$L(w, b; \lambda, \mu) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \lambda_i [1 - y_i(w^T x_i + b)]$$

$$\frac{\partial L}{\partial w} = 0 \implies w = \sum_{i=1}^{n} \lambda_i y_i x_i$$

$$\frac{\partial L}{\partial b} = 0 \implies \sum_{i=1}^{n} \lambda_i y_i = 0$$

(D)

$$\min_{\lambda} \left(-\sum_{i=1}^{n} \lambda_i + \frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j x_i^T x_j \right)$$

subject to

$$\lambda_i \geq 0, \quad i \in [n], \quad \sum_{i=1}^{n} \lambda_i y_i = 0$$

1. $w^* = \sum_{i=1}^{n} \lambda_i^* y_i x_i$, i.e. w^* is a linear combination of training data.
2. $\lambda_i^* [y_i((w^*)^T x_i + b) - 1] = 0, \forall i$. $\lambda_i^* = 0$ only if x_i is not the closest to the hyperplane.

Thus,

$$w^* = \sum_{i=1}^{n} \lambda_i^* y_i x_i$$

where x_i are the closest points (Support Vectors).

Optimal Linear Classifier

Support Vector Machine (SVM): Large-margin classifier.

SVMs construct the separating hyperplane with maximum margin between classes. Only support vectors—points exactly on the margin boundaries—determine the decision function, enabling kernel extensions that implicitly map inputs to high-dimensional feature spaces while preserving convex optimization structure.