VOLCANOLAKE: SISTEMA DE APRENDIZAJE POR REFUERZO

Documento Funcional

Asignatura: Inteligencia Artificial
Curso: 3º de Ingeniería Informática
Grupo: 3
Autores: Eduardo Estefanía Ovejero, Álvaro Martín García

Control de Documento

Versión	Fecha	Autores	Descripción / Cambios
0.1	19/11/2025	Eduardo Estefanía, Álvaro Martín	Borrador inicial y esqueleto (Versión 1 - FrozenLake).
0.5	20/11/2025	Eduardo Estefanía, Álvaro Martín	Añadidos detalles de la Versión 2 (Wrappers).
1.0	22/11/2025	Eduardo Estefanía, Álvaro Martín	Versión Final (Incluye Versión 3 y resultados).

Índice

1. Introducción

El proyecto **VolcanoLake** es un entorno de simulación diseñado para el entrenamiento y evaluación de agentes de Inteligencia Artificial mediante Aprendizaje por Refuerzo (Reinforcement Learning).

El sistema simula un entorno hostil (un terreno volcánico) donde un agente autónomo debe aprender a navegar desde un punto de inicio hasta una meta, evitando peligros mortales y obstáculos, mientras optimiza su ruta para recoger recompensas adicionales.

2. Objetivos del Sistema

El objetivo principal del sistema es proporcionar un entorno escalable y con dificultad incremental para estudiar el comportamiento de aprendizaje del agente.

Las metas funcionales del agente dentro de la simulación son:

1. **Navegación:** Encontrar la ruta desde la casilla de Salida (**S**) hasta la Meta (**G**).
2. **Supervivencia:** Evitar caer en casillas de Lava (**L**) que terminan la partida inmediatamente.
3. **Optimización:** Maximizar la puntuación total recogiendo Tesoros (**T**) y minimizando el número de pasos dados.

3. Descripción de las Versiones (Modos de Juego)

El sistema se ha desarrollado en tres fases incrementales, cada una añadiendo nuevas capas de complejidad funcional.

Versión 1: Entorno Base (Línea Base)

- **Descripción:** Configuración estándar basada en el problema clásico "FrozenLake".
- **Mapa:** Cuadrícula pequeña (4x4).
- **Movimiento:** 4 direcciones básicas (Arriba, Abajo, Izquierda, Derecha).
- **Física:** Suelo resbaladizo estocástico (el agente puede no moverse a donde desea).
- **Propósito:** Establecer una línea base de rendimiento para el algoritmo de aprendizaje.

Versión 2: Entorno Dinámico y Visión Limitada

- **Descripción:** Introduce reglas que modifican el entorno en tiempo de ejecución.
- **Dinámica de Hoyos (Increasing Holes):** El mapa no es estático. Durante el episodio, pueden aparecer nuevos agujeros en casillas previamente seguras, obligando al agente a readaptar su ruta.
- **Visión Limitada:** El agente posee un campo de visión restringido. Recibe penalizaciones adicionales si "mira" hacia una casilla peligrosa, simulando la percepción de peligro antes de caer.

Versión 3: Entorno Avanzado (Personalizado)

- **Descripción:** Un entorno completamente personalizado con físicas y reglas complejas.
- **Mapa:** Mapas de gran tamaño (5x5, 25x25, 50x50, 100x100) cargados desde archivos externos, permitiendo diseños complejos y aleatorios.
- **Movimiento Ampliado:** El agente tiene **8 grados de libertad**, pudiendo moverse en direcciones cardinales y **diagonales**.
- **Física de Fluidos (Agua):** Las casillas de agua (**W**) introducen una mecánica de deslizamiento específica:
    - 80% de probabilidad de movimiento exitoso.
    - 20% de probabilidad de desvío lateral (resbalón).
- **Objetos Consumibles:** Los tesoros (**T**) desaparecen del mapa una vez recolectados, impidiendo la acumulación infinita de recompensas en un mismo punto.

3.1 Matriz Comparativa de Versiones

La siguiente tabla resume las diferencias funcionales clave entre las tres entregas del sistema:

Funcionalidad / Característica	**Versión 1 (Línea Base)**	**Versión 2 (Dinámica)**	**Versión 3 (Avanzada)**
Entorno Base	Gym FrozenLake	Gym FrozenLake	VolcanoLakeEnv (Propio)
Tamaño del Mapa	Pequeño (4x4)	Pequeño (4x4)	**Grande (25x25, 50x50...)**
Movimientos (Acciones)	4 (Cardinal)	4 (Cardinal)	**8 (Cardinal + Diagonal)**
Física del Agua	Deslizamiento simple	Deslizamiento simple	**Física de Fluidos (80/20%)**
Dinámica del Terreno	Estático	**Inestabilidad Persistente**	**Inestabilidad Persistente**
Sistema de Visión	Ciego	**Visión Frontal (4 dir)**	**Visión Avanzada (8 dir)**
Límites del Mundo	Muro (Choca)	Muro (Choca)	**Mundo Toroide**
Objetivos Secundarios	Ninguno	Ninguno	**Recolección de Tesoros**

4. Elementos del Entorno y Sistema de Puntuación

El entorno se compone de una cuadrícula donde cada celda tiene propiedades y recompensas específicas.

Símbolo	Tipo de Terreno	Efecto Funcional	Recompensa (Puntos)
S	Inicio (Start)	Punto de partida del agente.	-0.001 (Coste por paso)
.	Tierra Firme	Terreno seguro. Movimiento determinista.	-0.001 (Coste por paso)
G	Meta (Goal)	Objetivo final. Termina el episodio exitosamente.	**+10**
L	Lava	Obstáculo mortal. Termina el episodio con fracaso.	**-10**
W	Agua	Terreno inestable. Provoca deslizamientos aleatorios.	**-1**
T	Tesoro	Bonificación opcional. Se consume al pisarlo.	**+5**

*Nota: Se aplica una penalización constante de **-0.001** por cada paso dado para incentivar al agente a encontrar el camino más corto posible.*

5. Reglas de Interacción

1. **Límites del Mapa:** Si el agente intenta moverse fuera de los límites de la cuadrícula, permanecerá en la misma casilla (choca contra un muro invisible).
2. **Consumo de Tesoros (V3):** Un tesoro solo otorga puntos la primera vez que se pisa en un episodio. La casilla se transforma visual y funcionalmente en "Tierra Firme" (.) tras la recolección.
3. **Finalización del Episodio:** La simulación termina cuando:
    - El agente llega a la Meta (**G**).
    - El agente cae en Lava (**L**).
    - Se supera el límite máximo de pasos permitidos (Time Limit), para evitar bucles infinitos.
4. **Reglas Especiales:**
    - Conexión de bordes (Mundo Toroide): El entorno simula una superficie continua. Si el agente se mueve hacia un borde del mapa (ej. extremo derecho), no choca, sino que aparece en el lado opuesto (extremo izquierdo), permitiendo estrategias de movimiento avanzadas.
    - Fallo de Hardware (Control defectuoso): Se simula que el sistema de control del agente está dañado. Ocasionalmente, cuando el agente intenta realizar una acción, el “mando” envía una señal errónea o aleatoria (flickering), provocando movimientos no deseados o espasmódicos, lo que obliga al agente a aprender a corregir su trayectoria constantemente.
    - Dinámica de hoyos de Lava: El entorno es más inestable y se vuelve más hostil. A diferencia de la versión anterior, los mapas pueden ser más grandes y por tanto, más cantidad de hoyos de lava posibles incluso sustituyendo casillas de tesoros.
    - Visión limitada: El sistema de percepción se ha actualizado para soportar los **8 grados de libertad**. Ahora, el agente no solo detecta peligros en las direcciones cardinales sino que también recibe alertas si su trayectoria diagonal apunta hacia un obstáculo, aumentado la precisión sensorial necesaria.

6. Métricas y Visualización de Resultados

Para evaluar el desempeño y la estrategia aprendida por el agente, el sistema genera dos tipos de reportes al finalizar el entrenamiento:

6.1 Métricas de Rendimiento (Numéricas)

- **Tasa de Éxito:** Porcentaje de episodios en los que el agente llega a la Meta.
- **Recompensa Acumulada:** Puntuación total media obtenida (incluyendo tesoros y penalizaciones).
- **Eficiencia de Ruta:** Número de pasos necesarios para llegar a la meta (menor es mejor).
- **Tesoros Recolectados:** Cantidad media de objetivos secundarios alcanzados antes de llegar a la meta.

6.2 Mapas de Estrategia (Visuales)

Además de datos numéricos, el sistema genera mapas visuales que permiten al usuario "ver" el conocimiento adquirido por el agente:

- **Mapa de Calor de Valor (Value Heatmap):**
    - **Qué muestra:** Un gráfico de colores sobre la cuadrícula del mapa.
    - **Interpretación Funcional:** Las zonas brillantes (amarillas) indican áreas que el agente considera "altamente deseables" o cercanas a la meta. Las zonas oscuras indican áreas percibidas como peligrosas (lava) o callejones sin salida. Esto permite verificar si el agente ha identificado correctamente los peligros.
- **Mapa de Navegación Óptima (Policy Map):**
    - **Qué muestra:** Una cuadrícula con flechas direccionales. A diferencia de las anteriores versiones, en esta se genera una imagen.
    - **Interpretación Funcional:** Representa la "mejor decisión" que tomaría el agente en cada casilla posible. Al observar las flechas, se puede trazar visualmente la ruta ideal que el agente ha planeado desde la Salida hasta la Meta, y cómo planea recuperarse si es desviado a una zona alejada.