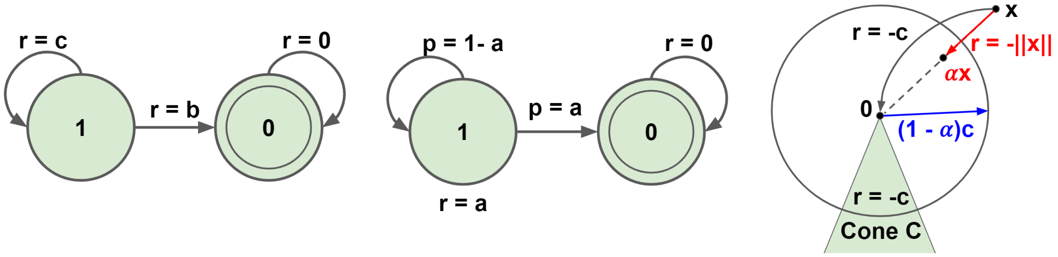


(a) Shortest path problem (b) Blackmailer's problem (c) Optimal stopping problem

Figure 6: Revisiting Fig. 1 for the discounted cases where $\gamma \in (0, 1)$.

A More Examples with Multiple Fixed Points

A.1 Cases of $\gamma = 1$

Consider MDP examples with an terminal state 0, as shown in Fig. 1 (we adapt dynamic programming examples [5, 39] into reinforcement learning settings),

- **Shortest path problem (deterministic)** in Fig. 1(a): At state 1, an agent transits to either state 1 or 0 with reward 0 or b, respectively. Assume the value function for state 0 is $V(0) = 0$. The Bellman's optimality equation for state 1 is $V(1) = \max\{V(1), b\}$, where any $V(1) \geq b$ is a feasible solution. If initialize $V_0(1) \geq b$, a resulting policy is that an agent at state 1 always transits back to state 1; otherwise, drives to terminal state 0 (always returns back to itself with reward 0).
- **Blackmailer's problem (stochastic)** in Fig. 1(b): At state 1, a profit maximizing blackmailer demands a cash amount $a \in (0, 1]$; a victim transits to state 1 with probability a or state 0 with probability $1 - a$, respectively. At state 0, a victim always refuses to yield, i.e., $V(0) = 0$. The Bellman's optimality equation for state 1 is $V(1) = \max_a\{a + (1 - a)V(1)\}$, where any $V(1) \geq 1$ is a feasible solution. If initialize $V_0(1) > 1$, the blackmailer's policy is demanding $a \to 0$ to keep the victim at state 1; otherwise, demanding $a = 1$ that drives the victim to terminal state 0.
- **Optimal stopping problem (terminating policies)** in Fig. 1(c): In a space $\mathbb{R}^2$ with terminal state of point 0, an agent at point $x \neq 0$ moves to either point 0 with negative reward $-c$ or point αx with reward $-||x||$, respectively, where $\alpha \in (0, 1)$. The Bellman's optimality equation is $V(x) = \max\{-c, -||x|| + V(\alpha x)\}$ and the optimal policy is to continue inside the sphere of radius $(1 - \alpha)c$ and to stop outside. If add a cone region C within which an agent always receives a reward $-c$, a second policy is jumping to point 0 at any point in region C.

A.2 Cases of $\gamma \in (0.1)$

First, we consider the discounted formulations of the three examples (shown in Fig. 1), as shown in Fig. 6 where $\gamma \in (0, 1)$. The differences are marked in red.

- (a) **Shortest path problem (deterministic, discounted case)**: Given two states 1 and 0, an agent at state 1 transits to either state 1 or 0 with rewards $r = c$ or $r = b$, respectively. $c > (1 - \gamma) \cdot b$. At state 0, the value function is $V(0) = 0$. At state 1, the Bellman's optimality equation is $V(1) = \max\{c + \gamma \cdot V(1), b\}$, where any $V(1) \geq (b - c)/\gamma$ is a solution. If initialize $V_0(1) \geq b$, an agent obtains a policy that always transits back to state 1; otherwise, a result policy drives to terminal state 0.
- (b) **Blackmailer's problem (stochastic, discounted case)**: Different from (a), a profit maximizing blackmailer/agent at 1 demands a cash amount $a \in (0, 1]$ (an action), while a victim transits to state 1 with probability a or to state 0 with probability $1 - a$, respectively. At state 0, a victim always refuses to yield to the blackmailer's demand, i.e., $V(0) = 0$. The Bellman's optimality equation is $V(1) = \max_a\{a + \gamma \cdot (1 - a)V(1)\}$ for state 1, where any $V(1) \geq 1$ is a feasible solution. If initialize $V_0(1) = c > 1$, the blackmailer's policy is demanding $a \to 0$ at the k-th step to keep the victim stay at state 1, for any $k \leq K_0 = -\lfloor \log_\gamma c \rfloor$; and taking $a = 1$ to transit to terminal

state 0 at the k-th step, for any $k \geq K_0 + 1$; otherwise initialize $V_0(1) = c \leq 1$, the result policy is demanding the maximum $a = 1$ that drives the victim to a refusal state 0 (a terminal state).

- (c) **Optimal stopping problem (terminating policies, discounted case)**: In a space $\mathbb{R}^2$ with terminating state at point 0, at point $x \neq 0$ an agent moves to either point 0 with negative reward $-c$ or point αx with reward $-||x||$, respectively, where $\alpha \in (0, 1)$. The Bellman's optimality equation is $V(x) = \max\{-c, -||x|| + \gamma \cdot V(\alpha x)\}$ and the optimal policy is to continue inside the sphere of radius $(1 - \alpha)c$ and to stop outside. If add a cone region C within which an agent always receives a reward $-c$, a second policy is jumping to point 0 at any point in region C.

Then, we elaborate how the proposed H-term fixes the problems in Fig. 6.

(a) Shortest path problem (deterministic, discounted case)

Assume $V_0(1) \geq b$ and $c > (1-\gamma)b$, we have

$$\begin{aligned}
&V_1(1) = c + \gamma \cdot V_0(1) \geq c + \gamma \cdot b > b \\
&V_2(1) = c + \gamma \cdot c + \gamma^2 \cdot V_0(1) \geq (1+\gamma)c + \gamma^2 b > b \\
&V_3(1) = c + \gamma \cdot c + \gamma^2 c + \gamma^3 \cdot V_0(1) \geq (1+\gamma+\gamma^2)c + \gamma^3 b > b \\
&\cdots \\
&V_k(1) = \sum_{i=0}^{k-1} \gamma^i \cdot c + \gamma^k \cdot V_0(1) \geq \sum_{i=0}^{k-1} \gamma^i \cdot c + \gamma^k b > b \\
&\cdots \\
&V^*(1) = \sum_{i=0}^{\infty} \gamma^i \cdot c = \frac{1}{1-\gamma} c > b
\end{aligned} \tag{9}$$

The values of $H(0)$ and $H(1)$ are as follows:

$$H(0) = 0, \quad H(1) = -b - \sum_{k=2}^{\infty} (\sum_{i=1}^{k-1} \gamma^{i-1} \cdot c + \gamma^k b) = -\infty. \tag{10}$$

Adding the above H-values to state 1 and 0, respectively, we have

$$\begin{aligned}
&V^*(1) + H(1) = \sum_{i=0}^{\infty} \gamma^i \cdot c - \infty = -\infty \\
&V^*(0) + H(0) = b.
\end{aligned} \tag{11}$$

Therefore, $V^*(1) + H(1) < V^*(0) + H(0)$, independent of the initial value $V_0(1)$. That is, an agent always obtains a policy that drives to terminal state 0 at step 1.

(b) **Blackmailer's problem (stochastic, discounted case)**

If initialize $V_0(1) = c > 1$, the blackmailer's policy is demanding $a \to 0$ at the k-th step to keep the victim stay at state 1, for any $k \leq K_0 = -\lfloor \log_\gamma c \rfloor$; and taking $a = 1$ to transit to terminal state 0 at the k-th step, for any $k \geq K_0 + 1$; otherwise initialize $V_0(1) = c \leq 1$, the result policy is demanding the maximum $a = 1$ that drives the victim to a refusal state 0 (a terminal state).

The values of $H(0)$ and $H(1)$ are as follows:

$$H(0) = 0, \quad H(1) = -\sum_{k=1}^{\infty} \sum_{i=1}^{k-1} \gamma^{i-1} \cdot a = -\infty. \tag{12}$$

For arbitrary initial value of $V_0(1)$, $V_1(1) = a + (1-a) \cdot \gamma (V_0(1) + H(1))$ take maximum $V_1(1) = 1$ when $a = 1$. Therefore, the policy always drives to terminal state 0 at step 1.

(c) **Optimal stopping problem (terminating policies, discounted case)**

Any policy that takes infinite steps will have

$$H(x) = -c - \sum_{k=2}^{\infty} \left[\sum_{i=1}^{k-1} \gamma^i \cdot \alpha^i \cdot \|x\| + \gamma^k \cdot (-c) \right] = -\infty \tag{13}$$

and a direct jumping policy will have $H(x) = -c$. Therefore, the H-term drives to a terminating policy.

B MUJOCO TASKS WITH MULTIPLE POLICIES

B.1 DESCRIPTION OF MUJOCO TAKS

We selected six challenging robotic locomotion tasks from MuJoCo, namely, Swimmer-v3, Hopper-v3, Walker2D-v3, HalfCheetah-v3, Ant-v3, Humanoid-v3. Table 3 lists the action space and state space of each task.

Table 3: The state and action spaces of six challenging MuJoCo tasks.

Tasks	Agent	Action Space	State Space
Swimmer-v3	Three-link swimming robot	2	8
Hopper-v3	Two-dimensional one-legged robot	3	11
Walker2d-v3	Two-dimensional bipedal robot	6	17
HalfCheetah-v3	Two-dimensional robot	6	17
Ant-v3	Four-legged creature	8	111
Humanoid-v3	Three-dimensional bipedal robot	17	376

B.2 MULTIPLE POLICIES IN MUJOCO TASKS

In the supplementary files, we includes rendered videos of different policies, as given in Table 4.

- Different policies are obtained over 20 runs of the PPO algorithm. We rendered theses polices and classified them by physical gaits.
- The policies in bold texts are physically stationary.

Table 4: List of video files for different policies.

Task	Different Policies	Video Name
Hopper	**hopping**	hopper_hopping.mp4
	diving	hopper_diving.mp4
	standing	hopper_standing.mp4
Ant	**running**	ant_running.mp4
	standing	ant_standing.mp4
	flipping	ant_flipping.mp4
Walker	**walking**	walker_walking.mp4
	diving	walker_diving.mp4
	standing	walker_standing.mp4
Humanoid	**two-legs**	humanoid_two_legs.mp4
	one-leg	humanoid_one_leg.mp4
	backward	humanoid_backward.mp4
HalfCheetah	**running**	halfcheetah_running.mp4
	diving	halfcheetah_diving.mp4
	flipping	halfcheetah_flipping.mp4
	standing	halfcheetah_standing.mp4
Swimmer	**moving**	swimmer_moving.mp4
	standing	swimmer_standing.mp4

C REINFORCEMENT LEARNING AND BELLMAN EQUATION

A reinforcement learning (RL) [44] agent interacts with an unknown environment and learns an optimal policy that maximizes the cumulative reward. Mathematically, the environment can be formulated as a Markov Decision Process (MDP) with the five-tuple $\langle \mathcal{S}, \mathcal{A}, \mathbb{P}, R, \gamma \rangle$. Here $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces; $\mathbb{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ denotes a transition probability function, where Δ is a probability simplex; $R : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ denotes a reward function; and $\gamma \in (0, 1]$ denotes a discount factor. The objective is to find an optimal policy $\pi^* : \mathcal{S} \to \Delta(\mathcal{A})$ that maximizes (discounted) expected reward.

Consider a discrete, finite, discounted MDP with infinite horizon, one can define the Q-value function of a state-action pair (s, a) under policy π as follows

$$Q^{\pi}(s,a) = \mathbb{E}_{S_{k+1} \sim \mathbb{P}(\cdot|S_k, A_k), A_{k+1} \sim \pi(S_{k+1}, \cdot)} \left[\sum_{k=0}^{\infty} \gamma^k \cdot R(S_k, A_k, S_{k+1}) | S_0 = s, A_0 = a \right], \quad (14)$$

where $R(S_k, A_k, S_{k+1})$ denotes the immediate reward when taking action A_k at state S_k and arriving at state S_{k+1}, capital letters denote random variables and lowercase letters denote values. The conventional objective function $J(\theta)$ of reinforcement learning [44] takes the following form

$$J(\theta) \triangleq \mathbb{E}_{S_0, A_0}[Q^{\pi_\theta}(S_0, A_0)] = \mathbb{E}_{\tau \sim \pi}\left[R(\tau) \cdot P(\tau|\pi_\theta)\right], \quad (15)$$

where τ is a trajectory, i.e., $\tau = (S_0, A_0, \cdots)$, and
$P(\tau|\pi_\theta) = d_0(s_0) \cdot \prod_{k=0}^{T} \mathbb{P}(s_{k+1}|s_k, a_k)\pi_\theta(a_k|s_k)$.

The Bellman equation [44] converts (14) into a recursive form as follows

$$\begin{aligned} Q^{\pi}(s,a) &= \sum_{s' \in \mathcal{S}} \mathbb{P}(s' \mid s, a) \left[R(s, a, s') + \gamma \sum_{a' \in \mathcal{A}} \pi(s', a') Q^{\pi}(s', a') \right] \\ &= R(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathbb{P}(s' \mid s, a) \sum_{a' \in \mathcal{A}} \pi(s', a') Q^{\pi}(s', a'), \end{aligned} \quad (16)$$

which expresses the expected reward as a summation of immediate reward $R(s, a)$ and discounted future rewards, and the immediate reward $R(s, a)$ is defined as $R(s, a) = \sum_{s' \in \mathcal{S}} \mathbb{P}(s' \mid s, a) R(s, a, s')$.

The Bellman's optimality equation [44] is

$$Q^*(s,a) = \sum_{s' \in \mathcal{S}} \mathbb{P}(s' \mid s, a) \left[R(s, a, s') + \gamma \max_{a'} Q^*(s', a') \right]. \quad (17)$$

The optimal policy π^* is given by a greedy strategy such that $\pi^* = \arg\max_{\pi} \; Q^{\pi}(s, a)$.

D QUANTUM K-SPIN HAMILTONIAN FORMULATION OF REINFORCEMENT LEARNING

We provide the detailed steps of reformulating (14) into a K-spin Hamiltonian equation

$$
\begin{aligned}
H(\theta) &\triangleq -\mathbb{E}_{S_0,A_0}\left[Q^{\pi_\theta}(S_0,A_0)\right] \\
&= -\mathbb{E}_{S_0,A_k\sim\pi_\theta(S_k,\cdot),S_{k+1}\sim\mathbb{P}(\cdot|S_k,A_k)}\left[\sum_{k=0}^{\infty}\gamma^k\cdot R(S_k,A_k)\right] \\
&= -\sum_{k=0}^{K-1}\mathbb{E}_{S_0,A_0,\cdots,S_k\sim\mathbb{P}(\cdot|S_{k-1},A_{k-1}),A_k\sim\pi_\theta(S_k,\cdot)}\left[\gamma^k\cdot R(S_k,A_k)\right] \\
&= -\sum_{k=0}^{K-1}\sum_{\mu_0}^{\mathcal{S}\times\mathcal{A}}\cdots\sum_{\mu_k}^{\mathcal{S}\times\mathcal{A}}\gamma^k\cdot R(\mu_k)\cdot d_0(S_0)\cdot\pi_\theta(\mu_0)\prod_{i=0}^{k-1}\left[\mathbb{P}(S_{i+1}|\mu_i)\cdot\pi_\theta(\mu_{i+1})\right] \\
&= -\sum_{k=0}^{K-1}\sum_{\mu_0}^{\mathcal{S}\times\mathcal{A}}\cdots\sum_{\mu_k}^{\mathcal{S}\times\mathcal{A}}\left[\gamma^k\cdot R(\mu_k)\cdot d_0(S_0)\cdot\prod_{i=0}^{k-1}\mathbb{P}(S_{i+1}|\mu_i)\right]\cdot\pi_\theta(\mu_0)\cdots\pi_\theta(\mu_k) \\
&= -\sum_{k=0}^{K-1}\sum_{\mu_0}^{\mathcal{S}\times\mathcal{A}}\cdots\sum_{\mu_k}^{\mathcal{S}\times\mathcal{A}}L_{\mu_0,\ldots,\mu_k}\cdot\pi_\theta(\mu_0)\cdots\pi_\theta(\mu_k),
\end{aligned}
\tag{18}
$$

where $K\to\infty$, and the density function is

$$
L_{\mu_0,\ldots,\mu_k} = \gamma^k\cdot R(\mu_k)\cdot d_0(S_0)\cdot\prod_{i=0}^{k-1}\mathbb{P}(S_{i+1}|\mu_i). \tag{19}
$$

E DERIVATION STEPS FOR HAMILTONIAN'S POLICY GRADIENTS

We provide the policy gradient of the quantum K-spin Hamiltonian equation in (3) for both stochastic and deterministic cases, which are variants of the well-known policy gradient theorem [44].

Theorem 3. *(**Hamiltonian's stochastic policy gradient**) The stochastic gradient of the K-spin Hamiltonian equation (3) w.r.t. parameter θ is*

$$\nabla_\theta H(\theta) = -\mathbb{E}_{\mu_0,\ldots,\mu_{K-1}}\left[\sum_{k=0}^{K-1} \gamma^k \cdot R(\mu_k) \cdot \nabla_\theta \log\left(\pi_\theta(\mu_0)\cdot\pi_\theta(\mu_1)\cdots\pi_\theta(\mu_k)\right)\right]. \tag{20}$$

Corollary 1. *When $K \to \infty$, the Hamiltonian's stochastic policy gradient $\nabla_\theta H(\theta)$ in (20) is equal to the stochastic policy gradient $\nabla_\theta J(\theta)$ in [45],*

$$\lim_{K\to\infty} \nabla_\theta H(\theta) = -\nabla_\theta J(\theta) = -\mathbb{E}_{s\sim d_\theta, a\sim\pi_\theta}\left[Q^{\pi_\theta}(s,a)\nabla_\theta \log \pi_\theta(s,a)\right]. \tag{21}$$

Let $\eta_\theta(\cdot) : \mathcal{S} \to \mathcal{A}$ denote a deterministic policy, while we use $\widetilde{\pi}_{\theta,\delta}(\mu)$ to represent that a Gaussian noise (a.k.a, an exploration noise) with standard deviation $\delta > 0$ is added in the exploration process.

Theorem 4. *(**Hamiltonian's deterministic policy gradient**) The deterministic gradient of the K-spin Hamiltonian equation (3) w.r.t. parameter θ is*

$$\nabla_\theta H'(\theta) = -\mathbb{E}_{\mu_0,\ldots,\mu_{K-1}}\left[\sum_{k=0}^{K-1} \gamma^k \cdot R(\mu_k) \cdot \nabla_\theta \log\left(\widetilde{\pi}_{\theta,\delta}(\mu_0)\cdot\widetilde{\pi}_{\theta,\delta}(\mu_1)\cdots\widetilde{\pi}_{\theta,\delta}(\mu_k)\right)\right]. \tag{22}$$

Corollary 2. *When $K \to \infty$, the Hamiltonian's deterministic policy gradient $\nabla_\theta H'(\theta)$ in (22) is equal to the deterministic policy gradient $\nabla_\theta J'(\theta)$ in [43],*

$$\lim_{K\to\infty} \nabla_\theta H'(\theta) = -\nabla_\theta J'(\theta) = -\mathbb{E}_{s\sim d_\theta}\left[\nabla_a Q^{\widetilde{\pi}_{\theta,\delta}}(s,a)|_{a=\eta_\theta}\nabla_\theta \eta_\theta(s)\right]. \tag{23}$$

Corollary 3. *When the variance of the exploration noise approaches zero, i.e., $\delta \to 0$, the deterministic policy gradient $\nabla_\theta H'(\theta)$ is the limiting case of the stochastic policy gradient $\nabla_\theta H(\theta)$,*

$$\nabla_\theta H'(\theta) = \lim_{\delta\to 0} \nabla_\theta H(\theta). \tag{24}$$

E.1 PROOF OF THEOREM 3: HAMILTONIAN'S STOCHASTIC POLICY GRADIENT

Proof.

$$\begin{aligned}
\nabla_\theta H(\theta) &= -\sum_{k=0}^{K-1}\sum_{\mu_0}^{\mathcal{S}\times\mathcal{A}}\cdots\sum_{\mu_k}^{\mathcal{S}\times\mathcal{A}} L_{\mu_0,\ldots,\mu_k}\nabla_\theta\left[\pi_\theta(\mu_0)\cdots\pi_\theta(\mu_k)\right]\\
&= -\sum_{k=0}^{K-1}\sum_{\mu_0}^{\mathcal{S}\times\mathcal{A}}\cdots\sum_{\mu_k}^{\mathcal{S}\times\mathcal{A}} L_{\mu_0,\ldots,\mu_k}\left[\pi_\theta(\mu_0)\cdots\pi_\theta(\mu_k)\right]\nabla_\theta \log\left[\pi_\theta(\mu_0)\cdots\pi_\theta(\mu_k)\right]\\
&= -\sum_{k=0}^{K-1}\sum_{\mu_0}^{\mathcal{S}\times\mathcal{A}}\cdots\sum_{\mu_k}^{\mathcal{S}\times\mathcal{A}} \gamma^k\cdot R(\mu_k)\cdot d_0(S_0)\cdot\pi_\theta(\mu_0)\prod_{i=0}^{k-1}\left[\mathbb{P}(S_{i+1}|\mu_i)\cdot\pi_\theta(\mu_{i+1})\right]\cdot\nabla_\theta\log\left[\pi_\theta(\mu_0)\cdots\pi_\theta(\mu_k)\right]\\
&= -\mathbb{E}_{\mu_0,\ldots,\mu_{K-1}}\left[\sum_{k=0}^{K-1}\gamma^k\cdot R(\mu_k)\cdot\nabla_\theta\log\left[\pi_\theta(\mu_0)\cdots\pi_\theta(\mu_k)\right]\right],
\end{aligned} \tag{25}$$

where $\mu_k = (S_k, A_k)$, $S_0 \sim d_0(\cdot)$, $A_k \sim \pi_\theta(S_k,\cdot)$, $S_{k+1} \sim \mathbb{P}(\cdot \mid S_k, A_k)$ for $k = 0\cdots K$. □

E.2 PROOF OF COROLLARY 1

Proof.

$$
\begin{aligned}
&\nabla_\theta H(\theta) \overset{(a)}{=} -\sum_{k=0}^{K-1}\sum_{\mu_0}^{\mathcal{S}\times\mathcal{A}}\cdots\sum_{\mu_k}^{\mathcal{S}\times\mathcal{A}} L_{\mu_0,\ldots,\mu_k}\nabla_\theta\left[\pi_\theta(\mu_0)\cdots\pi_\theta(\mu_k)\right]\\
&\overset{(b)}{=} -\sum_{k=0}^{K-1}\sum_{\mu_0}^{\mathcal{S}\times\mathcal{A}}\cdots\sum_{\mu_k}^{\mathcal{S}\times\mathcal{A}} L_{\mu_0,\ldots,\mu_k}\sum_{i=0}^{k}\pi_\theta(\mu_0)\cdots\pi_\theta(\mu_{i-1})\pi_\theta(\mu_{i+1})\cdots\pi_\theta(\mu_k)\nabla_\theta\pi_\theta(\mu_i)\\
&\overset{(c)}{=} -\sum_{k=0}^{K-1}\sum_{\mu_0}^{\mathcal{S}\times\mathcal{A}}\cdots\sum_{\mu_k}^{\mathcal{S}\times\mathcal{A}}\gamma^k\cdot R(\mu_k)\cdot d_0(S_0)\prod_{i=0}^{k-1}\mathbb{P}(S_{i+1}|\mu_i)\sum_{i=0}^{k}\left[\prod_{j=0}^{i-1}\pi_\theta(\mu_j)\cdot\nabla_\theta\pi_\theta(\mu_i)\cdot\prod_{j=i+1}^{k}\pi_\theta(\mu_j)\right]\\
&\overset{(d)}{=} -\sum_{k=0}^{K-1}\sum_{\mu_0}^{\mathcal{S}\times\mathcal{A}}\cdots\sum_{\mu_k}^{\mathcal{S}\times\mathcal{A}}\sum_{i=0}^{k} d_0(S_0)\left[\gamma^i\prod_{j=0}^{i-1}\pi_\theta(\mu_j)\mathbb{P}(S_{j+1}|\mu_j)\right]\nabla_\theta\pi_\theta(\mu_i)\left[\prod_{j=i+1}^{k-1}\pi_\theta(\mu_j)\mathbb{P}(S_{j+1}|\mu_j)\pi_\theta(\mu_k)\gamma^{k-i}R(\mu_k)\right]\\
&\overset{(e)}{=} -\sum_{k=0}^{K-1}\sum_{i=0}^{k}\sum_{S_0}^{\mathcal{S}} d_0(S_0)\sum_{S_i}^{\mathcal{S}}\rho(S_0,S_i,i)\sum_{A_i}^{\mathcal{A}}\nabla_\theta\pi_\theta(S_i,A_i)\cdot\sum_{\mu_k}^{\mathcal{S}\times\mathcal{A}}\rho(S_i,S_k,k-i)\cdot\pi_\theta(\mu_k)\cdot R(\mu_k)\\
&\overset{(f)}{=} -\sum_{S_0}^{\mathcal{S}} d_0(S_0)\sum_{S}^{\mathcal{S}}\sum_{i=0}^{K-1}\rho(S_0,S,i)\sum_{A}^{\mathcal{A}}\nabla_\theta\pi_\theta(S,A)\cdot\left[\sum_{S'}^{\mathcal{S}}\sum_{k=i}^{K-1}\rho(S,S',k-i)\cdot\sum_{A'}^{\mathcal{A}}\pi_\theta(S',A')\cdot R(S',A')\right]\\
&\overset{(g)}{=} -\sum_{S_0}^{\mathcal{S}} d_0(S_0)\sum_{S}^{\mathcal{S}}\sum_{i=0}^{\infty}\rho(S_0,S,i)\sum_{A}^{\mathcal{A}}\nabla_\theta\pi_\theta(S,A)\cdot Q^{\pi_\theta}(S,A)\\
&\overset{(h)}{=} -\left[\sum_{S}^{\mathcal{S}}\sum_{S_0}^{\mathcal{S}} d_0(S_0)\sum_{i=0}^{\infty}\rho(S_0,S,i)\right]\cdot\sum_{S}^{\mathcal{S}}\frac{\sum_{S_0}^{\mathcal{S}} d_0(S_0)\sum_{i=0}^{\infty}\rho(S_0,S,i)}{\sum_{s}^{\mathcal{S}}\sum_{S_0}^{\mathcal{S}} d_0(S_0)\sum_{i=0}^{\infty}\rho(S_0,S,i)}\sum_{A}^{\mathcal{A}}\nabla_\theta\pi_\theta(S,A)\cdot Q_\theta(S,A)\\
&\overset{(i)}{\propto} -\sum_{S}^{\mathcal{S}} d_{\pi_\theta}(S)\sum_{A}^{\mathcal{A}}\nabla_\theta\pi_\theta(S,A)\cdot Q^{\pi_\theta}(S,A)\\
&\overset{(j)}{=} -\mathbb{E}_{S\sim d_\theta,A\sim\pi_\theta(S,\cdot)}[Q^{\pi_\theta}(S,A)\nabla_\theta\log\pi_\theta(S,A)],
\end{aligned}
\tag{26}
$$

where $\rho(S,S',i)$ denotes the probability of state S transfer to S' in i steps.

We provide detailed explanations step-by-step:

- Equality (a) holds by definition.
- In equality (b), using the chain rule, we take derivative of $\nabla_\theta[\pi_\theta(\mu_0)\cdots\pi_\theta(\mu_k)]$ with respect to $\pi_\theta(\mu_i)$, $i=1,...,k$.
- In equality (c), we plug in $L_{\mu_0,\cdots,\mu_k}$ in (2).
- In equality (d), we insert $\mathbb{P}(S_{i+1}|\mu_i)$ $\mathbb{P}(S_{i+1}|\mu_i)$ between $\pi_\theta(\mu_i)$ and $\pi_\theta(\mu_{i+1})$, $i=1,...,k$.
- In equality (e), we split trajectory $\mu_0,\cdots,\mu_i,\cdots,\mu_k$ into two trajectories $\mu_0,\cdots,\mu_i$ and $\mu_i,\cdots,\mu_k$. Therefore, we can classify all trajectories $\mu_0,\cdots,\mu_k$ by μ_0,μ_i,μ_k, and i.
- In equality (f), we reorganize $\sum_{k=0}^{K-1}\sum_{i=0}^{k}$ into $\sum_{i=0}^{K-1}\sum_{k=i}^{K-1}$. The former one first traverses the length k of a trajeoctory, and then traverses the i-th step on it.The latter one first traverses the i-th step of a trajectory, and then traverses the length k of it.
- In equality (g), we calculate the limit of (f) when K approaches ∞.
- In equality (h), we normalize $\sum_{S_0}^{\mathcal{S}} d_0(S_0)\sum_{i=0}^{\infty}\rho(S_0,S,i)$ to be a probability distribution.

- In equality (i), we remove the constant $\sum_S^{\mathcal{S}} \sum_{S_0}^{\mathcal{S}} d_0(S_0) \sum_{i=0}^{\infty} \rho(S_0, S, i)$ and replace the fraction with $d_{\pi_\theta}(S)$, the stationary distribution of state S under policy π_θ.
- In equality (j), we reformulate (i) as expectation.

□

E.3 PROOF OF THEOREM 4: HAMILTONIAN'S DETERMINISTIC POLICY GRADIENT

Proof. Let $\eta_\theta(\cdot) : \mathcal{S} \to \mathcal{A}$ denote a deterministic policy, while we use $\widetilde{\pi}_{\theta,\delta}(\mu)$ to represent that a Gaussian noise (a.k.a, an exploration noise) with standard deviation $\delta > 0$ is added in the exploration process. In the inference stage, there is no exploration noise, the policy is deterministic, i.e., $\delta = 0$ and $A_k = \eta_\theta(S_k)$.

$$\begin{aligned}
H'(\theta) &\triangleq -\mathbb{E}_{S_0 \sim d_0, A_0 \sim \widetilde{\pi}_{\theta,\delta}} \left[Q^{\widetilde{\pi}_{\theta,\delta}}(S_0, A_0) \right] \\
&= -\mathbb{E}_{S_0, A_k \sim \widetilde{\pi}_{\theta,\delta}(S_k,\cdot), S_{k+1} \sim \mathbb{P}(\cdot|S_k, A_k)} \left[\sum_{k=0}^{\infty} \gamma^k \cdot R(S_k, A_k) \right] \\
&= -\sum_{k=0}^{K} \mathbb{E}_{S_0, A_k \sim \widetilde{\pi}_{\theta,\delta}(S_k,\cdot), S_{k+1} \sim \mathbb{P}(\cdot|S_k, A_k)} \left[\gamma^k \cdot R(S_k, A_k) \right] \\
&= -\sum_{k=0}^{K} \sum_{\mu_0}^{\mathcal{S}\times\mathcal{A}} \cdots \sum_{\mu_k}^{\mathcal{S}\times\mathcal{A}} \gamma^k \cdot R(\mu_k) \cdot d_0(S_0) \cdot \widetilde{\pi}_{\theta,\delta}(\mu_0) \prod_{i=0}^{k-1} \left[\mathbb{P}(S_{i+1}|\mu_i) \cdot \widetilde{\pi}_{\theta,\delta}(\mu_{i+1}) \right] \\
&= -\sum_{k=0}^{K} \sum_{\mu_0}^{\mathcal{S}\times\mathcal{A}} \cdots \sum_{\mu_k}^{\mathcal{S}\times\mathcal{A}} \left[\gamma^k \cdot R(\mu_k) \cdot d_0(S_0) \cdot \prod_{i=0}^{k-1} \mathbb{P}(S_{i+1}|\mu_i) \right] \cdot \widetilde{\pi}_{\theta,\delta}(\mu_0) \cdots \widetilde{\pi}_{\theta,\delta}(\mu_k) \\
&= -\sum_{k=0}^{K} \sum_{\mu_0}^{\mathcal{S}\times\mathcal{A}} \cdots \sum_{\mu_k}^{\mathcal{S}\times\mathcal{A}} L_{\mu_0,\ldots,\mu_k} \cdot \widetilde{\pi}_{\theta,\delta}(\mu_0) \cdots \widetilde{\pi}_{\theta,\delta}(\mu_k),
\end{aligned} \tag{27}$$

where $K \to \infty$, and

$$L_{\mu_0,\ldots,\mu_k} = \gamma^k \cdot R(\mu_k) \cdot d_0(S_0) \cdot \prod_{i=0}^{k-1} \mathbb{P}(S_{i+1}|\mu_i). \tag{28}$$

□

E.4 PROOF OF COROLLARY 2

Proof.

$$
\begin{aligned}
\nabla_\theta H'(\pi_\theta) &= -\sum_{k=0}^{K}\sum_{\mu_0}^{\mathcal{S}\times\mathcal{A}}\cdots\sum_{\mu_k}^{\mathcal{S}\times\mathcal{A}}\left(L_{\mu_0,\ldots,\mu_k}\cdot\nabla_\theta\left[\widetilde{\pi}_\theta(\mu_0)\cdots\widetilde{\pi}_\theta(\mu_k)\right]+\nabla_\theta L_{\mu_0,\cdots,\mu_k}\cdot\widetilde{\pi}_\theta(\mu_0)\cdots\widetilde{\pi}_\theta(\mu_k)\right)\\
&= -\sum_{k=0}^{K}\sum_{\mu_0}^{\mathcal{S}\times\mathcal{A}}\cdots\sum_{\mu_k}^{\mathcal{S}\times\mathcal{A}}\left[\widetilde{\pi}_\theta(\mu_0)\cdots\widetilde{\pi}_\theta(\mu_k)\right]\cdot\nabla_\theta L_{\mu_0,\ldots,\mu_k}\\
&= -\sum_{k=0}^{K}\sum_{\mu_0}^{\mathcal{S}\times\mathcal{A}}\cdots\sum_{\mu_k}^{\mathcal{S}\times\mathcal{A}}\nabla_\theta\left[\gamma^k\cdot R(\mu_k)\cdot d_0(S_0)\cdot\prod_{i=0}^{k-1}\mathbb{P}(S_{i+1}|\mu_i))\right]\\
&= -\sum_{k=0}^{K}\sum_{\mu_0}^{\mathcal{S}\times\mathcal{A}}\cdots\sum_{\mu_k}^{\mathcal{S}\times\mathcal{A}}\nabla_A\left[\gamma^k\cdot R(\mu_k)\cdot d_0(S_0)\cdot\prod_{i=0}^{k-1}\mathbb{P}(S_{i+1}|\mu_i))\right]\nabla_\theta\eta_\theta(S)\\
&= -\sum_{S_0}^{\mathcal{S}} d_0(S_0)\nabla_A\mathbb{E}_{S_{t+1}\sim\mathbb{P}(\cdot|S_t,A_t)}\left[\sum_{t=0}^{\infty}\gamma^k R(S_t,A_t)\right]\cdot\nabla_\theta\eta_\theta(S)\\
&= -\sum_{S_0}^{\mathcal{S}} d_0(S_0)\nabla_A Q(S_0,A_0)\cdot\nabla_\theta\eta_\theta(S)\\
&= -\mathbb{E}_{S_0\sim d_0(\cdot)}\left[\nabla_A Q(S_0,A_0)\cdot\nabla_\theta\eta_\theta(S)\right]
\end{aligned}
\tag{29}
$$

where $\mu_k = (S_k, A_k)$, $S_0 \sim d_0(\cdot)$, $A_k \sim \pi_\theta(S_k, \cdot)$, $S_{k+1} \sim \mathbb{P}(\cdot \mid S_k, A_k)$, for $k = 0 \cdots K$. ☐

E.5 PROOF OF COROLLARY 3

Proof. In Corollary 2 and Corollary 1, we have

$$
\begin{aligned}
\nabla_\theta H'(\theta) &= -\nabla_\theta J'(\theta),\\
\nabla_\theta H(\theta) &= -\nabla_\theta J(\theta),
\end{aligned}
\tag{30}
$$

when $K \to \infty$.

[43] proved that

$$\nabla_\theta J'(\theta) = \lim_{\delta\to 0}\nabla_\theta J(\theta), \tag{31}$$

where δ is the standard deviation of the Gaussian noise of stochastic policy π_θ.

Therefore,

$$\nabla_\theta H'(\theta) = \lim_{\delta\to 0}\nabla_\theta H(\theta) \tag{32}$$

☐

F VARIANCE REDUCTION WITH QUANTUM K-SPIN HAMILTONIAN EQUATION

F.1 MONTE CARLO ESTIMATOR OF QUANTUM K-SPIN HAMILTONIAN EQUATION

Monte Carlo Estimator [38]: Consider a general probabilistic objective $\mathcal{F}$ of the form:

$$\mathcal{F} \triangleq \mathbb{E}_{p(\boldsymbol{x};\theta)}[f(\boldsymbol{x};\phi)], \tag{33}$$

in which a function f of an input variable $\boldsymbol{x}$ with *structural parameters* ϕ is evaluated on average with respect to an input distribution $p(\boldsymbol{x};\theta)$ with *distributional parameters* θ.

A Monte Carlo method evaluates the function by first drawing independent samples $\hat{\boldsymbol{x}}^{(1)}, ..., \hat{\boldsymbol{x}}^{(N)}$ from the distribution $p(\boldsymbol{x};\theta)$, and then computing the average:

$$\widehat{\mathcal{F}}_N = \frac{1}{N}\sum_{i=1}^{N} f(\hat{\boldsymbol{x}}^{(i)}), \text{ where } \hat{\boldsymbol{x}}^{(i)} \sim p(\boldsymbol{x};\theta) \text{ for } i = 1, ..., N. \tag{34}$$

The Monte Carlo estimator for (15) is

$$\widehat{J}(\theta) = \frac{1}{N}\sum_{i=1}^{N} R(\tau^{(i)}), \text{ where } \tau^{(i)} \sim P(\tau^{(i)}|\pi_\theta) \text{ for } i = 1, ..., N, \tag{35}$$

and

$$P(\tau^{(i)}|\pi_\theta) = d_0(s_0^{(i)}) \cdot \prod_{k=0}^{T} \mathbb{P}(s_{k+1}^{(i)}|s_k^{(i)}, a_k^{(i)})\pi_\theta(a_k^{(i)}|s_k^{(i)}). \tag{36}$$

The Monte Carlo estimator for (3) is

$$\widehat{H}(\theta) = \frac{1}{N'}\sum_{i=1}^{N'}\sum_{k=0}^{K-1} L_{\mu_0^{(i)},...,\mu_k^{(i)}}, \text{ for } i = 1, ..., N', \tag{37}$$

and

$$L_{\mu_0^{(i)},...,\mu_k^{(i)}} = \gamma^k \cdot R(\mu_k^{(i)}) \cdot d_0(s_0^{(i)}) \cdot \prod_{\ell=0}^{k-1} \mathbb{P}(s_{\ell+1}^{(i)}|\mu_\ell^{(i)}). \tag{38}$$

Remark: The above two Monte Carlo estimators are quite different in the simulation process. (36) samples a random trajectory by following an environment's stochastic transition and a policy. In contrast, (38) measures a random path's discounted reward (the "energy") without following any policy, and the Hamiltonian equation (3) combinatorially enumerates all possible paths of length K over the state-action space. In other words, the simulation process of the Hamiltonian term does not rely on any policy. Therefore, the Hamiltonian term is a suitable regularizer for both on-policy and off-policy algorithms.

This fundamental difference is due to the Ising model in (1), which combinatorially enumerates all paths and separates the environment and the policy.

F.2 VARIANCE REDUCTION

For the general function in (33), one simple but effective variance reduction technique is to subtract a baseline term as follows:

$$\mathbb{E}_{p(\boldsymbol{x};\theta)}\left[(f(\boldsymbol{x}) - \beta)\nabla_\theta \log p(\boldsymbol{x};\theta)\right], \tag{39}$$

where β is the baseline term.

Our reasoning logic:

1). We first briefly describe a high-level idea [22] that adding a baseline term, like the proposed H-term, will help reduce the gradient variance.

2). We sketch the steps to show how the proposed H-term will mathematically reduce the gradient variance, following the framework in Section 5.2 of [23].

High-level IDEA. One generic approach to reduce the variance of Monte Carlo estimates is to use an additive control variate. Suppose we wish to estimate the integral of the function $f : \mathcal{X} \to \mathbb{R}$, and we know the value of the integral of another function on the same space $\phi : \mathcal{X} \to \mathbb{R}$.
We have

$$\int_{\mathcal{X}} f(x) = \int_{\mathcal{X}} (f(x) - \phi(x)) + \int_{\mathcal{X}} \phi(x), \tag{40}$$

and the integral of $f(x) - \phi(x)$ can be estimated. If $\phi(x) = f(x)$, meaning that , then we have managed to reduce our variance to zero [22]. More generally,

$$\text{Var}(f - \phi) = \text{Var}(f) - 2\text{Cov}(f, \phi) + \text{Var}(\phi). \tag{41}$$

If ϕ and f are strongly correlated, so that the covariance term on the right hand side is greater than the variance of ϕ, i.e., $-2\text{Cov}(f, \phi) + \text{Var}(\phi) \leq 0$. then a variance reduction has been made over the original estimation problem [22], i.e., $\text{Var}(f - \phi) \leq \text{Var}(f)$.

Our reasoning. Then, we present our reasoning.

Note that the gradient of the new objective function of the actor network in (8) consists of two components, namely $\nabla_\theta J(\theta)$ and $\nabla_\theta H(\theta)$. Here, we consider

$$\nabla_\theta J(\theta) - \lambda \, \nabla_\theta H(\theta), \quad \text{where } \lambda > 0 \text{ is a temperature parameter,} \tag{42}$$

where $\nabla_\theta J(\theta)$ in (21) is the above function $f(\cdot)$ and $\lambda \, \nabla_\theta H(\theta)$ in (4) is the above function $\phi(\cdot)$.

The Hamiltonian stochastic gradient in (4) has the optimal value

$$\nabla_\theta H^*(\theta) = - \lim_{K \to \infty} \mathbb{E}_{\mu_0, \ldots, \mu_{K-1}} \left[\sum_{k=0}^{K-1} \gamma^k \cdot R(\mu_k) \cdot \nabla_\theta \log \left(\pi_\theta(\mu_0) \cdot \pi_\theta(\mu_1) \cdots \pi_\theta(\mu_k) \right) \right]. \tag{43}$$

According to Theorem 8 of [22] that is proved via (41), we have

$$\begin{aligned} \text{Var}\left[\nabla_\theta J(\theta) - \lambda \, \nabla_\theta H^*(\theta)\right] &= \text{Var}[\nabla_\theta J(\theta)] - \frac{1}{\lambda} \mathbb{E}_{s \sim d_\theta, a \sim \pi_\theta} \left[\frac{\left(\mathbb{E}_{s \sim d_\theta, a \sim \pi_\theta} \left[(\nabla_\theta \log \pi_\theta(s,a))^2 \nabla_\theta J(\theta) \right]\right)^2}{\mathbb{E}_{s \sim d_\theta, a \sim \pi_\theta} \left[(\nabla_\theta \log \pi_\theta(s,a))^2 \right]} \right] \\ &\leq \text{Var}[\nabla_\theta J(\theta)], \end{aligned} \tag{44}$$

where the second term is positive and

$$\nabla_\theta H^*(\theta) = \frac{\mathbb{E}_{s \sim d_\theta, a \sim \pi_\theta} (\left[\nabla_\theta \log \pi_\theta(s,a))^2 \nabla_\theta J(\theta) \right]}{\mathbb{E}_{s \sim d_\theta, a \sim \pi_\theta} \left[(\nabla_\theta \log \pi_\theta(s,a))^2 \right]}. \tag{45}$$

In Alg. 1 and Alg. 2, we used a general H-term $\nabla_\theta H(\theta)$, not the optimal one in (43). Next, we provide a general characterization for this case.

According to Theorem 10 of [22], we have

$$\begin{aligned} &\text{Var}\left[\nabla_\theta J(\theta) - \lambda \, \nabla_\theta H(\theta)\right] - \text{Var}\left[\nabla_\theta J(\theta) - \lambda \, \nabla_\theta H^*(\theta)\right] \\ &= \lambda^2 \, \mathbb{E}_{s \sim d_\theta, a \sim \pi_\theta} \left[(\nabla_\theta \log \pi_\theta(s,a))^2 (\nabla_\theta H(\theta) - \nabla_\theta H^*(\theta))^2 \right] \end{aligned} \tag{46}$$

Assume Lipschiz continuity of the graident $\nabla_\theta H(\theta)$ such that

$$||\nabla_\theta H(\theta) - \nabla_\theta H^*(\theta)||_2 \leq 2L(H(\theta) - H^*(\theta)) \leq 2L\epsilon, \tag{47}$$

given $K \geq \log_\gamma \epsilon$ with $L > 0, \epsilon > 0$, as pointed out in the end of Section 3.2.

Therefore, combining (46), (48) with (48), we obtain that

$$
\begin{aligned}
&\text{Var}\left[\nabla_\theta J(\theta) - \lambda\, \nabla_\theta H(\theta)\right] = \text{Var}[\nabla_\theta J(\theta)] - \frac{1}{\lambda}\mathbb{E}_{s\sim d_\theta, a\sim\pi_\theta}\left[\frac{\left(\mathbb{E}_{s\sim d_\theta, a\sim\pi_\theta}\left[(\nabla_\theta \log \pi_\theta(s,a))^2 \nabla_\theta J(\theta)\right]\right)^2}{\mathbb{E}_{s\sim d_\theta, a\sim\pi_\theta}\left[(\nabla_\theta \log \pi_\theta(s,a))^2\right]}\right] \\
&\quad + \lambda^2 (2L\epsilon)^2\, \mathbb{E}_{s\sim d_\theta, a\sim\pi_\theta}\left[(\nabla_\theta \log \pi_\theta(s,a))^2\right] \\
&\leq \text{Var}[\nabla_\theta J(\theta)],
\end{aligned}
\tag{48}
$$

when both $|\nabla_\theta \log \pi_\theta(s,a)|$ and $|\nabla_\theta J(\theta)|$ are upper bounded, e.g., $|\nabla_\theta \log \pi_\theta(s,a)| < C_1$ and $|\nabla_\theta J(\theta)| < C_2$; and we set ϵ, λ properly such that

$$
\begin{aligned}
&-\frac{1}{\lambda}C_2^2 + 4\lambda^2 L^2 \epsilon^2 C_1^2 < 0 \\
&\lambda^3 \epsilon^2 < \frac{C_2^2}{4L^2 C_1^2},
\end{aligned}
\tag{49}
$$

which can be easily satisfied by properly selecting λ and $K \geq \log_\gamma \epsilon$.

Conclusion:

To sum up, we show that it is easy to achieve $\text{Var}\left[\nabla_\theta J(\theta) - \lambda\, \nabla_\theta H(\theta)\right] \leq \text{Var}[\nabla_\theta J(\theta)]$, which means adding the H-term can lead to smaller variance than that of the conventional gradient.

G ACTOR-CRITIC ALGORITHMS FOR DEEP REINFORCEMENT LEARNING

The gradient of (15) is [44]

$$\nabla_\theta J(\theta) \triangleq \sum_S^{\mathcal{S}} d_{\mathcal{S},\theta}(S) \sum_A^{\mathcal{A}} Q_\theta(S,A)\; \nabla_\theta \pi_\theta(S,A). \tag{50}$$

Since Q_θ in (50) is unknown [49] (the stationary distribution d_θ is unknown), one can plug in a critic network with parameter ϕ as an estimator of Q_θ and obtain

$$\nabla_\theta^\phi J(\theta,\phi) = \sum_S^{\mathcal{S}} d_{\mathcal{S},\theta}(S) \sum_A^{\mathcal{A}} Q_\phi(S,A)\; \nabla_\theta \pi_\theta(S,A), \tag{51}$$

where $d_{\mathcal{S},\theta} \in \mathbb{R}_+^{|\mathcal{S}||\mathcal{A}|\times 1}$ denotes the stationary distribution over the states instead of state-action pairs.

(51) is a bi-level optimization problem [9], and a natural solution is an iterative algorithm that alternates between estimating Q_ϕ with parameter ϕ and improving policy π_θ with parameter θ. Therefore, a family of actor-critic algorithms are proposed with following objective functions:

$$\begin{cases} \text{Actor} : \max\limits_\theta J_\pi(\theta,\phi) = (1-\gamma)\mathbb{E}_{S_0\sim d_0, A_0\sim\pi_\theta(S_0,\cdot)}\left[Q_\phi(S_0,A_0)\right] \\ \text{Critic} : \max\limits_\phi J_Q(\theta,\phi) = \dfrac{1}{2}\mathbb{E}_{S\sim d_\theta(\cdot), A\sim\pi_\theta(S,\cdot)}\left[(Q_\phi(S,A) - y(S,A))^2\right]. \end{cases} \tag{52}$$

The gradient of (52) can be estimated as follows

$$\begin{aligned} \nabla_\theta \widehat{J}_\pi(\theta,\phi) &= \frac{1}{N}\sum_{i=1}^N Q_\phi(\mu)\cdot\nabla_\theta \log\pi_\theta(\mu) \\ \nabla_\phi \widehat{J}_Q(\theta,\phi) &= \frac{1}{N}\sum_{i=1}^N \left[Q_\phi(S,A) - y(S,A)\right]\cdot\nabla_\phi Q_\phi(S,A) \end{aligned} \tag{53}$$

The parameters ϕ and θ are updated as follows:

$$\begin{cases} \text{Actor} : & \theta \leftarrow \theta + \alpha\; \nabla_\theta^\phi \widehat{J}_\pi, \\ \text{Critic} : & \phi \leftarrow \phi - \alpha\; \nabla_\phi \widehat{J}_Q. \end{cases} \tag{54}$$

Algorithm 2 Stationary Actor-Critic Algorithm with H-term (Deterministic Version)

```
Input: learning rate $\alpha$, temperature $\lambda$, look-ahead step $K$, and parameters $\delta, M, T, G, B, B'$
Initialize actor network $\eta$ and critic network $Q$ with parameters $\theta$, $\phi$, and replay buffers $\mathcal{D}_1$, $\mathcal{D}_2$
for episode $= 1, \cdots, M$ do
    Initialize state $s_0$
    for $t = 0, \cdots, T-1$ do
        Take action $a_t = \eta_\theta(s_t) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \delta^2)$
        Execute action $a_t$, receive reward $r_t$, and observe new state $s_{t+1}$
        Store a transition $(s_t, a_t, r_t, s_{t+1})$ in $\mathcal{D}_1$
    end
    Store a trajectory $\tau$ of length $T$ in $\mathcal{D}_2$
    for $g = 1, \cdots, G$ do
        Randomly sample a mini-batch of $B$ transitions $\{(s_i, a_i, r_i, s_{i+1})\}_{i=1}^{B}$ from $\mathcal{D}_1$
        Randomly sample a mini-batch of $B'$ trajectories (of length $K$) $\{\tau_j\}_{j=1}^{B'}$ from $\mathcal{D}_2$
        Update critic network using a conventional method
        Update actor network as $\theta \leftarrow \theta + \alpha \left(\nabla_\theta \widehat{J}'(\theta) - \lambda \, \nabla_\theta \widehat{H}'(\theta)\right)$.
    end
end
```

H Deterministic Policy Gradient Algorithm with H-term

For completeness, we present the details of the deterministic actor-critic algorithm with H-term.

We apply the proposed Hamiltonian equation (3) to regularize the actor network. Specifically, $H'(\theta)$ in (3) is added to the actor's objective with weight $\lambda > 0$. The objective functions of actor and critic networks become:

$$\begin{cases} \text{Actor} : \max\limits_{\theta} J'_\pi(\theta, \phi) = (1-\gamma)\mathbb{E}_{S_0 \sim d_0, A_0 = \eta_\theta(S_0)} \left[Q_\phi\left(S_0, A_0\right)\right] - \lambda H'(\theta), \\ \text{Critic} : \min\limits_{\phi} J_Q(\theta, \phi) = \dfrac{1}{2}\mathbb{E}_{S \sim d_\theta(\cdot), A = \eta_\theta(S)} \left[\left(Q_\phi(S, A) - y(S, A)\right)^2\right]. \end{cases} \tag{55}$$

The gradient of (55) is

$$\nabla_\theta J'_\pi(\theta, \phi) = (1-\gamma) \sum_{S}^{\mathcal{S}} d_{\mathcal{S},\theta}(S) \nabla_A Q_\phi(S, A) \cdot \nabla_\theta \eta_\theta(S) - \lambda \nabla_\theta H'(\theta), \tag{56}$$

$$\nabla_\phi J_Q(\theta, \phi) = \sum_{S}^{\mathcal{S}} d_{\mathcal{S},\theta}(S) \cdot \left[Q_\phi(S, A) - y(S, A)\right] \cdot \nabla_\phi Q_\phi(S, A)|_{A = \eta_\theta(S)}. \tag{57}$$

To estimate $\nabla_\theta H'(\theta)$, the Monte Carlo gradient estimator in (6) is used. Therefore, (56) and (57) can be estimated as follows:

$$\nabla_\theta \widehat{J}'_\pi(\theta, \phi) = \frac{1}{N} \sum_{i=1}^{N} \left[\nabla_A Q_\phi(S, A)|_{A=\eta_\theta(S)} \nabla_\theta \eta_\theta(S)\right] - \frac{1}{N'} \sum_{i=1}^{N'} \left[\lambda \sum_{k=0}^{K} \gamma^k R(\mu_k) \nabla_\theta \log \left[\widetilde{\pi}_\theta(\mu_0) \cdots \widetilde{\pi}_\theta(\mu_k)\right]\right], \tag{58}$$

$$\nabla_\phi \widehat{J}_Q(\theta, \phi) = \frac{1}{N} \sum_{i=1}^{N} \left[Q_\phi(S, A) - y(S, A)\right] \cdot \nabla_\phi Q_\phi(S, A)|_{A=\eta_\theta(S)}. \tag{59}$$

I EXPERIMENTS: HYPERPARAMETERS AND MORE RESULTS

I.1 HYPERPARAMETERS IN EXPERIMENTS

Table 5: Hyperparameters used for the PPO and PPO + H in MuJoCo tasks

Parameters	Values
Optimizer	Adam
Learning rate	$3 \cdot 10^{-4}$
Discount (γ)	0.99
GAE parameter	0.95
Number of hidden layers for all networks	3
Number of hidden units per layer	256
Mini-batch size	32
Importance rate of H-term (λ)	2^{-3}
Truncation step of H-term (K)	16

Table 6: Hyperparameters used for the DDPG and DDPG + H in MuJoCo tasks

Parameters	Values
Optimizer	Adam
Learning rate	$5 \cdot 10^{-4}$
Target Update Rate (τ)	10^{-3}
Discount (γ)	0.995
Replay buffer size	10^6
Number of hidden layers for all networks	3
Number of hidden units per layer	256
Batch size	64
Importance rate of H-term (λ)	2^{-3}
Truncation step of H-term (K)	16

Table 7: Hyperparameters used for the REINFORCE and REINFORCE + H in TSP, graph maxcut, MIMO beamforming, and non-convex deep learning classifier tasks

Parameters	Values
Optimizer	Adam
Learning rate	$5 \cdot 10^{-5}$
Discount (γ)	0.99
Number of hidden layers for all networks	3
Number of hidden units per layer	512
Batch size	4096
Importance rate of H-term (λ)	2^{-3}
Truncation step of H-term (K)	6

I.2 MORE RESULTS

Fig. 7 shows the H-value (average over 20 runs) during the training process, which verified that the trained agents have converged to policies with small H-values.

Fig. 8 shows more performance of the PPO+H algorithm, for $K = 8, 16, 24$. We run each experiment with 20 random seeds and each run we test 100 episodes.

To verify the hypothesize that smaller replay buffer hurts the performance, we rerun the trials of $K = 8, 16$ with a replay buffer size 800.

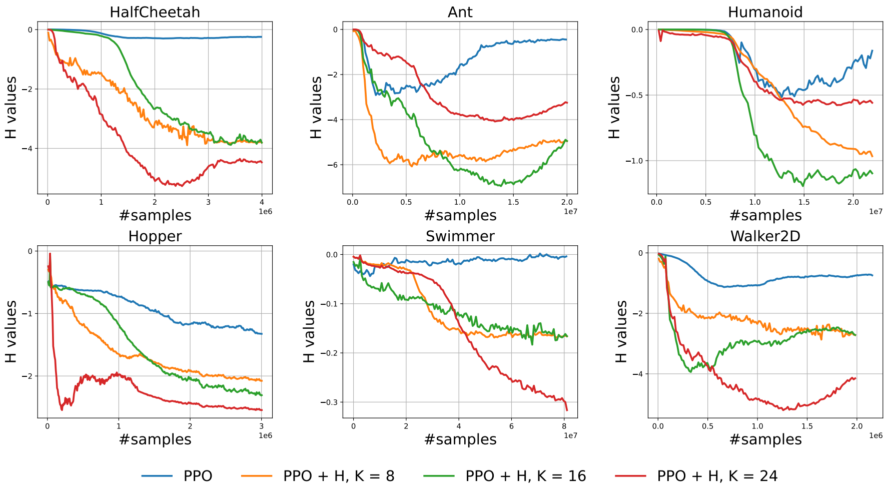


Figure 7: H values during the training process.

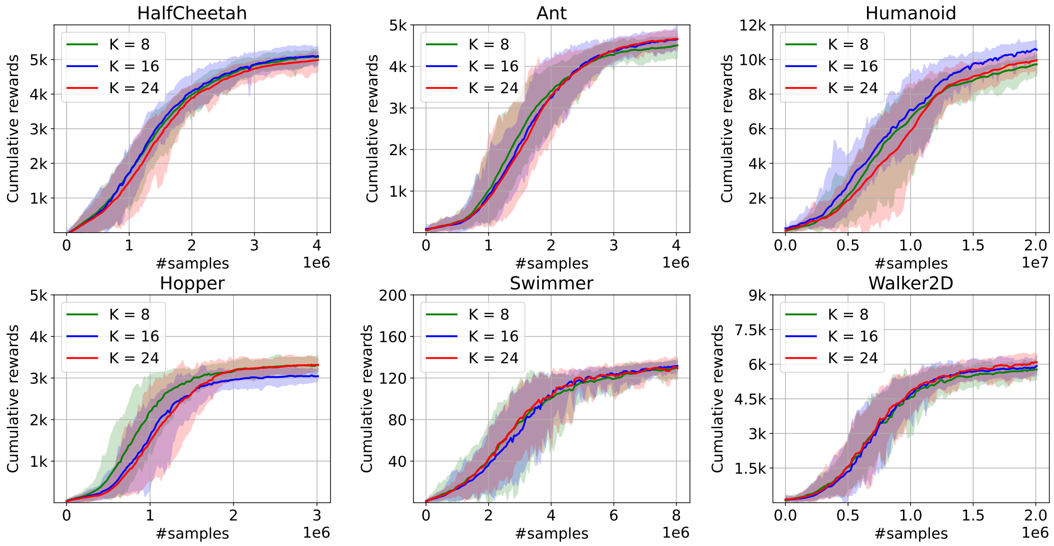


Figure 8: For the proposed PPO+H algorithm, the performance with different K values.

J HAMILTONIAN POLICY NETWORK

J.1 HAMILTONIAN POLICY NETWORK

Since Hamiltonian equation in (3) is a functional of policy π_θ, a natural question would be: can we use the Hamiltonian equation replace existing Bellman's equation (16) or the policy gradient's objective function (15)?

As a verification, we test the capability of Hamiltonian equation in (3) as a loss function to train a policy network. The algorithm is first given as follows.

In Alg. 3, an agent interacts with an environment and updates its policy network. The algorithm has M episodes and each episode consists of a (Monte Carlo) simulation process and a learning process (gradient estimation) as follows :

Algorithm 3 Hamiltonian Policy Network

```
1:  Input: learning rate $\alpha$, look-ahead step $K$, and parameters $M, T, G, B$
2:  Initialize policy network with parameters $\theta$, and replay buffer $\mathcal{D}$
3:  for episode $= 1, \cdots, M$ do
4:      Initialize state $s_0$
5:      for $t = 0, \cdots, T-1$ do
6:          Select action $a_t \sim \pi_\theta(\cdot|s_t)$
7:          Execute action $a_t$, receive reward $r_t$, and observe new state $s_{t+1}$
8:      end
9:      Store a trajectory $\tau$ of length $T$ in $\mathcal{D}$
10:     for $g = 1, \cdots, G$ do
11:         Randomly sample a mini-batch of $B$ trajectories (of length $K$) $\{\tau_j\}_{j=1}^{B}$ from $\mathcal{D}$
12:         Update pocliy network as $\theta \leftarrow \theta - \alpha\ \nabla_\theta \widehat{H}(\theta)$.
13:     end
14: end
```

- During the (Monte Carlo) simulation process (lines 5-9 of Alg. 3), an agent takes action a_t according to a policy $\pi_\theta(\cdot|s_t)$, $t = 0, \cdots, T-1$, generating a trajectory of T steps/transitions. Then, the full trajectory $\tau = (s_0, a_0, r_0, s_1, \cdots, s_{T-1}, a_{T-1}, r_{T-1}, s_T)$ is stored in replay buffer $\mathcal{D}$.
- During the learning process ($G \geq 1$ updates in one episode) (lines 10-12 of Alg. 1), a mini-batch of B trajectories (of length K) $\{\tau_j = (s_0^j, a_0^j, r_0^j, s_1^j, \cdots, s_{K-1}^j, a_{K-1}^j, r_{K-1}^j, s_K^j)\}_{j=1}^{B}$ are sampled from $\mathcal{D}$, respectively. The policy network is updated by a Monte Carlo gradient estimator over B trajectories.

Implementation of replay buffer $\mathcal{D}$. After a full trajectory τ of length T is generated, it is partitioned into $T - K + 1$ trajectories of length K. We rank them according to the cumulative reward and store the top portion, say 80%, into a new replay buffer $\mathcal{D}$ (line 9 of Alg. 3). We randomly sample a mini-batch of B trajectories from $\mathcal{D}$ (line 11 of Alg. 3) to compute the H-term.

J.2 FROZENLAKE TASK

Environment: Frozenlake 8×8, a game in OpenAI Gym.

Rules: As shown in Fig. 9 (left), the Frozenlake task has 8×8 states with 4 optional actions to move around. The agent needs to go from the start point and find the way to the destination in limited steps. There are 8 holes which can cause the agent to fail the game.

Experiment settings: We take Deep Q-learning (DQN) [37] as our baseline and use the implementation from the ElegantRL library. We use a 4-layer fully connected neural network as the deep policy network both in DQN and DHN. We use the Adam optimizer with a learning rate 1×10^{-3} and a batch size 100.

Evaluation: We evaluate the performance of policy by computing the success rate, in which we use 50 agents to walk 100 steps and compute the rates of agents who successfully arrive the destination.

Results for the Frozenlake task: Fig. 10 (left) shows the success rate of agents with increasing the number of transitions learned by the network. compared with DQN, DHN has a more stable training process. It is easy for DQN to quickly obtain a good policy to win the game. But with increasing the number of transitions fed to the network, the performance of DQN shows a large and frequent shock while the performance of DHN shows the strong stability.

J.3 GRIDWORLD TASK

Environment: a Gridworld of size 10×10, a game available in our code.

Rules: As shown in the Fig. 9 (right), the Gridworld has 10×10 states with 4 optional actions to move around. The agent will initialize at a random locations and it needs to find the smiley as many

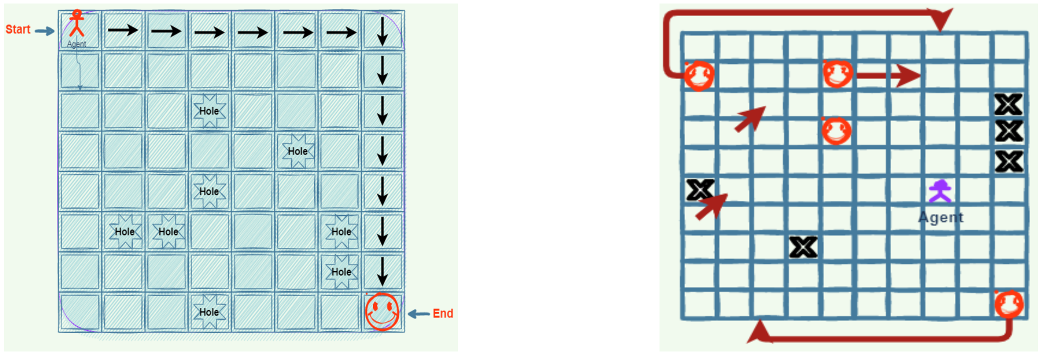


Figure 9: The Frozenlake task (left) and Gridworld task (right).

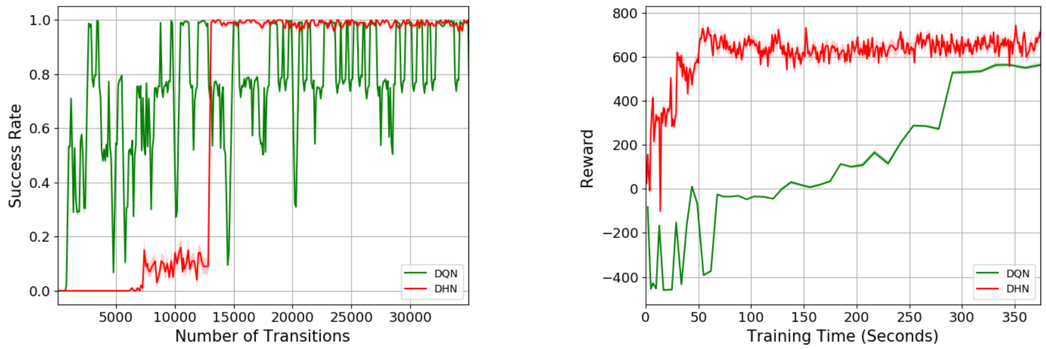
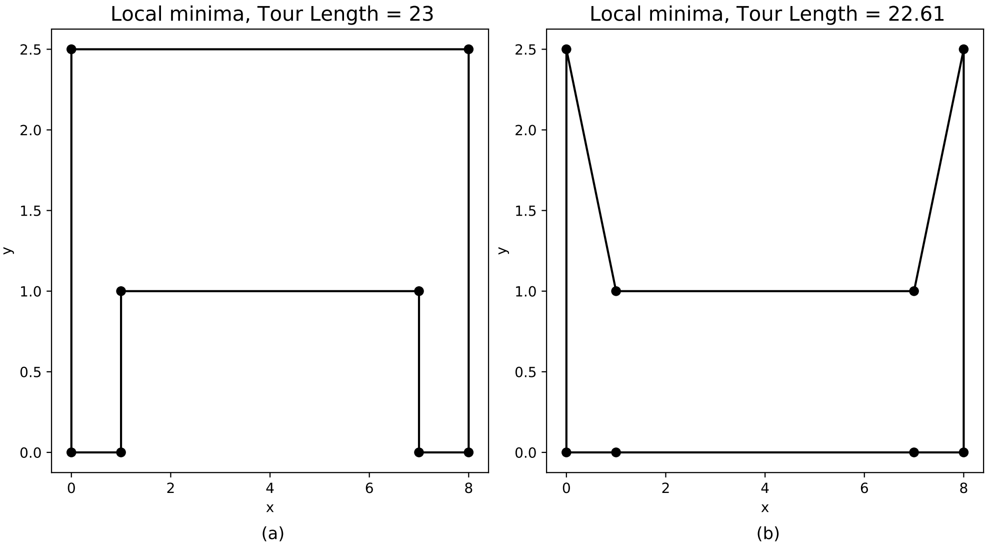


Figure 10: Comparison between the DQN and DHN algorithms. The Frozenlake task (left) and Gridworld task (right).

as possible which has 10 reward in turn. It should be noted that there are some endpoints which may cause the agent game over and some transfer-points which transfer the agent to certain location.

Experiment settings and evaluation: Both the experiment settings and evaluation method are the same with that on Frozenlake 8×8 game.

Results for the Gridworld task: Fig. 10 (rigt) shows the mean reward obtained by the agents with increasing the training time. Compared with DQN, DHN has a faster training process. It only needs massive random parallel samples of trajectories and do not need any policy for guided sampling while DQN needs guided exploration in the training process which costs a large time consumption.



(a) (b)

Figure 11: TSP problem has two local minimas in this case.

K Combinatorial Optimization Problems

K.1 Problem Formulation and MDP formulation

- **Graph maxcut**: Given a graph $G = (V, E)$, where V is the set of nodes and E the set of edges, find a subset $S \subseteq V$ that maxmizes the weight of the cut-set $\sum_{u \in S, v \in V \setminus S, (u,v) \in E} w(u, v)$.
- **Traveling salesman problems (TSP)**: Given a fully connected graph $G = (V, E)$, find a tour J that minimizes the edge weights $\sum_{e \in J} w(e)$. A tour starts and ends at a specific node after having visited each node exactly once.
- **MDP formulation**: The MDP formulations of graph maxcut and TSP are given in Table 8. Note, J is a partially ordered set, (J, v) means add node v to the end of J.

K.2 Existence of Many Local Minimas

- Travelling salesman problem (TSP): the case of 8 cities (Fig. 11) has 2 local minimas.
- Graph maxcut: a fully connected graph of 20 nodes (Appx. K) has 390 local minimas.

K.3 Experimental Settings

- **Environments (tasks)**. We train the proposed algorithms on TSP and graph maxcut tasks with number of nodes $N = 100$. For TSP, all nodes in training instances are generated from an uniform distribution within an unit square. For graph maxcut, training instances are randomly generated from Erdős–Rényi model. We use the same test cases in [14, 31].
- **Compared algorithms**. To evaluate policy gradient algorithm, we choose REINFORCE [45]. Since the H-term is compatible with existing variance reduction techniques, we implement the REINFORCE algorithm with baseline reduction. For a fair comparison, we keep the hyperparameters (listed in Appx. I) the same and make sure that the obtained results reproduce existing benchmark tests.
- **Performance metrics**. We employ the metric: Frequency vs. Approximation Ratio. We run each experiment with 20 random seeds and in each run we test 100 episodes.The approximation ratio ϵ is defined as

$$\epsilon = \max(\frac{Optimal}{Obj}, \frac{Obj}{Optimal}), \tag{60}$$

where $Optimal$ is the optimal objective value by Branch and Bound [32], Obj is the objective value of the best solution obtained from RL.

Table 8: MDP formulation of combinatorial optimization: Graph maxcut and TSP

Problem	State	Action	Transition	Reward	Termination
Graph maxcut	S	select $v \in V \backslash S$	$S' \leftarrow S \cup \{v\}$	$\text{cut}(S') - \text{cut}(S)$	$\forall S', \text{cut}(S') < \text{cut}(S)$
TSP	J	select $v \in V \backslash J$	$J' \leftarrow (J, v)$	$cost(J') - cost(J)$	$\forall v \in V, v \in J$

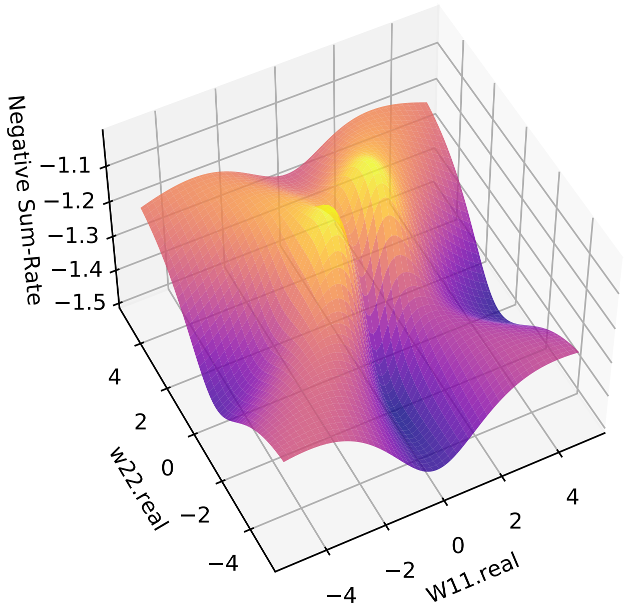


Figure 12: Geometry of sum-rate (N = 2, M = 2, SNR = 10) with respect to the real part of W_{11} and W_{22}. It has three local minimas.

L NON-CONVEX OPTIMIZATION PROBLEMS

- MIMO beamforming [7]: Given a channel $[\boldsymbol{h}_1, \cdots, \boldsymbol{h}_M] \in \mathbb{C}^{N \times M}$, and a power constraint $P \in \mathbb{R}$, find a beamformer $W = [\boldsymbol{w}_1, \cdots, \boldsymbol{w}_M] \in \mathbb{C}^{N \times M}$ that maximizes the summation of rates:

$$\begin{aligned} &\max_{\boldsymbol{W}} \quad \sum_{m=1}^{M} \log_2 \left(1 + \text{SINR}_m\right) \\ &s.t. \quad \operatorname{Tr}\left(\boldsymbol{W}\boldsymbol{W}^H\right) \leq P, \\ &\text{where} \quad \text{SINR}_m = \frac{\left|\boldsymbol{h}_m^H \boldsymbol{w}_m\right|^2}{\sum_{j=1, j \neq m}^{M} \left|\boldsymbol{h}_m^H \boldsymbol{w}_j\right|^2 + \sigma^2}. \end{aligned} \tag{61}$$

- Non-convex deep learning classifier [33]: Consider a 2-layer binary classifier, using ReLU as activation layer and softmax as output layer. The task is to find weights W_1, W_2 and biaes b_1, b_2 that maximize the cross entropy loss:

$$\min_{W_1, W_2, \mathbf{b_1}, \mathbf{b_2}} -\frac{1}{n} \sum_{i=1}^{n} \log \left(\frac{\exp\left(\left(W_2 \max\left(W_1 \mathbf{x}_i + \mathbf{b_1}, 0\right) + \mathbf{b_2}\right)_{y_i} \right)}{\sum_j \exp\left(\left(W_2 \max\left(W_1 \mathbf{x}_i + \mathbf{b_1}, 0\right) + \mathbf{b_2}\right)_j \right)} \right), \tag{62}$$

where W_i and b_i are the weight and bias of i-th layer, n is the batch size, x_i is the input, and y_i is the target class.