Ising Networks for Deep Hierarchical Reinforcement Learning

1 Notation

Reinforcement Learning: We review the RL setup wherein an agent interacts with the environment in order to transition to new states and observe rewards by following a sequence of actions. The problem is modeled as a finite-horizon Markov Decision Process(MDP) [**?**] defined by the tuple $(\mathcal{S}, \mathcal{A}, r, P, \gamma)$ where the state space $\mathcal{S}$ and action space $\mathcal{A}$ are continuous, r presents the reward observed by agent such that $r : \mathcal{S} \times \mathcal{A} \to [r_{min}, r_{max}]$, $P : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, \infty)$ presents the unknown transition model consisting of the transition probability to the next state $s_{t+1} \in \mathcal{S}$ given the current state $s_t \in \mathcal{S}$ and action $a_t \in \mathcal{A}$ at time step t and γ is the discount factor. We consider a policy $\pi_\theta(a_t|s_t)$ as a function of model parameters θ. Standard RL defines the agent's objective to maximize the expected discounted reward $\mathbb{E}_{\pi_\theta}[\sum_{t=0}^{T} \gamma^t r(s_t, a_t)]$ as a function of the parameters θ. The action-value function for an agent is represented as $Q(a, s; \theta) = \mathbb{E}_{\pi_\theta}[\sum_{t=1}^{T} \gamma^t r(s, a)|s = s_t, a = a_t]$ which is the expected sum of payoffs obtained in state s upon performing action a by following the policy π_θ. We denote the optimal policy π^*_θ such that $Q(a, s; \theta^*) \geq Q(a, s; \theta) \forall s \in S, a \in A$.

Multi-Agent Learning: We now review the cooperative MARL setup. The problem is modeled as a Partially Observable Markov Decision Process (POMDP) [**?**] defined by the tuple $(\mathcal{S}, \mathcal{A}, r, C, P, \mathcal{Z}, O, \gamma)$ wherein the notations are consistent with the single-agent RL MDP setup with the state space $\mathcal{S}$ and action space $\mathcal{A}$ being discrete and r presenting the reward observed by agents $u \in C$ where C is the collective set of all agents. Note that we use alternate notations of $s' \in \mathcal{S}$ for s_{t+1} and $a' \in \mathcal{S}$ for a_{t+1}. We consider a partially observable setting in which each agent u draws individual observations $z \in \mathcal{Z}$ according to the observation function $O(s, a) : \mathcal{S} \times \mathcal{A} \to \mathcal{Z}$. Similar to the single-agent RL case, we denote the optimal policy π^*_θ such that $Q(a, s; \theta^*) \geq Q(a, s; \theta) \forall s \in S, a \in A$. In the case of multiple agents, the joint optimal policy can be expressed as the Nash Equilibrium [**?**] of the Stochastic Markov Game as $\pi^* = (\pi^{1,*}, \pi^{2,*}, ...\pi^{C,*})$ such that $Q(a^u, s; \theta^*) \geq Q(a^u, s; \theta) \forall s \in S, a \in A, u \in C$.

2 Introduction

3 Related Work

3.1 Hierarchical Reinforcement Learning

3.1.1 Temporal Abstraction

Various methods have devised hierarchical techniques to abstract the temporal correlation in MDPs [**?**, **?**, **?**, **?**]. The Hierarchical Deep Q-Network (hDQN) [**?**] proposes an integration of temporal abstraction and intrinsic motivation in order to solve long-horizon problems. hDQN demonstrates improved long-horizon optimality at the cost of manual feature engineering of the states and reward function. This is addressed by making use of deep successor networks [**?**] which extract relevant goal embeddings for the agent and direct its behavior towards the goal. A more sophisticated temporal scheme consists of learning abstraction of hierarchies based on expert demonstrations and regularized latent space constraints [**?**]. However, sub-policies rely on memory-based controllers for learning fine-grained abstraction and often cripple the lower levels of the hierarchy once the task has been accomplished. This leads to limited transfer of skills between policies and hence, results in sub-optimal convergence [**?**]. Transfer of skills between different policies operating in sub-MDPs is extensively studied using exploration [**?**]. [**?**] presents MAVEN, a hierarchical architecture for variational exploration in the case of multi-agent settings. Policies carry out exploration in the latent space using variational inference. MAVEN demonstrates improved coordination between agents in the long-horizon. However, policy behavior distributed across agents which forces sub-policies to rely on other agent policies for optimal convergence. [**?**] proposes the trajectory autoencoder which enables the agent to exhibit self-consistent behavior and tackle sparse rewards in the long-horizon. Long-horizon suitability can further be improved by jointly training all policies and utilize the skills learned individually [**?**]. Such kind of sub-policy adaptation stabilizes the behavior of agent and provides efficient training of hierarchies at all levels.

Abstraction may be carried out among states [**?**] or using a different medium such as language [**?**] which allows the agent to solve temporally-extended tasks in a diverse manner. [**?**] presents the MAXQ algorithm which proposes an efficient method for training state-based abstractions in the form of hierarchies. Q-values corresponding to each lower-level task are aggregated at the higher-level policy nodes to yield a modular framework for abstraction. While the MAXQ demonstrates suitability for low-dimensional tasks, its performance cannot be assessed in high-dimensional state and action spaces due to the larger number of abstractions produced by the framework. [**?**] tackles the problem of high-dimensional and long-horizon learning by presenting a theoretical perspective of abstracting states using transitive behavior which helps in optimal computations. Such techniques have motivated active research in the form of modified abstraction

algorithms for performance improvement and schemes leading to richer abstract representations [**?**].

3.1.2 Option-based Hierarchies

Hierarchies can be constructed on the basis of skills or possible options of policies which the agent may choose from during its interactions in the environment [**?**, **?**]. [**?**] presents the Options-Critic Architecture which is a collection of various policy options. Agent can select from the collection of sub-policies at various timesteps to yield a policy which is robust and optimal in the long-horizon. The provision of options enables the agent to acquire a wide variety of skills during exploration and obtain a policy which makes best use of these skills. Furthermore, the architecture incorporates learning of termination conditions of a particular option which is essential for switching option policies [**?**]. The architecture is extended to the Double Actor-Critic architecture which comprises of two parallel MDPs [**?**]. A dual architecture enables the learning of intra-option policies with termination conditions of each option. While, option-based learning provides diversity in skills and control of temporal extension to the agent, frequent fluctuations between options results in instability in updates and hence, sample-inefficient learning [**?**]. [**?**] addresses the instability in training by combining the options framework with off-policy maximum entropy reinforcement learning [**?**, **?**]. Maximization of entropy allows the agent to effectively explore all options and consistently select the optimal sub-policies based on selected options. However, usage of different skills and compactness remains an open problem in the case of real-world tasks consisting of large action spaces. [**?**] addresses the problem of compact options by improvising communication between policies using binary vectors. Modular policies using binary vectors enable richer communication and mixing of skills. Techniques related to the relative optimization of policy can be extended using advantage-weighted importance sampling [**?**] in options based on option-value functions.

3.1.3 Hierarchical Control

Hierarchical Reinforcement Learning borrows from multiple approaches in control [**?**] and adaptive policy optimization [**?**]. Multiplicative Compositional Policies (MCPs) [**?**] present the composable nature of hierarchies which arises from coordination of multiple control-based skills. MCPs improve the hierarchical coordination between sub-policies which results in adaptive motor skills in the case of robotic control such as locomotion and manipulation. While MPC demonstrates the practical usage of hierarchical reinforcement learning motivated by adaptive control, [**?**] extends the hierarchical framework to Quadruped Locomotion. The lower-level policy uses latent commands for robot actuator control and the higher-level policy operates at different time intervals. Abstraction of control between lower and higher-level policies allows the agent to acquire a diverse set of task-related skills. Moreover, policies can be efficiently adapted on

a different task from the same domain. Additionally, diversity can be achieved in a scalable manner by training policies levelwise [**?**]. However, levelwise training of sub-policies becomes intractable in cases of a large number of hierarchies [**?**]. [**?**] tackles the problem of a large number of hierarchies and levelwise training by making use of latent space policies. Each latent space policy is trained using the maximum entropy learning framework in order to directly solve the task. Hierarchies trained in latent space demonstrate suitable initializations on complex control tasks which enable the agent to maintain consistent behavior. Consistency can also be achieved using goal embeddings which motivate goal-directed behavior in the case of navigation tasks [**?**]. [**?**] additionally improves consistency in hierarchical agents by making use of information theoretic regularizations which pose a reward penalty on the policy distribution. Hierarchical control in reinforcement learning has been successful in tackling knowledge transfer as well as data-efficiency [**?**]. Various levels of the hierarchy make use of inductive-biases for sharing off-policy data. Sharing knowledge across policies in a compositional manner results in positive transfer of policies. Such techniques are additionally used to solve numerically challenging control problems consisting of unstructured environments [**?**].

3.2 Ising Models

3.2.1 Energy-based Reinforcement Learning

Recent work on energy-based learning [**?**] has demonstrated significant success in the arena of reinforcement learning [**?**, **?**]. [**?**] highlights the use of energy-based policies capable of learning large action spaces by making use of factored approximations. Such approximations can be thought of as inference mechanisms [**?**] using energy-based distributions such as the boltzmann distribution [**?**]. These have given rise to energy-based extensions of models with different architectures such as actor-critic frameworks [**?**]. Another suitable method for training energy-based policies in reinforcement learning is by making use of the Helmholtz free energy [**?**]. [**?**] presents the suitability of improved helmholtz energy operator under various reinforcement learning scenarios with theoretical guarantees on robustness and improved inference in comparison to boltzmann energy. Applications of the energy operator include surprise minimization in the case of multi-agent learning [**?**]. [**?**] highlights the practical use of energy in multi-agent learning across agents which improves the joint policy to tackle environments consisting of high variance. The use of energy-based policies can further be generalized to inverse reinforcement learning [**?**] wherein the maximum entropy framework can be realized as a special case of energy-based models in the generative setting. Although energy plays a key role the training dynamics of reinforcement learning and optimization of the agent to an optimal policy, it does little to improve the computational framework of learning hierarchical relationships in agent's policy [**?**]. This indicates the requirement for a hierarchical framework capable of mapping relationships between sub-policies in the hierarchy.

3.2.2 Ising Model Learning

Ising models [?] are computational frameworks in physics which consist of interactions between particles in a system or lattice. Particles in the model comprise of finite spin states [?] which governs their relationship to neighboring particles in the model. Ising models obey an energy function [?] which is minimized over time as the model reaches its equilibrium state. Most energy-based models are generalizations of Ising models with their objective function as the loss function [?]. The motivation behind developing Ising models as computational frameworks stems from the literature of physical lattices and spins [?, ?, ?, ?]. [?] presents the first major demonstration of behavior of Ising models near criticality. Assessing the model near criticality is essential as it describes convergence properties corresponding to free energy of the system. This is further extended in [?] wherein the non-equilibrium critical relaxation of the model is assessed under varying dynamics. Estimation of high-dimensional criticality provides suitable scope of the model towards large-scale computational problems. To further investigate the properties of Ising models, [?] highlights the theoretical guarantees of spin-spin correlations under zero field susceptibility. [?] makes use of simplified elliptical transformations to demonstrate criticality of the model. While the Ising model presents suitable convergence properties [?] in the number of particles and time evolution, the model is itself computationally expensive [?]. Manually estimating the state of the model leads to accuracy and efficient inference techniques at the cost of time complexity to assess the state of the system. [?] presents a fast and efficient method for the simulation of Ising models which relies on non-linear relaxation. However, assumptions on spin states and their neighbors renders the state estimates inaccurate and provides a significant need for sampling. [?] addresses the computational problem in Ising structures using maximum-margin parameter estimation and graphical models.

Most works in literature aim to learn the behavior [?] and parameters [?] of the Ising model for particle spin state estimation. [?] presents the learning of Ising model behavior at criticality near phase transitions using generative networks. The approach highlights the requirement of various architectural choices and draws a comparison with shallow counterparts of Restricted Boltzmann Machine (RBM). [?] demonstrates the learning of optimal structure and parameters of the Ising model using interactive screening. Interactive screening is presented as a tractable and optima estimation method which solves the inverse Ising problem universally. [?] presents the application of RBM to reconstruct long-range Ising model configurations in one dimension. Configuration quality in generative scenarios for long-range structures can be assessed by making use of field susceptibility. While learning the Ising model is essential from a structural perspective, the process does not benefit from the computational properties of spin states [?] and their arrangement at thermal equilibrium. One of the successful examples demonstrating the use of Ising models in learning algorithms is that of correlating spiking activities in neurons [?]. The model, consisting of maximum entropy, can be used to reproduce pairwise correlations between

particle spins which is found to be analogous to correlations in the spiking activity of neurons. Moreover, the networks are found to operate closer to critical point and demonstrate behavior similar to spin glasses. [**?**] further highlights the use of Ising models as learning mechanisms by transforming spin structures to Boltzmann Machines. The method is generalized to many-spin interactions and exact mapping which can be improved to fewer spin states by compromising on the number of degrees of freedom. [**?**] and [**?**] present some of the few instances of learning capabilities of Ising models in the case of unstructured environments. Although the methods are suitable for supervised tasks and modular structures, no concrete evidence of spin-based learning in hierarchical structures is found.

4 The Ising Model

The Ising Model [**?**] is an energy-based model which consists of physical particles in a lattice structure. Particles acquire spin states in the lattice which affect the states of neighbouring particles as well as the overall configuration of the system [**?**]. Consider an Ising model consisting of N particles with $N \geq 1$ and $N \neq \infty$. The set of all particles can then be represented as $\{1, 2, ..., N\}$. Mathematically, let σ_i be the spin state of particle i in the system, then σ_i may take a value from the set S which comprises of all the possible spin states. Most Ising models are bivariate which consist of only two spin states $S = \{-1, +1\}$ [**?**]. In physics, the two spin states correspond to clockwise and counter-clockwise spins of particles of a ferromagnetic material [**?**]. However, modern Ising models are often presented to contain more than two spin states [**?**]. Particles in an Ising model interact with each other by means of their spin states. Consider two particles i and j with spin states σ_i and σ_j. The interaction between the two particles $< \sigma_i, \sigma_j >= \mu_{ij}$ can be mathematically expressed as $\mu_{ij} = \sigma_i.\sigma_j$. These interactions are also called spin-spin interactions since they take place between spin states of the two particles [**?**]. Considering an external magnetic field h_j [**?**] on particle j, we can express the total energy E of the Ising model using the Hamiltonian function [**?**] as followed-

$$\begin{aligned} E(\sigma_i, \sigma_j) &= -\sum_{<i,j>} J_{ij}\sigma_i\sigma_j - \sum_j h_j\sigma_j, \qquad \forall i, j \in \{1, 2, ..., N\} \\ &= \sum_{<i,j>} J_{ij}\mu_{ij} - \sum_j h_j\sigma_j \qquad \forall i, j \in \{1, 2, ..., N\} \end{aligned}$$

Here, $< i, j >$ represent the indices of all particles in the Ising model, J_{ij} indicates an interaction parameter which quantifies the probability of interaction between i and j particles. Note that the total energy $E(\sigma_i, \sigma_j)$ is a function of only the spin states of particles indicating that only the internal composition of the lattice is responsible for minimizing the energy of the system [**?**]. Moreover, the partition function Z_β [**?**] corresponding to the energy function is given by

??. Here $\beta = \frac{1}{T}$ represents the inverse of temperature T.

$$Z_\beta = \sum_{\sigma_i, \sigma_j} \exp\left(-\beta E(\sigma_i, \sigma_j)\right) \qquad \forall i, j \in \{1, 2, ..., N\} \tag{1}$$

Given the energy $E(\sigma_i, \sigma_j)$ and partition function Z_β corresponding to the states of the system, we can express the configuration probability $P_\beta(\sigma_i, \sigma_j)$ of the model using the Boltzmann distribution [**?**] with $\beta \geq 0$. $P_\beta(\sigma_i, \sigma_j)$ for a given configuration is presented in **??** as the ratio between the configuration $\exp\left(-\beta E(\sigma_i, \sigma_j)\right)$ and the partition function Z_β denoting the sum over all possible configurations. The negative sign in the exponent denotes assigns a higher probability to low energy states. Similarly, a high β value accentuates the probabilities of the low energy states. Upon carrying out simulated annealing [**?**] of T closer to the low energy states, the system tends to reach thermal equilibrium.

$$P_\beta(\sigma_i, \sigma_j) = \frac{\exp\left(-\beta E(\sigma_i, \sigma_j)\right)}{Z_\beta} \qquad \forall i, j \in \{1, 2, ..., N\} \tag{2}$$

The partition function aids in the assessment of necessary thermodynamical and computational quantities such as internal energy U which is defined as the negative derivative of the partition function Z_β [**?**]. U is mathematically expressed in .

$$U = -\frac{\partial Z}{\partial \beta} = -\frac{1}{Z} \sum_{\sigma_i \sigma_j} \exp\left(-\beta E(\sigma_i, \sigma_j)\right) \qquad \forall i, j \in \{1, 2, ..., N\} \tag{3}$$

Similarly, the free energy per particle F can be obtained as the average of log partition $\log Z_\beta$ in the limit of the number of particles in the model [**?**]. This is mathematically expressed in **??**.

$$F = F(\beta, E, N) = \lim N \to \infty \frac{1}{N} \log Z_\beta(E, N) \tag{4}$$

The limit $N \to \infty$ in **??** is called the thermodynamic limit. In practical scenarios, computation of the free energy F is non-trivial as there is no guarantee of the existence of thermodynamic limit $N \to \infty$ as the lattice system can grow at different rates and in different directions. Such constraints on the growth and size of the system hinder a closed form expression for F and hence pose restrictions on the solution of the lattice. Moreover, the structure and complexity of the model increases with the number dimensions [**?**] which further makes simulations intractable and inaccurate by making use of approximations. As a result of this, most modern computational frameworks [**?**] make use of the energy and partition functions in an indirect manner by making use of structures which can be realized and yield probabilistically approximate estimates [**?**].

5 Ising Networks

5.1 The Spin-based Objective

5.2 Learning Spin Values of Hierarchies

6 Intuition for Spin Values

7 Implementation Details

This section highlights the implementation details for the Ising networks framework when combined with different RL methods. Corresponind to each setting, we first discuss the baseline implementations and their hyperparameters followed by the details of the hierarchical Ising networks.

7.1 Continuous Control

We first look at the continuous control setting wherein our action space $\mathcal{A}$ is continuous. We carry out our experiments in the MuJoCo [**?**] and DM Control Suite [**?**] domains. Each action in these domains is bounded in the $[-1, 1]$ which depicts the force or torque require to control the agent. We divide our experiments into two schemes- (1) state-based learning and (2) learning from pixels. In the first scheme the state input provided to the agent is a feature vector consisting of the agent's position and velocity while in the second scheme, the agent is presented with an image of its current setting. Learning from pixels has proven to be difficult in literature [**?**, **?**] and is an active area of work [**?**, **?**].

7.1.1 State-based Learning

In the case of state-based learning we combine the IS framework with SAC [**?**] which is a state-of-the-art algorithm for off-policy continuous control RL. Various extensions of SAC have proven to be sample-efficient [**?**] which we consider in our experiments for comparison. We compare our IS-SAC implementation with ESAC [**?**], SAC [**?**], TD3 [**?**] which is an improved version of DDPG [**?**], PPO [**?**] and ES [**?**]. Note that we compare our algorithm to on-policy, off-policy and evolutionary methods to assess the capability of IS-SAC w.r.t each type of RL agent. All agents were trained on a total of 21 environments (9 MuJoCo tasks and 12 DM Control tasks) in the OpenAI Gym Suite [**?**] for a total of 5 random seeds.

Baselines: Our experiments make use of author-provided baseline implementations for ES [**?**], [**?**], [**?**] and [**?**]. In the case of PPO [**?**] we use the implementation provided in OpenAI baselines [**?**] as it demonstrates consistent results. Moreover, we omit the Virtual BatchNorm [**?**] component from ES as it is often found to hinder scalability and does not affect the performance. All implementations make use of the same architecture consisting of two hidden

layers of 1024 units with ReLU non-linearity [**?**]. The output layer makes use of a Tanh activation in order to bound the actions in the $[-1, 1]$ range.

IS-SAC:

7.1.2 Learning from Pixels

In the case of learning from pixels we combine our approach with SAC+AE [**?**] and denote it with IS-SAC+AE for our experiments. We compare between IS-SAC+AE and SAC+AE on the 100k benchmark [**?**] for a total of 5 random seeds on 12 DM control suit environments. The 100k benchmark evaluates the algorithm for only 100k timesteps and assesses data-efficiency in pixel-based learning.

Baseline: Our baseline implementation of SAC+AE makes use of the author-provided implementation [**?**] with tuned hyperparameter.

IS-SAC+AE:

7.2 Discrete Control

Our experiments in discrete control make use of two learning domains- (1) Atari 2600 suite in OpenAI Gym [**?**] and (2) Mazelab [**?**]. We make use of the Atari 2600 suite to present the large-scale generalization of hierarchies utilizing spin-spin alignments in the case of long-horizon environments. On the other hand, we use Mazelab to demonstrate their success in sparse environments consisting of limited optimal behavior.

7.2.1 Atari 2600 Games

In the case of Atari 2600 tasks, we train our agents on 28 games for a total of 5 random seeds. All agents were trained for 5 million steps with learning starting after 10k steps. We combine our framework with Rainbow [**?**] which is a state-of-the-art model-free off-policy RL algorithm utilizing the original DQN [**?**] implementation. and denote it with IS-Rainbow for our experiments. IS-Rainbow is compared to Rainbow [**?**], A2C [**?**], PPO [**?**] and ACKTR [**?**]. Additionally, we present the scores of an average human subject and a random agent for comparison and reference.

Baselines: Our implementation of Rainbow is based on [**?**]. However, we limit the replay memory to the most recent 10^6 steps for a fair comparison with other methods. In the case of A2C, PPO and ACKTR, we make use of the implementations provided in [**?**] as these are found to be stable with respect to random seeds in comparison to OpenAI baseline implementations [**?**]. Agents are tuned to their optimal hyperparameter values as per author-provided details.

IS-Rainbow:

7.2.2 Mazelab

Baselines:

IS-hDQN:

7.3 Multi-Agent Learning

We select StarCraft II scenarios particularly for three reasons. Firstly, micromanagement scenarios consist of a larger number of agents with different action spaces. This requires a greater deal of coordination. Secondly, micromanagement scenarios consist of partial observability wherein agents are restricted from responding to enemy fire and attacking enemies when they are in range. This allows agents to explore the environment effectively and find an optimal strategy purely based on collaboration rather than built-in game utilities. Lastly, micromanagement scenarios in StarCraft II consist of multiple opponents which introduce a greater degree of surprise within consecutive states. Irrespective of the time evolution of an episode, environment dynamics of each scenario change rapidly as the agents need to respond to enemy's behavior. Agents were trained for a total of 5 random seeds consisting of 2 million steps in each environment. A total of 32 validation episodes carried out at every 10,000 step intervals were interleaved during agent's interactions. All baselines implementation consist of a Recurrent Neural Network (RNN) agent having memory consisting of past states and actions. We use an ϵ-greedy exploration scheme wherein ϵ is annealed from 1 to 0.01 during the initial stages of training.

Baselines: Our baselines consist of state-of-the-art MARL methods for cooperative control. We compare our framework to EMIX [**?**], QMIX [**?**], [**?**], [**?**] and [**?**]. Additionally, we use a model-free implementation of SMiRL [**?**] combined with QMIX to assess surprise-robust behavior of our framework. We denote this implementation as SMiRL-QMIX for our experiments. All implementations were adopted from the PyMARL [**?**] framework with hyperparameters tuned to their optimal parameters using author-provided notes. In the case of EMIX, we tune the the temperature parameter β to 0.03 as it provides us with consistent results across different seeds. We make use of the generalized implementation of SMiRL [**?**] in SMiRL-QMIX wherein we utilize the standard deviations of states across batches in order to reduce variance in rewards. Moreover, the temperature parameter in SMiRL-QMIX is tuned to 0.1 in steps of 0.02.

IS-EMIX:

8 Propositions

Proposition 1. *The Ising system can consist of M^N possible spin states of levels in the hierarchy.*

Proof. Consider a level in the hierarchy $h_i \in H, \forall i \in \{1, 2, ..., N\}$. Clearly, the level h_i can acquire a total of M possible spin states. Applying this to all N levels, we get $M \times M...Ntimes$ as $|H| = N$, which is equal to M^N. □

Proposition 2. *In the case of infinite levels of a hierarchy $|H| \to \infty$ representing a continuous chain of particles, the Ising model does not converge to a thermal equilibrium and a state of stable alignment between levels is not achieved.*

Proof. Consider the Hamiltonian function $H(\sigma) = -\sum_{<i,j>} \sigma_i \sigma_j - \sum_i \sigma_i$,

$$H(\sigma) = -\sum_{i,j\in\{1,..,\infty\}} \sigma_i\sigma_j - \sum_{i\in\{1,..,\infty\}} \sigma_i$$

$$H(\sigma) = -(\sum_{i,j\in\{1,..,\infty\}} \sigma_i\sigma_j + \sum_{i\in\{1,..,\infty\}} \sigma_i) \tag{5}$$

Considering $i = 1$ in **??**, we get,

$$H(\sigma) = -(\sum_{i=1,j\in\{1,..,\infty\}} \sigma_i\sigma_j + \sigma_1) \tag{6}$$

On comparing **??** and **??** we observe the following

$$-(\sum_{i,j\in\{1,..,\infty\}} \sigma_i\sigma_j + \sum_{i\in\{1,..,\infty\}} \sigma_i) \leq -(\sum_{i=1,j\in\{1,..,\infty\}} \sigma_i\sigma_j + \sigma_1)$$

$$\sum_{i,j\in\{1,..,\infty\}} \sigma_i\sigma_j + \sum_{i\in\{1,..,\infty\}} \sigma_i \geq \sum_{i=1,j\in\{1,..,\infty\}} \sigma_i\sigma_j + \sigma_1$$

Thus, adding a new level in the hierarchy H increases the energy of the system and the model does not converge to a fixed point equilibrium when $|H| \to \infty$. □

Proposition 3. *For a given agent with N levels in the hierarchy $H = \{h_1, h_2, .., h_N\}$ and a policy $\pi(a_t|s_t)$, define a new policy $\tilde{\pi}(a_t|s_t)$ such that $\tilde{\pi}(a_t|s_t) \geq \pi(a_t|s_t)$ with $\pi_i(a_t|s_t)$ and $\pi_j(a_t|s_t)$ as policies for levels i and j in the hierarchy.*

Proof. The proof is mostly based on the policy improvement in SQL[**?**]. We consider the intrinsic motivation case among levels of hierarchy H. By definition of $Q^\pi(s_t, a_t)$,

$$Q^\pi(s_t, a_t) = \mathbf{E}_{s_1}[r_0 + \beta(\sum_{i,j} \sigma_0^{\pi_i}\sigma_0^{\pi_j}) + \gamma\mathbf{E}_{a_i\sim\pi}[Q^\pi(s_1, a_1)]]$$

$$\leq \mathbf{E}_{s_1}[r_0 + \beta(\sum_{i,j} \sigma_0^{\pi_i}\sigma_0^{\pi_j})^2 + \gamma\mathbf{E}_{a_i\sim\tilde{\pi}}[Q^\pi(s_1, a_1)]]$$

$$= \mathbf{E}_{s_1}[r_0 + \beta(\sum_{i,j} \sigma_0^{\pi_i}\sigma_0^{\pi_j})^2 + \gamma(r_1 + \beta(\sum_{i,j} \sigma_1^{\pi_i}\sigma_1^{\pi_j})^2)] + \gamma^2\mathbf{E}_{s_2}[\mathbf{E}_{a_2\sim\pi}[Q^\pi(s_2, a_2)]]$$

$$\leq \mathbf{E}_{s_1}[r_0 + \beta(\sum_{i,j} \sigma_0^{\pi_i}\sigma_0^{\pi_j})^2 + \gamma(r_1 + \beta(\sum_{i,j} \sigma_1^{\pi_i}\sigma_1^{\pi_j})^2)] + \gamma^2\mathbf{E}_{s_2}[\mathbf{E}_{a_2\sim\pi}[Q^{\tilde{\pi}}(s_2, a_2)]]$$

$$\vdots$$

$$\leq \mathbf{E}_{\tau\sim\tilde{\pi}}[r_0 + \beta(\sum_{i,j} \sigma_0^{\pi_i}\sigma_0^{\pi_j})^2 + \sum_{t=1}^{\infty} \gamma^t(r_t + \beta(\sum_{i,j} \sigma_t^{\pi_i}\sigma_t^{\pi_j})^2)]$$

$$= Q^{\tilde{\pi}}(s_t, a_t)$$

This indicates that $\tilde{\pi}(a_t|s_t) \geq \pi(a_t|s_t)$ which completes the proof. □

Proposition 4. *Gradient of the Ising spin σ^{π_i} for any two levels h_i and h_j in the hierarchy H simplifies the policy gradient to*

$$\nabla_\theta J(\theta) = \underset{s_t,a_t}{\boldsymbol{E}}[\nabla_\theta(\log \pi_i^\theta(a_t|s_t)(Q^\pi(s_t,a_t) + b^\theta(s_t)) + \underset{a_t\sim\pi}{\boldsymbol{E}}[\sigma^{\pi_i}\sigma^{\pi_j}])]$$

Proof. The proof begins with the policy gradient consisting of the spin-spin interactions,

$$\nabla_\theta J(\theta) = \underset{s_t,a_t}{\mathbf{E}}[\nabla_\theta(\log \pi_i^\theta(a_t|s_t)(Q^\pi(s_t,a_t) + b^\theta(s_t)))] + \mathbf{E}[\sum_{i,j} \sigma^{\pi_i}\sigma^{\pi_j}]$$

Evaluating the gradient of the spin-spin interaction term, we get,

□