Quantum Machine Intelligence manuscript No.
(will be inserted by the editor)

Supplementary Material: K-spin Hamiltonian for quantum-resolvable Markov decision processes

Eric B. Jones · Peter Graf · Eliot Kapit · Wesley Jones

1 Introduction

Note that all equations and figures referenced in this document refer to those in the main mauscript.

2 Superextensivity of the K-spin Hamiltonian

Consider the k^{th} order of Eq. 10 when the policy variables are aligned $\pi^{on} = (1, \ldots, 1)$ (Note that when $\pi = (0, \ldots, 0)$ the energy at all orders except 0^{th} vanishes):

$$\begin{aligned} H_k[\pi^{on}] &= -\sum_{\mu_1 \cdots \mu_k} J_{\mu_1 \cdots \mu_k} \\ &= -\gamma^k \sum_{s_0, a_0} \cdots \sum_{s_{k+1}} P^{a_0}_{s_0 s_1} \cdots P^{a_k}_{s_k s_{k+1}} R^{a_k}_{s_k s_{k+1}}. \end{aligned}$$

If we assume for simplicity that all reward functions, though potentially different from one another, are of order $R^a_{ss'} \approx 1$ and carry out the resulting sums in the order $s_{k+1}, a_k, s_k, a_{k-1}, \ldots, s_0$ while using the conditional probability distribution normalization condition, we find

$$\begin{aligned} H_k[\pi^{on}] &\approx -\gamma^k \sum_{s_0} \sum_{a_0, s_1} P^{a_0}_{s_0 s_1} \cdots \sum_{a_{k-1}, s_k} P^{a_{k-1}}_{s_{k-1} s_k} \sum_{a_k, s_{k+1}} P^{a_k}_{s_k s_{k+1}} \\ &\approx -\gamma^k \sum_{s_0} \sum_{a_0, s_1} P^{a_0}_{s_0 s_1} \cdots \sum_{a_{k-1}, s_k} P^{a_{k-1}}_{s_{k-1} s_k} |A| \\ &\approx \cdots \\ &\approx -(\gamma |A|)^k |S \times A|, \end{aligned}$$

where $|A|$ is the magnitude of the action space and $|S \times A|$ is the magnitude of the state-action space. Therefore, under these assumptions $H[\pi]$ displays superextensivity, the severity of which is given by

$$\frac{H_{k+1}[\pi^{on}]}{H_k[\pi^{on}]} \approx \gamma |A|.$$

It is infeasible for this to be rectified by setting $\gamma \leq 1/|A|$ because in realistic MDPs one requires $\gamma \approx 0.9$ or above to find longer-term rewards while $A \geq 2$ is typically taken to be an integer. However, in contrast to the requirements of statistical mechanics applications where one might want to include an additional factor of $1/|A|$ (or potentially $1/|S \times A|$), this superextensivity of the K-spin Hamiltonian is ideal for the determination of optimal policies since it is precisely what allows determination of optimal policies based on long-term rewards. In addition, including an additional $1/|A|$ correction factor at each subsequent order in the Hamiltonian appears to be inconsistent with the definition of the Hamiltonian in terms of the Q functional derived in Eq. 6.

As a result of its superextensivity, the truncation of Eq. 10 is potentially not well controlled in many instances. Developing heuristics for understanding how

Eric B. Jones
National Renewable Energy Laboratory, Golden, CO 80401, USA
Department of Physics, Colorado School of Mines, Golden, CO 80401, USA
E-mail: Eric.Jones@nrel.gov, ebjones@mymail.mines.edu

Peter Graf
National Renewable Energy Laboratory, Golden, CO 80401, USA

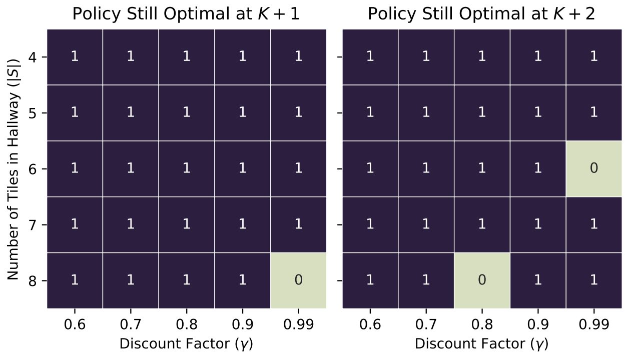
Eliot Kapit
Department of Physics, Colorado School of Mines, Golden, CO 80401, USA

Wesley Jones
National Renewable Energy Laboratory, Golden, CO 80401, USA

best to truncate Eq. 10 and to interpret the resulting ground state policy will be an important direction for future work. We demonstrate a prototypical heuristic here. In the caption to Fig. 3, we show that the minimal truncation order scales roughly as $K_{heur} \sim |S|/2$ based on results for some small system sizes. One could then imagine using Eq. 10 to find the ground state policy for much larger environment sizes with the similar transition and reward structures using the established small-size truncation heuristic.

In addition, it is important to note that there exists an in-built "backstop" to Eq. 10 for fixed state-action space size $|S \times A|$, namely, that including terms at $K > |S \times A|$ makes little sense from the perspective of either spin Hamiltonian physics or of the Markov decision process since the term at $K = |S \times A|$ couples all state-action policy variables together. This backstop creates a well-defined and finite regime $K_{heur} \leq K \leq |S \times A|$ wherein one can study the effects of Hamiltonian superextensivity and methods for controlling the truncation.

Finally, in order to see that the series truncation for the hallway environment in the main text is reasonably well-controlled for nearly all environmental parameters $(|S|, \gamma)$, consider the following plot.



Policy Still Optimal at $K+1$

|S|	0.6	0.7	0.8	0.9	0.99
4	1	1	1	1	1
5	1	1	1	1	1
6	1	1	1	1	1
7	1	1	1	1	1
8	1	1	1	1	0

Policy Still Optimal at $K+2$

|S|	0.6	0.7	0.8	0.9	0.99
4	1	1	1	1	1
5	1	1	1	1	1
6	1	1	1	1	0
7	1	1	1	1	1
8	1	1	0	1	1

Heat map squares annotated to 1 (purple) indicate that the optimal ground state policy at K remains so at $K + n$, where here we show results for $n = 1, 2$. Annotation to 0 (cream) means that the optimal policy at K is no longer the unique ground state policy at $K+n$. These results were obtained by simulated annealing since the physical qubit counts grew too large to run on the 2000Q quantum processor at these orders of K. At $K+1$, the only environment setting that moves away from the optimal policy is $(|S| = 8, \gamma = 0.99)$. This would appear to be a reasonable result since superextensivity of the Hamiltonian is more severe at both large $|S|$ and large γ. However, at $K + 2$, the ground state of $(|S| = 8, \gamma = 0.99)$ returns to the optimal policy. It may be the case that such a ground state fluctuation is caused by non-inclusion of terms from $K+2$ that resolve dynamical frustration at $K+1$. At $K+2$, two points in the environmental parameter space $(|S| = 6, \gamma = 0.99)$ and $(|S| = 8, \gamma = 0.8)$ deviate from the optimal policy. Given the previous discussion surrounding the fluctuation at $(|S| = 8, \gamma = 0.99)$, we suspect that these isolated fluctuations in the ground state policy may be rectified at $K+3$, although difficulty in converging simulated annealing results as a function of total annealing time due to the large numbers of variables at this order prevents us from showing this rigorously. An interesting future study could use high-performance computing resources to study the convergence in the optimal policy further past the minimal truncation order (K).

Nevertheless, even if imperfect, the convergent behavior of the ground state policy at and past the minimal truncation order (K) is patently distinct from the $< K$ regime where the ground state policy is not only not optimal, but also not converged for any environmental parameter settings.

3 Hamiltonian Construction Without an Initial Model

In our formulation, in order to calculate the couplings one must have on-hand transition and reward functions. When an agent does not have a model of its environment, i.e. the sets P and R are unknown, one typically resorts to reinforcement learning algorithms in order to determine an optimal policy. Q-Learning is one such example of a reinforcement learning algorithm. One performs Q-Learning by initializing the agent many times in a simulated environment and updating and tabulating the Q function on the fly directly by its immediate experience with discovered transition and reward functions. In a similar fashion, one could imagine initializing the agent many times in a simulated environment so that it could discover and tabulate transition and reward functions to use in an eventual (perhaps coarse-grained) policy determination via minimizing Eq. 10. However, it is unclear that such an approach would match the efficiency of dedicated reinforcement learning algorithms such as Q-Learning. We also note that both model-free and model-based formulations of reinforcement learning exist, and so it is likely that some of the techniques from the model-based branch of the field might inform construction and minimization of Eq. 10 where transition and reward functions are initially unknown (Janner et al. 2019).

4 Variational Definitions and Proofs

Let η_μ be an arbitrary member of the policy function space and ϵ a small number close to zero. As the field π is varied $\pi \to \pi + \epsilon\eta$, the first variation of the Hamiltonian may be equivalently defined in two ways as

$$\begin{aligned}\delta H[\pi;\eta] &\equiv \sum_{\bar{\mu}} \eta_{\bar{\mu}} \frac{\delta H[\pi]}{\delta \pi_{\bar{\mu}}} \\ &\equiv \frac{d}{d\epsilon} H[\pi + \epsilon\eta]|_{\epsilon=0}\end{aligned}$$

The quantity $\frac{\delta H[\pi]}{\delta \pi_{\bar{\mu}}}$ in the first line is typically called the "variational derivative" or "functional derivative" and is defined at a particular point $\bar{\mu}$. If H is stationary at the field configuration π then $\delta H[\pi;\eta] = 0$, and the variational derivative must also vanish at every point $\left(\frac{\delta H[\pi]}{\delta \pi_{\bar{\mu}}} = 0\right)$ because the field variation was arbitrary. This result is known as the fundamental lemma of the calculus of variations.

If a functional H is minimized (maximized) at a particular field configuration π, then the corresponding variational conditions are

$$\begin{aligned}\delta H[\pi;\eta] &= 0 \\ \delta^2 H[\pi;\eta] &> 0 \\ &\qquad (< 0).\end{aligned}$$

We first show that the definition of the Hamiltonian is consistent with the aim of an MDP agent, which is to maximize $Q_\mu[\pi]\ \forall\mu$. Maximization of each $Q_\mu[\pi]$ corresponds to the variational conditions $\delta Q_\mu[\pi;\eta] = 0$ and $\delta^2 Q_\mu[\pi;\eta] < 0$. Then, employing the linearity of the functional derivative,

$$\begin{aligned}\delta H[\pi;\eta] &= -\sum_\mu \delta Q_\mu[\pi;\eta] = \sum_\mu 0 = 0 \\ \delta^2 H[\pi;\eta] &= -\sum_\mu \delta^2 Q_\mu[\pi;\eta] = -\left[\sum_\mu \# < 0\right] > 0,\end{aligned}$$

showing that the Hamiltonian will be minimized as a result. On the other hand, for our formulation to be useful, minimization of the Hamiltonian should (at least approximately) correspond to maximizing each Q function individually. To this end consider,

$$\begin{aligned}0 &= \delta H[\pi;\eta] \\ &= \sum_{\bar{\mu}} \eta_{\bar{\mu}} \frac{\delta H[\pi]}{\delta \pi_{\bar{\mu}}} \\ &= -\sum_\mu \sum_{\bar{\mu}} \eta_{\bar{\mu}} \frac{\delta Q_\mu[\pi]}{\delta \pi_{\bar{\mu}}}.\end{aligned}$$

It is difficult to argue in the general case that each μ-indexed summand identically must be zero in the equation above. We note however, that the situation here vis-à-vis energetic tradeoffs between different terms in the Hamiltonian is of a somewhat different nature than energetic tradeoffs between the one- and two-body terms in, say, the 2D Ising model that can variously lead to paramagnetic, ferromagnetic, and antiferromagnetic states. If it were the case, then we could factor a general spin Hamiltonian into sums of Q functions and recursively factor those in terms of spin variables. However, if *that* were the case, then we could solve a general spin Hamiltonian via the dynamic programming method typically used for solving MDPs (a polynomial-time algorithm), presenting a serious challenge to the widely-believed complexity-theoretic statement that $P \neq NP$.

Rather, we point out that each $Q_\mu[\pi]$ may be, in and of itself, considered a spin Hamiltonian, which includes k-body policy interactions at all orders of k (see Eq. 6). The loose state-action index $\mu = (s, a)$ on each $Q_\mu[\pi]$ refers only to loose indices on the initial transition probability in each if the k-body couplings and not to a loose index on a policy variable. Hence, each Q function differs from the others only in its local transition and reward environment out to K^{th} order (upon truncation) and not strictly on which many-body interactions it incorporates or neglects. One could therefore imagine maximizing each of the $|S \times A|$ copies of Eq. 6 separately using simulated or quantum annealing and then post-selecting which policy configurations are consistent across the entire state-action space in order to stitch together an optimal solution. The intuition behind the Hamiltonian in Eq. 10 is that this consistency requirement is definitionally built-in, requiring a single minimization routine.

We argue that each $Q_\mu[\pi]$ must individually be maximized for a class of environments, of which our hallway example and typical GridWorld environments are members, where there are a few well-isolated terminal states (e.g. dirt piles) at the extremities of the environment separated by a relatively larger number of "bulk" states.

In the interior of the environment there is a homogeneity in the transition and reward functions. Because of this, the energy landscape of each $Q_\mu[\pi]$ looks the same in the interior. That is,

$$\begin{aligned} H[\pi] &= -\sum_{\mu} Q_\mu[\pi] \\ &= -\sum_{\mu\in bulk} Q_\mu[\pi] - \sum_{\mu\in boundary} Q_\mu[\pi] \\ &= -N_{bulk}Q_{bulk}[\pi] - \sum_{\mu\in boundary} Q_\mu[\pi]. \end{aligned}$$

In the second line we separate the sum into two sums, one over Q functions that only couple policy variables within the bulk and one that couples bulk policy variables into terminal states with transition and reward functions different from the bulk. We note that there is actually a third sum as well, which couples policy variables from the terminal tiles to bulk policy variables, but these terms can be trivially set to zero since by virtue of being a terminal state all policy variables so indexed must vanish (see for example the terminal state transition and reward structure in Sec. 2). In the third line we use the homogeneity of transition and reward functions in the bulk to replace the sum with a multiplicity factor (N_{bulk}).

We now consider two limiting cases. First, if $N_{bulk} >> N_{boundary}$ and all reward functions are of similar magnitude, then the contribution to the Hamiltonian from the bulk Q functions outweighs those of the boundary functions

$$0 = \delta H[\pi;\eta] \approx -N_{bulk}\delta Q_{bulk}[\pi;\eta].$$

Thus each bulk Q function is approximately extremized. Further, due to the opposite sign of H and Q_{bulk}, minimization of the former leads to maximization of the latter.

In the opposite limit, we may still have that $N_{bulk} > N_{boundary}$, but here we assume that the reward functions leading into the boundaries are of much larger magnitude than the reward functions between bulk states $|R^a_{bulk\to boundary}| >> |R^a_{bulk\to bulk}|$. Then, the boundary summation dominates the energetics:

$$H[\pi] \approx -\sum_{\mu\in boundary} Q_\mu[\pi].$$

Since each boundary function is well-separated from the others, we assume that it acts on a disjoint subset of the bulk policy variables $\tilde{\pi}^{(\mu)}$. Then, we can perform arbitrary variations only among each of the disjoint bulk policy subsets $\tilde{\pi}^{(\nu)} \to \tilde{\pi}^{(\nu)} + \epsilon\tilde{\eta}^{(\nu)}$. The corresponding variational derivative of H with respect to the disjoint policy subset variable at a point $\bar{\nu}$ results in

$$\begin{aligned} 0 = \frac{\delta H[\pi]}{\delta \tilde{\pi}^{(\nu)}_{\bar{\nu}}} &= -\sum_{\mu\in boundary} \frac{\delta Q_\mu[\tilde{\pi}^{(\mu)}]}{\delta \tilde{\pi}^{(\nu)}_{\bar{\nu}}} \\ &= -\sum_{\mu\in boundary} \delta_{(\mu)(\nu)} \frac{\delta Q_\mu[\tilde{\pi}^{(\mu)}]}{\delta \tilde{\pi}^{(\nu)}_{\bar{\nu}}} \\ &= -\frac{\delta Q_\nu[\tilde{\pi}^{(\nu)}]}{\delta \tilde{\pi}^{(\nu)}_{\bar{\nu}}}. \end{aligned}$$

Therefore, since each boundary Q function only depends on a disjoint subset of policy variables, an arbitrary variation within the disjoint policy variable function space only affects the Q function that depends on that disjoint function space (line 2 above). Because variation of the disjoint policy field also constitutes an arbitrary variation of the full Hamiltonian, minimization of the Hamiltonian corresponds to maximization of each boundary Q function independently (line 3).

We note that the numerical calculations in the main text (e.g. Fig. 3) do not generally abide by the assumptions of the two limiting cases made here, while still yielding ground state policies that are Bellman optimal. This is encouraging since it indicates that the equivalence between Hamiltonian minimization and Bellman optimality may be a more general (if not completely general) characteristic of Markov decision processes.

5 Definition of Long-Range, k-Local Hamiltonian

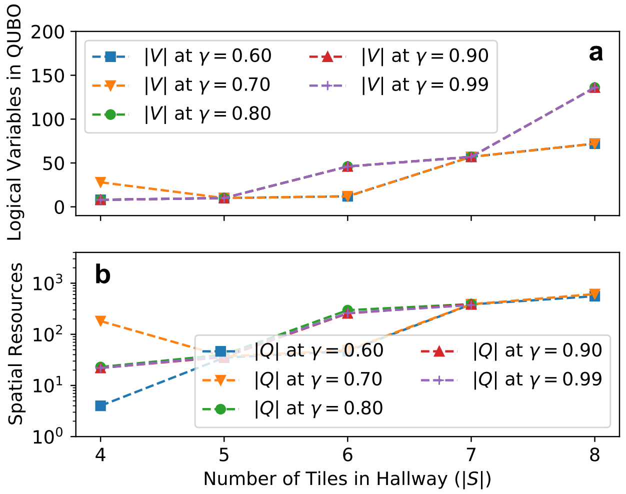
At each order k, Eq. 10 is precisely a k-local Hamiltonian: a sum of terms each of which being a Hermitian operator acting on at most k qubits (i.e. policy variables) (see for example Definition 2 in (Kempe et al. 2006)). If one is able to truncate Eq. 10 to finite order K, which we have shown is a viable approach for finding optimal policies in Fig. 3, then Eq. 10 formally becomes a K-local Hamiltonian for which $K < |S \times A|$, the total number of qubits that would be involved in the computation were order-reduction not performed. For example, even at $\gamma = 0.99$ and $|S| = 8$, $K = 4$ while $|S \times A| = 16$. It is true that such interactions between qubits will in general be long-range. However, nothing in the definition of a spin Hamiltonian prohibits the inclusion of long-range interactions (Richerme et al. 2014).

6 K-induced Ground State Transitions

The $K = 2 \to 3$ ground state transition in Fig. 2 of the main text is definitely not a thermally-induced phase transition as is observed in classical statistical mechanics, since we assume that the ground state of our K-spin Hamiltonian (not to be confused with the full adiabatic transverse field Ising Hamiltonian of quantum annealing) evolves at very low, constant temperature as a function of its parameters. Whether the ground state transition rigorously constitutes a phase transition as a function of the "parameter" K is an interesting question. When the truncation order of Eq. 10 is incremented $K \to K+1$, a number of new couplings $\gamma^{K+1} J_{\mu_1 \ldots \mu_{K+1}}$ are changed from zero to non-zero values. All couplings also change when γ is varied. The transition of a ground state as a function of Hamiltonian parameters is highly suggestive of a quantum phase transition. However, given that the degrees of freedom in Eq. 10 are purely classical (the mapping to qubit operators being only necessary for quantum optimization heuristics), it appears untenable to argue that such a transition is mediated by quantum fluctuations.

7 Parameter Selection in Fig. 4

We note that the scaling of spatial resources in Fig. 4 is relatively insensitive to the parameter (γ) setting in the sense described here. The purpose of Fig. 4 is to show that the asymptotic scaling of logical variables ($|V|$) as a function of problem size likely follows the polynomial $O(|S \times A|K)$ result given by (Fix et al. 2011) and that the number of physical qubits ($|Q|$) also follows a polynomial scaling given that only small-size ($|S| \leq 8$) problem instances of the hallway can be treated on the D-Wave 2000Q annealer. As can be seen in panel **a** of the plot below, the scaling in $|V|$ is nearly identical for $\gamma = 0.6$ and 0.7 (For $|S| \geq 5$) and again for $\gamma = 0.8 - 0.99$. Meanwhile, in panel **b** the large $|S|$ behavior of $|Q|$ is again equivalent for $\gamma = 0.6$ and 0.7 and again for $\gamma = 0.8 - 0.99$. Therefore, in terms of spatial resource scaling (and temporal complexity for that matter), $\gamma = 0.6 - 0.7$ constitutes one approximate equivalence class and $\gamma = 0.8 - 0.99$ another. The selection of equivalence class representative lines $\gamma = 0.6$ and $\gamma = 0.9$ in Fig. 4 is simply for visual clarity.



Acknowledgements This work was authored in part by the National Renewable Energy Laboratory (NREL), operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. This work was supported by the Laboratory Directed Research and Development (LDRD) Program at NREL. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes. This research used Ising, Los Alamos National Laboratory's D-Wave quantum annealer. Ising is supported by NNSA's Advanced Simulation and Computing program. The authors would like to thank Scott Pakin and Denny Dahl. This material is based in-part upon work supported by the National Science Foundation under Grant No. PHY-1653820.

8 Declarations

Funding

This work was supported by the Laboratory Directed Research and Development (LDRD) Program at NREL. This material is based in-part upon work supported by the National Science Foundation under Grant No. PHY-1653820.

Conflict of Interests

The Authors declare no competing Financial or Non-Financial Interests.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Code Availability

The code that supports the findings of this study are available from the corresponding author upon reasonable request.

References

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. In *Advances in Neural Information Processing Systems*, pages 12519–12530, 2019.

Julia Kempe, Alexei Kitaev, and Oded Regev. The complexity of the local hamiltonian problem. *SIAM Journal on Computing*, 35(5):1070–1097, 2006.

Philip Richerme, Zhe-Xuan Gong, Aaron Lee, Crystal Senko, Jacob Smith, Michael Foss-Feig, Spyridon Michalakis, Alexey V Gorshkov, and Christopher Monroe. Non-local propagation of correlations in long-range interacting quantum systems. *arXiv preprint arXiv:1401.5088*, 2014.