

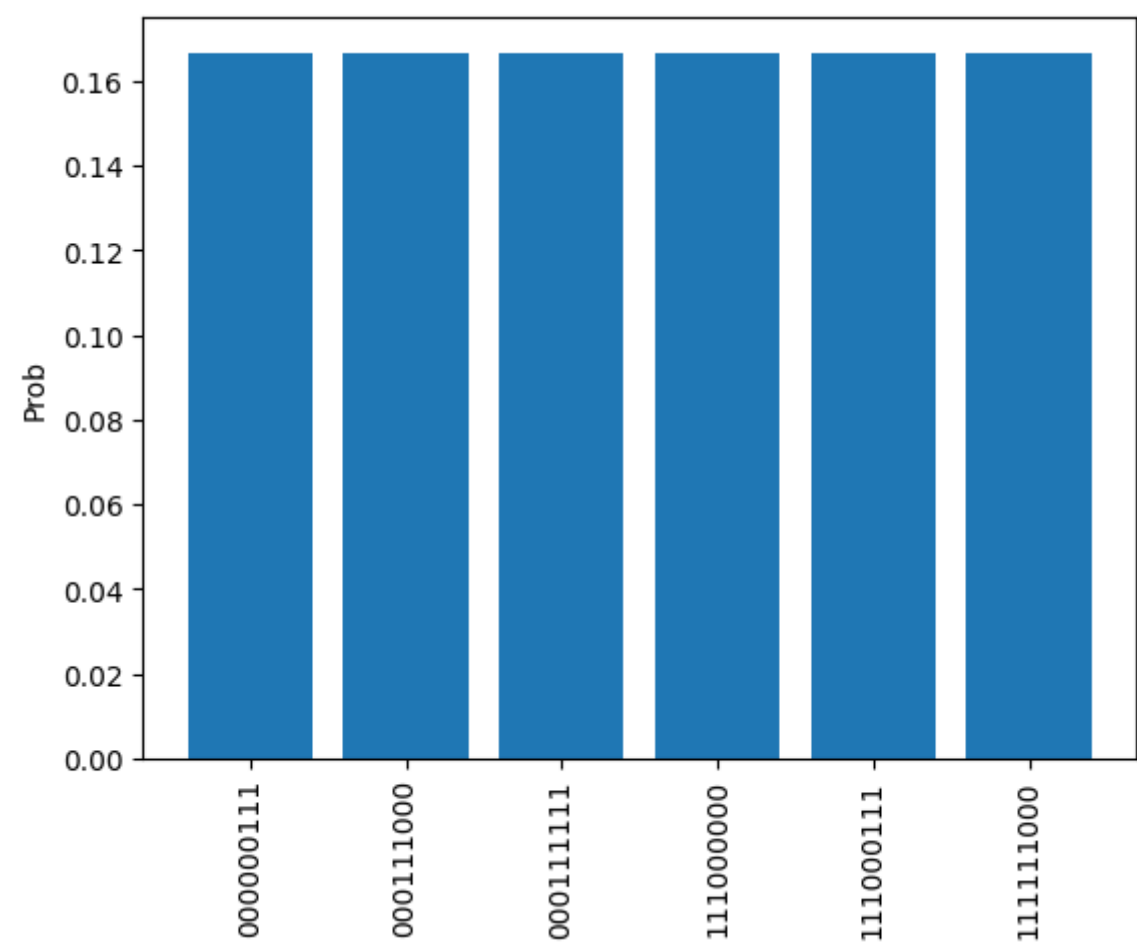
Note on MIXER CONSTRUCTION & EFFECTS

So here I will sum up the effects of mixers for the purpose of sampling (and training) from ising-type models. We have already seen the effect of using 'vanilla' mixers for sampling from

MIXER CONSTRUCTION from DATA

Observations

For the sake of simplicity, I will base my reasoning about the case of BAS (just bars for now) dataset. The dataset is given by \mathcal{D}^{b3} :

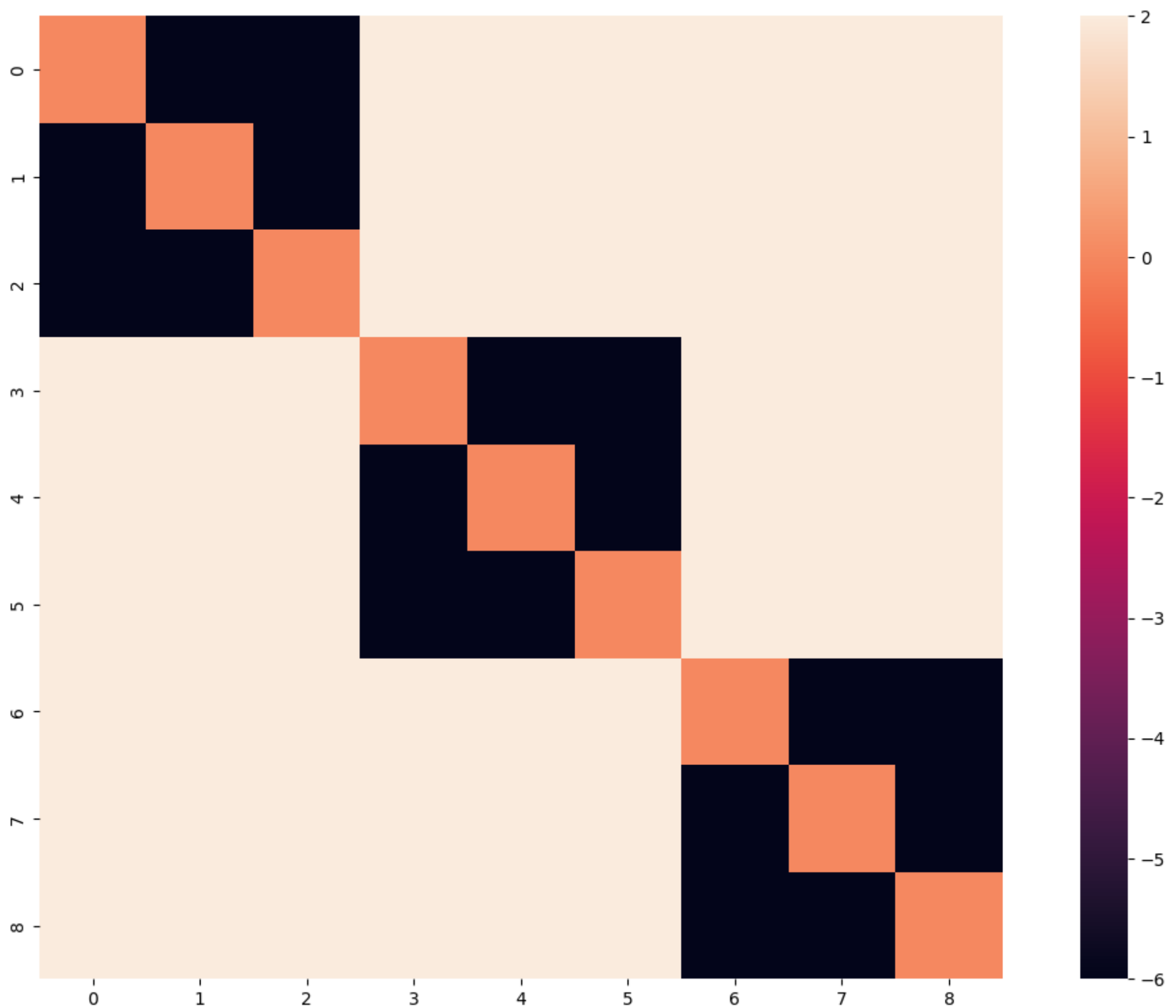


It should be noted that such distributions have equal probability distribution over all elements of the set, though its a trivial fact it might be useful for some deductions later.

In training 'Boltzmann Machines' we are aiming to find a suitable binary interactions and biases for our model ; $\mathcal{M}(J, h) = \mathcal{M}(J_{ij}, h_i)$ so that the corresponding boltzman distribution corresponds to the data set.

However for datasets of small size we can use methods like 'Hebbian Learning' to construct a model in similar fashion, which is also something we used to get a baseline model for comparing the effectivity of the sampling routines.

The model M_{hebb}^{b3} constructed is depicted here:



which when expanded mathematically have the structure

$$M_{hebb}^{b3} = (-6) ((Z_0Z_1 + Z_1Z_2 + Z_0Z_2) + (Z_3Z_4 + Z_4Z_5 + Z_5Z_3) + (Z_6Z_7 + Z_7Z_8 + Z_8Z_6)) + 2 (\text{..rest of the interactions..}) + 0 (\text{biases})$$

We can draw a close analogy with *Stabilizer Hamiltonians*, i.e hamiltonians which have as their ground states the positive eigenspace of the elements in *Stabilizer Group*.

Its easy to recognise that the pauli terms with -6 as their coefficient are all part of the stabilizer group of the well known *3-qubit repetition QEC* concatenated 3 times.

Ideally the stabilizer hamiltonian should be constructed as the negative sum of all the generators of the stabilizer group i.e

$\mathcal{S}^{b3} = \langle Z_0Z_1, Z_1Z_2, Z_3Z_4, Z_4Z_5, Z_6Z_7, Z_7Z_8 \rangle$, thus the hamiltonian is

$$M_{stab}^{b3} = -(Z_0Z_1 + Z_1Z_2 + Z_3Z_4 + Z_4Z_5 + Z_6Z_7 + Z_7Z_8)$$

We can verify that both M_{hebb}^{b3} , M_{stab}^{b3} have the same ground states (except states with all zeros or all ones), however they have different i energy level distribution of excited levels. The additional terms (with coefficient +2) in M_{hebb}^{b3} have the effect to introduce larger energy gap between the ground-states and non ground states of the system, which in turn makes it easier for sampling algorithms to distinguish between them.

Coming to better mixer constructions, ideally we would like to have mixer hamiltonians T which only allow for transitions between the elements of the dataset (or more generally elements that have a significantly higher probability than the rest).

Though under general circumstances it is hard to create such operators, we can use some ideas from *stabilizer codes* for cases where the dataset can be related to ground states of a stabiliser hamiltonian. In QEC any non-trivial operator that maps a stabilized state to another stabilized state is simply a 'logical operator', it has the property that it commutes with the elements in the stabilizer group.

Thus for the elements of \mathcal{D}^{b3} , the ideal mixer would be the one generated by $\mathcal{L}_X^{b3} = \langle X_0X_1X_2, X_3X_4X_5, X_6X_7X_8 \rangle$, which are simply the 'logical X' of the *3-qubit repetition QEC*. As expected the operators in \mathcal{L}_X^{b3} commute with those in \mathcal{S}^{b3} and thus with the corresponding hamiltonian M_{stab}^{b3} . However, note that the mixer terms doesn't commute with M_{hebb}^{b3} , because of the *non-stabilizer* terms in the hamiltonian.

It should be noted that in generic QAOA type circuits are built with interspersed layers of Problem (P) and Mixer Hamiltonian (T) that do not mutually commute with each other, $[T, P] \neq 0$.

The main purpose of having non-commuting layers is to facilitate transitions between states with different energies, thus allowing for the algorithm to get an overview of the global energy landscape of the system which is necessary for optimisation / sampling tasks.

Note that in case $[T, P] = 0$, the corresponding time-evolution unitary would fail to facilitate transitions between states of different energy levels. Usually such choice of mixer hamiltonians would be disadvantageous, except if all the states of interest have the same energy, which is the case here with hamiltonians like M_{stab}^{b3} , M_{hebb}^{b3} . Thus constructing T out of the elements in \mathcal{L}_X^{b3} would be sufficient to induce necessary transitions.

Generalization

These ideas could be generalised as follows. Say we are interested in a classical sampling / optimisation process where the problem hamiltonian is P , which is by requirement a diagonal hamiltonian (i.e composed of only Pauli Z terms) but we are interested in just the states (computational basis states) within a dataset \mathcal{D} ($\mathcal{D} \subset \{1, 0\}^n$).

We can assume there exists a an oracle $O_{\mathcal{D}}$, with the property

$$O_{\mathcal{D}}|\psi\rangle = \begin{cases} |\psi\rangle ; |\psi\rangle \in \mathcal{D} \\ \gamma_{\psi} |\psi\rangle ; |\psi\rangle \notin \mathcal{D}, \gamma_{\psi} \neq 1, \end{cases}$$

that the elements from the dataset are in +1 eigenspace of the $O_{\mathcal{D}}$. The fact that such an operator would exist and would be a diagonal unitary is easy to realise.

Given this, we can say that the ideal choice of the mixer operator would be T , such that

$$[T, O_{\mathcal{D}}] = 0$$

which makes sure that all transitions induced by T preserves the dataset \mathcal{D} . But note also that under general circumstances we would also like to have

$$[T, P] \neq 0$$

to ensure that T is able to induce transitions between different energy levels. Of course this would not be a requirement anymore if all states in \mathcal{D} have the same energy, as in the case with stabilizer hamiltonians.

For example with \mathcal{D}^{b3} , we can pick the oracle to be

$$O_{\mathcal{D}^{b3}} = \left(\frac{-1}{6} \right) M_{stab}^{b3}$$

and $T = \mathcal{L}_X^{b3} = \langle X_0 X_1 X_2, X_3 X_4 X_5, X_6 X_7 X_8 \rangle$, as for the the problem hamiltonian P we can either set

- $P = M_{stab}^{b3}$; Then we will have $[T, P] = 0$, since all elements of the dataset have same probability it shouldn't be a big issue.
- $P = M_{hebb}^{b3}$; The presence of non-stabilizer terms (like 2 ($Z_0 Z_3 + Z_1 Z_7 + \dots$)) will lead to $[T, P] \neq 0$ since this terms would anti-commute. Since the elements of the dataset still have same energy we do not gain in the criterion, however the sampling might benefit from the larger energy gap between non-ground states.

Speaking of generality, we should note that for most practical datasets it is might be highly troublesome to find suitable operators for T and $O_{\mathcal{D}}$. We have been able to deal with our example only because its related to an instance of stabilizer-hamiltonian, but in general case it might take exponential amount of computation.

EFFECTS of DATA BASED MIXERS

So next we can study the effects of using mixers constructed out of the properties of Datasets, for problems like sampling and training boltzman machines. The example are all based on the experiments run on the bars dataset \mathcal{D}^{b3} .

For the purpose of simulation we have used the M_{hebb}^{b3} as our problem hamiltonian. It was indicated in the previous section that this hamiltonian has the same ground states as M_{stab}^{b3} but had larger energy difference between ground-states and non ground states.

Following the discussion in the last section we know how to construct ideal mixer that would restrict transitions to those that preserve the Dataset constraint. For the sake of comparison we will also consider mixers that do not preserve the Dataset, since we don for example

- Mixers hamiltonian of weight 2, T_{w_2} : Operators of form $X_i X_j$
 - Mixers hamiltonian of wight 3, T_{w_3} : Operators of form $X_i X_j X_k$
- and so on. Since we do not have any specific choice over the configuration (i.e choice of the indices i, j, k . .) of those operators, we construct the final mixer by summing over all possible configurations, i.e

$$T_{w_3} = \sum_{\substack{i,j,k=1 \\ i \neq j \neq k}}^n X_i X_j X_k$$

thus in general any mixer of weight w , T_w is a sum over $\binom{n}{w}$ pauli strings of weight w .

Though such mixers do not satisfy $[T, O_{\mathcal{D}}] = 0$, we will see from our experiments that they might be more helpful for practical purposes.

NB : To denote the type of MCMC used we will stick to the following convention:

mcmc type -> <cl/qu>-<mixer type and weight>

Eg. qu-wt1 -> quantum MCMC where the mixer is T_{w_1} ;

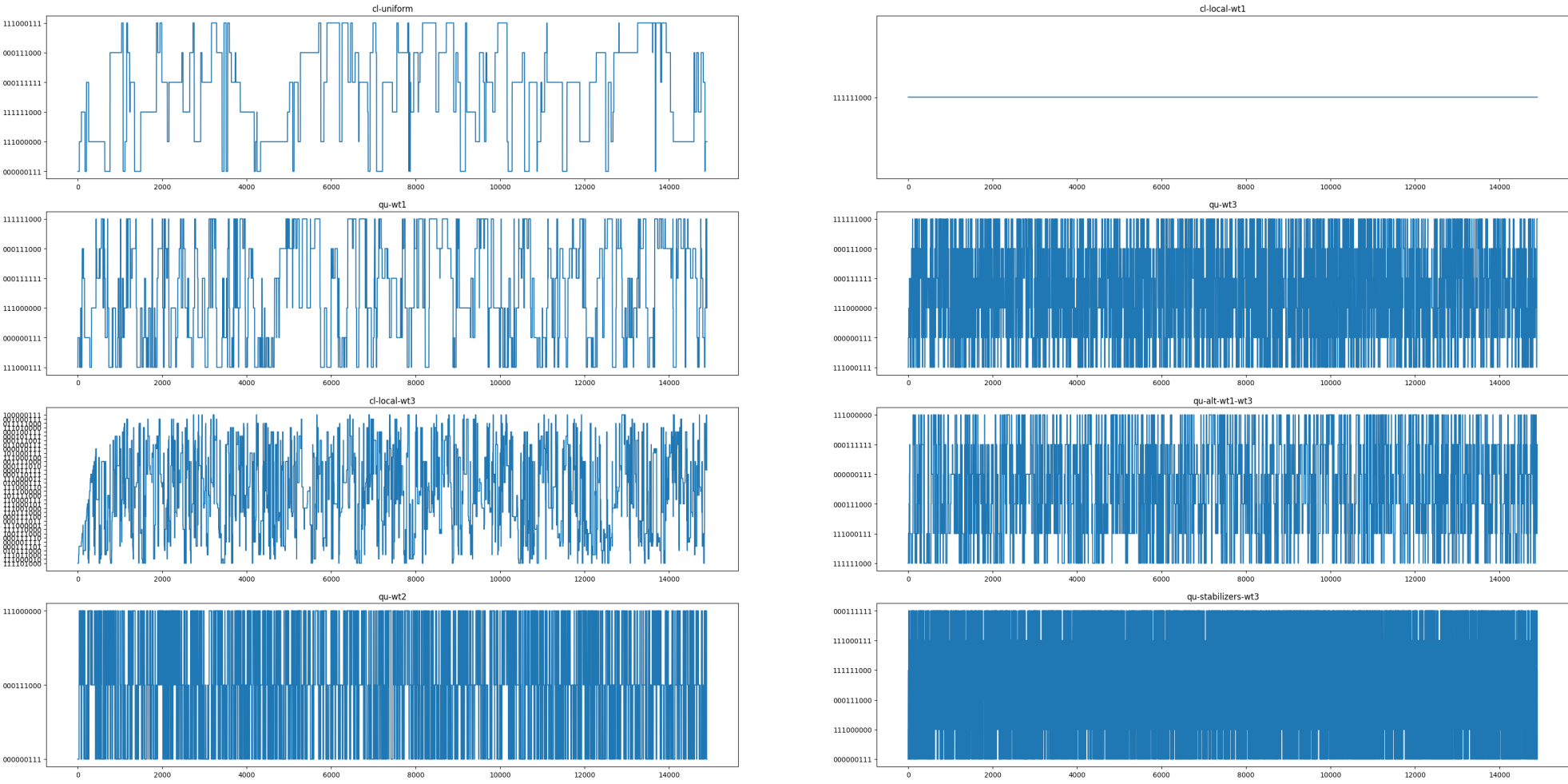
qu-stabilizers-wt3 -> quantum MCMC where the mixers are based on the stabilizer hamiltonians and are of weight 3; i.e the mixer hamiltonian is $T_{\mathcal{D}^{w3}} = X_0X_1X_2 + X_3X_4X_5 + X_6X_7X_8$.

Sampling

Here we run 10 independent instances for each of the MCMC types : ['cl-uniform', 'cl-local-wt1', 'qu-wt1', 'qu-wt3', 'cl-local-wt3', 'qu-alt-wt1-wt3', 'qu-wt2', 'qu-stabilizers-wt3']

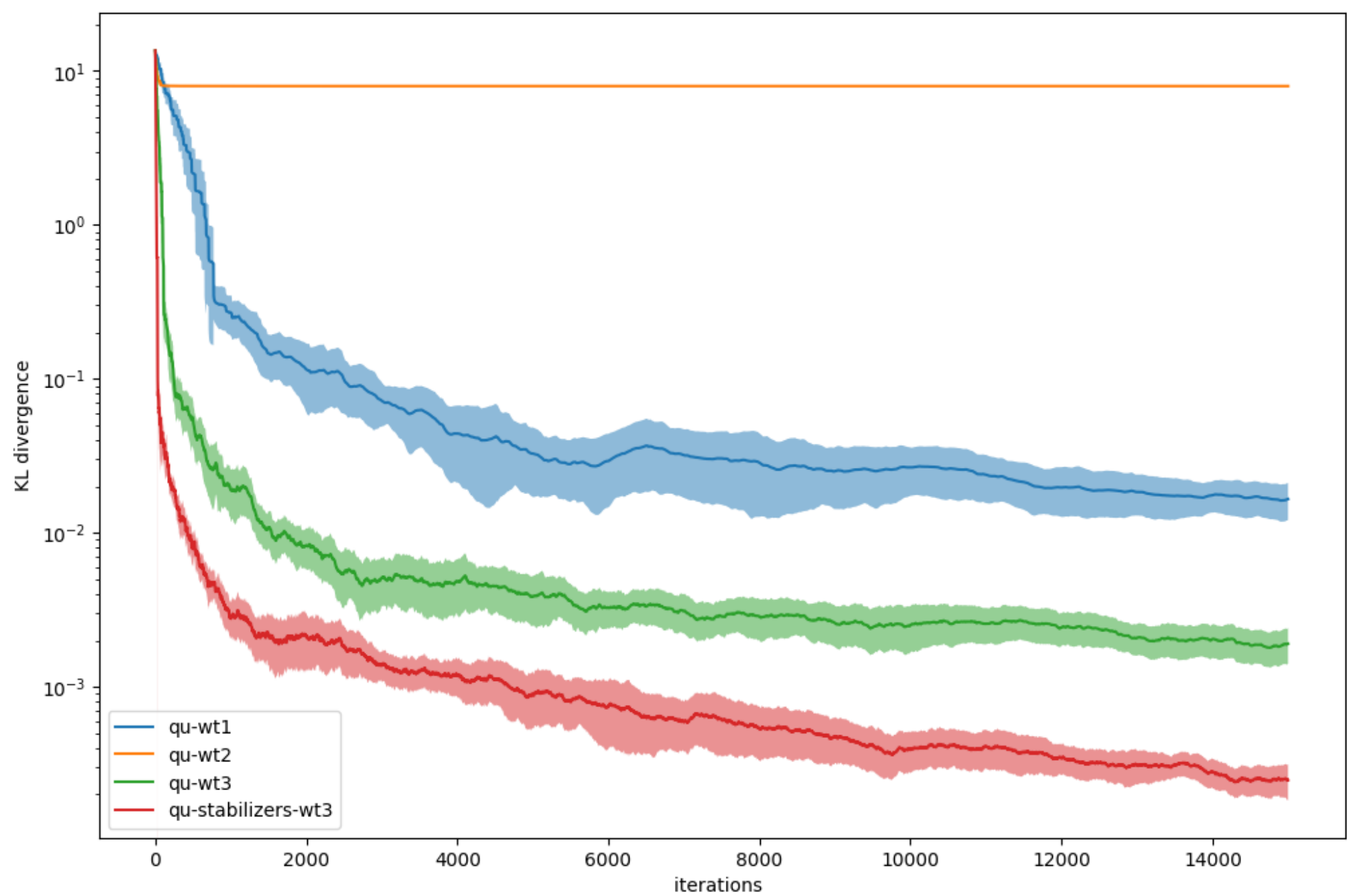
NB. : For all experiments we initiate the MCMC chain from an element of the dataset. It will be clear later that the former is a stringent requirement for some mixers to perform sampling.

Trajectory Visualisation



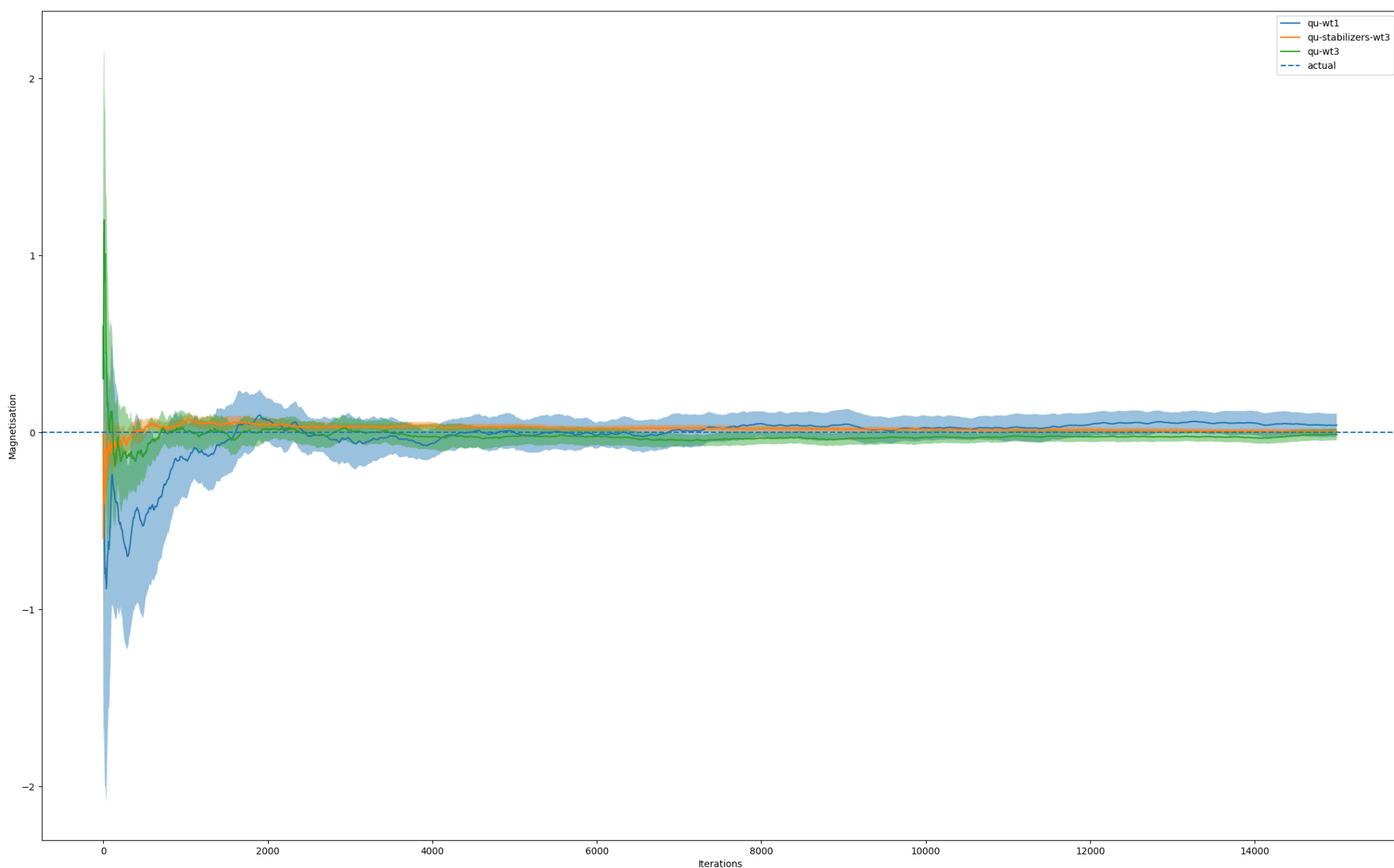
Despite the differences, most mcmc types manage to sample the all the ground states. Except `cl-local-wt1` which fails to make any transitions and `qu-wt2` which stays transitioning between just 3 states. All except the mcmc chain for `cl-local-wt3` remains confined to the elements of the dataset \mathcal{D}^{b3}

KL Divergence

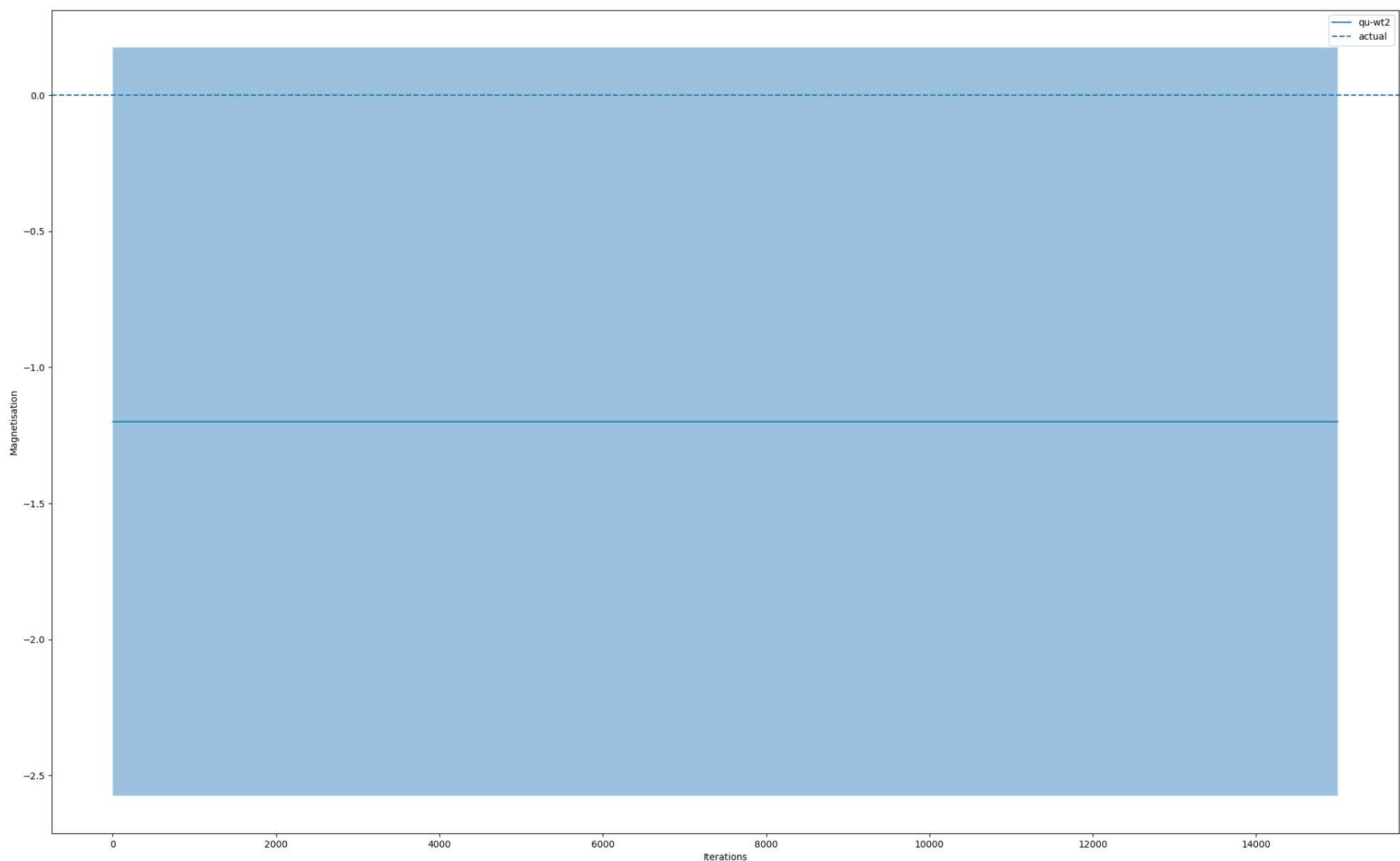


Here `qu-stabilizers-wt3` is clearly the winner. It also shows that `qu-wt2` fails to sample the hamiltonian even to the worst accuracy, this also apparent from the last plot, nonetheless we have seen that it preserves the dataset constraint.

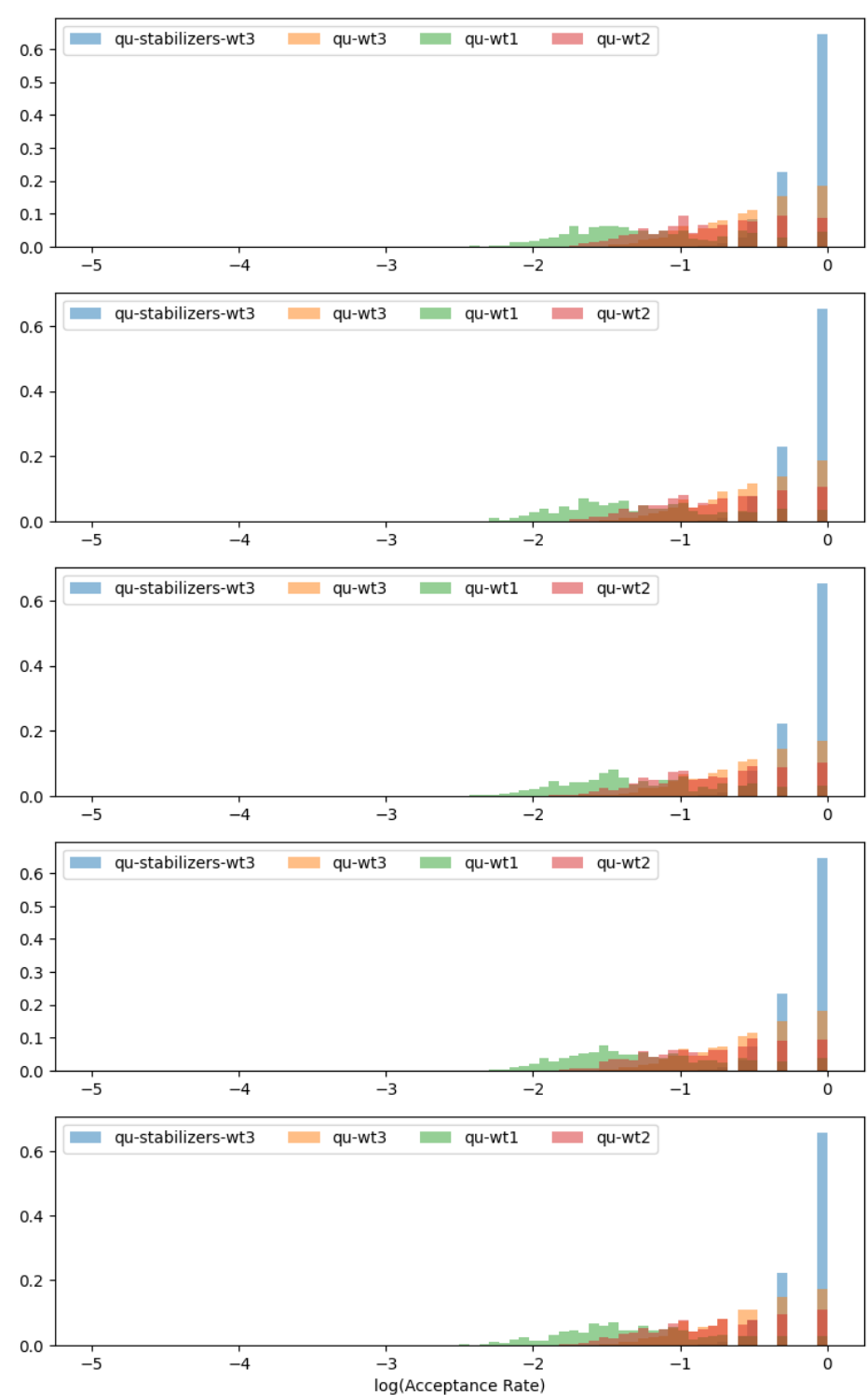
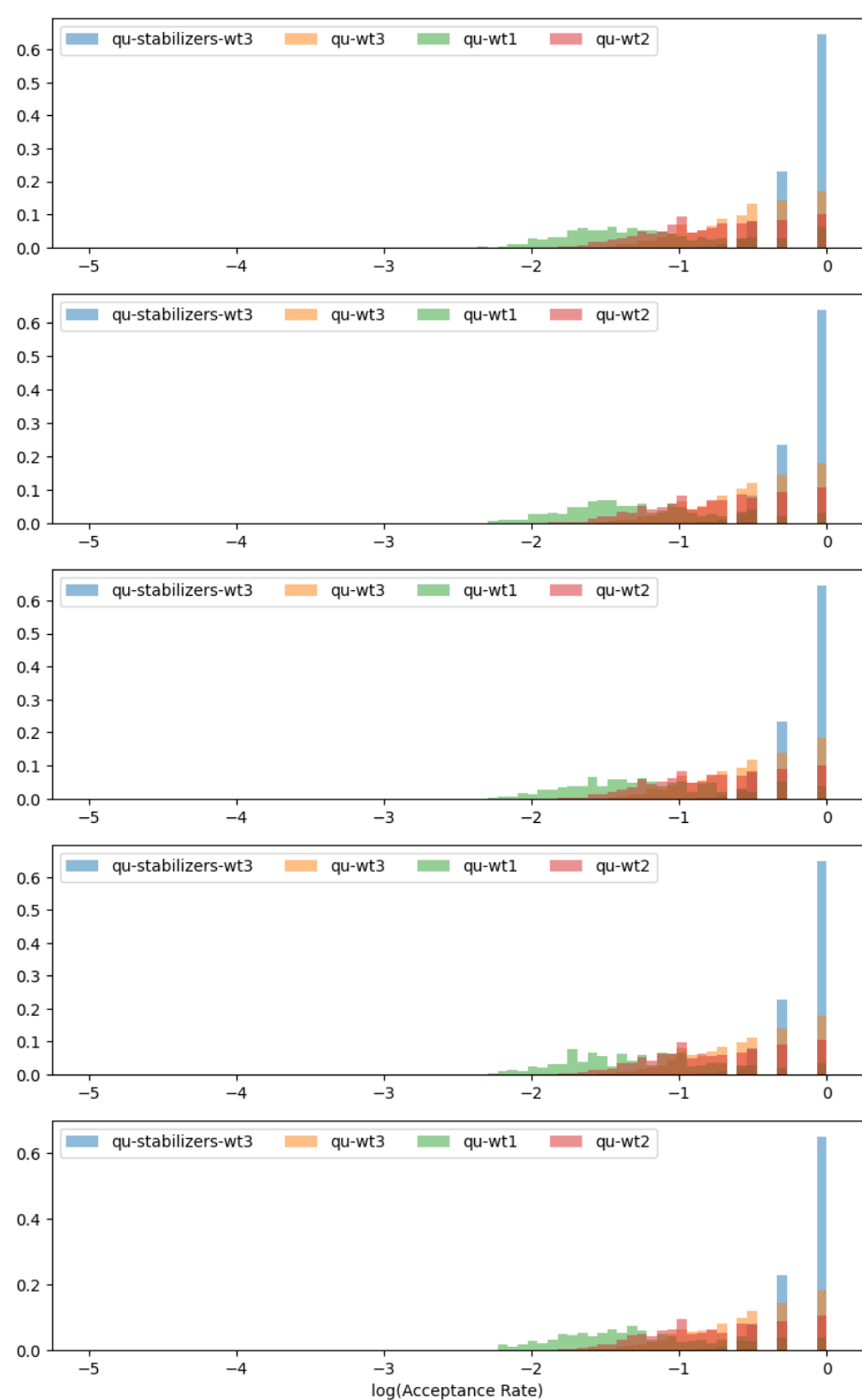
Magnetisation

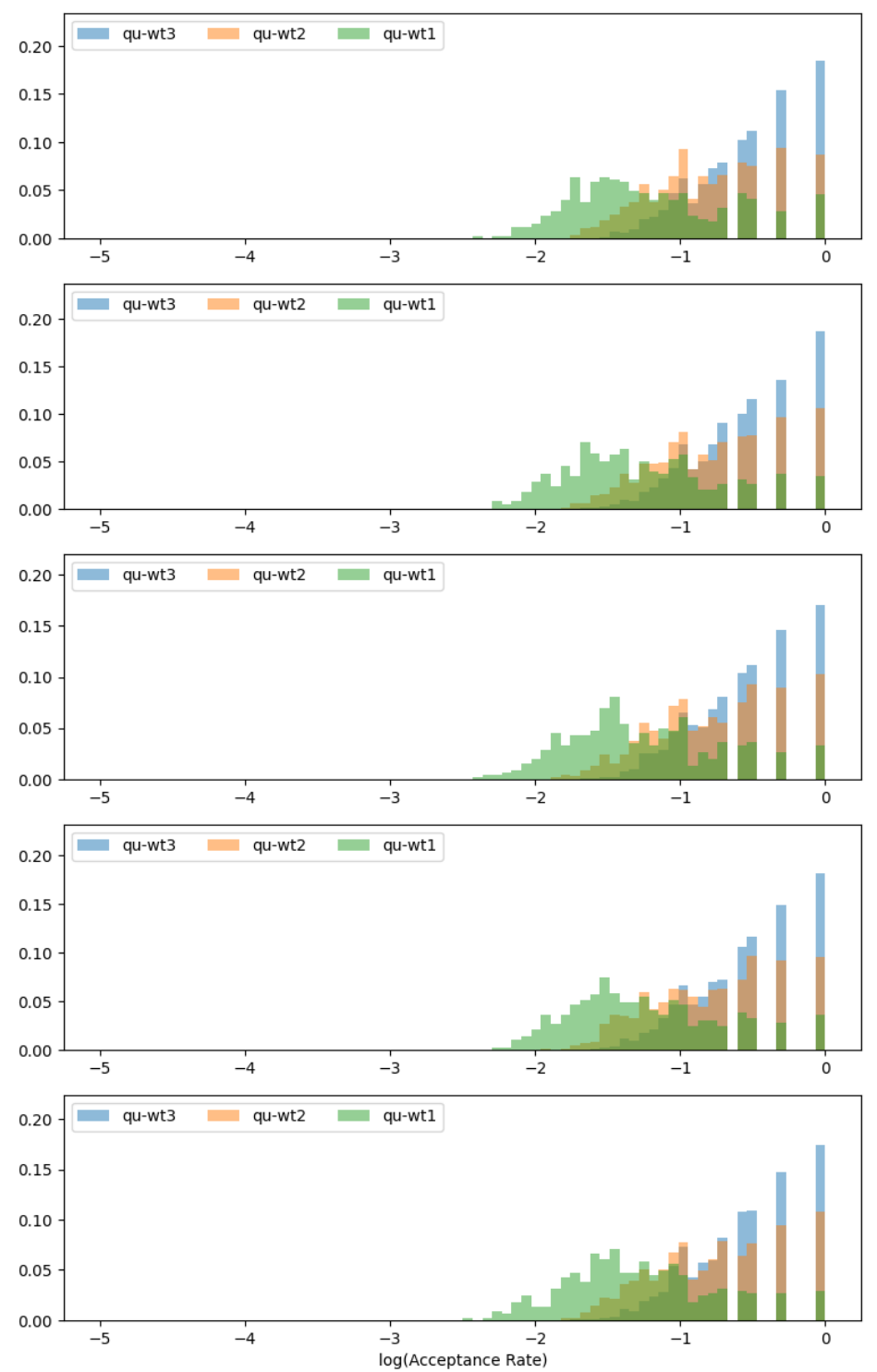
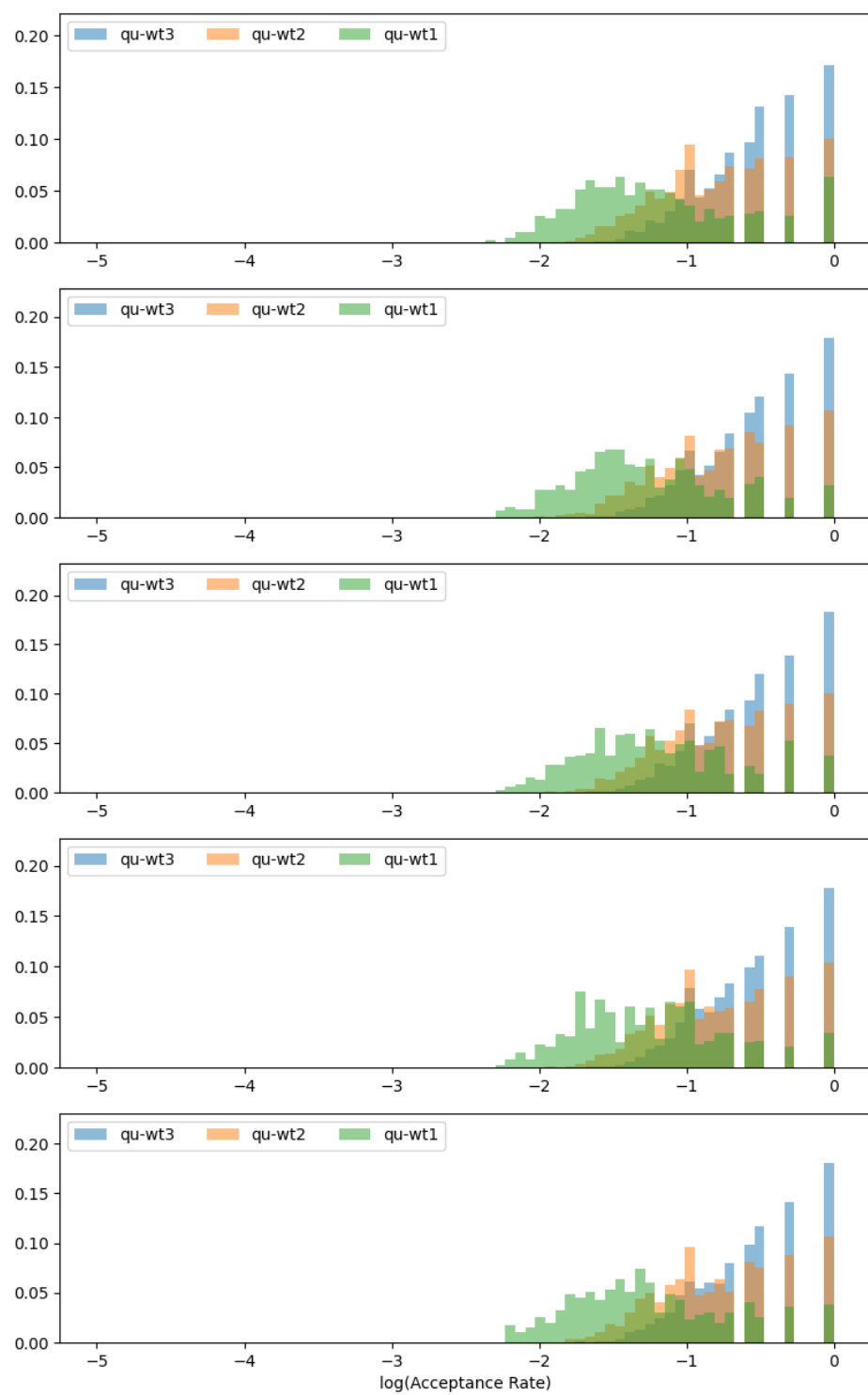


The performance in this case as similar to the previous one. The case for `qu-wt2` fails disastrously as expected, and doesn't even approach the actual magnetisation.



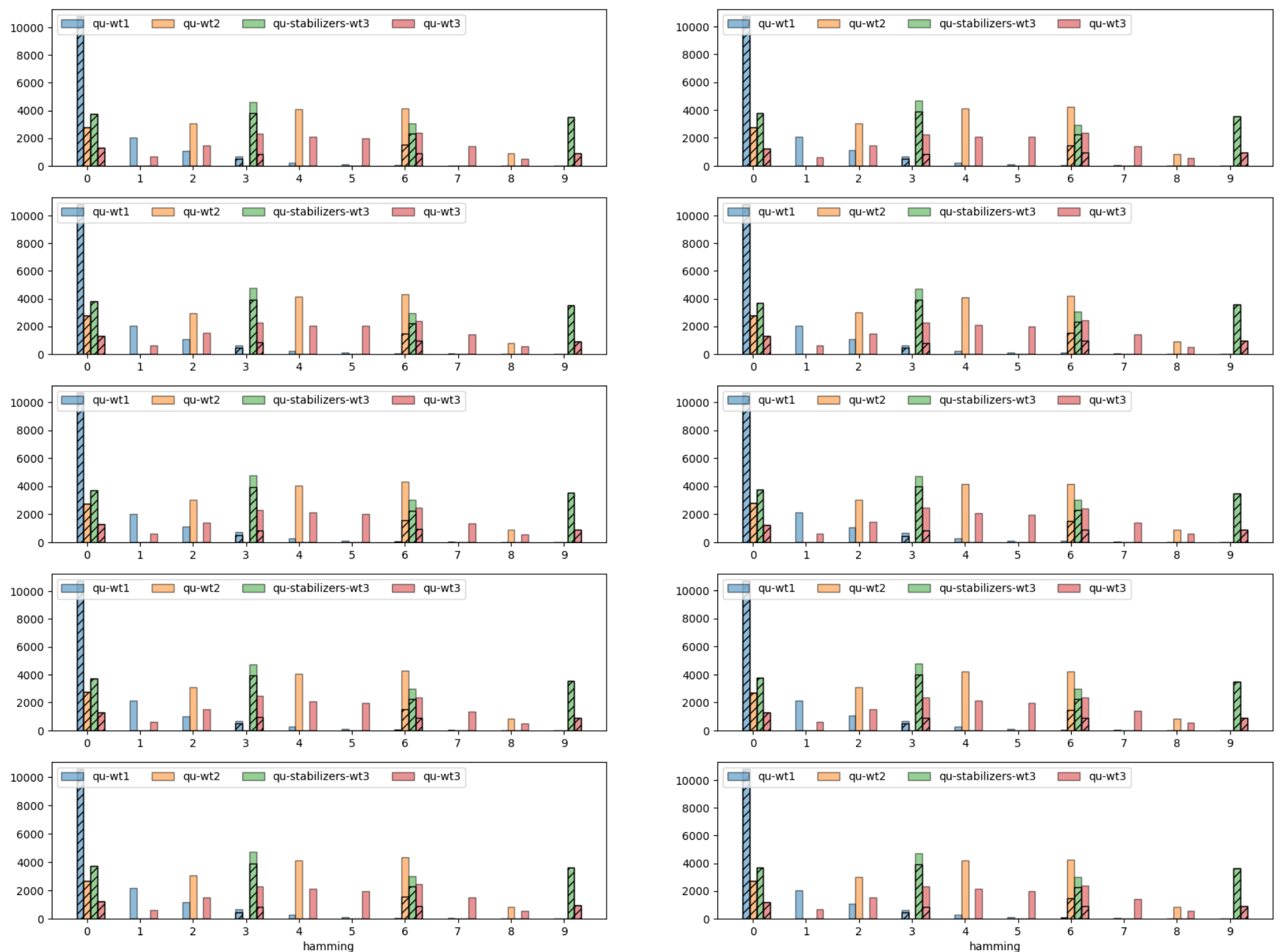
Acceptance Probability





- `qu-stabilizers-wt3` has the best acceptability as expected.
- The distribution of `qu-wt2` is rather surprising, and highlights one of the key properties about customised mixer hamiltonians.
 - The plot shows that propositions made by `qu-wt2` are well accepted but yet we know from the KL-Div plot that it barely manages to sample the distribution,
 - This happens because the mixer in `qu-wt2` restricts the propositions made by the sampler to a few elements within the data set which are likely to be accepted but fails to propose most elements in the dataset, thus reducing its effectivity as a sampler.

Hamming Distance



NB. : The shaded portions indicate the fraction of the propositions that were accepted

Here we see that

- That only transitions of h_d in multiple of 3 are accepted in significant proportions.
- For `qu-stabilizers-wt3` ; the proposed transitions are restricted to states differing by multiples of 3 h_d , and most of the m are accepted.
- For `qu-wt2` ; the proposed transitions are restricted to states differing by multiple of 2 h_d , but only propositions of 6 h_d are accepted.
- For `qu-wt3` ; the proposed transitions are spread over all h_d

To some extent, I think the data can be explained by observing the terms in the mixer unitary (i.e unitary corresponding to mixer hamiltonian).

- For `qu-stabilizers-wt3` :

$$\exp(i\gamma T_{D^{w3}}) = (c_\gamma I + is_\gamma X_0 X_1 X_2) (c_\gamma I + is_\gamma X_3 X_4 X_5) (c_\gamma I + is_\gamma X_6 X_7 X_8)$$

where $c_\gamma = \cos(\gamma)$; $s_\gamma = \sin(\gamma)$ (for convenience we use γ to represent γt_{ev} in the theory).

The expansion will only have pauli strings of weights 0,3,6,9 because none of the three terms overlap against each other. Of course the relative weight of the pauli terms would depend on the coefficient, but it still explains our previous observation.

- For `qu-wt1` :

$$\exp(i\gamma T_{w_1}) = \prod_i (c_\gamma I + s_\gamma X_i)$$

Which upon expansion will lead to operators of all possible weights, with an operator of weight w having a coefficient proportional to s_γ^w . Since we have chosen a small gamma, operators of higher weights will have lower coefficient and thus a lower effect in inducing corresponding transitions

- For `qu-wt2` :

$$\exp(i\gamma T_{w_2}) = \prod_{\substack{i,j \\ i \neq j}} (c_\gamma I + s_\gamma X_i X_j)$$

Unlike the previous case here the individual terms do overlap, since the pauli strings are comprised of two elements there could be overlap on either on

- a single element ; like $X_i X_j \cdot X_j X_k = X_i X_k$
 - or both the elements ; in which case the resultant operator is identity I . (This case is unlikely since every pair appears just once)
- But in both the cases the resultant operator has an even weight of the pauli string, this explains why every transition induced by the mixer

were of $2h_d$. In case there are no overlaps it follows trivially that the strings will be of weight in multiples of $2h_d$.

- For `qu-wt3`:

$$\exp(i\gamma T_{w_3}) = \prod_{\substack{i,j,k \\ i \neq j \neq k}} (c_\gamma I + s_\gamma X_i X_j X_k)$$

Here the overlaps could be of either 2 or 1 element which creates to an overall operator of weight 2 and 4, and without overlaps it creates operators of weight 6, 9, etc. which will then interact with each other. Moreover since operators of higher weights can be created via the interaction of few terms, there relative strength doesn't attenuate as in the case of `qu-wt1`, atleast for 'not-to-bad' choices of γ .

The fact that most of the accepted states are of $h_d = 3, 6, 9$ reflects the effect of the classical-loop that accept/rejects a transitions rather than effect of the mixer that induces the transition.

Now the major aim here is certainly to design mixer such that proposed transitions are accepted with higher probability, and to that end it is apparent that `qu-stabilizers-wt3` outperforms rest of the schemes. But one should realise that we have only been able to design such a mixer using explicit information about the design, to the extent that the sampling itself doesn't add much practical value to it. Moreover, we will see in the Boltzmann Training section that using such well-designed mixers would have rather adverse effects on the training process