

# Minimizing couplings in renormalization by preserving short-range mutual information

Christian Bertoni<sup>1,\*</sup>  and Joseph M Renes<sup>2</sup> 

<sup>1</sup> Dahlem Center for Complex Quantum Systems, Freie Universität Berlin, Germany

<sup>2</sup> Institute for Theoretical Physics, ETH Zürich, Switzerland

E-mail: [bertoni@physik.fu-berlin.de](mailto:bertoni@physik.fu-berlin.de) and [renes@phys.ethz.ch](mailto:renes@phys.ethz.ch)

Received 1 April 2022, revised 13 July 2022

Accepted for publication 22 July 2022

Published 18 August 2022



CrossMark

## Abstract

The connections between renormalization in statistical mechanics and information theory are intuitively evident, but a satisfactory theoretical treatment remains elusive. We show that the real space renormalization map that minimizes long range couplings in the renormalized Hamiltonian is, somewhat counterintuitively, the one that minimizes the loss of short-range mutual information between a block and its boundary. Moreover, we show that a previously proposed minimization focusing on preserving long-range mutual information is a relaxation of this approach, which indicates that the aims of preserving long-range physics and eliminating short-range couplings are related in a nontrivial way.

Keywords: renormalization, real space mutual information, Markov network

(Some figures may appear in colour only in the online journal)

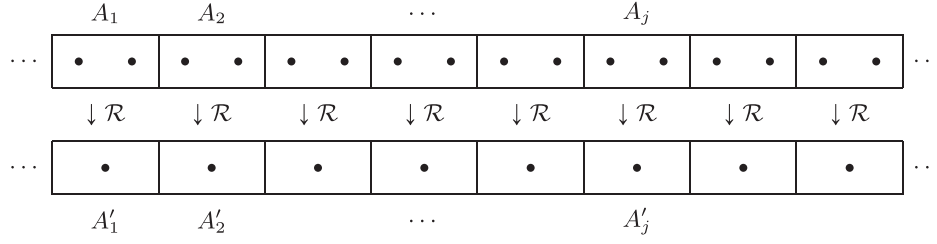
## 1. Introduction

A renormalization process progressively removes degrees of freedom from a physical system, mapping it to an effective system having the same physics at large scales [1, 2]. One may regard the renormalization map as removing unimportant short-range information while leaving long-range information intact, and therefore possible connections to information theory have been explored in several different approaches [3–9]. One difficulty in the renormalization procedure

\* Author to whom any correspondence should be addressed.



Original content from this work may be used under the terms of the [Creative Commons Attribution 4.0 licence](https://creativecommons.org/licenses/by/4.0/). Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.



**Figure 1.** Illustration of a block-spin renormalization procedure on a 1D lattice: the map  $\mathcal{R}$  is applied to each block to get a renormalized system at larger scale.

is finding an appropriate renormalization map. In real space renormalization [10], for example, there is no unique way to remove degrees of freedom, and several maps can plausibly be used. In particular in the context of block-spin renormalization, one divides the system into blocks and tries to map all the degrees of freedom inside the block to a single degree of freedom that nevertheless captures the overall behavior of the block. A common choice of map in spin systems is the so called ‘majority vote’ map, a new spin is produced from each block by choosing the value of the majority of the spins in the block, but an apparently equally sensible choice is to produce a new spin by choosing the value of any of the spins in the block. Both of these choices, as well as others, look sensible in the sense that they aim to capture in a single spin the overall behavior of a larger block, nevertheless, work noticeably better than others [11], and there is no clear criterion for choosing the best map. Intuitively, the best map should be the one that best preserves the long range effective physics of the system. Recently, Koch-Janusz and Ringel [12] proposed choosing real-space renormalization maps based on an information-theoretic criterion, as follows. Consider a spin model on a lattice  $\Lambda$ , at every vertex of the lattice  $i$  there is a random variable  $S_i$ , and the physical state of the system is described by the overall probability distribution  $P$ . In particular the system is described by some nearest neighbor Hamiltonian  $H$ , and we take  $P$  to be the Gibbs thermal state of  $H$  at inverse temperature  $\beta$

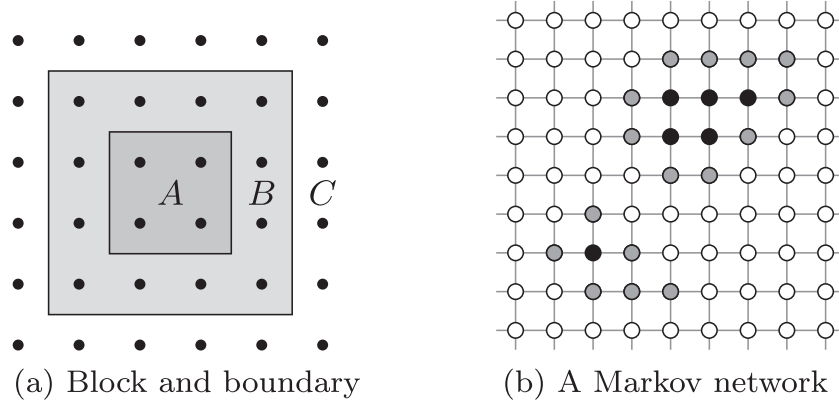
$$P(s_1, \dots, s_{|\Lambda|}) = \frac{1}{Z} e^{-\beta H(s_1, \dots, s_{|\Lambda|})}, \quad Z = \sum_{s_1, \dots, s_{|\Lambda|}} e^{-\beta H(s_1, \dots, s_{|\Lambda|})}. \quad (1)$$

Divide the lattice into non overlapping blocks  $A_j$ . Let  $\mathcal{R}$  be a renormalization map on a single block specifically a stochastic transformation on the random variables describing the spins in the block, and call its output on the  $j$ th block  $A'_j$ . This procedure is illustrated in figure 1. In the renormalization procedure  $\mathcal{R}$  is applied to each  $A_j$ , but here we need only focus on a single block  $A$  with output  $A' = \mathcal{R}(A)$ . In particular, dividing the lattice into the block in question,  $A$ , its neighboring spins  $B$  within some distance, and the remainder of the spins  $C$ , as illustrated in figure 2(a), Koch-Janusz and Ringel propose choosing

$$\mathcal{R}_{\text{KJR}} = \operatorname{argmax}_{\mathcal{R}} I(A' : C)_{\mathcal{R}(P)}, \quad (2)$$

where  $I(A : C)_P$  is the mutual information of random variables  $A$  and  $C$  under the distribution  $P$ , defined as

$$I(A : C)_P = \sum_{a,c} P_{AC}(a, c) \log \left( \frac{P_{AC}(a, c)}{P_A(a)P_C(c)} \right), \quad (3)$$



**Figure 2.** (a) Division of a 2D lattice system into the block to be renormalized  $A$ , its boundary  $B$ , and the rest of the lattice  $C$ . (b) The random variables in the black region are conditionally independent of the those in the white region given the gray region, as the gray region shields the former from the latter in the Markov network. The regions need not be connected.

where  $a$  and  $c$  label the possible values of the random variables  $A$  and  $C$ , here the combined configurations of the spins inside the respective sections of the lattice. In loose terms, this quantity measures the amount of information shared between the two random variables. Two independent variables have vanishing mutual information, while correlated value have higher mutual information. It also satisfies the *data processing inequality*, that is, it must decrease under the application of a stochastic map to either of the entries. It follows that  $I(A : C)_P \geq I(A' : C)_{\mathcal{R}(P)}$ , and hence  $\mathcal{R}_{\text{KJR}}$  retains the most mutual information between the block and the long range parts of the lattice (here and in what follows, we use the name of a region of the lattice as shorthand for the random variable corresponding to the set of spins in said region). Koch-Janusz and Ringel argue that it therefore extracts the relevant degrees of freedom and that it results in a renormalized Hamiltonian with short-range couplings, that is, this method prevents the appearance of out of control long-range interactions, i.e. interactions between far away parts of  $B$  and  $C$ , or interactions between the renormalized block  $A'$  and  $C$  not mediated by the boundary. They also propose a machine-learning algorithm to determine  $\mathcal{R}_{\text{KJR}}$  on a parametrized subset of all possible maps. The resulting real space mutual information (RSMI) algorithm produces good results when benchmarked on various physical models. Lenggenhager *et al* [13] further showed that  $\mathcal{R}_{\text{KJR}}$  does not create any long-range couplings within  $C$  when  $I(A : C)_P = I(A' : C)_{\mathcal{R}(P)}$ . Their theoretical work was expanded to field theory [14] and their algorithm improved by using deep learning techniques [15].

Here we argue that, contrary to the above intuition, to minimize long-range couplings one should instead choose the renormalization map to retain *short-range* mutual information:

$$\mathcal{R}^* = \operatorname{argmax}_{\mathcal{R}} I(A' : B)_{\mathcal{R}(P)}. \quad (4)$$

As we show in detail below, in fact no map  $\mathcal{R}$  can result in long-range couplings within  $C$  or from  $A$  to  $C$ , and  $\mathcal{R}^*$  additionally minimizes coupling within the boundary  $B$ . This approach has several other advantages. For one, the optimization is considerably simpler, as it only involves the block in question and its boundary. Moreover, it is the case that  $I(A' : B)_{\mathcal{R}(P)} \geq I(A' : C)_{\mathcal{R}(P)}$  for every map  $\mathcal{R}$ , and hence the optimization in (2) is a relaxation of the optimization in (4), in the sense that optimizing (2) yields a lower bound for (4). It should be noted that while

we present the main results in the context of nearest neighbor interaction Hamiltonians for simplicity, in the presence of longer range interactions it suffices to define effective degrees of freedom in the interaction lattice by grouping together spins such that the interaction is nearest neighbor for these effective spins. For instance, in a 1D next to nearest neighbor models, one can group together pairs of spins that interact with their neighboring pairs. In this case one can form the renormalization blocks by grouping together said effective degrees of freedom.

We emphasize here that these two optimizations are born out of two different motivations: (2) identifies the degrees of freedom that are most relevant to the long range physics, while (4) aims to control the proliferation of couplings. It is not expected that these two motivations yield the same optimization problem, and the relaxation described above relates the two. Finally, the optimizer of (4) (as well as of (2)) is a deterministic map, which makes brute-force optimization feasible for small blocks by searching the entire map space directly on the probability distribution, rather than by using sampling techniques. We illustrate how the optimization can be performed for  $2 \times 2$  maps using tensor network representations for the 2D Ising model. The rest of this article is structured as follows: in section 2, we introduce the main tools we use for our proof, in section 3 we prove our main result, in section 4 we show that the optimization must be a deterministic map, in section 5 we illustrate our results on the 2D Ising model, and finally we draw conclusions in section 6.

## 2. Gibbs states as Markov networks

To prove our claims we make use of the Hammersley–Clifford theorem of probability theory, which states that every Gibbs state of a local Hamiltonian is a *Markov network*. A Markov network is a (probability distribution on a) collection of random variables with conditional independence relations that are captured by an undirected graph. Consider a collection of random variables  $V = (V_1, \dots, V_n)$  associated to vertices of a graph  $\mathcal{G}$  and having a joint probability distribution  $P(V)$ . Vertices  $V_j$  and  $V_k$  connected by an edge in  $\mathcal{G}$  correspond to dependent random variables, for which  $I(V_j : V_k) \neq 0$ . Given three regions of the graph  $A$ ,  $B$ , and  $C$ , corresponding to disjoint collections of the random variables,  $B$  is said to shield  $A$  from  $C$  if all paths connecting  $A$  to  $C$  pass through  $B$ . An example is depicted in figure 2(b), where the grey regions shield the black regions from the white regions. The region themselves need not be connected, in figure 2(b) we may treat the black spins as a single region  $A$ , the grey spins as a single region  $B$ , and the white spins as a single region  $C$ .

Then  $(\mathcal{G}, P)$  is a Markov network if every two regions shielded by a third are conditionally independent, i.e.  $A$  and  $C$  are independent given the value of  $B$ . Put yet differently, the correlations between  $A$  and  $C$  are mediated entirely by  $B$ . Conditional independence can be succinctly expressed using the conditional mutual information (CMI) as  $I(A : C|B)_P = 0$ , where

$$I(A : C|B)_P := I(A : BC)_P - I(A : B)_P. \quad (5)$$

The Hammersley–Clifford theorem [16, 17] then states that  $(\mathcal{G}, P)$  is a Markov network if and only if  $P(V) = e^{h(V)}$  for some local function  $h$ , meaning  $h = \sum_{c \in \mathcal{C}} h_c - \log Z$ , where  $\mathcal{C}$  is the set of cliques of the graph (the fully-connected subgraphs) and each  $h_c$  is a function only of the variables involved in the clique  $c$ . Here  $\log Z$  is simply a normalization constant to ensure that  $\sum_V P(V) = 1$ .

The renormalization procedure begins with the Gibbs state of a local Hamiltonian  $P \propto e^H$ . Renormalizing a block  $A$  with map  $\mathcal{R}$  results in a new probability  $P' = \mathcal{R}(P) = e^{H'}$ , where

we define  $h' = \log P'$ . Renormalizing all blocks results in some distribution  $P''$ , and the corresponding  $h''$  is just the renormalized Hamiltonian, up to the inverse temperature  $\beta$  and normalization constant factors. By the Hammersley–Clifford theorem,  $h''$  will not contain any couplings between random variables which are conditionally independent, and this property can be established by showing that the CMI vanishes. And by data processing, it is sufficient to consider just  $h'$  to determine where new couplings may arise.

### 3. Ruling out couplings

Let us summarize our claims: the only couplings that can be created by a single block renormalization are within the boundary of the block,  $B$ . In one dimension, these couplings are minimized by picking a map that minimizes the loss of short ranged mutual information  $I(A : B) - I(A' : B)$ . The division of the lattice into blocks in one dimension is illustrated in figure 3, where we additionally divided the boundary and environment into left and right parts, so that  $B = B_R B_L$ ,  $C = C_R C_L$ .

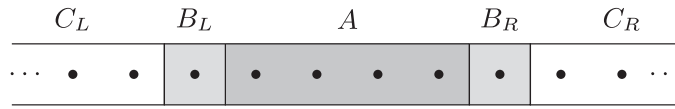
We first observe that due to the boundary  $B$  around the block  $A$  no renormalization map  $\mathcal{R}$  can create couplings within the environment  $C$  or between the new block  $A'$  and the environment  $C$ . Consider two parts  $C_1$  and  $C_2$  of  $C$  which are not already coupled, thus they are conditionally independent given the remainder  $R$  of the random variables comprising the system. Region  $A$  is a part of  $R$ , and the rest we can call  $D$  so that  $R = AD$ . This is illustrated in figure 4. Since  $B$  bounds  $A$ , it must be the case that  $D$  shields  $C_1$  from  $C_2$  and therefore  $I(C_1 : C_2 | D)_P = 0$ . This does not change under application of any map  $\mathcal{R}$ ,  $I(C_1 : C_2 | D)_{\mathcal{R}(P)} = 0$ , and therefore  $C_1$  and  $C_2$  are not coupled in  $h'$ . To show the same thing, the authors of [13] prove instead that  $I(C_1 : C_2 | A') = 0$  by assuming that long range mutual information is preserved, i.e.  $I(A : C)_P = I(A' : C)_{\mathcal{R}(P)}$ . That  $A'$  will not become coupled to anything in  $C$  follows because all the correlations are mediated by  $B$ . Using the positivity of CMI and data processing, we have  $0 \leq I(A' : C | B)_{\mathcal{R}(P)} \leq I(A : C | B)_P = 0$ .

Hence, the main concern is couplings between parts of  $B$  which may be induced by  $\mathcal{R}$ . In one-dimensional systems, as depicted in figure 3, it turns out that coupling between  $B_L$  and  $B_R$  is related to the change in mutual information between the block  $A$  and the boundary  $B = B_L B_R$ . If the mutual information is unchanged after  $\mathcal{R}$ , then  $B_L$  and  $B_R$  are uncoupled in  $h'$ . This is a consequence of the following more general statement.

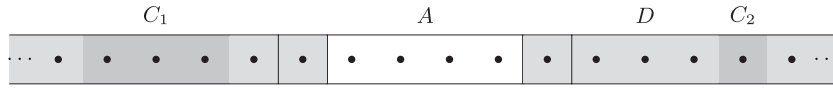
**Theorem 1.** *Consider a one-dimensional lattice model with nearest-neighbor Hamiltonian  $H$  in a Gibbs state, divided into subregions as in figure 3. For any renormalization map  $\mathcal{R} : A \rightarrow A'$ ,  $I(B_L : B_R | A')_{\mathcal{R}(P)} \leq I(A : B)_P - I(A' : B)_{\mathcal{R}(P)}$ .*

**Proof.** Start from  $I(B_L : B_R | A') = I(B_L : B_R A') - I(B_L : A')$  and apply data processing to the first term to obtain  $I(B_L : B_R | A') \leq I(B_L : B_R A) - I(B_L : A')$ . We omitted the probability distribution in the subscript of the CMI, the renormalized distribution is signalled by the presence of a prime in the variables' names. Now note that  $A$  and  $C$  can be swapped in (5), i.e.  $I(A : C | B) = I(C : AB) - I(C : B)$ , and therefore  $I(AB : C) - I(A : BC) = I(B : C) - I(A : B)$ . Using this property for each term in the expression above gives  $I(B_L : B_R | A') \leq I(A : B) - I(A' : B) + I(B_R : B_L A') - I(B_R : A)$ . Another application of data processing to the third term and the CMI definition gives  $I(B_L : B_R | A') \leq I(A : B) - I(A' : B) + I(B_R : B_L | A)$ . The final term is zero by assumption.  $\square$

Typically, no nontrivial map  $\mathcal{R}$  will precisely preserve the mutual information for reasons we shall explain in a moment. Nevertheless, minimizing the change in mutual information, by maximizing  $I(A' : B)_{\mathcal{R}(P)}$  as in (4), minimizes the coupling between  $B_L$  and  $B_R$ . This is because



**Figure 3.** Division of a 1D lattice into block and boundary.

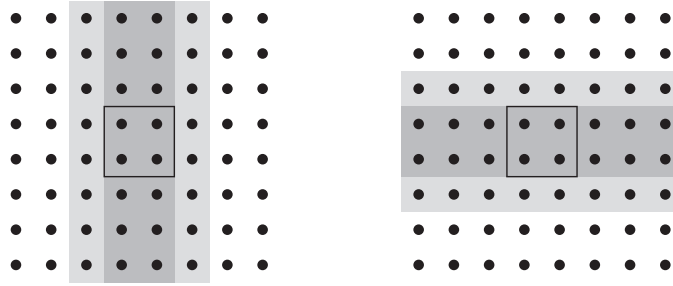


**Figure 4.** Division of a 1D lattice into far away regions  $C_1, C_2$ , the block  $A$ , and the remainder  $D$ . The vertical lines represent the earlier division into blocks  $ABC$ .

the smaller the CMI, the closer the distribution  $\mathcal{R}(P)$  is to some  $P'$  in which  $B_L$  and  $B_R$  are conditionally independent, as measured by the total variational distance between distributions (see [18, lemma 1]). That is, if we have  $I(B_L : B_R | A)_{\mathcal{R}(P)} = \epsilon$ , there exists a distribution  $P'$  where  $B_L$  and  $B_R$  are decoupled such that  $\|\mathcal{R}(P) - P'\|_1 \sim \epsilon$ . Hence smaller CMI leads to an associated  $h'$  with weaker couplings. Somewhat counterintuitively, then, to minimize couplings it is more important to preserve mutual information between a block and its boundary rather than between a block and distant spins.

For isotropic systems, we can translate the 1D argument to multiple dimensions by treating a  $D$  dimensional isotropic lattice as a 1D system in every direction, as proposed by Leggett *et al* [13]. The lattice can be separated into disconnected regions by hyperplanes creating effectively a 1D system (figure 5) and the argument of theorem 1 carries over, so that no couplings will appear between the spins in the boundary strips  $B_L$  and  $B_R$ . Couplings might still appear inside the central strip, but if the system is isotropic we can repeat the same argument with hyperplanes separating the renormalized block from the rest in a different dimension and expect that if a map maximized  $I(A' : B)$  in one dimension, it will do so also in the other dimension. This argument breaks down for non isotropic systems as the different directions may have different optimal maps. For disordered systems in which the disorder is isotropic, the optimal map is different in different directions for individual disorder realizations, but it is easy to see that the map which maximizes the average drop of mutual information is the same for both directions, as the maximization will only depend on the distribution of the couplings, which is the same in both cases.

Before proceeding to examine the two optimizations in more detail, let us remark that a renormalization map which precisely preserves the mutual information can actually be undone by a suitable stochastic map. This accords with the idea that no information is lost along the renormalization flow in this case by assumption, but one does not typically expect renormalization to be reversible. Starting from  $I(A' : B)_{\mathcal{R}(P)} = I(A : B)_P$  and using the fact that  $I(A : C|B)_P = I(A' : C|B)_{\mathcal{R}(P)} = 0$ , it follows that the total mutual information is preserved,  $I(A : BC)_P = I(A' : BC)_{\mathcal{R}(P)}$ . Then we can appeal to lemma 1 of [18], which ensures that the so-called ‘transpose’ map or Petz recovery map  $\hat{\mathcal{R}}$  is such that  $\hat{\mathcal{R}} \circ \mathcal{R}(P) = P$  [19]. The transpose map depends on  $\mathcal{R}$  and the marginal distribution of  $A$  under  $P$ , but we shall not go into further details here.



**Figure 5.** The dark and light gray strips indicate the blocks that are used when treating the system as one dimensional in each direction, while the square indicates a block to be renormalized. If the renormalization map is optimal, the light gray strips are uncoupled. If the system is isotropic, the optimal maps for the two directions are the same.

#### 4. The optimal map must be deterministic

Computing  $I(A : B)$  does not require handling the whole probability distribution, but only the marginal distribution on the  $AB$  subsystem. This simplifies the optimization relative to Koch-Janusz and Ringel’s proposal, where the distribution on the entire spin system must be treated somehow. As mentioned above, (2) is a relaxation of (4) in that  $I(A' : C)_{\mathcal{R}(P)} \leq I(A' : B)_{\mathcal{R}(P)}$ . This follows directly from the definition of the CMI and the Markov condition:  $I(A' : BC) = I(A' : B)$  since  $I(A' : C|B) = 0$ , but then  $I(A' : C) \leq I(A' : B)$  by data processing. The equality  $I(A' : C|B) = 0$  reflects the fact that all correlations between  $A'$  and  $C$  are mediated through  $B$ . Therefore, maximizing the mutual information of the former sets a lower bound on the mutual information of the latter. We expect the gap  $I(A' : B) - I(A' : C)$  to become small as the system flows towards an infinite temperature fixed point. In particular by the chain rule we have  $I(A' : B|C) = I(A' : C|B) + I(A' : B) - I(A' : C)$ . That is,  $I(A' : B) = I(A' : C)$  implies  $I(A' : B|C) = I(A' : C|B) = 0$ , then by the Hammersley–Clifford theorem there are no couplings between  $A$  and  $B$ , and thus the system is at infinite temperature and uncorrelated i.e.  $I(A' : B) = I(A' : C) = 0$ . It might also be that the probability distribution is not strictly positive and the Hammersley–Clifford theorem does not apply, but in a Hamiltonian system this implies that the system is at zero temperature, where the probability distribution is uniform on the ground states and vanishes everywhere else. If the configuration of  $A'$  and  $B$  in a given ground state is entirely determined by the configuration of  $C$ , then  $A'$  and  $B$  are independent given  $C$  (as they are deterministic), hence at least in some systems this gap might close approaching a zero temperature fixed point.

In both (2) and (4) the optimal map  $\mathcal{R}^*$  is necessarily deterministic, i.e. all its transition probabilities are either zero or one. This follows because the objective function, the mutual information, is a convex function of the optimization variable, the map  $\mathcal{R}$ , and the extreme points of stochastic maps are deterministic maps.

**Theorem 2.** *Let  $\mathcal{C}$  be the space of stochastic maps from  $A$  to  $A'$ . For a fixed probability distribution  $P_{AB}$  the function  $\mathcal{C} \rightarrow \mathbb{R}_+, W \mapsto I(A' : B)_{W(P)}$  is convex.*

**Proof.** Consider a collection of stochastic maps  $\{W_z\}_{z \in \mathcal{Z}}$  indexed by the values of a finite random variable  $Z$  with distribution  $Q$ . In the following the alphabet of a random variable is denoted by the corresponding curly letter. The average map  $W_Z$  is given by  $W_Z(P_{AB}) = \sum_{z \in \mathcal{Z}} Q(z)W_z(P_{AB})$  for any  $P_{AB}$ . For simplicity, denote  $W_Z(P)$  just by  $P'$ . The probability distribution of  $A'$  and  $B$  is then obtained by averaging over the corresponding distribution, knowing

that we applied a given map,

$$P'_{A'B}(a, b) = \sum_z Q(z) P'_{A'B|Z=z}(a, b) = \sum_{z \in \mathcal{Z}} Q(z) W_z(P_{AB})(a, b). \quad (6)$$

Meanwhile, the average mutual information is given by the CMI  $I(A' : B|Z)_{P'}$ ,

$$\sum_{z \in \mathcal{Z}} Q(z) I(A' : B)_{W_z(P)} = \sum_{z \in \mathcal{Z}} Q(z) I(A' : B|Z = z)_{W_z(P)} = I(A' : B|Z)_{P'},$$

where

$$I(A' : B|Z = z)_{W_z(P)} = \sum_{a \in A', b \in B} P'_{A'B|Z=z}(a, b) \log \left( \frac{P'_{A'B|Z=z}(a, b)}{P'_{A'|Z=z}(a) P'_{B|Z=z}(b)} \right). \quad (7)$$

But then, since  $B$  and  $Z$  are uncorrelated, we obtain

$$I(A' : B|Z)_{P'} = I(A'Z : B)_{P'} - I(Z : B)_{P'} = I(A'Z : B)_{P'} \geq I(A' : B)_{P'}, \quad (8)$$

and therefore the mapping is convex.  $\square$

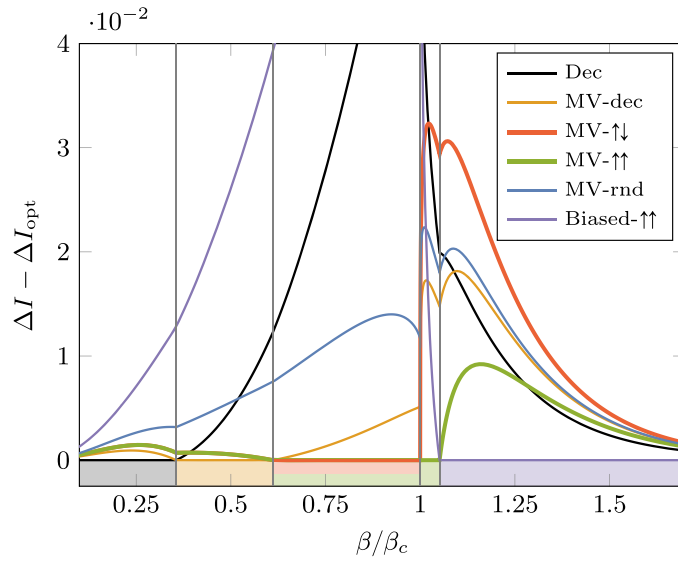
When maximizing a convex function over a convex set, the optimum will occur at one of the extreme points [20, theorem 32.2], which in this case are the deterministic maps [21, theorem 1]. This simplifies the optimization by making the search space finite. While a brute force search might still be out of reach for interesting systems, more sophisticated methods such as machine learning techniques can be informed by this fact.

## 5. The Ising model

Consider renormalization maps on  $2 \times 2$  blocks in the 2D square-lattice ferromagnetic Ising model, with Hamiltonian  $-\sum_{\langle i,j \rangle} \sigma_i \sigma_j$ ,  $\sigma_i \in \{-1, 1\}$ . To investigate which maps are optimal according to (4), we use the corner transfer matrix algorithm [22] with open boundary condition to extract the marginal distribution of a  $4 \times 4$  block, and we measure the change in mutual information between the central  $2 \times 2$  block and its boundary after each of the possible  $2^{16}$  deterministic maps mapping this block to a single spin. We iterated the algorithm until  $\sum_k |\sigma(C_i)_k - \sigma(C_{i-1})_k| \leq 10^{-11}$ , where  $\sigma(C_i)_k$  are the singular values of the corner tensor at the  $i$ th iteration. The results in this section are obtained using a maximum bond dimension  $\chi = 35$ . A higher bond dimension did not affect the results on a coarser temperature range. We then compute the change in mutual information for each map over the range of temperatures  $\beta \in [0.1\beta_c, 1.9\beta_c]$  where  $\beta_c = \log(1 + \sqrt{2})/2$  is the critical inverse temperature and find the optimal map at each temperature. In figure 6 we show the change in mutual information compared with the minimum value for some common maps:

- (a) Decimation: the value of the renormalized spin is simply the value of one of the 4 spins in the block.
- (b) Majority vote: the renormalized spin is assigned a value +1 if the majority of the spins in the block are +1, and vice versa. Ties must be broken with a  $2 \times 2$  block, we do this in 4 possible ways: using a predetermined fixed value (i.e. the ties are always resolved with +1 or -1), using one of the spins in the block (hence the map becomes decimation in case of ties), or choosing a value at random.
- (c) Biased: the all configurations are mapped to +1 except for  $(-1, -1, -1, -1)$  or, vice versa, to -1 except for  $(+1, +1, +1, +1)$ .





**Figure 6.** Difference of the mutual information change for each map above the optimal change, as a function of inverse temperature. MV stands for majority vote and Dec for decimation.  $\uparrow\uparrow$  and  $\uparrow\downarrow$  stand for the aligned or unaligned tie breaks. Each shaded region indicates which map is optimal in the corresponding interval. Note that while both majority vote maps which break ties aligned (MV- $\uparrow\uparrow$ ) and antialigned (MV- $\uparrow\downarrow$ ) with the overall magnetization are optimal in the interval (0.6109, 1), the random tiebreaker map (MV-rnd) is far from optimal.

Some of these maps are not symmetric under spin flips, namely the majority vote with fixed value tie breaker and the biased maps. At low temperature, the spin flip symmetry is broken in the CTM algorithm and one of two symmetry breaking sectors is selected, which version is optimal depends on this selected sector. We call the tie breaker or the biased map ‘aligned’ (denoted  $\uparrow\uparrow$  in the figure) if the relevant fixed value for the renormalized spin is aligned with the magnetization in the symmetry-breaking state, and ‘antialigned’ ( $\uparrow\downarrow$ ) otherwise. It was shown that for the 2D Ising model the best performing map majority vote [11], and it is known that the decimation map leads to uncontrolled long range couplings. At high temperature ( $\beta/\beta_c \lesssim 0.3554$ ), the optimal map is decimation. This is a symptom of the fact that in this regime, the probability distribution over the spins is close to uniform, i.e. the infinite temperature trivial fixed point, where there is no short or long ranged information to preserve. The distribution on a subset of the spin is still the uniform distribution, hence the infinite temperature state is trivially a fixed point of this map. At lower temperatures, for  $0.3554 \lesssim \beta/\beta_c \lesssim 0.6109$ , majority vote with tie breaks decided by decimation is optimal. From that point up to the critical temperature, both version of fixed tie breaker majority vote are optimal, the aligned version remains so up to  $\beta/\beta_c \approx 1.0509$ , after which the probability distribution is close to a deterministic distribution peaked at one of the ground states, the low temperature symmetry breaking prevails and the best map is the aligned biased map. We can also see that decimation starts performing better again in this regime, as the zero temperature distribution is also a fixed point for this map. Interestingly, majority vote with random tie breaker is rather far from optimal (it cannot be optimal as it is not deterministic) and fares worse of all other tie breakers except the antialigned one at low temperature. It can also be seen that decimation performs poorly, especially around the critical point. This is consistent with the observations of [11].

## 6. Conclusions

We argued that maximizing the short-range mutual information between a block and its boundary yields a renormalized system with reduced long-range couplings. In particular, couplings are never introduced beyond the boundary region of the renormalization map, and are suppressed when more of the short-range mutual information is preserved. This gives an information-theoretic account of some aspects of renormalization. The optimization suggested by this approach leads to a simple brute-force algorithm for finding the optimal renormalization map which requires only the probability distribution of the input region of the map and its boundary. It is efficient enough for small systems, as demonstrated in the 2D Ising model. Further work is required to explore the robustness of this result when information is only approximately preserved, perhaps by using an approximate generalization of the Hammersely–Clifford theorem.

Our approach contrasts with the focus of [12, 13], which maximizes the long-range mutual information with the dual goals of capturing the relevant degrees of freedom and reducing long-range couplings. The fact that their long-range mutual information optimization is a relaxation of our short-range optimization implies some connection between these goals: If we view extracting the relevant information as the primary justification for the long-range optimization (an intuitively very plausible statement), then it will necessarily do this by minimizing long-range couplings in the renormalized Hamiltonian to some extent. The open question is how much. It would therefore be interesting to investigate under what conditions or in which models the optimal renormalization maps of the two approaches actually coincide. To this end it would also be interesting to modify the RSMI algorithm to focus on short-range mutual information, as exact optimization is computationally difficult for more complicated models. In either scenario one may also be able to take into account the fact that the optimal renormalization map is necessarily deterministic.

Finally, it should be noted that the focus on short-range versus long-range information here is reminiscent of the relation between the tensor renormalization group [23] and the tensor network renormalization (TNR) [24] algorithms. The latter is a refinement of the former in which the additional steps are meant to remove short-range correlations, improving the algorithm near the critical point. The key difference between our approach and TNR is that while we consider purely stochastic maps, the TNR algorithm involves approximating the tensors locally with different tensors, and optimizing the resulting error. This procedure turns out to be neither positive nor trace-preserving. The first problem was solved in the  $\text{TNR}_+$  algorithm [25], where the flow is constrained to be positive by replacing the singular value decomposition by a more appropriate decomposition, nevertheless the map is still not trace-preserving. This is explained by noting that the interactions in the effective Hamiltonian are constrained to be local, while a stochastic map might increase the range of interactions. It should furthermore be noted that while locally disentangling the tensor network and minimizing the error does morally achieve the same goal, there is no general reason for this minimization to be equivalent to minimizing a quantity as complicated as the mutual information. We might nevertheless speculate that some of our results survive if we replace the mutual information by a simpler quantity such as another Rényi divergence, and that this might offer a more direct comparison to the aforementioned tensor network methods.

## Acknowledgments

We thank Doruk Efe Gökmen and Maciej Koch-Janusz for useful discussions and the anonymous referees for their valuable comments and suggestions. CB acknowledges support from the Deutsche Forschungsgemeinschaft via grant CRC183, JMR the Swiss National Science Foundation via the National Center for Competence in Research for Quantum Science and Technology (QSIT).

## Data availability statement

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

## ORCID iDs

Christian Bertoni  <https://orcid.org/0000-0001-7715-4591>

Joseph M Renes  <https://orcid.org/0000-0003-2302-8025>

## References

- [1] Wilson K and Kogut J 1974 The renormalization group and the  $\varepsilon$  expansion *Phys. Rep.* **12** 75–199
- [2] Wilson K G 1975 The renormalization group: critical phenomena and the Kondo problem *Rev. Mod. Phys.* **47** 773–840
- [3] Gaite J and O'Connor D 1996 Field theory entropy, the  $h$  theorem, and the renormalization group *Phys. Rev. D* **54** 5163–73
- [4] Gaite J 1998 Relative entropy in 2d quantum field theory, finite-size corrections, and irreversibility of the renormalization group *Phys. Rev. Lett.* **81** 3587–90
- [5] Apenko S M 2012 Information theory and renormalization group flows *Physica A* **391** 62–77
- [6] Bény C and Osborne T J 2013 Renormalisation as an inference problem (arXiv:1310.3188 [quant-ph])
- [7] Bény C and Osborne T J 2015 The renormalization group via statistical inference *New J. Phys.* **17** 083005
- [8] Bény C and Osborne T J 2015 Information-geometric approach to the renormalization group *Phys. Rev. A* **92** 022330
- [9] Machta B B, Chachra R, Transtrum M K and Sethna J P 2013 Parameter space compression underlies emergent theories and predictive models *Science* **342** 604–7
- [10] Kadanoff L P 1966 Scaling laws for ising models near  $T_c$  *Phys. Phys. Fiz.* **2** 263–72
- [11] Swendsen R H 1979 Monte Carlo renormalization-group studies of the  $d = 2$  Ising model *Phys. Rev. B* **20** 2080–7
- [12] Koch-Janusz M and Ringel Z 2018 Mutual information, neural networks and the renormalization group *Nat. Phys.* **14** 578–82
- [13] Lenggenhager P M, Gökmen D E, Ringel Z, Huber S D and Koch-Janusz M 2020 Optimal renormalization group transformation from information theory *Phys. Rev. X* **10** 011037
- [14] Gökmen D E, Ringel Z, Huber S D and Koch-Janusz M 2021 Statistical physics through the lens of real-space mutual information (arXiv:2101.11633 [cond-mat.stat-mech])
- [15] Gökmen D E, Ringel Z, Huber S D and Koch-Janusz M 2021 Phase diagrams with real-space mutual information neural estimation (arXiv:2103.16887 [cond-mat.stat-mech])
- [16] Hammersley J M and Clifford P 1971 *Markov Fields on Finite Graphs and Lattices*
- [17] Koller D and Friedman N 2009 *Probabilistic Graphical Models: Principles and Techniques* (Cambridge, MA: MIT Press)
- [18] Li K and Winter A 2018 Squashed entanglement,  $k$ -extendibility, quantum Markov chains, and recovery maps *Found. Phys.* **48** 910–24

- [19] This gives a simple proof of the statement of [13], as  $I(A : C) = I(A' : C)$  implies we can use data processing both from  $A$  to  $A'$  and from  $A'$  to  $A$ . Thus,  $I(C_1 : C_2 A') \leq I(C_1 : C_2 A) = I(C_1 : A) \leq I(C_1 : A')$  and hence  $I(C_1 : C_2 | A') = 0$ . The equality is  $I(C_1 : C_2 | A) = 0$  and the inequalities are data processing.
- [20] Rockafellar T R 1970 *Convex Analysis* (Princeton, NJ: Princeton University Press)
- [21] Davis A S 1961 Markov chains as random input automata *Am. Math. Mon.* **68** 264–7
- [22] Nishino T and Okunishi K 1996 Corner transfer matrix renormalization group method *J. Phys. Soc. Japan* **65** 891–4
- [23] Levin M and Nave C P 2007 Tensor renormalization group approach to two-dimensional classical lattice models *Phys. Rev. Lett.* **99** 120601
- [24] Evenbly G and Vidal G 2015 Tensor network renormalization *Phys. Rev. Lett.* **115** 180405
- [25] Bal M, Mariën M, Haegeman J and Verstraete F 2017 Renormalization group flows of Hamiltonians using tensor networks *Phys. Rev. Lett.* **118** 250602