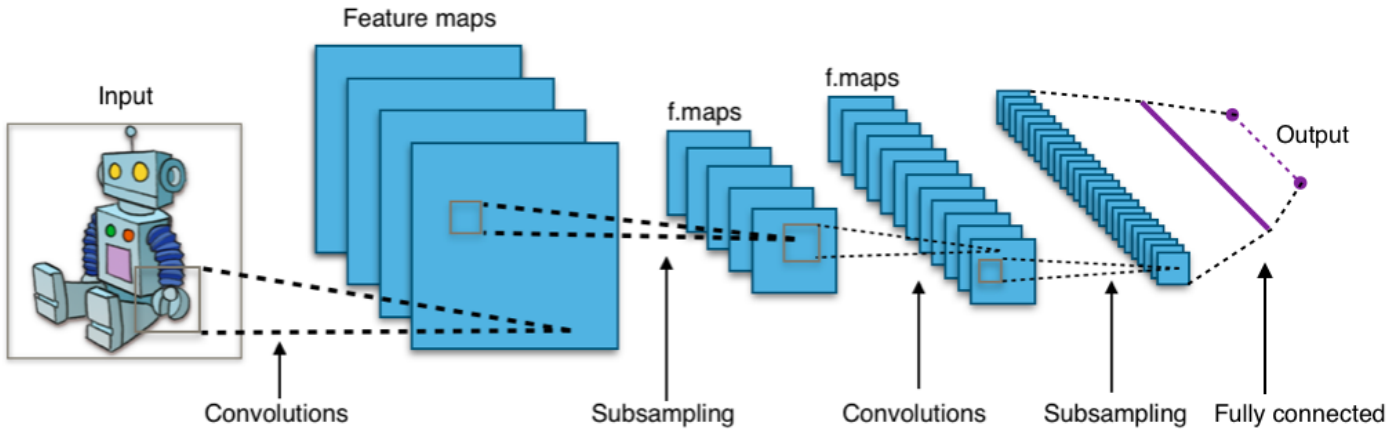
Hierarchical Flows

Dmitry Yarotsky

Motivation: convnets

Practical neural networks work with complex multi-dimensional data, and have appropriate architectures

Convnets:



https://en.wikipedia.org/wiki/Convolutional_neural_network

- Weight-sharing (translation-equivariance)
- Locality
- Pooling
- Growing "feature dimension" (vs. decreasing "geometry resolution")

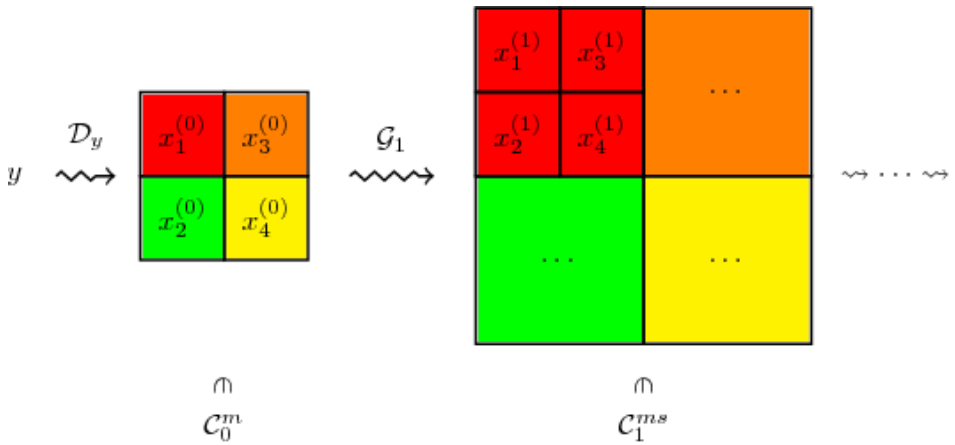
Difficulties of a theoretical analysis

- Practical datasets and maps of interest are hard to characterize theoretically
- The problem is potentially infinite-dimensional (the input object – image – is a function on a rectangle)

In these slides: view the task of target prediction by convnet as reconstruction of hierarchically transformed signal

Hierarchical generative models: an example[1]

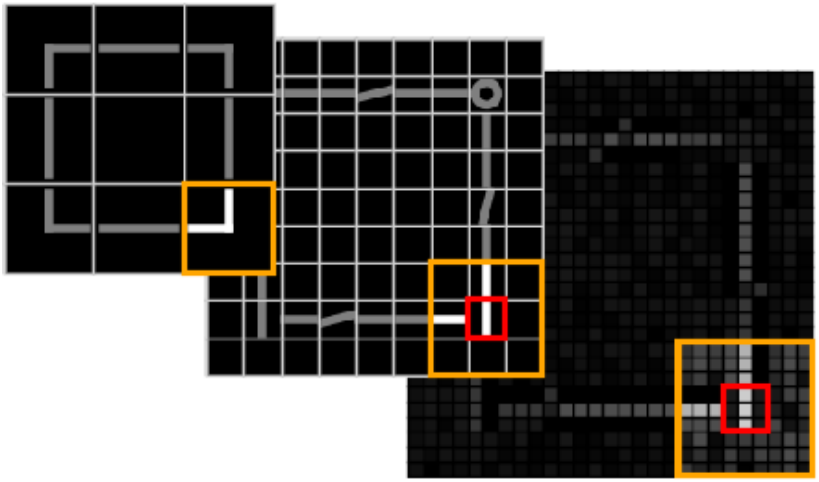
- Consider the image as generated by a hierarchical process
- Pixels subdivided into smaller pixels
- Pixels have states; the states of successor pixels are random but depend on the predecessors

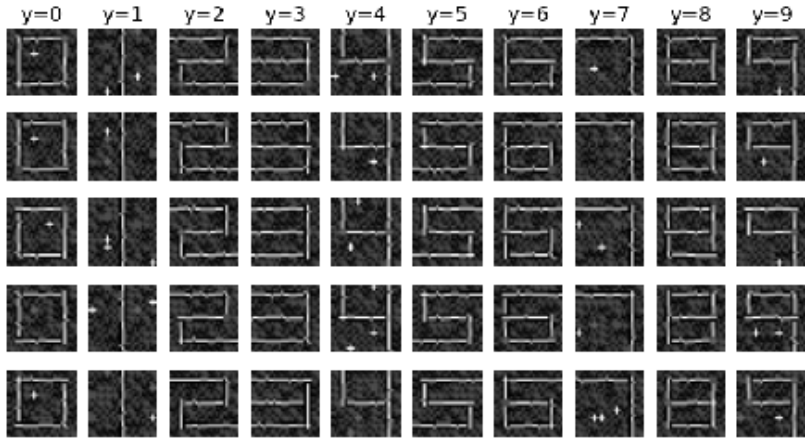


[1] E. Malach, Sh. Shalev-Shwartz, A Provably Correct Algorithm for Deep Learning that Actually Works, arXiv:1803.09522

Example: generation of MNIST-like images

A hierarchy with 3 levels:





Task: reconstruct the original label given the final pixel representation

Simple hierarchical models

Assume level-independent:

- state space S
- branching number b
- transition probabilities $P(\sigma_n|\sigma_{n-1})$

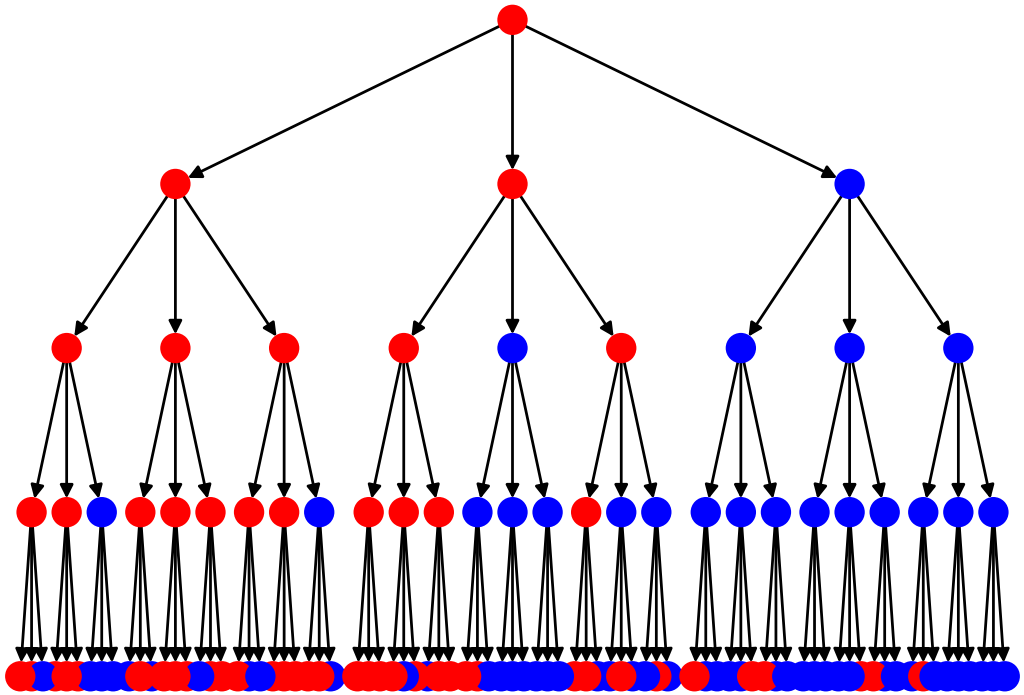
Key advantage: analytically tractable in the infinite-depth limit

The task: reconstruct initial object from a distant hierarchy level

Main example: binary symmetric channel[2]

States ("spins"): ± 1

Transition probabilities: $M = \begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix}$, where ϵ is the error rate



[2] E. Mossel, Survey: Information flow on trees, arXiv:0406446

Optimal (maximum likelihood) reconstruction

σ_0: the spin at the root
$\boldsymbol{\sigma}_n = \{\sigma_{n,p}\}_{p=1}^{N_n}$: the spins at the n'th level

Let $\mathcal{A}$ be a reconstruction algorithm: $\widehat{\sigma}_0 = \mathcal{A}(\boldsymbol{\sigma}_n)$

The probability of correct reconstruction:

$$\begin{aligned} P(\widehat{\sigma}_0 = \sigma_0) &= \sum_{\boldsymbol{\sigma}_n} P(\mathcal{A}(\boldsymbol{\sigma}_n) = \sigma_0 | \boldsymbol{\sigma}_n) P(\boldsymbol{\sigma}_n) \\ &\leq \sum_{\boldsymbol{\sigma}_n} P(\mathcal{A}_{\mathrm{ML}}(\boldsymbol{\sigma}_n) = \sigma_0 | \boldsymbol{\sigma}_n) P(\boldsymbol{\sigma}_n), \end{aligned}$$

where

$$\mathcal{A}_{\mathrm{ML}}(\boldsymbol{\sigma}_n) = \arg\max_{\sigma_0} P(\sigma_0 | \boldsymbol{\sigma}_n)$$

How to efficiently compute $\arg\max_{\sigma_0} P(\sigma_0|\boldsymbol{\sigma}_n)$?

Theorem: For any $\boldsymbol{\sigma}_n$, $\arg\max_{\sigma_0} P(\sigma_0|\boldsymbol{\sigma}_n)$ can be computed with $|S|^2 \cdot O(b^n)$ elementary operations (as $n \to \infty$).

Computation of $\mathcal{A}_{\mathrm{ML}}(\boldsymbol{\sigma}_n)$: Bayes and Markov reductions

Bayes:

$$P(\sigma_0|\boldsymbol{\sigma}_n) = c_{\boldsymbol{\sigma}_n} P(\boldsymbol{\sigma}_n|\sigma_0) P(\sigma_0)$$

Then

$$\mathcal{A}_{\mathrm{ML}}(\boldsymbol{\sigma}_n) = \arg\max_{\sigma_0} (P(\boldsymbol{\sigma}_n|\sigma_0) P(\sigma_0))$$

Markov:

$$P(\boldsymbol{\sigma}_n|\sigma_0) = \sum_{\boldsymbol{\sigma}_1 \in S^{N_1}} \cdots \sum_{\boldsymbol{\sigma}_{n-1} \in S^{N_{n-1}}} \prod_{k=1}^{n} P(\boldsymbol{\sigma}_k|\boldsymbol{\sigma}_{k-1})$$

Here, S is the set of spin states ($|S| = 2$ for binary channel), and N_k is the number of spins in the k'th level of the hierarchy ($N_k = b^k$)

Huge complexity of direct summation: $O(\prod_{k=1}^{n} |S|^{N_k})$, i.e. $O(|S|^{\sum_{k=1}^{n} b^k})$

Transfer matrix reduction

Consider the $|S|^{N_k} \times |S|^{N_{k'}}$ transfer matrices

$$T_{kk'} = [P(\boldsymbol{\sigma}_k|\boldsymbol{\sigma}_{k'})] \qquad (k' < k)$$

Then for any $k'' < k' < k$

$$T_{kk''} = T_{kk'} T_{k'k''}$$

(in the usual matrix product sense), and in particular

$$T_{n,0} = T_{n,n-1} T_{n-1,n-2} \cdots T_{1,0}$$

Collect the desired probabilities $(P(\boldsymbol{\sigma}_n|\sigma_0 = s))_{s\in S}$ into a vector

$$\mathbf{v}_0 = (P(\boldsymbol{\sigma}_n|\sigma_0 = s_1), \ldots, P(\boldsymbol{\sigma}_n|\sigma_0 = s_{|S|}))^t$$

Then

$$\mathbf{v}_0^t = \mathbf{v}_n^t T_{n,n-1} T_{n-1,n-2} \cdots T_{1,0},$$

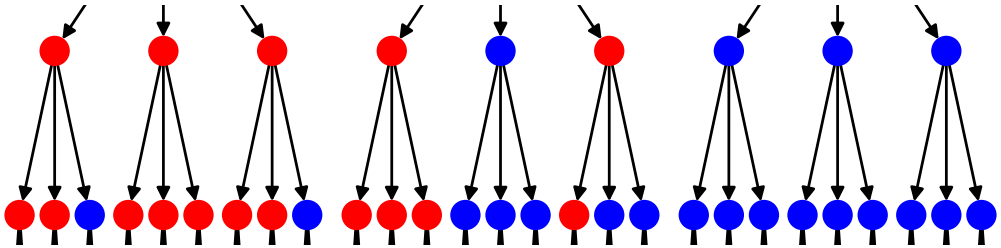
where

$$\mathbf{v}_n^t = (0, \ldots, 0, \underset{\boldsymbol{\sigma}_n}{1}, 0, \ldots, 0) \in \mathbb{R}^{|S|^{N_n}}$$

Factorization of $T_{k,k-1}$

Thanks to independence of parallel tree branches, $T_{k+1,k}$ decouples into a tensor product of N_k transfer matrices of size $|S|^b \times |S|$ acting in respective subtrees:

$$T_{k+1,k} = \otimes_{i=1}^{N_k} T^{(i)}_{k+1,k}$$



$(N_k = 9, S = \{\pm 1\}, b = 3)$

Let $\boldsymbol{\sigma}^{(i)}_{k+1} = (\sigma^{(j_1)}_{k+1}, \ldots, \sigma^{(j_b)}_{k+1}) \in S^b$ be the successors of the spin $\sigma^{(i)}_k$. Then the matrix elements of $T^{(i)}_{k+1,k}$ indexed by $\boldsymbol{\sigma}^{(i)}_{k+1}$ and $\sigma^{(i)}_k$ equals

$$(T^{(i)}_{k+1,k})_{\boldsymbol{\sigma}^i_{k+1},\sigma^i_k} = \prod_{p=1}^{b} P(\sigma^{(j_p)}_{k+1}|\sigma^{(i)}_k)$$

Factorization of intermediate vectors

The vector $\mathbf{v}_n = (0, \ldots, 0, 1, 0, \ldots, 0)^t \in \mathbb{R}^{|S|^{N_n}}$ defining a particular final configuration $\boldsymbol{\sigma}_n$ is itself factorizable as a tensor product:

$$\mathbf{v}_n = \otimes_{i=1}^{N_n} \mathbf{v}_n^{(i)}, \quad \mathbf{v}_n^{(i)} = \mathbf{e}_{\sigma_n^{(i)}} \in \mathbb{R}^{|S|}$$

Then, thanks to factorization of $T_{k,k-1}$, all intermediate products $\mathbf{v}_k^t = \mathbf{v}_n^t T_{n,n-1} T_{n-1,n-2} \cdots T_{k+1,k}$ are also factorizable:

$$\mathbf{v}_k = \otimes_{i=1}^{N_k} \mathbf{v}_k^{(i)}, \quad \mathbf{v}_k^{(i)} \in \mathbb{R}^{|S|}$$

Then computation of $\mathbf{v}_0 = (P(\boldsymbol{\sigma}_n|\sigma_0 = s_1), \ldots, P(\boldsymbol{\sigma}_n|\sigma_0 = s_{|S|}))$ reduces to computation of all the factor vectors $\mathbf{v}_k^{(i)}$ on the tree:

$$\mathbf{v}_k^{(i)} = \left[\bigotimes_{j \in \text{successors}(i)} \mathbf{v}_{k+1}^{(j)} \right] T_{k+1,k}^{(i)}$$

Computation of $\mathbf{v}_k^{(i)}$

Let $\mathbf{v}_k^{(i)} = (v_{k;1}^{(i)}, \ldots, v_{k;|S|}^{(i)})$. Then

$$
\begin{aligned}
v_{k;s}^{(i)} &= \left(\left[\bigotimes_{j \in \text{successors}(i)} \mathbf{v}_{k+1}^{(j)} \right] T_{k+1,k}^{(i)} \right)_s \\
&= \sum_{s_1=1}^{|S|} \cdots \sum_{s_b=1}^{|S|} \prod_{p=1}^{b} \left[P(s_p|s) v_{k+1;s_p}^{(j_p)} \right] \\
&= \prod_{p=1}^{b} \left[\sum_{s_p=1}^{|S|} P(s_p|s) v_{k+1;s_p}^{(j_p)} \right]
\end{aligned}
$$

Thus, the total computation of all factor vectors $\mathbf{v}_k^{(i)}$ up to $\mathbf{v}_0$ (and hence of $\mathcal{A}_{\mathrm{ML}}(\boldsymbol{\sigma}_n)$) requires just $|S|^2 \cdot O(b^n)$ operations.

Remark: intermediate rescaling

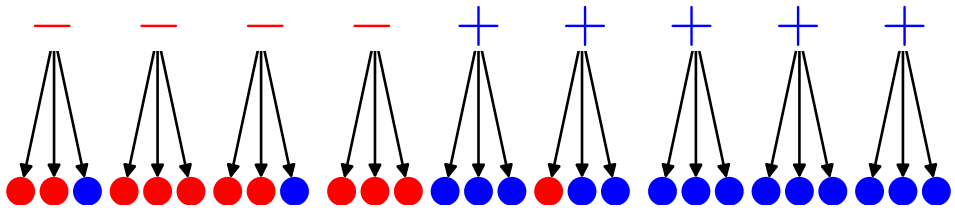
In ML reconstruction we need to know the vector
$\mathbf{v}_0 = (P(\boldsymbol{\sigma}_n|\sigma_0 = s_1), \ldots, P(\boldsymbol{\sigma}_n|\sigma_0 = s_{|S|}))^t$ only up to a constant factor.

Hence, the matrix $T^{(i)}_{k+1,k}$ can be rescaled by any positive constant without affecting the result (e.g., for numerical reasons, to ensure that the magnitude of intermediate vectors does not vanish).

The "local majority" reconstruction

Assume: the binary symmetric channel, odd branching number b.

Reconstruct spins hierarchically, from level n to 0, by choosing the most popular spin value among immediate successors.



Exercise: Local majority reconstruction is the limit of ML reconstruction as error rate $\epsilon \to 0$.

"Global majority" reconstruction

Assume binary symmetric channel.

Global majority: Define $\widehat{\sigma}_0$ as the most popular value among spins of $\boldsymbol{\sigma}_n$

Exercise: Global majority reconstruction is the $\epsilon \nearrow \frac{1}{2}$ limit of ML reconstructions

"Census" reconstructions

Global majority is a special case of "census" reconstructions

"Census" reconstructions: a general class of reconstructions only depending on the number of different states appearing in $\boldsymbol{\sigma}_n$ (i.e., on $(\#(i : \sigma_n^{(i)} = s))_{s \in S}$)

Example: Symmetric binary channel with $\epsilon = 1$: global majority yields a "perfectly bad" reconstruction if depth n is odd, but can be turned into a "perfectly good" census reconstruction by reversing the predicted state.

Reconstruction complexities

Amount of computation required by each reconstruction algorithm at a single node of depth-n tree, assuming S and b fixed:

ML: $O(1)$ arithmetic operations

Local majority: $O(1)$ boolean operations

Global majority: $O(1)$ arithmetic or $O(n)$ boolean operations (to compute the partial sums for census)

Optimal $\boldsymbol{\sigma}_n$-independent (trivial) reconstruction

Suppose we know the distribution of the root spin, $(P(\sigma_0))_{\sigma_0 \in S}$, but the reconstruction algorithm $\mathcal{A}$ has no information about spins in a particular flow realization.

Then the optimal reconstruction is

$$\mathcal{A}_{\text{opt}} = \arg\max_{\sigma_0} P(\sigma_0)$$

and has success probability $P(\widehat{\sigma}_0 = \sigma_0) = \max_{\sigma_0} P(\sigma_0)$.

If $P(\sigma_0) \equiv \frac{1}{|S|}$ for all σ_0, then all $\boldsymbol{\sigma}_n$-independent $\mathcal{A}$ are equally efficient.

The reconstruction solvability problem

Does success probability $P(\mathcal{A}_{\mathrm{ML}}(\boldsymbol{\sigma}_n) = \sigma_0)$ of ML reconstruction converge to the success probability of the (trivial) optimal $\boldsymbol{\sigma}_n$-independent reconstruction as $n \to \infty$?

(We will see that for binary symmetric channel the answer is:

- No (i.e, reconstruction is **solvable**), if error rate ϵ is small and branching number b large,
- Yes (i.e, reconstruction is **unsolvable**), if error rate is large and branching number is small.)

Exercise: The success probability of $\mathcal{A}_{\mathrm{ML}}$ is monotone non-increasing in n.

Some equivalent formulations of reconstruction solvability

- Conditional distributions $(P(\boldsymbol{\sigma}_n|\sigma_0 = s))_{s\in S}$ remain separated in total variation distance as $n \to \infty$
- Mutual information between σ_0 and $\boldsymbol{\sigma}_n$ remain separated from 0 as $n \to \infty$

Reconstruction solvability in terms of total variation distance

Let $\mathbf{P}_n^s$ be the distribution on the set of level-n spin configurations $\{\boldsymbol{\sigma}_n\}$ defined by

$$\mathbf{P}_n^s(\boldsymbol{\sigma}_n) = P(\boldsymbol{\sigma}_n | \sigma_0 = s)$$

Exercise: Let $S = \{s_1, s_2\}$ and $P(\sigma_0 = s_1) = P(\sigma_0 = s_2) = \frac{1}{2}$. Then

$$P(\mathcal{A}_{\mathrm{ML}}(\boldsymbol{\sigma}_n) = \sigma_0) = \tfrac{1}{2} + \tfrac{1}{4}\|\mathbf{P}_n^{s_1} - \mathbf{P}_n^{s_2}\|_1,$$

where

$$\|\mathbf{P}_n^{s_1} - \mathbf{P}_n^{s_2}\|_1 = \sum_{\boldsymbol{\sigma}_n} |\mathbf{P}_n^{s_1}(\boldsymbol{\sigma}_n) - \mathbf{P}_n^{s_2}(\boldsymbol{\sigma}_n)|$$

is (twice) the total variation distance between $\mathbf{P}_n^{s_1}$ and $\mathbf{P}_n^{s_2}$.

Therefore, reconstruction is not solvable $\Leftrightarrow \|\mathbf{P}_n^{s_1} - \mathbf{P}_n^{s_2}\|_1 \to 0$ as $n \to \infty$.

Reconstruction solvability in terms of total variation distance

Exercise: Derive a similar relation between success probability and variational distance in the case of the initial spin distribution with $P(\sigma_0 = s_1) \neq P(\sigma_0 = s_2)$.

Reconstruction solvability in terms of mutual information

Entropy of a discrete random variable X taking values x with probabilities $p(x)$:

$$H(X) = -\sum_x p(x) \log_2 p(x)$$

Conditional entropy:

$$H(X|Y) = \sum_y p_Y(y) H(X|Y=y) = -\sum_y p_Y(y) \sum_x p_X(x|y) \log_2 p_X(x|y)$$

Exercise: $H(X|Y) = H(X,Y) - H(Y)$
Exercise: $H(X|Y) \leq H(X)$ and

$$\max(H(X), H(Y)) \leq H(X,Y) \leq H(X) + H(Y)$$

Mutual information:

$$I(X;Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X,Y)$$

Exercise: What is $I(X;Y)$ if X is independent of Y? If $X = Y$? Given X, which Y's maximize $I(X;Y)$?

Reconstruction solvability in terms of mutual information

Assume that $S = \{s_1, s_2\}$ and $P(\sigma_0 = s_1) = P(\sigma_0 = s_2) = \frac{1}{2}$.
Let $X = \sigma_0$ and $Y = \boldsymbol{\sigma}_n$. Then

$$I(\sigma_0; \boldsymbol{\sigma}_n) = H(\sigma_0) - H(\sigma_0|\boldsymbol{\sigma}_n) = \sum_{\boldsymbol{\sigma}'_n \in S^{N_n}} P(\boldsymbol{\sigma}'_n)(H(\sigma_0) - H(\sigma_0|\boldsymbol{\sigma}_n = \boldsymbol{\sigma}'_n)),$$

where

$$\begin{aligned} H(\sigma_0) &= -P(\sigma_0 = s_1)\log_2 P(\sigma_0 = s_1) - P(\sigma_0 = s_2)\log_2 P(\sigma_0 = s_2) \\ &= 1 \\ H(\sigma_0|\boldsymbol{\sigma}_n = \boldsymbol{\sigma}'_n) &= -P(\sigma_0 = s_1|\boldsymbol{\sigma}_n = \boldsymbol{\sigma}'_n)\log_2 P(\sigma_0 = s_1|\boldsymbol{\sigma}_n = \boldsymbol{\sigma}'_n) \\ &\quad - P(\sigma_0 = s_2|\boldsymbol{\sigma}_n = \boldsymbol{\sigma}'_n)\log_2 P(\sigma_0 = s_2|\boldsymbol{\sigma}_n = \boldsymbol{\sigma}'_n) \\ &\leq 1 \end{aligned}$$

Exercise: $I(\sigma_0; \boldsymbol{\sigma}_n) \to 0 \Leftrightarrow \|\mathbf{P}_n^{s_1} - \mathbf{P}_n^{s_2}\|_1 \to 0$ as $n \to \infty$

The q-state Potts channel

q-state **Potts channel**: $|S| = q$ and

$$P(\sigma_1|\sigma_0) = \begin{cases} 1-(q-1)\epsilon, & \sigma_1 = \sigma_0 \\ \epsilon, & \sigma_1 \neq \sigma_0 \end{cases}$$

Binary symmetric channel: Potts with $q = 2$

The eigenvalues of matrix of transition probabilities

Consider binary symmetric channel with the transition matrix

$$M = \begin{pmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{pmatrix}$$

Eigenvalues of M :

$$\lambda_1 = 1: \qquad \begin{pmatrix} 1 & 1 \end{pmatrix} M = \lambda_1 \begin{pmatrix} 1 & 1 \end{pmatrix}$$

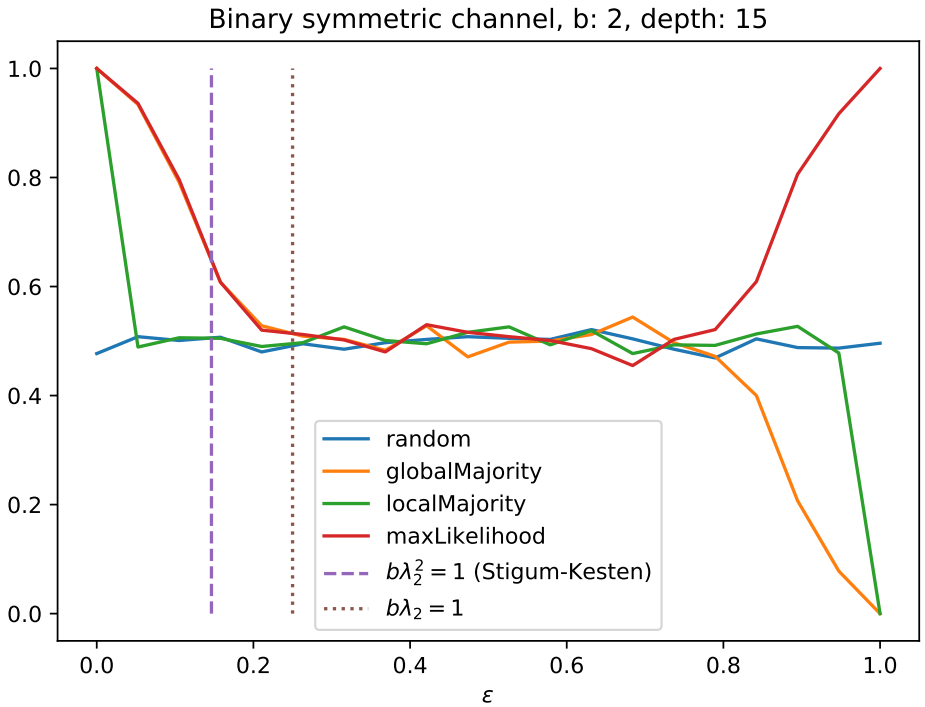
$$\lambda_2 = 1-2\epsilon: \qquad \begin{pmatrix} 1 & -1 \end{pmatrix} M = \lambda_2 \begin{pmatrix} 1 & -1 \end{pmatrix}$$

Exercise: Any matrix of transition probabilities has eigenvalue $\lambda_1 = 1$; for any other eigenvalue λ we have $|\lambda| \leq 1$.

Exercise: The q-state Potts channel has only two eigenvalues, $\lambda_1 = 1$ and $\lambda_2 = 1 - q\epsilon$.

The solvability of binary symmetric and Potts flows

	Binary ($q = 2$)	General Potts
$b < \frac{1}{\lvert\lambda_2\rvert}$	non-solvable	
$\frac{1}{\lvert\lambda_2\rvert} < b < \frac{1}{\lvert\lambda_2\rvert^2}$	non-solvable	solvable for large q
$\frac{1}{\lvert\lambda_2\rvert^2} < b$	solvable by global majority	



Census-solvable reconstruction

Let $\lambda_2 = \lambda_2(M)$ be the second largest eigenvalue of M in absolute value.

Theorem (Kesten-Stigum bound): Under certain nondegeneracy conditions on transition matrix M (it must correspond to an ergodic Markov chain), the reconstruction problem is census-solvable if and only if $b|\lambda_2|^2 > 1$. In particular, for binary symmetric channel, if $b\lambda_2^2 > 1$, then the reconstruction problem is solvable by global majority.

Proof of solvability for binary symmetric channel

Let $R_n = R_n(\boldsymbol{\sigma}_n)$ be the number of $+$ spins minus the number of $-$ spins in $\boldsymbol{\sigma}_n$. We compute the first two moments of R_n w.r.t. $\mathbf{P}_n^{\pm}$. Let $\mathsf{E}^{\pm}$ denote expectation w.r.t. $\mathbf{P}_n^{\pm}$.

$$\begin{aligned}
\mathsf{E}^{+}(R_n) &= \sum_{i=1}^{N_n} \mathsf{E}^{+}(\sigma_n^{(i)}) \\
&= b^n \mathsf{E}^{+}(\sigma_n^{(1)}) \\
&= b^n \begin{pmatrix} 1 & 0 \end{pmatrix} M^n \begin{pmatrix} 1 \\ -1 \end{pmatrix} \\
&= b^n \tfrac{1}{2} \big(\begin{pmatrix} 1 & 1 \end{pmatrix} + \begin{pmatrix} 1 & -1 \end{pmatrix} \big) M^n \begin{pmatrix} 1 \\ -1 \end{pmatrix} \\
&= b^n \lambda_2^n
\end{aligned}$$

Similarly,

$$\mathsf{E}^{-}(R_n) = -\mathsf{E}^{+}(R_n) = -b^n \lambda_2^n$$

The second moment

Given $\sigma_n^{(i)}, \sigma_n^{(j)}$, let m be the distance to their closest common predecessor

$$\begin{aligned}
\mathsf{E}^+(R_n^2) &= \sum_{i=1}^{N_n}\sum_{j=1}^{N_n} \mathsf{E}^+(\sigma_n^{(i)}\sigma_n^{(j)}) \\
&= b^n \sum_{m=0}^{n} (b^m - \delta_{m>0} b^{m-1}) \times \\
&\quad \times \Bigg(P(\sigma_{n-m}=1 | \sigma_0 = 1) \Big[\begin{pmatrix} 1 & 0 \end{pmatrix} M^m \begin{pmatrix} 1 \\ -1 \end{pmatrix} \Big]^2 \\
&\quad + P(\sigma_{n-m}=-1 | \sigma_0 = 1) \Big[\begin{pmatrix} 0 & 1 \end{pmatrix} M^m \begin{pmatrix} 1 \\ -1 \end{pmatrix} \Big]^2 \Bigg) \\
&= b^n \sum_{m=0}^{n} (b^m - \delta_{m>0} b^{m-1}) \lambda_2^{2m} \\
&= \begin{cases} (b\lambda_2)^{2n}(c + o(1)), & b\lambda_2^2 > 1 \\ b^n (c + o(1)), & b\lambda_2^2 < 1 \end{cases} \qquad (0 < c < \infty)
\end{aligned}$$

Lower bound on $\|\mathbf{P}^+ - \mathbf{P}^-\|_1$

Assuming $b\lambda_2^2 > 1$,

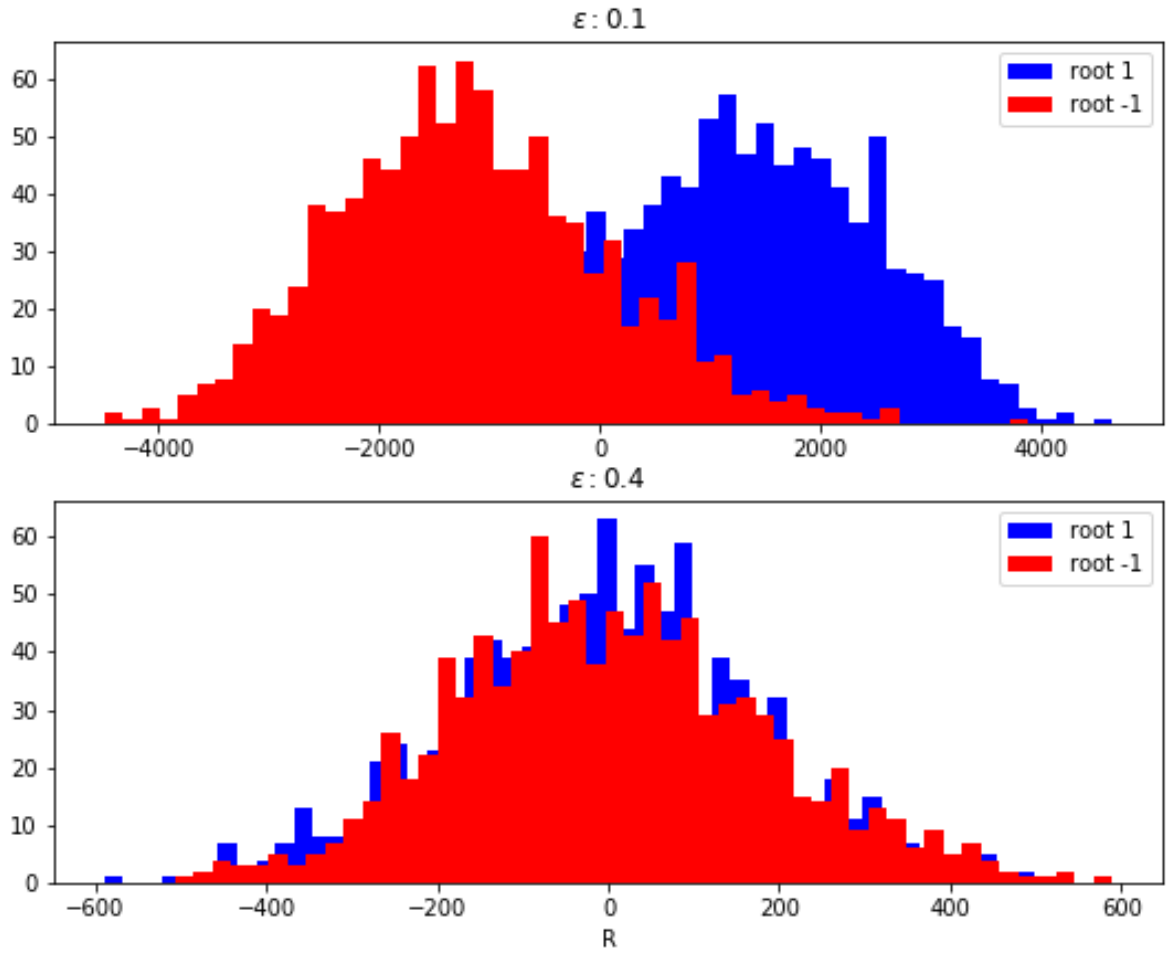
$$
\begin{aligned}
2b^n\lambda_2^n &= \mathsf{E}^+(R_n) - \mathsf{E}^-(R_n) \\
&= \sum_{\boldsymbol{\sigma}_n} R_n(\boldsymbol{\sigma}_n)\big(\mathbf{P}^+(\boldsymbol{\sigma}_n) - \mathbf{P}^-(\boldsymbol{\sigma}_n)\big) \\
&\overset{Cauchy}{\leq} \Big(\sum_{\boldsymbol{\sigma}_n} \big|\mathbf{P}^+(\boldsymbol{\sigma}_n) - \mathbf{P}^-(\boldsymbol{\sigma}_n)\big|\Big)^{1/2} \Big(\sum_{\boldsymbol{\sigma}_n} R_n^2 \big|\mathbf{P}^+(\boldsymbol{\sigma}_n) - \mathbf{P}^-(\boldsymbol{\sigma}_n)\big|\Big)^{1/2} \\
&\leq \|\mathbf{P}^+ - \mathbf{P}^-\|_1^{1/2} (\mathsf{E}^+(R_n^2) + \mathsf{E}^-(R_n^2))^{1/2} \\
&\leq \|\mathbf{P}^+ - \mathbf{P}^-\|_1^{1/2} (c + o(1)) b^n \lambda_2^n,
\end{aligned}
$$

implying

$$\|\mathbf{P}^+ - \mathbf{P}^-\|_1 \geq c + o(1), \quad c > 0$$

Illustration

With $b = 2$, depth 15:



Reconstruction unsolvability for small b/large ϵ

Theorem: For binary symmetric channel, reconstruction is unsolvable if $b|\lambda_2| < 1$.

Transfer matrix: change of variables

The elementary step $T^{(i)}_{k+1,k}$ from ML reconstruction (in a simplified notation):

$$\mathbf{v}_i = \Big[\bigotimes_{j \in \text{successors}(i)} \mathbf{v}_j \Big] T^{(i)}_{k+1,k}, \quad v_{i;s} = \prod_{j=1}^{b} \Big[\sum_{s_j=1}^{|S|} P(s_j|s) v_{j;s_j} \Big]$$

For binary channel: $\mathbf{v} = (v_+, v_-)^t$, with $v_+, v_- \geq 0$. Let us define h by

$$e^{2h} = \frac{v_+}{v_-}$$

Proposition. For the binary symmetric channel, if $\mathbf{v}_i$ is described by h_i and $\mathbf{v}_j$ by h_j, then $T^{(i)}_{k+1,k}$ can be written as

$$h_i = \sum_{j=1}^{b} \operatorname{arctanh}(\lambda_2 \tanh(h_j))$$

Proof of proposition

Using $\operatorname{arctanh} z = \frac{1}{2} \ln \frac{1+z}{1-z}$ and $\lambda_2 = 1 - 2\epsilon$:

$$
\begin{aligned}
h_i &= \frac{1}{2} \ln \frac{v_{i,+}}{v_{i,-}} \\
&= \frac{1}{2} \sum_{j=1}^{b} \ln \frac{(1-\epsilon) v_{j,+} + \epsilon v_{j,-}}{(1-\epsilon) v_{j,-} + \epsilon v_{j,+}} \\
&= \frac{1}{2} \sum_{p=1}^{b} \ln \frac{(1-\epsilon) e^{h_j} + \epsilon e^{-h_j}}{(1-\epsilon) e^{-h_j} + \epsilon e^{+h_j}} \\
&= \frac{1}{2} \sum_{p=1}^{b} \ln \frac{1 + (1-2\epsilon) \tanh(h_j)}{1 - (1-2\epsilon) \tanh(h_j)} \\
&= \sum_{p=1}^{b} \operatorname{arctanh}(\lambda_2 \tanh(h_j))
\end{aligned}
$$

Contractiveness of transfer matrix

Exercise: $|h_i| \le b|\lambda_2| \max_j |h_j|$ (Hint: use $|\operatorname{arctanh}'(\lambda_2 x)| \le |\operatorname{arctanh}'(x)|$ for $|\lambda_2| \le 1$)

Consider any spin configuration in the final n'th level and the corresponding ML reconstruction. Denote h values in the k'th level by $h_k^{(j)}$.

In the final configuration $h_n^{(j)} = \pm\infty$ (why?), but already for the next, $(n-1)$'th level (after one step of transfer matrix)

$$|h_{n-1}^{(j)}| \le b \operatorname{arctanh}(|\lambda_2|) < \infty,$$

since $|\tanh(h)| \le 1$ and $|\lambda_2| < 1$.

Then for the k'th level, by the Exercise,

$$\max_j |h_k^{(j)}| \le |b\lambda_2|^{n-1-k} b \operatorname{arctanh}(|\lambda_2|)$$

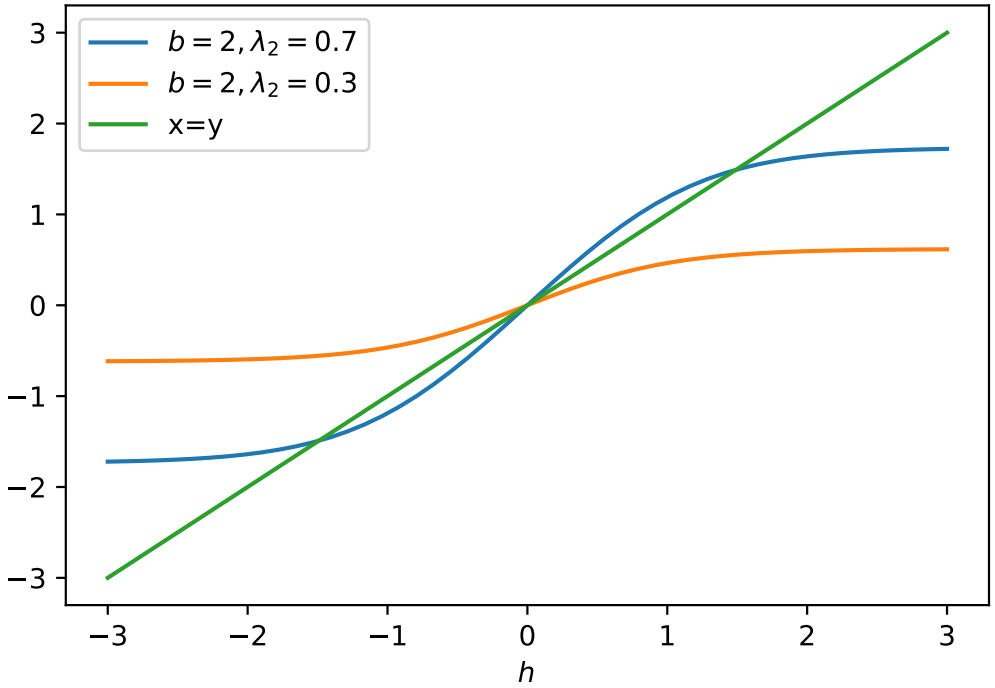
In particular, $|h_0| = O(|b\lambda_2|^n) \overset{n\to\infty}{\longrightarrow} 0$, so $\frac{P(\boldsymbol{\sigma}_n|\sigma_0=+)}{P(\boldsymbol{\sigma}_n|\sigma_0=-)} \to 1$ i.e., reconstruction is unsolvable. □

Nontrivial fixed points

Remark: if $b\lambda_2 > 1$ and all the spins in $\boldsymbol{\sigma}_n$ are ordered so that h_j are equal, then the map

$$h_j \mapsto b \operatorname{arctanh}(\lambda_2 \tanh(h_j))$$

has nontrivial fixed points, and so h can have a nonzero limit. However, this does not hold in general for disordered $\boldsymbol{\sigma}_n$, and so does not imply reconstruction solvability for $b\lambda_2 > 1$.



Solvability at $b|\lambda_2| > 1$ for Potts with $q = \infty$

Theorem (Mossel): For Potts channel, if b and λ_2 are fixed and $b\lambda_2 > 1$, then reconstruction is solvable if q is large enough.

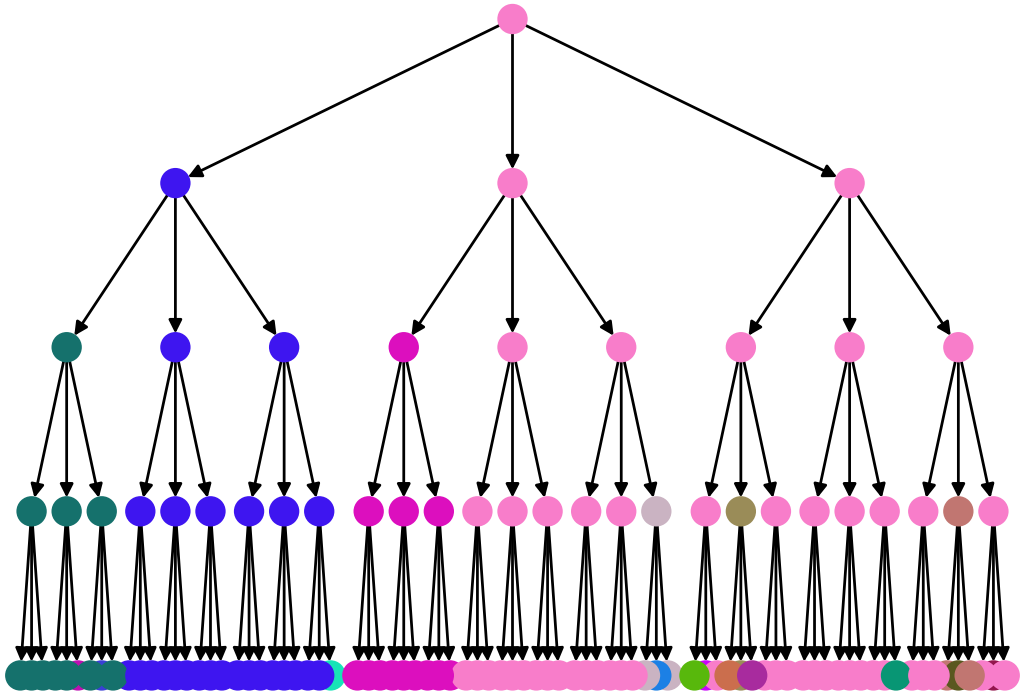
Key idea: consider the limit $q \to \infty$ of Potts channels:

- With probability λ_2, current state is preserved
- Otherwise, the state changes to one of infinitely many alternatives
- If the channel left a state, it cannot reappear later (probability 0)
- Newly appearing states are always different

Reconstruction idea: divide the spins in given σ_n into b subsets A_p descending from the b level-1 spins. Try to find two spins from different A_p having the same state s. Then the root spin state must be s.

To prove: a nonvanishing probability of success. Sufficient to prove: the root spin state indefinitely survives with positive probability.

Example: a tree for $q = \infty$



Proof

Let $U_n = \sum_{i=1}^{N_n} \mathbf{1}(\sigma_n^{(i)} = \sigma_0)$. Sufficient to prove: $P(U_n \geq 1) > c$ for some n-independent $c > 0$.

Similarly to the proof of Kesten-Stigum bound,

$$\mathrm{E}(U_n) = b^n \lambda_2^n$$

$$\mathrm{E}(U_n^2) = b^n \sum_{m=0}^{n} (b^m - \delta_{m>0} b^{m-1}) \lambda_2^{n+m}$$

$$= \begin{cases} (b\lambda_2)^{2n}(c + o(1)), & b\lambda_2 > 1 \\ (b\lambda_2)^n (c + o(1)), & b\lambda_2 < 1 \end{cases} \qquad (0 < c < \infty)$$

From $\mathrm{E}(U_n) = \mathrm{E}(\mathbf{1}(U_n \geq 1) \cdot U_n) \leq (P(U_n \geq 1))^{1/2} (\mathrm{E}(U_n^2))^{1/2}$:

$$P(U_n \geq 1) \geq c > 0, \quad (b\lambda_2 > 1)$$

Another proof

Exercise:

- Give another proof (that the survival probability is not vanishing iff $b|\lambda_2| > 1$) using the transfer matrix approach and the factorization $\mathbf{1}(\sigma_n^{(i)} \neq \sigma_0 \ \forall i) = \otimes_i \mathbf{1}(\sigma_n^{(i)} \neq \sigma_0)$ in the last layer.
- Compute explicitly the $n \to \infty$ limit of the survival probability for $b = 2$, by finding the nontrivial fixed point of the layer-to-layer transformation.

We have seen that recurrence relations answer the solvability problem for a fixed constant final layer configuration.

Can we extend this approach to final layer configurations randomly generated according to transition probabilities?

Yes, but at the cost of infinitely dimensional objects in the recurrence.

The distributions Q_n^s

Define the distributions[3]

$$Q_n^s = \sum_{\boldsymbol{\sigma}_n} P(\boldsymbol{\sigma}_n | \sigma_0 = s) \delta_{\mathbf{p}_{\boldsymbol{\sigma}_n}},$$

where

$$\begin{aligned} \mathbf{p}_{\boldsymbol{\sigma}_n} &= \left(P(\sigma_0 = r | \boldsymbol{\sigma}_n)\right)_{r=1}^{|S|} \\ &= \left(Z^{-1} P(\boldsymbol{\sigma}_n | \sigma_0 = r)\right)_{r=1}^{|S|} \in \mathbb{R}^{|S|} \end{aligned}$$

(assuming the uniform initial spin distribution $P(\sigma_0 = r) = \frac{1}{|S|}$)

The vectors $\mathbf{p}_{\boldsymbol{\sigma}_n}$ live in the $(|S| - 1)$-dimensional simplex

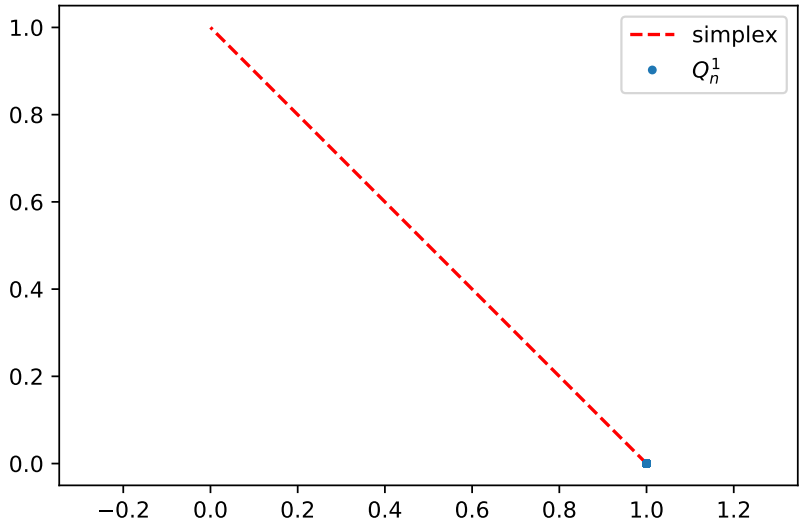
$$\mathbf{p} : \sum_{s=1}^{|S|} p_s = 1, p_s \geq 0$$

[3] M. Mezard, A. Montanari, Reconstruction on trees and spin glass transition, arXiv:cond-mat/0512295

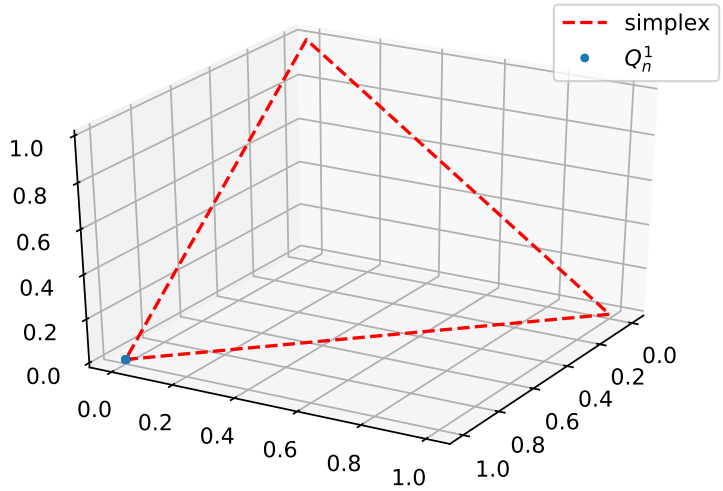
Examples: initial distribution for q-state Potts

Initial distribution $Q^s_{n=0}$ is concentrated at a single point $(1, 0, \ldots, 0)$:

$q = 2$



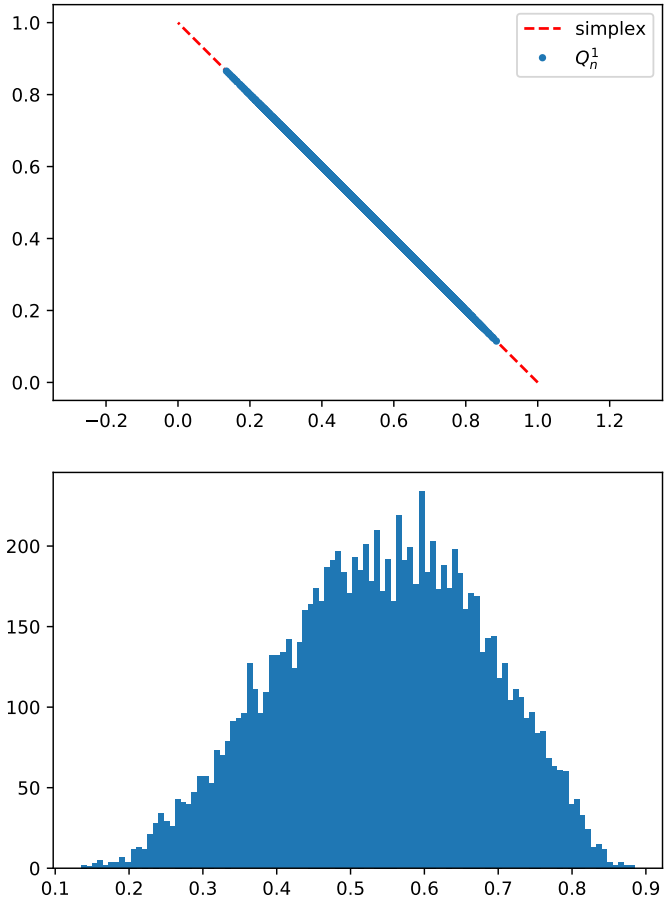
$q = 3$



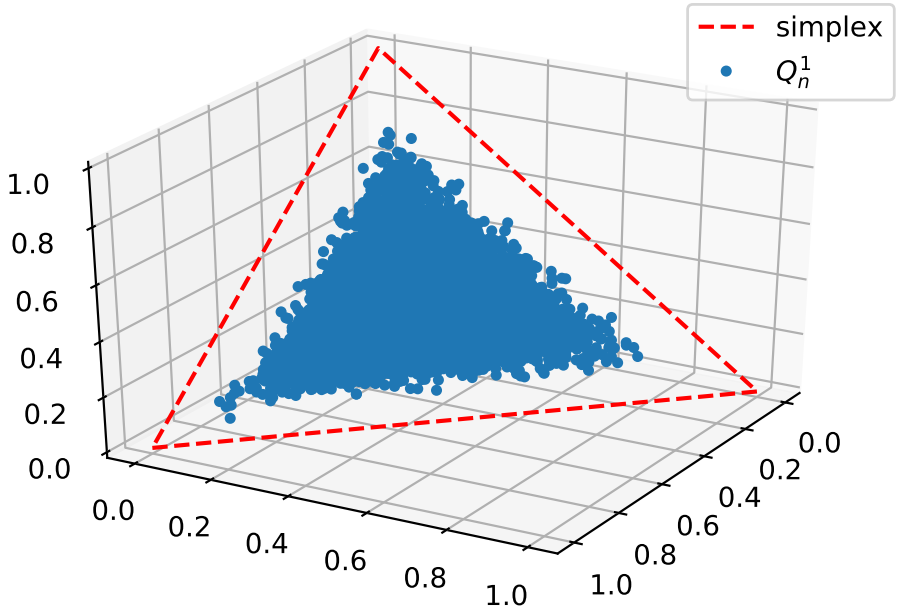
Examples: limit distributions for q-state Potts

Limiting distributions $\lim_{n\to\infty} Q_n^s$ for ϵ **below** critical KS value:

$q = 2$



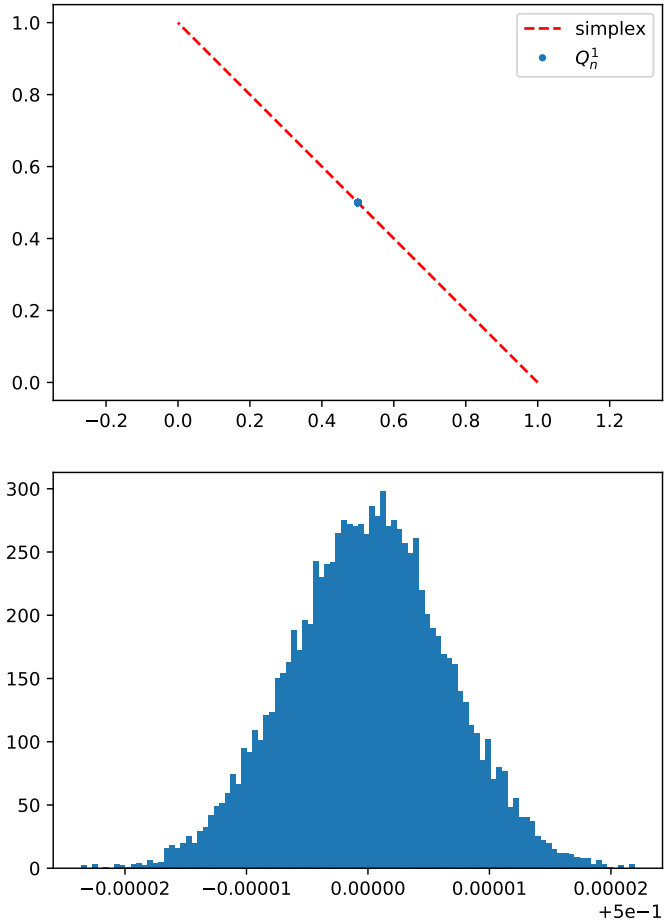
$q = 3$



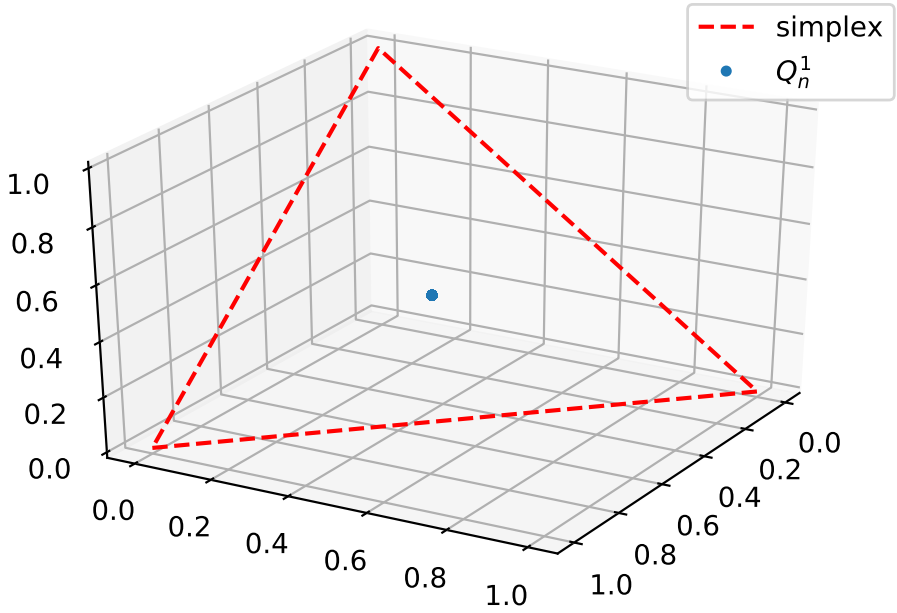
Examples: limit distributions for q-state Potts

Limiting distributions $\lim_{n\to\infty} Q_n^s$ for ϵ **above** critical KS value:

$q = 2$



$q = 3$



Properties of Q_n^s

Exercise: Let $Q_n = \frac{1}{|S|}\sum_{s=1}^{|S} Q_n^s$.

- Show how to reconstruct Q_n^s for each s from Q_n.
- Prove that $\int \mathbf{p} dQ_n = \mathbf{p}_* = (\frac{1}{|S|}, \ldots, \frac{1}{|S|})$

Exercise: Reconstruction non-solvability is equivalent to $\lim_{n\to\infty} Q_n = \delta_{\mathbf{p}_*}$ and to $\lim_{n\to\infty} Q_n^s = \delta_{\mathbf{p}_*}$[4]

[4] Convergence $\lim_{n\to\infty} Q_n = \delta_{\mathbf{p}_*}$ is understood in the sense that for any open $U \ni \mathbf{p}_*$ we have $\lim_{n\to\infty} Q_n(U) = 1$ and $\lim_{n\to\infty} Q_n(\mathbb{R}^{|S|} \setminus U) = 0$

The recurrence relation

Theorem. There is map Φ such that

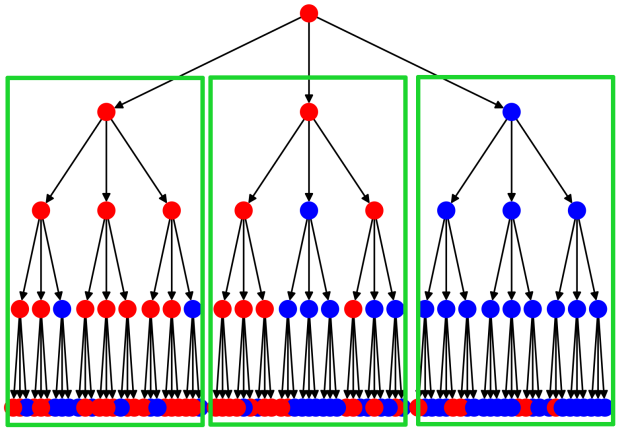
$$\{Q^s_{n+1}\}_{s\in S} = \Phi(\{Q^s_n\}_{s\in S})$$

Fixed points $\{Q^s\}$ of Φ :

- $Q^s = \delta_{\mathbf{p}_*}$ with $\mathbf{p}_* = (\frac{1}{|S|}, \ldots, \frac{1}{|S|})$ is a (trivial) fixed point of Φ corresponding to non-solvable reconstruction
- Non-trivial fixed points of Φ correspond to solvable reconstructions

The recurrence relation: first step

Represent $\boldsymbol{\sigma}_{n+1} = (\boldsymbol{\sigma}_{n+1;1}, \ldots, \boldsymbol{\sigma}_{n+1;b})$



$$
\begin{aligned}
Q^s_{n+1} &= \sum_{\boldsymbol{\sigma}_{n+1}} P(\boldsymbol{\sigma}_{n+1}|\sigma_0 = s)\delta_{\mathbf{p}_{\boldsymbol{\sigma}_{n+1}}} \\
&= \sum_{\mathbf{s}_1} P(\boldsymbol{\sigma}_1 = \mathbf{s}_1|\sigma_0 = s) \sum_{\boldsymbol{\sigma}_{n+1}} P(\boldsymbol{\sigma}_{n+1}|\boldsymbol{\sigma}_1 = \mathbf{s}_1)\delta_{\mathbf{p}_{\boldsymbol{\sigma}_{n+1}}} \\
&= \sum_{s_{1;1}=1}^{|S|} \cdots \sum_{s_{1;b}=1}^{|S|} \prod_{i=1}^{b} P(s_{1;i}|s) \times \\
&\quad \times \sum_{\boldsymbol{\sigma}_{n+1;1}} \cdots \sum_{\boldsymbol{\sigma}_{n+1;b}} \prod_{i=1}^{b} P(\boldsymbol{\sigma}_{n+1;i}|\sigma_{1;i} = s_{1;i})\delta_{\mathbf{p}_{\boldsymbol{\sigma}_{n+1}}}
\end{aligned}
$$

The recurrence relation for $\mathbf{p}_{\boldsymbol{\sigma}_{n+1}}$

$$
\begin{aligned}
(\mathbf{p}_{\boldsymbol{\sigma}_{n+1}})_s &= P(\sigma_0 = s|\boldsymbol{\sigma}_{n+1}) \\
&= \frac{P(\boldsymbol{\sigma}_{n+1}|\sigma_0 = s)}{\sum_{s'} P(\boldsymbol{\sigma}_{n+1}|\sigma_0 = s')} \\
&= \frac{\sum_{\mathbf{r}_1} P(\boldsymbol{\sigma}_1 = \mathbf{r}_1|\sigma_0 = s) P(\boldsymbol{\sigma}_{n+1}|\boldsymbol{\sigma}_1 = \mathbf{r}_1)}{\sum_{s'} \sum_{\mathbf{r}_1} P(\boldsymbol{\sigma}_1 = \mathbf{r}_1|\sigma_0 = s') P(\boldsymbol{\sigma}_{n+1}|\boldsymbol{\sigma}_1 = \mathbf{r}_1)} \\
&= \frac{\sum_{r_{1;1}=1}^{|S|} \cdots \sum_{r_{1;b}=1}^{|S|} \prod_{i=1}^{b} P(r_{1;i}|s) (\mathbf{p}_{\boldsymbol{\sigma}_{n+1;i}})_{r_{1;i}}}{\sum_{s'} \sum_{r_{1;1}=1}^{|S|} \cdots \sum_{r_{1;b}=1}^{|S|} \prod_{i=1}^{b} P(r_{1;i}|s') (\mathbf{p}_{\boldsymbol{\sigma}_{n+1;i}})_{r_{1;i}}} \\
&= F_s(\mathbf{p}_{\boldsymbol{\sigma}_{n+1;1}}, \ldots, \mathbf{p}_{\boldsymbol{\sigma}_{n+1;b}})
\end{aligned}
$$

The recurrence relation: final form

Summarizing:

$$Q^s_{n+1} = \sum_{s_{1;1}=1}^{|S|} \cdots \sum_{s_{1;b}=1}^{|S|} \prod_{i=1}^{b} P(s_{1;i}|s) \times$$

$$\times \sum_{\boldsymbol{\sigma}_{n+1;1}} \cdots \sum_{\boldsymbol{\sigma}_{n+1;b}} \prod_{i=1}^{b} P(\boldsymbol{\sigma}_{n+1;i}|\sigma_{1;i} = s_{1;i}) \delta_{F(\mathbf{p}_{\boldsymbol{\sigma}_{n+1;1}}, \ldots, \mathbf{p}_{\boldsymbol{\sigma}_{n+1;b}})}$$

$$= \sum_{s_{1;1}=1}^{|S|} \cdots \sum_{s_{1;b}=1}^{|S|} \int \cdots \int \Big[\prod_{i=1}^{b} P(s_{1;i}|s) \Big] \times$$

$$\times \delta_{F(\mathbf{p}_{\boldsymbol{\sigma}_{n+1;1}}, \ldots, \mathbf{p}_{\boldsymbol{\sigma}_{n+1;b}})} \prod_{i=1}^{b} dQ_n^{s_{1;i}}(\mathbf{p}_{\boldsymbol{\sigma}_{n+1;i}})$$

Some further interesting results

- (A partially proved conjecture of Mezard-Montanary.[5]) For solvability of Potts model, the Kesten–Stigum bound is tight
  - iff $q \leq 4$ in the ferromagnetic case ($\lambda_2 > 0$)
  - iff $q \leq 3$ in the anti-ferromagnetic case ($\lambda_2 < 0$)
- Any tree-based reconstruction algorithm restricted to $O(1)$ boolean computation at each tree node has critical solvability threshold ϵ strictly below the Kesten-Stigum value.[6]

[5] A. Sly, Reconstruction for the Potts model (2011)

[6] V. Jain et al., Accuracy-Memory Tradeoffs and Phase Transitions in Belief Propagation, arXiv:1905.10031

Connection to the Ising model

The Ising model (with spins ± 1) on a graph with vertices V and edges E: defined by the Gibbs measure

$$P(\boldsymbol{\sigma}) = Z^{-1} e^{\beta \sum_{(u,v)\in E} \sigma_u \sigma_v}, \quad \boldsymbol{\sigma} = (\sigma_u)_{u\in V}$$

Exercise: Show that the probability distribution of broadcasting with the binary symmetric channel with error ϵ is identical to the Ising model on the tree with empty boundary conditions, with $\frac{1-\epsilon}{\epsilon} = e^{2\beta}$ (i.e., $\epsilon = \frac{1}{1+e^{2\beta}}$)

Non-uniqueness of Gibbs measure

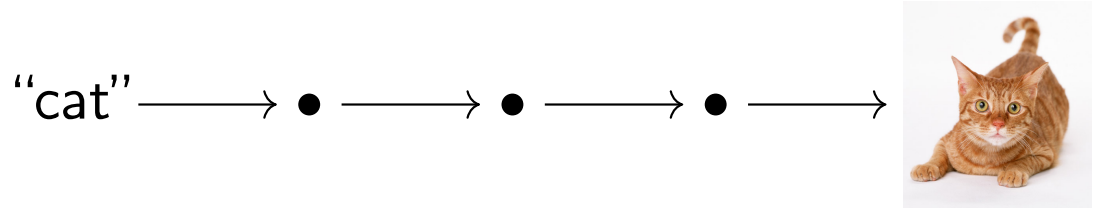
In general, the Gibbs measure is non-unique for different b.c.

Exercise: Consider finite-volume ground states (Gibbs measures corresponding to the limit $\beta \to +\infty$) of the Ising model with different boundary conditions. Perform the infinite-volume limit and show that, depending on boundary conditions, we obtain different infinite-volume ground states.

Theorem. Solvability of the hierarchical reconstruction problem is equivalent to the non-uniqueness of the Gibbs measure.

General lessons?

Assume objects are randomly generated from labels:



- If the distributions $P(\text{object}|\text{label})$ are known, then the optimal label prediction is $\mathcal{A}_{\text{ML}}(\text{object}) = \arg\max_{\text{label}} P(\text{label}|\text{object})$
- If the generation process is a Markov chain, then $\mathcal{A}_{\text{ML}}$ is a multi-stage computation retracing the chain backward (deep models are useful!)
- If the generation process is simple/low-noise, a shallow model may work well (recall "global majority")
- More complex/higher noise cases may require deeper methods (recall the spin position-aware reconstruction for $q \gg 1$, $\frac{1}{\lambda_2} < b < \frac{1}{\lambda_2^2}$)
- At even higher complexity/noise levels, prediction may be unfeasible

Some interesting topics for further research

- Learning the reconstruction algorithm? E.g.:
  - How efficiently does gradient descent learn ϵ in the ML reconstruction at large depths?
  - Any natural, practically relevant extensions of the hierarchical model by learnable parameters?
- Broadcasting with interacting spins and complex transition rules – phase diagrams?
- Critical behavior? (E.g., critical exponents for the reconstruction probability near the critical point?)