

Finite Likelihood Traps on Connected Boltzmann Trees

Alper Ferudun*

17 October 2026

Abstract

We construct finite nonglobal local maxima of exact empirical likelihood for a binary restricted Boltzmann machine with one hidden unit and three visible units. All seven weights and biases are free, the data have full support, and the corresponding two-component mixture has distinct components with equal positive mixing weights. We give the exact transition between a local maximum and a saddle along a uniform-law parameter fiber, including the degenerate threshold. A conditional likelihood argument joins the examples at a shared visible vertex: for every k , a connected allowed tree with k hidden vertices has at least 2^k distinct local-maximum visible laws, all represented by finite parameters. All but one of the displayed laws are non-global, with an explicit likelihood deficit. The local maxima are nonstrict and some allowed edge weights vanish. The construction separates exact tree inference from benign optimization in raw parameter space; it does not give a complexity lower bound or resolve the general sparse-RBM training question.

1 The model and the connected construction

A binary restricted Boltzmann machine (RBM) has visible and hidden spins in $\{-1, 1\}$ and a bipartite allowed connectivity graph. Each allowed edge weight and each vertex bias is an independent finite real parameter. Zero edge weights are permitted and are not removed from the parameter space. For an empirical law q , our objective is the exact, unregularized normalized log likelihood

$$\ell_q(\theta) = \mathbb{E}_q \log p_\theta(X),$$

where p_θ is the visible marginal. A local maximum always means a local maximum in the full Euclidean space of these raw parameters. This is different from local maximality in the set of represented probability distributions with its ambient simplex topology.

For $k \geq 1$, let T_k have a visible center X_0 and, for each $j \in \{1, \dots, k\}$, a hidden vertex H_j adjacent to X_0 and to two private visible leaves Y_j, Z_j . Thus T_k is a connected tree with k hidden vertices, $2k + 1$ visible vertices, and $3k$ allowed edges. Its parameter space has dimension $6k + 1$. Set

$$\rho = \frac{1}{4}, \quad t = \frac{\log 3}{2}, \quad q_k(x_0, y, z) = \frac{1}{2} \prod_{j=1}^k \frac{1 + \rho y_j z_j}{4}, \quad (1)$$

and define

$$\Delta(\rho) = \frac{(1 + \rho) \log(1 + \rho) + (1 - \rho) \log(1 - \rho)}{2}. \quad (2)$$

*Mercury Software GmbH. Email: alper@mercurycodelab.com. GitHub: [AlperTheKing](https://github.com/AlperTheKing). Unrefereed preprint. CC BY 4.0.

In particular $\Delta(1/4) > 0$.

Theorem 1. *For every subset $S \subseteq \{1, \dots, k\}$, set all biases to zero and assign the weights on $(X_0 H_j, Y_j H_j, Z_j H_j)$ to*

$$(0, t, t) \quad (j \in S), \quad (t, 0, 0) \quad (j \notin S).$$

Each resulting parameter is a local maximum of ℓ_{q_k} in $\mathbb{R}^{6k+1}$. Their 2^k visible laws are distinct. Exactly one of these displayed laws equals q_k , and every other one is nonglobal, with deficit $(k - |S|)\Delta(1/4)$ from the global maximum. The full-support law q_k is an empirical distribution of sample size $2 \cdot 16^k$.

The graph is connected as an *allowed architecture*; the theorem does not assert that the subgraph of nonzero weights is connected at these extrema. All those zero-weight coordinates remain free in the neighborhood assertion. Nor does the theorem classify every local maximum: 2^k is a lower bound on distinct displayed laws, not a count of hidden-label copies.

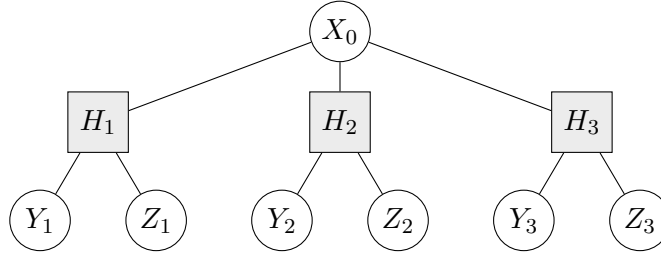


Figure 1: The allowed tree T_3 . The shared visible center makes this a connected construction. Conditional on that center, the three branches separate, which is the basis of the neighborhood proof.

2 The three-leaf threshold

Consider one hidden spin h and three visible spins $x = (x_1, x_2, x_3)$, with joint law proportional to

$$\exp\{a \cdot x + ch + h(rx_1 + ux_2 + vx_3)\}. \quad (3)$$

Write $\theta = (a_1, a_2, a_3, c, r, u, v) \in \mathbb{R}^7$. Summing out h gives

$$p_\theta(x) = \frac{e^{a \cdot x} 2 \cosh(c + rx_1 + ux_2 + vx_3)}{Z(\theta)}. \quad (4)$$

For $0 < |\rho| < 1$, use the full-support target

$$q_\rho(x) = \frac{1 + \rho x_2 x_3}{8}. \quad (5)$$

For rational ρ this is an ordinary empirical distribution; the analytic statement below also applies to arbitrary real ρ in this range. A strict saddle has both a positive and a negative Hessian eigenvalue. Inertia is listed as (n_+, n_-, n_0) .

Theorem 2. *At $a = c = u = v = 0$ and $r = t \geq 0$, the visible law is uniform and the parameter is critical for ℓ_{q_ρ} . For $t > 0$ its classification is*

$$\begin{aligned} |\rho| < \sinh^2 t & \quad (0, 5, 2) \quad \text{nonglobal nonstrict local maximum,} \\ |\rho| > \sinh^2 t & \quad (1, 4, 2) \quad \text{strict saddle,} \\ |\rho| = \sinh^2 t & \quad (0, 4, 3) \quad \text{nonstrict saddle.} \end{aligned}$$

At $t = 0$ it is a strict saddle.

2.1 A critical manifold and its normal Hessian

Expand $g(x) = \log \cosh(c + rx_1 + ux_2 + vx_3)$ in Walsh characters:

$$g(x) = g_\emptyset + \sum_i g_i x_i + \sum_{i < j} J_{ij} x_i x_j + T x_1 x_2 x_3. \quad (6)$$

Every coefficient is a smooth finite sum; for example, $J_{23} = 2^{-3} \sum_x x_2 x_3 g(x)$. Replace the visible fields by $\alpha_i = a_i + g_i$, while retaining (c, r, u, v) . This is a smooth global coordinate change with inverse $a_i = \alpha_i - g_i$. When $u = v = 0$, all interaction coefficients vanish. Hence

$$M = \{\alpha_1 = \alpha_2 = \alpha_3 = u = v = 0\}$$

is a smooth two-dimensional manifold of uniform visible laws, with coordinates (c, r) . The likelihood is $-\log 8$ on M .

Put

$$A = \frac{\tanh(c+r) - \tanh(c-r)}{2}, \quad B = \frac{\operatorname{sech}^2(c+r) + \operatorname{sech}^2(c-r)}{2}. \quad (7)$$

At $u = v = 0$, Walsh differentiation gives

$$dJ_{12} = A du, \quad dJ_{13} = A dv, \quad dJ_{23} = dT = 0, \quad (8)$$

$$\partial_u \partial_v J_{23} = B, \quad \partial_u^2 J_{23} = 0, \quad \partial_v^2 J_{23} = 0. \quad (9)$$

For instance $\partial_u g = x_2 \tanh(c + rx_1)$ and $\partial_u \partial_v g = x_2 x_3 \operatorname{sech}^2(c + rx_1)$, which immediately give these coefficients. The field contribution to $\partial_u g$ has already been absorbed into α_2 .

The only nonzero nonconstant Walsh moment of q_ρ is $\mathbb{E}_{q_\rho} x_2 x_3 = \rho$. The first-order features in the new normal coordinates $z = (\alpha_1, \alpha_2, \alpha_3, u, v)$ are

$$x_1, x_2, x_3, Ax_1 x_2, Ax_1 x_3.$$

Their expectations under q_ρ and the uniform law coincide. Thus every point of M is critical. Orthogonality of the Walsh characters under the uniform law, together with (9), yields

$$\nabla_z^2 \ell(c, r, 0) = -I_3 \oplus \begin{pmatrix} -A^2 & \rho B \\ \rho B & -A^2 \end{pmatrix}. \quad (10)$$

The (c, r) rows and columns vanish, since the full gradient vanishes on M . At $(c, r) = (0, t)$ the last two eigenvalues of (10) are

$$-\tanh^2 t \pm \rho \operatorname{sech}^2 t. \quad (11)$$

At a critical point Hessians in smooth invertible coordinates are congruent, so these signs give the stated raw-parameter inertia.

2.2 Why negative normal curvature gives a local maximum

Suppose $|\rho| < \sinh^2 t$. Equation (10) is strictly negative definite at $(0, t, 0)$. By continuity there are a neighborhood U of $(0, t)$, a ball V about $0 \in \mathbb{R}^5$, and $\delta > 0$ such that

$$\nabla_z^2 \ell(c, r, z) \preceq -\delta I_5 \quad ((c, r, z) \in U \times V).$$

For each fixed $(c, r) \in U$, both $\ell(c, r, 0) = -\log 8$ and $\nabla_z \ell(c, r, 0) = 0$. Taylor's formula integrated along the segment from 0 to z gives

$$\ell(c, r, z) \leq -\log 8 - \frac{\delta}{2} \|z\|^2. \quad (12)$$

This is a neighborhood in *all seven* parameters, not a symmetric slice with nuisance parameters fixed. Equality locally holds precisely on M , so the maximum is nonstrict. The reversed inequality in Theorem 2 gives both Hessian signs directly from (11); so does $t = 0$.

2.3 The degenerate threshold is a saddle

Suppose $|\rho| = \sinh^2 t$ and write $s = \text{sign } \rho$. Follow the raw-parameter curve

$$a = c = 0, \quad r = t - \varepsilon, \quad u = \varepsilon, \quad v = s\varepsilon.$$

For fixed r , its likelihood is even in ε , because simultaneously flipping x_2, x_3 preserves the target law. Its quadratic coefficient is

$$\lambda(r) = |\rho| \text{sech}^2 r - \tanh^2 r.$$

Here $\lambda(t) = 0$ and $\lambda'(t) = -2(1 + |\rho|) \text{sech}^2 t \tanh t = -2 \tanh t$. Smoothness makes the fourth-order remainder uniform for r near t , so substitution of $r = t - \varepsilon$ gives

$$\ell(\theta(\varepsilon)) - \ell(M) = 2 \tanh t \varepsilon^3 + O(\varepsilon^4) > 0$$

for sufficiently small positive ε . Varying only a_1 instead decreases likelihood by $\log \cosh a_1$. Hence the threshold point is a saddle even though its Hessian has no positive eigenvalue.

3 A finite improvement and an exact sample

For any $0 < |\rho| < 1$, set $b = \text{atanh } \sqrt{|\rho|}$. The finite point

$$a = c = r = 0, \quad u = b, \quad v = \text{sign}(\rho)b \tag{13}$$

has zero visible unary moments and pair correlation $\mathbb{E}[X_2 X_3] = \tanh u \tanh v = \rho$; its remaining variable is independent and uniform. Thus it represents q_ρ exactly. Its gain over the uniform law is $\text{KL}(q_\rho \| \text{Unif}) = \Delta(\rho) > 0$. This proves nonglobality in Theorem 2.

For the constants in (1), $\tanh t = 1/2$ and $\sinh^2 t = 1/3 > \rho$. Assign five observations to each of the four states with $x_2 x_3 = 1$ and three to each remaining state. These 32 observations give the full-support target with cell probabilities $5/32$ and $3/32$. The bad weights are $(t, 0, 0)$ and the good weights are $(0, t, t)$, all biases being zero. The seven raw spin-coordinate Hessian eigenvalues at the bad point are

$$-\frac{5}{4}, \quad -1, \quad -1, \quad -\frac{7}{16}, \quad -\frac{1}{16}, \quad 0, \quad 0.$$

The normalized likelihood gain is $\Delta(1/4) \simeq 0.0315839424$.

3.1 Interior mixture parameters

Conditioning on the hidden spin turns (3) into a mixture of two product Bernoulli laws. At the bad point its mixing weights are both $1/2$. The conditional probabilities of $X_1 = 1$ are $1/4$ and $3/4$, while those of $X_2 = 1, X_3 = 1$ are $1/2$ in both components. All mixture parameters are strictly interior, the components are distinct, and neither component is empty.

This correspondence preserves local extrema: conditional log odds recover a_i and the incident weight w_i from $2(a_i + w_i)$ and $2(a_i - w_i)$, and the positive mixture-weight ratio then recovers c through

$$\frac{\pi_+}{\pi_-} = e^{2c} \prod_{i=1}^3 \frac{\cosh(a_i + w_i)}{\cosh(a_i - w_i)}.$$

These formulas give a smooth inverse on the interior mixture parameter space.

For direct rational computation, changing spins to 0/1 variables gives incident weights $4w_i$, visible biases $2a_i - 2w_i$, and hidden bias $2c - 2\sum_i w_i$. Exponentiating these parameters produces

	Visible biases	Hidden bias	Incident weights
Bad	$(1/3, 1, 1)$	$1/3$	$(9, 1, 1)$
Good	$(1, 1/3, 1/3)$	$1/9$	$(1, 9, 9)$

The entries are exponentials, not the raw coefficients. In particular an entry 1 in the last column is a zero allowed weight.

4 Joining the local maxima through an observed center

We prove Theorem 1. Write a_0 for the center bias and β_j for the six parameters in branch j : its two leaf biases, hidden bias, and three incident weights. Given X_0 , the branches are independent and their conditional laws do not depend on a_0 or on the other branches. Therefore

$$\ell_{q_k}(\theta) = \frac{1}{2} \sum_{x_0=\pm 1} \log p_\theta(x_0) + \sum_{j=1}^k C_j(\beta_j), \quad (14)$$

where, with $q_{\text{pair}}(y, z) = (1 + \rho yz)/4$,

$$C_j(\beta_j) = \frac{1}{2} \sum_{x_0=\pm 1} \mathbb{E}_{q_{\text{pair}}} \log p_{\beta_j}(Y_j, Z_j \mid x_0).$$

Lemma 3. *In a neighborhood of the bad branch $(t, 0, 0)$ with zero branch biases, $C_j(\beta_j) \leq -\log 4$ for arbitrary perturbations of its six branch parameters.*

Proof. Take the isolated three-leaf RBM with this branch and let $B_+(\beta), B_-(\beta) > 0$ be its branch partition functions at $X_0 = 1, -1$, before including a center bias. Its unique center bias giving a uniform center marginal is

$$\hat{a}_0(\beta) = \frac{1}{2} \log \frac{B_-(\beta)}{B_+(\beta)}.$$

This is a smooth finite function, equal to zero at the bad branch. For β sufficiently close to that branch, the profiled seven-parameter point lies in the local maximum neighborhood of Theorem 2. Its likelihood is $-\log 2 + C_j(\beta) \leq -\log 8$, proving the assertion. $\square$

At a good branch, each conditional likelihood is at most $\mathbb{E}_{q_{\text{pair}}} \log q_{\text{pair}}$ by nonnegativity of KL divergence; this bound is global. The center term in (14) is at most $-\log 2$, also globally. For a fixed subset S , apply the lemma to every bad branch and the conditional KL bound to every good one. All bounds are attained at the displayed parameter: the center is uniform, each good pair has law q_{pair} , and each bad pair is uniform. Adding the bounds proves a local maximum in the full $6k + 1$ dimensional parameter space. The separate profiling biases in the proof are only used to bound C_j ; they are not imposed simultaneously on the shared center of the actual model.

The visible law at this parameter is

$$p_S(x_0, y, z) = \frac{1}{2} \prod_{j \in S} q_{\text{pair}}(y_j, z_j) \prod_{j \notin S} \frac{1}{4}.$$

Since $\rho \neq 0$, distinct subsets give distinct laws. The all-good law is exactly q_k and hence globally maximizes likelihood. Subtracting the objectives gives $(k - |S|)\Delta(\rho)$. Finally, assign to each state the positive integer count

$$n(x_0, y, z) = \prod_{j=1}^k \begin{cases} 5, & y_j z_j = 1, \\ 3, & y_j z_j = -1. \end{cases}$$

Their sum is $2 \cdot 16^k$ and their relative frequencies are (1), finishing the proof.

The construction does not require identical branches. One may replace ρ by nonzero branch-specific ρ_j with $|\rho_j| < 1$, choose bad weights t_j with $\sinh^2 t_j > |\rho_j|$, and good weights $(0, b_j, \text{sign}(\rho_j)b_j)$, where $b_j = \text{atanh} \sqrt{|\rho_j|}$. The same conditional argument applies, with deficit $\sum_{j \notin S} \Delta(\rho_j)$. Rational correlations again give full-support finite samples.

5 Scope, prior work and verification

Problem 3.1 in the AIM Boltzmann Machines list, attributed to Jason Rolfe, asks for efficient algorithms for training sparsely connected RBMs [1]. The present result, associated with the dataset record AIM-PROBABILITY-0007 [2], is an obstruction to inferring a benign raw-parameter likelihood landscape from sparse or acyclic connectivity. It is not an answer identifying the most efficient algorithm. Exact inference on T_k uses linear-time tree message passing; this does not remove the local maxima proved above.

Nonidentifiability, singularities and multiple optima in latent-class likelihoods are established phenomena; see Fienberg, Hersh, Rinaldo and Zhou [3]. Allman et al. [4] describe model boundaries and singular parameter loci and give exact global MLE calculations for the binary tripod. Our calculation concerns local maximality at a specified interior, singular parameter representative, not a new global-fitting description of that distribution model. Zhang, Poupart and Trimponias [5] exhibit bad Bernoulli-mixture optima in one-cluster regions with a zero mixing weight. The example here has two distinct components with weights 1/2, so its mechanism is different from that empty-component boundary construction.

The earlier degree-two RBM classification in [6] shows that every finite local maximum is global when all hidden degrees are at most two. That work also gives a three-input nonstrict saddle and a four-input nonglobal local maximum. Theorem 2 supplies a three-input nonglobal local maximum; in combination with the degree-two theorem, hidden degree three is the first degree at which such an example can occur. The explicit threshold and the shared-center amplification are the results developed here. No absolute-priority claim is made.

These maxima have flat fibers, so “trap” means local maximality, not a positive energy barrier. Along the uniform fiber of the basic example one can reduce r until a saddle permits improvement without first decreasing likelihood. In distribution space, the representable laws $q_{\varepsilon\rho}$ already improve on uniform for small positive ε ; useful representatives need not approach the bad parameter. The results do not assert robustness under arbitrary data perturbations, a lower bound on training complexity, or failure of every optimization algorithm.

The source package includes a supplementary checker. Conditional hidden moments give 60 exact raw-Hessian tests, while a separate hidden-state enumeration in 0/1 coordinates checks all 30 displayed connected configurations for $k = 1, 2, 3, 4$. In each normal and optimized run, 18,974 exact predicates and 1,507 floating-point checks pass. The latter include 1,500 seeded nearby perturbations and do not replace the analytic neighborhood proof. The checker uses Python and SymPy 1.14.0; neither it nor this self-audit is formal verification or independent peer review.

AI assistance was used in mathematical exploration, source comparison, symbolic checking and manuscript preparation. The author is responsible for the claims and their presentation.

References

- [1] American Institute of Mathematics, *Boltzmann Machines*, problem list, Section 3, Problem 3.1 (Jason Rolfe), <http://aimpl.org/boltzmann/3/>, accessed 17 October 2026.
- [2] ulamai/UnsolvedMath, version 1.6, record AIM-PROBABILITY-0007, <https://huggingface.co/datasets/ulamai/UnsolvedMath>.
- [3] S. E. Fienberg, P. Hersh, A. Rinaldo and Y. Zhou, *Maximum Likelihood Estimation in Latent Class Models for Contingency Table Data*, preprint, 2007, [arXiv:0709.3535](https://arxiv.org/abs/0709.3535).
- [4] E. S. Allman, H. Baños, R. Evans, S. Hoşten, K. Kubjas, D. Lemke, J. A. Rhodes and P. Zwiernik, *Maximum likelihood estimation of the latent class model through model boundary decomposition*, Journal of Algebraic Statistics **10** (2019), no. 1, 51–84; [arXiv:1710.01696](https://arxiv.org/abs/1710.01696).
- [5] G. Zhang, P. Poupart and G. Trimponias, *Comparing EM with GD in Mixture Models of Two Components*, Proceedings of Machine Learning Research **115** (2020), 164–174, <https://proceedings.mlr.press/v115/zhang20a.html>.
- [6] A. Ferudun, *Critical Points and Local Maxima of Sparse Binary Restricted Boltzmann Machines*, unrefereed preprint, 5 October 2026, [doi:10.5281/zenodo.23171795](https://doi.org/10.5281/zenodo.23171795).