

Critical Points and Local Maxima of Sparse Binary Restricted Boltzmann Machines

Alper Ferudun*

October 5, 2026

Abstract

We classify the finite critical points of exact empirical log likelihood for binary restricted Boltzmann machines in which every hidden unit has at most two visible neighbors. All allowed weights and biases are independent real parameters. A pairwise exponential-family representation, together with an analysis of its singular parametrization, gives the exact Hessian inertia at every critical point. Every finite local maximum is global, and every finite nonglobal critical point is a strict saddle. We describe finite attainment, the critical visible distributions through active-edge submodels, and both regular and singular maximizing fibers. Two samples of size six show that sample size and pairwise support patterns do not decide finite attainment. Higher hidden degrees have different behavior: a three-input hidden unit gives a nonstrict saddle, and a four-input hidden unit gives a finite spurious local maximum for full-support data. The latter example includes an explicit finite point of strictly higher likelihood. These are scoped results for the architecture-dependent likelihood question in the AIM Boltzmann Machines problem list, not a classification of arbitrary RBM architectures.

1 Model and scope

Let $V = \{1, \dots, n\}$, $n \geq 1$, and H be the visible and hidden vertices of a finite simple bipartite graph, with $m = |H|$ hidden vertices and q allowed edges. For binary vectors $x \in \{0, 1\}^n$ and $y \in \{0, 1\}^m$, the joint law is proportional to

$$\exp \left(\sum_i b_i x_i + \sum_j c_j y_j + \sum_{(i,j)} w_{ij} x_i y_j \right). \quad (1)$$

All $P = n + m + q$ parameters are finite and independent. Write p_θ for the visible marginal and, for any empirical probability law $\hat{p}$,

$$\ell(\theta) = \sum_x \hat{p}(x) \log p_\theta(x).$$

Zero empirical cells are allowed. Ordinary samples are represented by their relative frequencies. We study exact, unregularized likelihood, not a contrastive-divergence approximation, tied-parameter model or sign-constrained model. Local extrema are in the Euclidean space $\mathbb{R}^P$. For this maximization problem a *strict saddle* has both positive and negative Hessian eigenvalues; inertia is listed as (n_+, n_-, n_0) .

*Mercury Software GmbH. Email: alper@mercurycodelab.com. GitHub: <https://github.com/AlperTheKing>. Unrefereed preprint. CC BY 4.0.

Problem 5.2 in the 2018 AIM Boltzmann Machines list [1] asks for the dependence of local optima on architecture and empirical sample size. The corresponding frozen dataset record is AIM-PROBABILITY-0013 [2]. Sections 2–5 give a complete finite-parameter description when every hidden degree is at most two. Arbitrary visible cycles, repeated hidden neighborhoods, leaves and isolated hidden vertices are included. Sections 6 and 7 give two explicit limitations at higher degree.

The softplus interaction expansion and representation viewpoint are established tools; see Montúfar and Rauh [4], including their discussion of Younes’ earlier representation lemma. The interior criterion for a finite maximum-likelihood estimate in a full minimal exponential family is also standard [3, Theorem 13]. Bad local optima and parity-related likelihood examples in RBMs are already known [6, Section 6]; see also [5]. The leaf-hidden special case is derived in the frozen record’s upstream partial attempt [2] and is credited here. Our arguments address the singular raw-parameter map, give an exact critical-point and inertia description for degree two, and prove the stated higher-degree examples. No absolute priority claim is made for these results.

2 The pair map and its singularities

Throughout this section and the next three, assume every hidden degree is at most two. Let E be the set of distinct two-element visible neighborhoods. For $e \in E$ let H_e be its hidden vertices and put $r_e = |H_e|$. Set $s(t) = \log(1 + e^t)$ and $\sigma = s'$. A hidden vertex with neighborhood $e = \{i, k\}$, bias c and weights u, v contributes the interaction

$$J(c, u, v) = s(c + u + v) - s(c + u) - s(c + v) + s(c). \quad (2)$$

Indeed, on its four binary inputs its softplus contribution is exactly

$$s(c) + [s(c + u) - s(c)]x_i + [s(c + v) - s(c)]x_k + J(c, u, v)x_i x_k.$$

Degree-one units contribute only constants and unary terms, while degree-zero units contribute constants. Absorb all unary terms into the visible biases, calling the resulting vector h . The map $(b, c, w) \mapsto (h, c, w)$ is a global smooth diffeomorphism, whose inverse subtracts those same unary terms. With $J_e = \sum_{j \in H_e} J_j$, the visible family becomes

$$\begin{aligned} p_\theta(x) &= \exp\{\eta \cdot T(x) - \Psi(\eta)\}, \\ \eta &= (h, (J_e)_{e \in E}), \quad T(x) = ((x_i)_{i \in V}, (x_i x_k)_{\{i, k\} \in E}), \\ \Psi(\eta) &= \log \sum_{x \in \{0, 1\}^n} e^{\eta \cdot T(x)}. \end{aligned} \quad (3)$$

Lemma 2.1. *The differential of J vanishes exactly when $u = v = 0$. At such a point, the only nonzero second derivatives are $J_{uv} = J_{vu} = s''(c) > 0$. Moreover*

$$J(c, u, v) = uvK(c, u, v), \quad K(c, u, v) = \int_0^1 \int_0^1 s''(c + tu + rv) dt dr > 0. \quad (4)$$

Every real pair coefficient is attained by finite parameters.

Proof. Direct differentiation gives

$$J_u = \sigma(c + u + v) - \sigma(c + u), \quad J_v = \sigma(c + u + v) - \sigma(c + v).$$

Since σ is strictly increasing, $J_u \neq 0$ whenever $v \neq 0$ and $J_v \neq 0$ whenever $u \neq 0$. When both weights vanish all first derivatives vanish, and differentiating once more gives the asserted second

derivatives. Two applications of the fundamental theorem of calculus give (4), also for negative or zero weights. Finally, for $a \geq 0$,

$$J(-a, a, a) = 2 \log \cosh(a/2), \quad J(0, a, -a) = -2 \log \cosh(a/2).$$

Their magnitudes range continuously from zero to infinity. $\square$

One hidden vertex per pair can supply any desired J_e ; the others may have both weights zero. Free visible biases supply arbitrary h . Thus (3) has as its image the entire positive pairwise exponential family on (V, E) . Notice that $J = 0$ need not be singular: $u = 0$, $v \neq 0$ gives $dJ \neq 0$. Different hidden contributions to a pair can also cancel.

The statistics $1, T$ are linearly independent: evaluate a putative relation at the zero vector, at each singleton and at each pair. Consequently, with $\hat{t} = \mathbb{E}_{\hat{p}} T$, the effective score and Hessian are

$$g = \nabla_{\eta} \ell = \hat{t} - \mathbb{E}_{\eta} T, \quad \nabla_{\eta}^2 \ell = -\text{Cov}_{\eta}(T) \prec 0. \quad (5)$$

Positive definiteness follows from full support and the independence of $1, T$. Strict concavity in η alone does not settle local extrema in the singular RBM coordinates; the next theorem treats that issue.

3 All finite critical points

Call e *active* at a parameter point if at least one $j \in H_e$ has $(u_j, v_j) \neq (0, 0)$. Let $\mathcal{A}$ be the active set and $\mathcal{I} = E \setminus \mathcal{A}$. This definition is not the condition $J_e \neq 0$. Write g_i, g_e for the components of (5).

Theorem 3.1. *A finite parameter point is critical if and only if*

$$g_i = 0 \quad (i \in V), \quad g_e = 0 \quad (e \in \mathcal{A}). \quad (6)$$

At a critical point put

$$B = \{e \in \mathcal{I} : g_e \neq 0\}, \quad R = \sum_{e \in B} r_e.$$

The Hessian in the original RBM parameters has inertia

$$(R, n + |\mathcal{A}| + R, P - n - |\mathcal{A}| - 2R). \quad (7)$$

Every finite local maximum is global; every finite nonglobal critical point is a strict saddle. There is no finite local minimum.

Proof. In coordinates (h, c, w) the h derivatives are the node scores. Each hidden parameter on e has derivative g_e times the corresponding derivative of J_e . Leaf and isolated hidden parameters have no effect in these coordinates. Lemma 2.1 therefore gives (6), including singular points.

Let $F(h, c, w) = (h, J)$. Its differential has rank $n + |\mathcal{A}|$: the h columns give an identity block, each active pair gives one nonzero differential, and different pairs use disjoint hidden-parameter groups. Inactive rows vanish. The second derivative chain rule is

$$\nabla^2 \ell = -DF^{\top} \text{Cov}_{\eta}(T) DF + \sum_e g_e \nabla^2 J_e. \quad (8)$$

The first form is negative semidefinite of rank $n + |\mathcal{A}|$ and uses none of the hidden parameters on inactive pairs. At a critical point, the second term vanishes on active pairs. For each hidden vertex on an inactive pair its only possible nonzero block is

$$g_e s''(c_j) \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

on that vertex's two weights. There are no cross terms with the first form, with hidden biases, or between different hidden vertices. Every nonzero block contributes one positive and one negative eigenvalue; the remaining directions are null. This proves (7) in (h, c, w) coordinates. At a critical point a smooth coordinate change acts on the Hessian by congruence, so the original coordinates have the same inertia.

If $R > 0$ there are both signs. If $R = 0$, every pair score vanishes, including inactive ones, and all node scores vanish. Strict concavity in (5) then makes the visible law a global maximizer on the effective family and hence on the RBM parameter space. A local maximum is critical and cannot have $R > 0$. Finally, varying any one visible bias gives the strictly negative second derivative $-\text{Var}_\theta(x_i)$ at every finite point. Since $n \geq 1$, no finite local minimum is possible. $\square$

For $A \subseteq E$ let T_A consist of the node statistics and the pair statistics in A , and write

$$\mathcal{P}_A = \text{conv}\{T_A(x) : x \in \{0, 1\}^n\}, \quad \hat{t}_A = \mathbb{E}_{\hat{p}} T_A.$$

The polytope has full dimension by the preceding monomial argument.

Lemma 3.2. *The effective likelihood on A has a finite maximizer if and only if $\hat{t}_A \in \text{int } \mathcal{P}_A$. Its natural parameter and visible distribution are then unique.*

Proof. A finite model mean is a convex combination with positive coefficients of all $T_A(x)$, so cannot lie on a proper supporting face. Conversely, if $\hat{t}_A$ is interior, some ball of radius $\delta > 0$ about it lies in $\mathcal{P}_A$. Hence for every η ,

$$\Psi_A(\eta) - \eta \cdot \hat{t}_A \geq \max_x \eta \cdot (T_A(x) - \hat{t}_A) \geq \delta \|\eta\|.$$

The negative log likelihood is continuous and coercive, so attains a finite minimum. Its Hessian is positive definite, giving uniqueness. Every finite critical point satisfies the mean equation, which excludes one when the empirical moment is on the boundary. $\square$

This is the standard finite exponential-family criterion, included to make the sample conditions explicit. It is not a new MLE existence theorem.

Corollary 3.3. *For every $A \subseteq E$ with $\hat{t}_A \in \text{int } \mathcal{P}_A$, let p_A be the finite MLE in the submodel $J_e = 0$ for $e \notin A$. The full critical set is the union of the parameter points whose active set is A and whose visible law is p_A , over all such A . Each such stratum is nonempty. There are at most $2^{|E|}$ distinct critical visible laws and at most $2^{|E|}$ critical likelihood values.*

Proof. At a point with active set A , all inactive coefficients vanish and (6) is exactly the score equation in the submodel. Lemma 3.2 gives necessity and uniqueness. To realize p_A , use one hidden vertex on each pair in A . For a nonzero coefficient use Lemma 2.1; for a zero coefficient use $u = 0, v = 1$ so that the pair is still active. Set the weights on every inactive pair to zero and adjust the free visible biases. Equation (6) holds. There are only $2^{|E|}$ subsets, though different subsets can give the same law and the raw-parameter fibers need not be discrete. $\square$

Corollary 3.4. *A finite critical point exists exactly when every empirical node mean lies in $(0, 1)$. Finite local or global maxima exist exactly when $\hat{t}_E \in \text{int } \mathcal{P}_E$. If $\hat{t}_E$ is on the boundary, every finite critical point that exists is a strict saddle and the supremum is not attained at finite parameters.*

Proof. The node score equations and positivity give the first necessity. For sufficiency set all weights to zero and $h_i = \log(\mathbb{E}_{\hat{p}} x_i / (1 - \mathbb{E}_{\hat{p}} x_i))$; this product law is critical. The rest follows from surjectivity onto the effective family, Lemma 3.2 and Theorem 3.1. $\square$

4 Regular and singular maximizing fibers

Suppose the finite effective MLE η_* exists. At a maximum with every pair active, F is a submersion. Its fiber $F^{-1}(\eta_*)$ is locally a smooth manifold of dimension $P - n - |E|$. Equation (8) with $g = 0$ shows that its tangent space $\ker DF$ is exactly the Hessian kernel. It is therefore a Morse–Bott maximum manifold.

At a maximum with an inactive pair, each corresponding hidden vertex admits local coordinates

$$(c, u, v) \mapsto (c, U = u, V = vK(c, u, v)).$$

The Jacobian determinant at $u = v = 0$ is $s''(c) > 0$, and $J = UV$ exactly. For active pairs, each aggregate coefficient J_e can be chosen as a local coordinate by its nonzero differential, independently across pairs. Consequently the maximizing fiber germ is given by $h = h_*$, the fixed active coefficients, free nuisance coordinates, and the cones

$$\sum_{j \in H_e} U_j V_j = 0 \quad (e \in \mathcal{I}). \quad (9)$$

An inactive coefficient is necessarily zero at the point, so its target value is zero. In particular, one inactive hidden vertex gives the crossing $UV = 0$. The regular Morse–Bott description must not be extended across these singular fibers without qualification.

5 Equal sample sizes and different attainment

Take three visible vertices and one hidden vertex on each pair of their triangle. Consider first the sample containing each cube state except 000 and 111 once. Every node mean is $1/2$, and every pair table has cells 00, 11 of mass $1/6$ and cells 01, 10 of mass $1/3$. Yet the empirical moment lies on the proper supporting face

$$x_1 + x_2 + x_3 - x_1 x_2 - x_1 x_3 - x_2 x_3 \leq 1.$$

The left side is zero at the two omitted states and one at the other six. There is no finite maximum. The zero-parameter point is critical, with each pair score $1/6 - 1/4 = -1/12$ and inertia $(3, 6, 3)$.

Now take each cube state except 000 and 011 once. Add $(-1)^{x_1+x_2+x_3}/12$ to its empirical mass function. The resulting law has mass $1/12$ at the formerly missing states, $1/4$ at the other two even-parity states, and $1/12$ at each odd-parity state. It is positive and has the same moments of degree at most two: summing a parity term times such a monomial over an absent coordinate gives zero. The moment is therefore interior and has a finite MLE. Both samples have size and support size six, and all pairwise cells are positive in both samples. These counts and pairwise support patterns do not decide finite attainment.

6 A nonstrict saddle with three inputs

Here take one hidden unit adjacent to all three visible units, and set

$$\hat{p}(x) = \frac{1 + \varepsilon(-1)^{x_1+x_2+x_3}}{8}, \quad 0 < \varepsilon < 1.$$

At $b = w = 0$, $c = \log 2$, the visible law is uniform and the data agree with it through degree two. For $f_\theta(x) = b \cdot x + s(c + w \cdot x)$, the raw Hessian formula is

$$\nabla^2 \ell = (\mathbb{E}_{\hat{p}} - \mathbb{E}_p) \nabla^2 f_\theta - \text{Cov}_p(\nabla f_\theta). \quad (10)$$

The gradient is zero. The expected Hessian correction vanishes because its entries are constant, unary or pair statistics. The derivative features are x_i , $2/3$ and $(2/3)x_i$; their covariance has rank three. Thus the inertia is $(0, 3, 4)$.

Along the curve $w_i = t$, $c = \log 2$ and $b_i = -[s(c + t) - s(c)]$, all unary coefficients vanish. Pair coefficients are $O(t^2)$, while the triple coefficient is

$$s(c + 3t) - 3s(c + 2t) + 3s(c + t) - s(c) = s'''(c)t^3 + O(t^4) = -\frac{2}{27}t^3 + O(t^4).$$

At the uniform law the pair scores vanish and the triple score is $-\varepsilon/8$. The quadratic remainder in the saturated natural coordinates is $O(t^4)$, since the nonconstant coefficients are $O(t^2)$. It follows that

$$\ell(\theta(t)) - \ell(\theta(0)) = \frac{\varepsilon}{108}t^3 + O(t^4).$$

Both signs occur arbitrarily nearby. This is a saddle with no positive Hessian eigenvalue, so the strict-saddle conclusion fails at hidden degree three. It is not a spurious local maximum.

7 A finite spurious maximum with four inputs

Theorem 7.1. *For four visible units and one hidden unit adjacent to all four, let*

$$\hat{p}(x) = \frac{1 + \varepsilon(-1)^{x_1+x_2+x_3+x_4}}{16}, \quad 0 < \varepsilon < 1.$$

The point $b = w = 0$, $c = 0$ is a nonglobal, nonstrict local maximum of exact likelihood, with Hessian inertia $(0, 4, 5)$. The data have full support. For $\varepsilon = 1/2$ they are realized by a sample of size 32, with three observations at each even-parity state and one at each odd-parity state.

Proof. Let U be the uniform law on the four-cube, let $\chi_A(x) = (-1)^{\sum_{i \in A} x_i}$, and put

$$f(x) = b \cdot x + s(c + w \cdot x), \quad F = f - \mathbb{E}_U f.$$

The exact likelihood difference from the origin is

$$\ell(\theta) - \ell(0) = \varepsilon \mathbb{E}_U[\chi_{[4]} F] - \log \mathbb{E}_U e^F. \quad (11)$$

Local maximality. Consider the full neighborhood

$$|c| \leq 1/32, \quad \|w\|_1 \leq 1/32, \quad \|b\|_1 \leq 1/32. \quad (12)$$

Every intermediate argument in a finite difference of $s(c + w \cdot x)$ has absolute value at most $1/16$. Since $e^{1/16} \leq (1 - 1/16)^{-1} < 3/2$, its exponential is in $[2/3, 3/2]$, giving $s'' \geq 6/25$. Globally, with $D = s'' = \sigma(1 - \sigma) \in [0, 1/4]$,

$$s''' = D(1 - 6D), \quad |s''''| \leq 1/8.$$

The quadratic ranges from $-1/8$ to $1/24$ on that interval. Finite differences, averaged over unused inputs, now give

$$\left| \mathbb{E}_U[\chi_{\{i,j\}} F] \right| \geq \frac{3}{50} |w_i w_j|, \quad \left| \mathbb{E}_U[\chi_{[4]} F] \right| \leq \frac{|w_1 w_2 w_3 w_4|}{128}. \quad (13)$$

The first multiplier is $1/4$ times an average of s'' over the two-step rectangles, hence is positive and at least $6/(4 \cdot 25)$. The second has multiplier $1/16$ and four integrals of s'''' . Bias terms and centering contribute to neither coefficient. These formulas include negative and zero weights.

As $0 < s' < 1$, the oscillation of f is at most $\|b\|_1 + \|w\|_1 \leq 1/16$, so $\|F\|_\infty \leq 1/16$. For $K(t) = \log \mathbb{E}_U e^{tF}$, $0 \leq t \leq 1$, the tilted density relative to U is at least $e^{-1/8} \geq 7/8$. By the characterization of variance as $\min_a \mathbb{E}[(F - a)^2]$,

$$K''(t) = \text{Var}_t(F) \geq \frac{7}{8} \text{Var}_U(F).$$

Since $K(0) = K'(0) = 0$, integration and Parseval for the orthonormal Walsh characters yield

$$\log \mathbb{E}_U e^F \geq \frac{7}{16} \text{Var}_U(F) \geq \frac{63}{40000} \sum_{i < j} (w_i w_j)^2. \quad (14)$$

Pair the six terms into three complementary pairs and use $a^2 + b^2 \geq 2|ab|$ to obtain $\sum_{i < j} (w_i w_j)^2 \geq 6|w_1 w_2 w_3 w_4|$. Equations (11)–(14) imply

$$\ell(\theta) - \ell(0) \leq \left(\frac{\varepsilon}{768} - \frac{63}{40000} \right) \sum_{i < j} (w_i w_j)^2 \leq 0, \quad (15)$$

since $1/768 < 63/40000$. If two weights are nonzero this is strict. With at most one nonzero weight the parity coefficient vanishes, and Jensen's inequality gives $-\log \mathbb{E}_U e^F \leq 0$, with equality only for a uniform visible law. Thus the whole neighborhood (12), not merely a parameter slice, is covered. Varying c while $b = w = 0$ keeps the visible law uniform, so the maximum is nonstrict.

Nonglobality. Let q be the product Bernoulli($1/4$) law, and set

$$r = q/U, \quad d = r - 1, \quad M = (3/2)^4 - 1 = 65/16.$$

Then $|d| \leq M$ and

$$a := \mathbb{E}_{\hat{p}} d = \varepsilon \sum_x \chi_{[4]}(x) q(x) = \varepsilon/16 > 0.$$

Choose $\lambda = a/(4M^2) = 4\varepsilon/4225$ and $q_\lambda = (1 - \lambda)U + \lambda q$. As $|\lambda d| \leq 1/2$, the elementary inequality $\log(1 + t) \geq t - t^2$ on $[-1/2, 1/2]$ gives

$$\ell(q_\lambda) - \ell(U) \geq \lambda a - \lambda^2 M^2 = \frac{3}{4} \lambda a > 0. \quad (16)$$

This law is attained at finite RBM parameters: take $b = 0$, $e^{w_i} = 1/3$ for all i , and

$$e^c = \frac{\lambda}{1 - \lambda} (3/2)^4.$$

The two component partition functions are 16 and $(4/3)^4$, so the normalized hidden mixture has the desired weight λ . At $\varepsilon = 1/2$ the values are $\lambda = 2/4225$, $e^c = 81/33784$, and the lower bound in (16) is $3/270400$.

Finally, at the origin the derivative features of f are $x_i, 1/2, x_i/2$, whose covariance has rank four. The expected second-derivative correction in (10) vanishes because the data and uniform law agree through degree two. The gradient is zero and the Hessian inertia is $(0, 4, 5)$. $\square$

8 Consequences and verification

The finite degree-two landscape depends on empirical moments, not sample count alone. It can be described by finitely many effective submodels even though each critical law can have a positive-dimensional and singular raw-parameter fiber. Strict concavity in effective coordinates must be combined with the singular calculation in Theorem 3.1. These conclusions give no convergence guarantee for a training algorithm and do not classify paths with diverging parameters.

The degree-three example separates a nonstrict saddle from a local maximum. The degree-four example proves a spurious maximum using a full neighborhood and a finite improvement. Neither example classifies every architecture with the corresponding maximum degree. The unrestricted AIM question remains open under the present results; the existence of bad RBM optima and parity constructions is not claimed as a new discovery. Novelty of this precise collection of statements has not been established.

The source bundle includes an exact rational and symbolic checker with 65 assertions. It derives finite raw-coordinate Hessians from conditional hidden Bernoulli moments and compares their inertia using congruence elimination and a real-root characteristic-polynomial method. It also checks the sample examples and constants in the parity constructions. These finite checks support, but do not replace, the analytic arguments. They are not proof-assistant verification or independent peer review. AI-assisted tools supported the derivations, computational checks and manuscript preparation. The author remains responsible for the claims and final text; this preprint is self-audited and unrefereed.

References

- [1] American Institute of Mathematics, *Boltzmann Machines*, workshop problem list, 2018, Section 5, Problem 5.2. <http://aimpl.org/boltzmann/5/>.
- [2] ulamai, *UnsolvedMath*, dataset version 1.6.0, record AIM-PROBABILITY-0013 and its upstream partial attempt. Frozen commit `c6e7f41b4eb6ce876feb83e7eb626be1b243fd7e`. <https://huggingface.co/datasets/ulamai/UnsolvedMath>.
- [3] S. E. Fienberg and A. Rinaldo, Maximum likelihood estimation in log-linear models, *Annals of Statistics* **40** (2012), 996–1023. doi:10.1214/12-AOS986.
- [4] G. Montúfar and J. Rauh, Hierarchical models as marginals of hierarchical models, *International Journal of Approximate Reasoning* **88** (2017), 531–546. doi:10.1016/j.ijar.2016.09.003.
- [5] G. Montúfar, *Restricted Boltzmann Machines: Introduction and Review*, 2018. <https://arxiv.org/abs/1806.07066>.
- [6] A. Seigal and G. Montúfar, Mixtures and products in two graphical models, *Journal of Algebraic Statistics* **9** (2018), 1–20. doi:10.18409/jas.v9i1.90.