Relative covariance approximation in discrete Gibbs models

Mantis Discovery

Abstract

A covariance approximation can have small absolute error while badly distorting a direction of small variance. For a fixed finite Gibbs model, we identify the asymptotic covariance scales from the affine spans of energy sublevel sets and show that the final span-increasing energy determines the sampling threshold for relative covariance recovery. For distributions on matroid bases, single-element exchanges give a Laplacian comparison, and conditioning on bases within a fixed exchange distance gives a finite-temperature bound. For hard-core models, a cluster expansion yields an error bound controlled by local activity. Experiments confirm the sampling threshold, retain accuracy of the second-order hard-core approximation on growing strips at low activity, and compare exchange truncations in a small conditional subset-prediction problem. In the last setting, exact Newton is faster. The results concern relative covariance control and do not imply a general training-speed improvement.

1 Introduction

The Hessian of a log-partition function is a feature covariance. In a concentrated discrete model, variances can differ by exponentially large factors. An approximation that is accurate in absolute norm can therefore distort a low-variance direction, making its inverse unreliable. The same issue occurs in sampling: observations may describe common configurations while missing all configurations that vary in one direction.

We address two questions. First, which energy levels must an independent sample reach before its covariance is relatively accurate? Second, when can a known model yield a relative covariance approximation without sampling those configurations? The latter is a deterministic calculation from specified probabilities, not an estimator for an unknown distribution.

For a fixed finite Gibbs model, the affine spans of energy sublevel sets determine the covariance scales and the sampling threshold. For weighted matroid bases, the covariance is comparable with a Laplacian built from single-element exchanges; retaining bases at bounded exchange distance improves the approximation. For hard-core distributions, a cluster expansion gives a different approximation whose error depends on maximum neighborhood activity rather than total activity. This permits bounds that do not deteriorate with the number of variables.

The experiments measure the transition in empirical covariance error when rare configurations are missed, test the hard-core bound on growing grid strips, and compare exchange truncations in a conditional subset model built from handwritten images. Exact inference is available in every experiment and provides the reference covariance and likelihood.

Relation to prior work. The covariance identity for exponential families and the integral representation of relative entropy used below are standard [9]. We use second-order expectation propagation for the strip calculations [5]. The image experiment uses the fixed-cardinality distribution studied by Swersky et al. [7].

The finite-model calculation makes explicit the scales that vanish as a Gibbs distribution concentrates. Its upper sampling bound is an application of matrix Chernoff concentration [8, Corollary 5.2]; the lower bound follows from the chance of missing a proper affine span.

The matroid result uses the one-sided bijective exchange statement of Jukna and Seiwert [4, Proposition 1(b)]. The hard-core expansion, its tree bound, and the convergence condition are standard results for abstract polymer models [3]. Our contribution is to convert these inputs into relative matrix comparisons.

Relative covariance error. All matrix inequalities are in the positive-semidefinite order. Inverses and square roots are taken on the span of feature differences. For an approximation M to a covariance C on that span, define

$$\mathcal{E}(M, C) = \|C^{-1/2} M C^{-1/2} - I\|_{\mathrm{op}}. \tag{1}$$

The comparisons below often give $(1-\varepsilon)M \preceq C \preceq (1+\varepsilon)M$. If $\varepsilon < 1$, then $\mathcal{E}(M, C) \le \varepsilon/(1-\varepsilon)$.

For a conditional exponential family with linear score map $\theta \mapsto F_z^\top \theta$, the covariance contribution to likelihood curvature is $F_z C_z F_z^\top$. Congruence, averaging over inputs, and adding the same nonnegative quadratic regularizer preserve a covariance comparison. This is how the bounds enter the fitting experiments. Nonlinear score maps have additional Hessian terms.

2 Which configurations determine covariance?

Let Ω be finite, let $x_\omega \in \mathbb{R}^d$, and fix $a_\omega > 0$ and $g_\omega \ge 0$ with $\min_\omega g_\omega = 0$. Consider

$$p_\tau(\omega) = \frac{a_\omega e^{-g_\omega/\tau}}{Z_\tau}, \qquad C_\tau = \mathrm{Cov}_{p_\tau}(x_\omega), \qquad \tau > 0. \tag{2}$$

Choose a zero-energy state with feature x_0, and set

$$V_t = \mathrm{span}\{x_\omega - x_0 : g_\omega \le t\}, \quad V_{<t} = \mathrm{span}\{x_\omega - x_0 : g_\omega < t\}, \quad V = \mathrm{span}\{x_\omega - x_0 : \omega \in \Omega\}.$$

The spaces V_t do not depend on the chosen zero-energy state. Assume $\dim V > 0$. For each energy γ at which the dimension increases, let $W_\gamma = V_\gamma \cap V_{<\gamma}^\perp$, with $V_{<0} = \{0\}$, and let P_γ be the orthogonal projection onto W_γ.

Write $Z_0 = \sum_{g_\omega = 0} a_\omega$. On $W_0 = V_0$, let B_0 be the covariance of the ground-state features with probabilities a_ω / Z_0. For $\gamma > 0$, put

$$B_\gamma = \frac{1}{Z_0} \sum_{g_\omega = \gamma} a_\omega P_\gamma (x_\omega - x_0) P_\gamma (x_\omega - x_0)^\top. \tag{3}$$

Only such energies are included in the sum below.

Theorem 2.1. *Each B_γ is positive definite on W_γ, and*

$$B_\tau = \sum_\gamma e^{-\gamma/\tau} B_\gamma \quad \textit{satisfies} \quad \|B_\tau^{-1/2} C_\tau B_\tau^{-1/2} - I_V\|_{\mathrm{op}} \longrightarrow 0.$$

Thus the eigenvalues associated with W_γ have leading terms $e^{-\gamma/\tau}\lambda_j(B_\gamma|_{W_\gamma})$. In particular, with $\gamma_* = \min\{t : V_t = V\}$, $\lambda_{\min}(C_\tau|_V) = \Theta(e^{-\gamma_*/\tau})$.

Proof. Use the pairwise covariance identity

$$Z_\tau^2 C_\tau = \frac{1}{2} \sum_{\omega,\eta} a_\omega a_\eta e^{-(g_\omega + g_\eta)/\tau} (x_\omega - x_\eta)(x_\omega - x_\eta)^\top. \tag{4}$$

Group terms by the sum of their energies. The coefficient matrices are positive semidefinite, and the span of their ranges up to energy t is V_t. One inclusion is immediate; the other follows by pairing each state of energy at most t with the chosen zero-energy state.

Decompose V into the orthogonal spaces W_γ and multiply each W_γ block on both sides by $e^{\gamma/(2\tau)}$. A block joining W_γ and W_β receives contributions only from pair energies $\alpha \geq \max(\gamma, \beta)$. If $\gamma \neq \beta$, its normalized terms are therefore $O(e^{-|\gamma-\beta|/(2\tau)})$, and vanish. On a diagonal block, all energies above γ vanish after normalization.

At a positive diagonal energy γ, pairs with both energies below γ project to zero. The remaining leading terms pair a zero-energy state with a state of energy γ. Since all zero-energy features have the same projection onto W_γ, they contribute $Z_0^2 B_\gamma$. At energy zero, (4) gives $Z_0^2 B_0$. As $Z_\tau \to Z_0$, the scaled covariance converges to the block diagonal matrix with blocks B_γ.

The projected energy-γ differences span W_γ, proving positive definiteness in (3); the same statement for B_0 follows from the affine span of the ground-state features. Normalizing by these fixed positive blocks proves the relative limit. The eigenvalue conclusions follow from its two-sided matrix comparison and the min–max principle. □

2.1 Independent samples

Let $X_1, \ldots, X_{N_\tau}$ be independent feature vectors with law (2), and let $\widehat{C}_\tau = N_\tau^{-1} \sum_i (X_i - \overline{X})(X_i - \overline{X})^\top$.

Proposition 2.2. *For the fixed finite model* (2)*,*

$$\mathcal{E}(\widehat{C}_\tau, C_\tau) \xrightarrow{\mathbb{P}} 0 \quad \Longleftrightarrow \quad N_\tau e^{-\gamma_*/\tau} \longrightarrow \infty. \tag{5}$$

Proof. Whiten $Y = C_\tau^{-1/2}(X - \mathbb{E}X)$ on V, and let $r = \dim V$. By Theorem 2.1 and finiteness of the support, $R_\tau := \max_\omega \|Y(\omega)\|^2 = O(e^{\gamma_*/\tau})$. Applying the matrix Chernoff inequality [8, Corollary 5.2] to the independent matrices $Y_i Y_i^\top$, whose expectation is I_V, yields, for fixed $0 < \epsilon < 1$,

$$\mathbb{P}\left(\left\| N_\tau^{-1} \sum_i Y_i Y_i^\top - I_V \right\|_{\mathrm{op}} > \epsilon \right) \leq 2r \exp\left(-\frac{N_\tau \epsilon^2}{3R_\tau} \right).$$

The empirical centering correction is $\overline{Y}\overline{Y}^\top$, with $\mathbb{E}\|\overline{Y}\|^2 = r/N_\tau$. Thus the right-hand condition in (5) is sufficient.

For necessity, first suppose $\gamma_* > 0$. The affine space $x_0 + V_{<\gamma_*}$ is proper in $x_0 + V$, and the probability of falling outside it is $\rho_\tau = ce^{-\gamma_*/\tau}(1 + o(1))$ for some $c > 0$. If $N_\tau e^{-\gamma_*/\tau}$ stays bounded along a subsequence, the probability $(1 - \rho_\tau)^{N_\tau}$ that every observation lies in this proper affine space stays bounded away from zero. On that event $\widehat{C}_\tau$ is singular on V and its relative error is at least one. If $\gamma_* = 0$ and N_τ stays bounded, the event that every observation equals a fixed ground-state feature has the same consequence. □

For features $0, e_1, e_2$ with unnormalized probabilities $1, q, q^3$, relative recovery requires $Nq^3 \to \infty$. Even when Nq is large, the second direction remains unresolved. The proposition concerns independent samples and does not cover estimators that use model probabilities, importance sampling, or additional structure.

3 Approximation from feasible exchanges

A fixed-cardinality distribution assigns weight $\prod_{e \in B} w_e$ to each allowed subset B of a prescribed size. When every subset of that size is allowed, this is the probabilistic n-choose-k model [7]. Here the allowed subsets are the bases of a matroid, a class that includes fixed-size subsets and spanning trees.

Let $\mathcal{B}$ be the bases of a matroid on E, let $c(B) = \sum_{e \in B} c_e$, and define

$$p_\tau(B) \propto e^{-c(B)/\tau}, \qquad C_\tau = \mathrm{Cov}_{p_\tau}(\mathbf{1}_B).$$

Fix a basis B_0. Let $\mathcal{H}$ contain the single-element exchanges $h = (e, f)$ for which $e \in B_0$, $f \notin B_0$, and $B_0 - e + f \in \mathcal{B}$. Put

$$d_h = u_f - u_e, \quad q_h = e^{-(c_f - c_e)/\tau}, \quad L_\tau = \sum_{h \in \mathcal{H}} q_h d_h d_h^\top, \quad s_\tau = \sum_h q_h, \quad Q_\tau = \prod_h (1 + q_h), \quad (6)$$

where u_e is the coordinate vector at e. The matrix L_τ is the weighted Laplacian formed by these exchanges.

Theorem 3.1. *For every reference basis and every $\tau > 0$,*

$$Q_\tau^{-2} L_\tau \preceq C_\tau \preceq Q_\tau L_\tau. \quad (7)$$

These matrices have the same range. If B_0 is the unique minimum-cost basis, then $Q_\tau \to 1$ as $\tau \downarrow 0$.

The estimate does not require the partition function and needs at most one feasibility test for each pair in $B_0 \times (E \setminus B_0)$. It is most informative when Q_τ is close to one.

The proof combines the following covariance comparison with the cited exchange theorem.

Lemma 3.2. *Let $\mathcal{F} \subseteq 2^{\mathcal{H}}$ contain the empty set and every singleton. Give $F \in \mathcal{F}$ weight $q_F = \prod_{h \in F} q_h$, with $q_h > 0$, and feature $x_F = \sum_{h \in F} d_h$. Set $Z = \sum_{F \in \mathcal{F}} q_F$, $s = \sum_h q_h$, $Q = \prod_h (1 + q_h)$, and $L = \sum_h q_h d_h d_h^\top$. Then*

$$Z^{-2} L \preceq \mathrm{Cov}(x_F) \preceq QL.$$

Proof. Fix a direction v and write $a_h = v^\top d_h$, $a_F = \sum_{h \in F} a_h$, $M = \sum_F q_F a_F^2$, and $b = \sum_F q_F a_F$. Since $a_\varnothing = 0$, weighted Cauchy–Schwarz gives $b^2 \le (Z - 1)M$. Hence

$$\mathrm{Var}(v^\top x_F) = \frac{M}{Z} - \frac{b^2}{Z^2} \ge \frac{M}{Z^2} \ge \frac{1}{Z^2} \sum_h q_h a_h^2.$$

For the upper bound, use $a_F^2 \le |F| \sum_{h \in F} a_h^2$ and enlarge the sum to all subsets. The coefficient of a_h^2 is

$$\sum_{F \subseteq \mathcal{H}:\, h \in F} |F| q_F = q_h \prod_{j \ne h} (1 + q_j) \left(1 + \sum_{j \ne h} \frac{q_j}{1 + q_j} \right) \le q_h Q (1 + s).$$

Since $Z \ge 1 + s$, we obtain $\mathrm{Var}(v^\top x_F) \le M/Z \le Q v^\top L v$. □

Proof of Theorem 3.1. For each basis B, the one-sided bijective exchange theorem supplies a matching of $B_0 \setminus B$ with $B \setminus B_0$ such that every matched pair is in $\mathcal{H}$ [4, Proposition 1(b)]. After contracting $B \cap B_0$, common elements may be fixed. Choose one such matching and let F_B be its set of exchanges. Then

$$\mathbf{1}_B - \mathbf{1}_{B_0} = \sum_{h \in F_B} d_h, \qquad e^{-[c(B) - c(B_0)]/\tau} = \prod_{h \in F_B} q_h.$$

The map $B \mapsto F_B$ is injective, because the removed and added elements determine B. Its image contains the empty set and every singleton. Apply Lemma 3.2 and $Z \le Q_\tau$. The positive coefficients imply equality of ranges. At a unique optimum every feasible exchange has positive cost, so each $q_h \to 0$. □

3.1 Retaining more bases

Single-element exchanges need not give the best approximation. Let

$$\mathcal{B}_k = \{B \in \mathcal{B} : |B \setminus B_0| \le k\}, \qquad z_k = \sum_{B \in \mathcal{B}_k} e^{-[c(B)-c(B_0)]/\tau},$$

and let $C_{\tau,k}$ be the covariance conditional on $B \in \mathcal{B}_k$. For $k \ge 1$, define the scalar tails

$$b_k(s) = \sum_{j \ge k+1} \frac{s^j}{j!}, \qquad a_k(s) = \sum_{j \ge k} (j+1)\frac{s^j}{j!}. \tag{8}$$

Theorem 3.3. *For every $\tau > 0$ and $k \ge 1$, with $s = s_\tau$ and $z = z_k$,*

$$\frac{z}{z + b_k(s)} C_{\tau,k} \preceq C_\tau \preceq (1 + 2z a_k(s) + 2 b_k(s)) C_{\tau,k}. \tag{9}$$

If B_0 is uniquely optimal, then $\|C_{\tau,k}^{-1/2} C_\tau C_{\tau,k}^{-1/2} - I\|_{\mathrm{op}} = O(s_\tau^k)$.

Appendix A bounds the omitted mass and second moment separately, without a lower bound on the smallest covariance eigenvalue. In particular, if $b = \sum_h q_h d_h$, the one-exchange covariance is

$$C_{\tau,1} = \frac{L_\tau}{1 + s_\tau} - \frac{b b^\top}{(1 + s_\tau)^2}. \tag{10}$$

Enumerating $\mathcal{B}_k$ takes at most $\sum_{j=0}^{k} \binom{r}{j}\binom{|E|-r}{j}$ independence tests for rank r. These counts are polynomial only for fixed k.

This construction also identifies the scales in Section 2. If B_0 is optimal, exchange costs are nonnegative. The span of basis differences of cost at most t is therefore the span of d_h with $c_f - c_e \le t$: every exchange is a basis difference, and every summand in an exchange representation has cost at most the nonnegative total. For features $F\mathbf{1}_B$, add exchange vectors $F d_h$ in increasing cost order whenever they enlarge the span.

4 A local bound for hard-core models

The preceding comparison depends on total activity. A different bound applies when the admissible configurations are the compatible subsets of a finite incompatibility graph. Write $h \not\sim j$ for incompatibility, including $h \not\sim h$, and assign activity $q_h > 0$ and feature d_h to each vertex. For a compatible set I,

$$p(I) = Z^{-1} \prod_{h \in I} q_h, \qquad x_I = \sum_{h \in I} d_h.$$

Set

$$C = \mathrm{Cov}(x_I), \qquad L = \sum_h q_h d_h d_h^\top, \qquad \delta = \max_h \sum_{j:\, j \not\sim h} q_j. \tag{11}$$

The sum defining δ includes h itself. General matroid bases need not admit this hard-core representation.

Let $\phi(h_1, \dots, h_n)$ be the Ursell coefficient: the sum of $(-1)^{|A|}$ over connected spanning edge sets A of the incompatibility graph on the labeled positions $1, \dots, n$, with $\phi(h) = 1$. Repeated labels are allowed. Differentiating the Mayer expansion with respect to the feature tilt $q_h \mapsto q_h e^{t^\top d_h}$ gives the order-k Hessian [3, (2.3)–(2.4)]:

$$H_k = \sum_{n=1}^{k} \frac{1}{n!} \sum_{h_1, \dots, h_n} \phi(h_1, \dots, h_n) \prod_{i=1}^{n} q_{h_i} \left(\sum_i d_{h_i}\right)\left(\sum_i d_{h_i}\right)^\top. \tag{12}$$

The following remainder is expressed in the quadratic form L.

Theorem 4.1. *If $\delta < e^{-1}$, then, for every $k \geq 1$,*

$$-R_k(\delta)L \preceq C - H_k \preceq R_k(\delta)L, \qquad R_k(\delta) = \sum_{n \geq k+1} \frac{n^n}{n!}\delta^{n-1}. \tag{13}$$

In particular, $R_k(\delta) \leq e(e\delta)^k/(1-e\delta)$. If $\eta_k = \lambda_{\min}(L^{-1/2}H_kL^{-1/2}) > R_k(\delta)$, then

$$(1-\varepsilon_k)H_k \preceq C \preceq (1+\varepsilon_k)H_k, \qquad \varepsilon_k = R_k(\delta)/\eta_k < 1. \tag{14}$$

The constants are independent of the number of vertices, feature dimension, and activity ratios.

Proof. The tree bound for Ursell coefficients [3, Proposition 5 and (4.5)] bounds $|\phi(h_1, \dots, h_n)|$ by the sum, over labeled spanning trees T, of $\prod_{ij \in T} \mathbf{1}\{h_i \not\sim h_j\}$. Fix v and write $a_h = v^\top d_h$. For the nth matrix term A_n in (12), use $(\sum_i a_{h_i})^2 \leq n \sum_i a_{h_i}^2$. Symmetry of the labels then gives

$$|v^\top A_n v| \leq \frac{n^2}{n!} \sum_T \sum_{h_1,\dots,h_n} q_{h_1} a_{h_1}^2 \prod_{i=2}^{n} q_{h_i} \prod_{ij \in T} \mathbf{1}\{h_i \not\sim h_j\}.$$

Root T at position 1. Summing the labels from its leaves toward the root contributes at most δ for each of the $n-1$ nonroot vertices. The unsummed root contributes $\sum_h q_h a_h^2 = v^\top L v$. There are n^{n-2} labeled trees, so

$$|v^\top A_n v| \leq \frac{n^n}{n!}\delta^{n-1} v^\top L v. \tag{15}$$

The Kotecký–Preiss condition with constant control parameter 1 reduces to $e\delta < 1$ [3, (2.15)]. It remains valid in a sufficiently small feature-tilt neighborhood, so the expansion may be differentiated there. Equation (15) gives absolute convergence of the Hessian series, and summing from $n = k+1$ proves (13). The scalar estimate follows from $n! \geq (n/e)^n$. Finally, $H_k \succeq \eta_k L$ gives (14). □

The second-order expression is particularly simple:

$$H_2 = L - 2\sum_h q_h^2 d_h d_h^\top - \sum_{\{h,j\}:\, h \neq j,\ h \not\sim j} q_h q_j (d_h + d_j)(d_h + d_j)^\top. \tag{16}$$

The middle term arises from repeated labels. The approximation need not be positive semidefinite outside the stated regime.

If (14) holds with one $\varepsilon < 1$ along a parameter segment $\theta + tu$, it also bounds likelihood differences. The log-partition integral identity [9] gives

$$\mathrm{KL}(p_\theta \| p_{\theta+u}) = \int_0^1 (1-t) u^\top C_{\theta+tu} u \, dt.$$

Replacing C by H_k in this integral changes it by at most the factors $1 \pm \varepsilon$.

5 Empirical evaluation

We test the sampling threshold, the hard-core approximation on growing models, and the exchange approximation in conditional prediction. Exact inference supplies the population covariance and reference likelihood in every experiment. All fitting methods use the same objective and gradient; only the Hessian approximation changes.

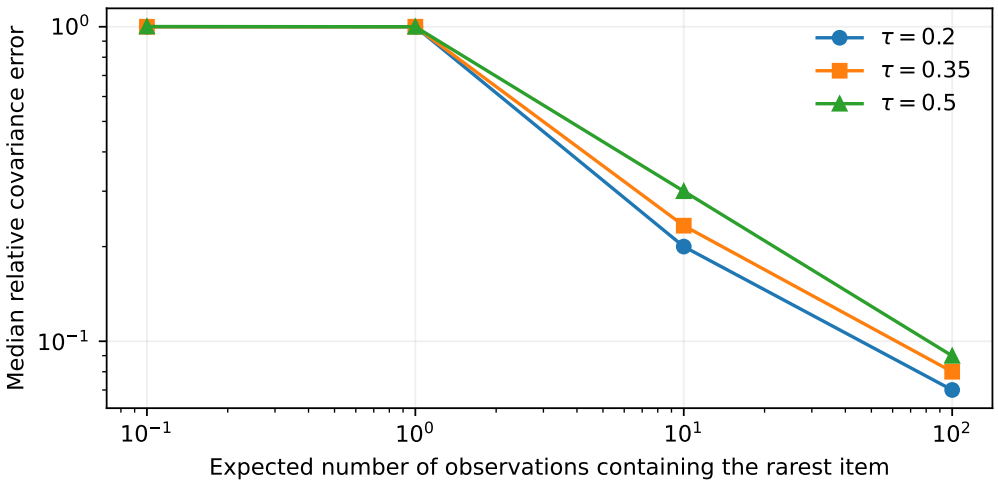


Figure 1: Relative error of the empirical covariance, with 2,000 replicates per point. Curves report medians, not fitted asymptotic rates. The three temperatures require very different absolute sample sizes to reach the same expected rare-item count.

5.1 Rare observations and uncertainty

We use a size-three subset model on ten items with costs $(0, 0, 0, 1, 1.5, 2, 2.5, 3, 3.5, 4)$ and temperatures $0.2, 0.35, 0.5$. The last item is needed to complete the feature span, so $\gamma_* = 4$. At each temperature, the sample sizes give expected counts $0.1, 1, 10, 100$ for that item. Each of the twelve settings has 2,000 independent multinomial replicates, for 24,000 repeated datasets. A multinomial draw over the 120 subsets generates their sufficient counts exactly, without materializing every observation.

Figure 1 shows the transition predicted by Proposition 2.2. At $\tau = 0.35$, $N = 35{,}023$ gives one expected occurrence of the rarest item, and the sample covariance is singular in 36.7% of replicates. Its median relative error falls from 1 to 0.233 and 0.080 as the expected count rises from 1 to 10 and 100.

For the rare-item marginal, we compare Wald and Clopper–Pearson intervals at nominal 95% coverage [2]. At $\tau = 0.35$ and one expected occurrence, their measured coverages are 64.05% and 97.80%; at 100 expected occurrences they are 95.15% and 95.55%. With 2,000 replicates, the Monte Carlo standard error of a coverage proportion is at most 1.12 percentage points.

At $\tau = 0.35$, the model-based relative errors of L, C_1, and C_2 are 0.437, 0.0805, and 0.000422. These values use the specified model probabilities and are not estimators of an unknown law.

5.2 Likelihood fitting on growing strips

Consider a hard-core model on a three-row grid strip of length 32, 128, or 512, with 96, 384, or 1,536 binary variables. Four shared parameters determine the activities; the last row has activities smaller by a factor of 100. The base activity is 0.01, 0.1, or 0.5. For each of the nine settings, five independent datasets contain 200 training and 5,000 held-out configurations. Exact transfer-matrix inference and moment propagation provide the reference [5]. The number of variables grows while the model retains four shared parameters.

At base activity 0.01, total activity grows from 0.686 to 10.856, while $\delta = 0.0554$ is unchanged. The relative error of H_2 stays below 0.00943, and the bound in the normalization (1) stays below 0.0181 (Figure 2). The leading approximation L has error about 0.103. The bound remains informative even as total activity increases with strip length.

We fit the penalized likelihood by exact Newton, Newton with L, Newton with H_2, and

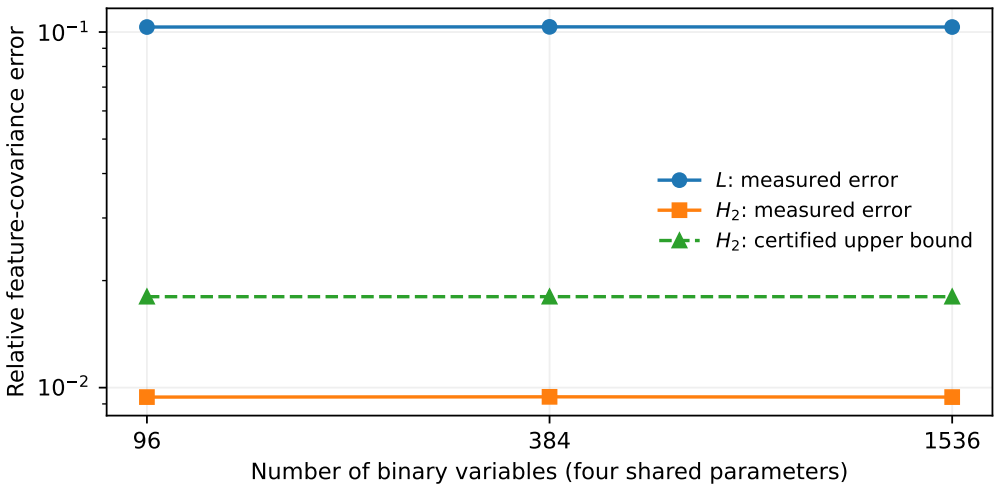


Figure 2: Growing strips at base activity 0.01. Covariances refer to the four shared features. The upper curve is $\varepsilon_2/(1-\varepsilon_2)$, in the same normalization as the measured errors.

Table 1: Mean parameter updates on grid strips, averaged over three sizes and five seeds. All Newton methods have a 50-update limit. The two H_2 columns differ only in their fallback test.

Base activity	Exact	L	H_2: positivity	H_2: certificate	L-BFGS
0.01	6.73	8.20	6.73	6.73	15.53
0.10	5.00	22.67	11.00	5.00	15.33
0.50	8.13	50.00	7.93	7.93	15.20

L-BFGS. The H_2 method uses exact curvature when H_2 is not positive definite. A second version uses H_2 only when $\varepsilon_2 \leq 0.1$. Table 1 summarizes the 15 datasets at each activity level.

At base activity 0.01, H_2 uses no exact-curvature fallback and reaches the same fitted likelihood and held-out loss as exact Newton. At 0.1, $\delta = 0.554 > 1/e$ and H_2 is indefinite at the generating parameter. The positivity-only version reaches the update limit on two of 15 datasets, whereas the $\varepsilon_2 \leq 0.1$ version reaches the gradient tolerance on all 15 by reverting to exact curvature. At 0.5, successful H_2 runs also use this fallback. The L method reaches the update limit in all 15 runs and has mean held-out loss 0.28570 per vertex, compared with 0.28563 for exact Newton.

5.3 Conditional subset prediction from images

The image experiment uses the 1,797 handwritten images supplied by `load_digits`, derived from the UCI optical-recognition data and distinct from MNIST [1, 6]. The upper half of an 8×8 image predicts the three brightest horizontally pooled cells in its lower half. The target is a size-three subset of 16 cells. The linear conditional model has 60 free parameters and 560 possible outputs.

For each of five seeds, we use 1,077 training, 360 validation, and 360 test images. Principal components are fitted on the training inputs, and the regularization coefficient is selected by validation likelihood. Every solver uses that coefficient. The comparison includes exact Newton, diagonal curvature, L, C_1, C_2, two Monte Carlo covariance estimates, a bound-based fallback, and L-BFGS. The Monte Carlo methods draw 32 or 128 independent model outputs per training context at each update.

The second-order truncation uses fewer updates than C_1, but exact Newton is faster than

Table 2: Image-derived prediction task. Values are means and standard deviations across five splits, except the last column, which is the maximum final penalized training-loss gap. Times are single-thread reference measurements.

Curvature / solver	Updates	Seconds	Maximum gap
Exact Newton	8.0 ± 0.0	0.200 ± 0.029	$3.1 \cdot 10^{-11}$
Diagonal	100.0 ± 0.0	2.222 ± 0.303	$3.8 \cdot 10^{-3}$
L	100.0 ± 0.0	3.686 ± 0.333	$2.1 \cdot 10^{-8}$
C_1	17.4 ± 0.5	0.781 ± 0.088	$8.9 \cdot 10^{-14}$
C_2	10.0 ± 0.0	0.412 ± 0.069	$4.9 \cdot 10^{-12}$
Bound-based fallback	8.0 ± 0.0	0.389 ± 0.035	$3.1 \cdot 10^{-11}$
Monte Carlo, 32	17.6 ± 1.5	0.568 ± 0.026	$2.3 \cdot 10^{-11}$
Monte Carlo, 128	12.0 ± 0.7	0.808 ± 0.058	$3.2 \cdot 10^{-11}$
L-BFGS	100.0 ± 0.0	1.326 ± 0.057	$3.8 \cdot 10^{-8}$

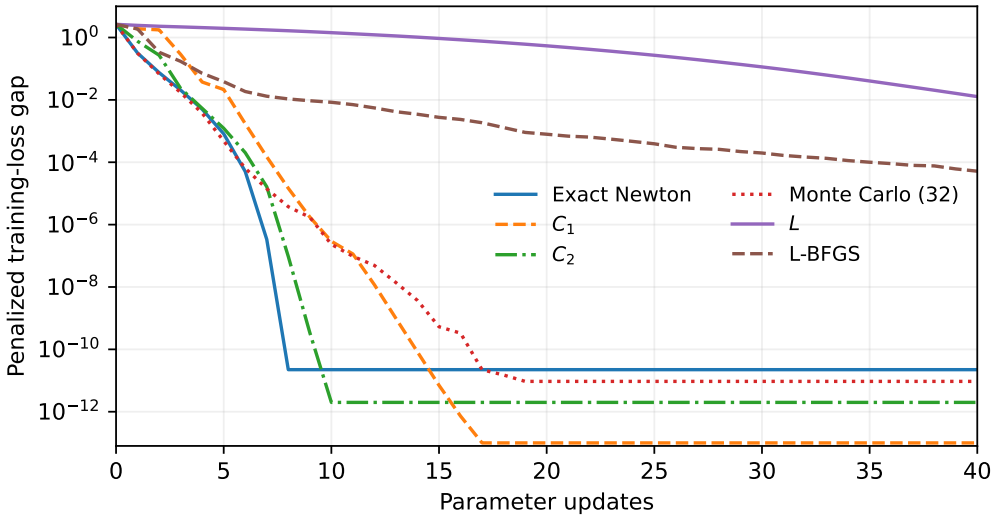


Figure 3: Median training-loss gap across the five image splits. The first 40 updates are shown; L and L-BFGS continue to 100. Traces are held constant after stopping and clipped below 10^{-13} for display. Different curves change the optimization method, not the fitted model.

both (Table 2, Figure 3). Exact Newton and C_2 have the same mean test negative log-likelihood to five decimal places, 3.70735, with split standard deviation 0.05301. Their mean Brier loss is 0.10481 and mean top-three recall is 0.54185.

On 64 fixed training contexts per split, the median sum of exchange weights lies between 5.98 and 6.96, and the median relative error of C_2 lies between 0.211 and 0.248. Neither truncation bound accepts any of these 320 contexts at tolerance 0.1, so the fallback uses exact curvature.

The prototype enumerates all 560 outputs when forming C_2 and computes exact covariances before applying the fallback mask. It therefore compares curvature approximations and solver behavior rather than end-to-end inference complexity.

6 Discussion

Relative covariance error reveals two failures that absolute error can obscure. An independent sample can miss a feature direction, while a deterministic approximation can preserve that direction but misestimate its scale. The support calculation identifies the sampling obstruction,

and the exchange and hard-core bounds give conditions for controlling the approximation error when model probabilities or activities are available.

On the grid strips, the local activity condition controls curvature as the number of variables grows; a positivity check alone is weaker. On the image-derived conditional model, the exchange bounds do not apply at the fitted distributions, and exact Newton is faster than the truncations.

The finite-support asymptotics concern a fixed model; only the hard-core bound has constants independent of system size. The hard-core graph is known, the strip model has four shared parameters, and the image task is small with overlapping random splits. Extending the computational conclusion would require an implementation that reduces total inference work.

A Proof of the truncation bound

Use the subset representation in the proof of Theorem 3.1, translating the reference feature to zero. A basis at distance j has a representation with j exchanges. Let t be the unnormalized mass omitted by truncation at k, and let

$$U = \sum_{F:\,|F|>k} q_F x_F x_F^\top$$

be its unnormalized second moment. Expanding $(\sum_h q_h)^j$ shows that the sum of products over distinct j-element subsets is at most $s^j/j!$. Hence $t \le b_k(s)$.

For any v, write $a_h = v^\top d_h$. The inequality $(\sum_{h\in F} a_h)^2 \le |F| \sum_{h\in F} a_h^2$ gives, after fixing an exchange h in the omitted sets,

$$v^\top U v \le \sum_h q_h a_h^2 \sum_{j\ge k} (j+1)\frac{s^j}{j!} = a_k(s) v^\top L v.$$

Thus $U \preceq a_k(s) L$. Let μ_k and C_k denote the retained mean and covariance, and let $z = z_k$. The retained distribution contains the empty set and every singleton, so Lemma 3.2 gives $L \preceq z^2 C_k$. Its atom of mass $1/z$ at zero also gives $\mu_k \mu_k^\top \preceq (z-1) C_k$: in any direction, weighted Cauchy–Schwarz bounds the squared mean by $(1 - 1/z)$ times the second moment.

The covariance formula for a mixture first yields

$$C \succeq \frac{z}{z+t} C_k \succeq \frac{z}{z + b_k(s)} C_k.$$

For the upper bound, center the full distribution at μ_k. Covariance is no larger than a second moment about any fixed center, and $(x-\mu_k)(x-\mu_k)^\top \preceq 2xx^\top + 2\mu_k\mu_k^\top$. Therefore

$$\begin{aligned} C &\preceq \frac{z C_k + 2U + 2t\mu_k\mu_k^\top}{z+t} \\ &\preceq \frac{z + 2a_k(s) z^2 + 2t(z-1)}{z+t} C_k \preceq \big(1 + 2z a_k(s) + 2 b_k(s)\big) C_k. \end{aligned}$$

This proves (9). At a unique optimum, $s \to 0$, $z \to 1$, $a_k(s) = O(s^k)$, and $b_k(s) = O(s^{k+1})$, which proves the relative rate.

B Experimental protocol and checks

Sampling experiment. We enumerate all size-three subsets and compute probabilities by log-sum-exp. Population covariances are centered at the most probable subset to avoid cancellation. One redundant coordinate, belonging to the reference basis, is removed before solving the generalized symmetric eigenvalue problems in (1); the rare-item coordinate remains. Empirical covariances use divisor N. Wald intervals are clipped to $[0, 1]$, and Clopper–Pearson endpoints use the usual beta quantiles. Seeds, sample sizes, coverage widths, and empirical-error quantiles are in `sampling.json`.

Strip model and fitting. Columns are indexed by $c = 0, \dots, \ell - 1$ and rows by $r = 0, 1, 2$. Vertex (c, r) has feature

$$d_{c,r} = (1, (-1)^c, \sin(2\pi c/17), \mathbf{1}\{r = 2\})^\top, \qquad q_{c,r} = e^{\theta^\top d_{c,r}}.$$

The generating parameter is $(\log q_0, 0.25, 0.3, -\log 100)$. The five allowed bitmasks within a column define the transfer matrix; neighboring columns are compatible when their masks are disjoint. First- and second-moment propagation computes likelihood derivatives [5], and backward sampling from forward messages generates training and held-out data. We minimize negative log-likelihood per vertex plus $10^{-6}\|\theta\|^2/2$, starting from $(\log 0.02, 0, 0, -2)$. Newton variants stop at gradient infinity norm 10^{-9} or 50 updates; L-BFGS allows 100 updates.

The bound for R_2 sums 500 positive terms and bounds the remaining tail geometrically. It is computed from the model, not from the measured covariance error. Full fitting traces record each use of exact curvature.

Image preprocessing and fitting. Each upper-half pixel is divided by 16. The lower half is pooled into 16 cells by summing adjacent horizontal pairs, and the three largest sums define the target with stable ordering for ties. Splits are stratified by digit identity, which is not an input feature. Three principal components are fitted on training inputs, scaled by their training standard deviations, and combined with an intercept. One of 16 item scores is fixed at zero, leaving $4 \times 15 = 60$ free parameters. The model assigns probability proportional to the exponential of summed item scores over a size-three subset.

The penalty is $\lambda\|\theta\|^2/2$ for $\lambda \in \{10^{-2}, 10^{-3}, 10^{-4}\}$, selected by validation negative log-likelihood; every split selects 10^{-4}. Solvers start at zero and stop at gradient infinity norm 10^{-7} or 100 updates. Newton steps use Armijo backtracking with sufficient-decrease constant 10^{-4} and step halving. The diagonal baseline uses the diagonal of the exact aggregate Hessian. Monte Carlo covariances use fresh independent output samples at every context and update.

The fallback accepts C_1 first when its bound is at most 0.1, then C_2 under the same condition, and otherwise uses exact curvature. Split indices, principal-component variance ratios, validation losses, fallback fractions, and traces are in `digits.json`. The code loads the public dataset through scikit-learn.

Numerical checks. The supplied scripts use Python 3.13.5, NumPy 2.3.5, SciPy 1.17.0, scikit-learn 1.8.0, and single-thread linear algebra. Numba compilation is excluded from the strip timing, and image timings include diagnostic callbacks and reference operations.

The verification script checks 236 matrix inequalities using exact rational arithmetic: 100 subset comparisons, 24 matroid comparisons, 12 truncation comparisons, and 100 cluster-remainder comparisons. A four-scale model with tied ground states and noncommensurate positive energies is evaluated at 100-digit precision. Its normalized covariance error decreases from 0.363 at $\tau = 0.1$ to $1.87 \cdot 10^{-8}$ at $\tau = 0.005$. Conditional-model finite-difference gradient and Hessian errors are below $7 \cdot 10^{-11}$. On a 3×3 grid, exact enumeration and transfer-matrix moments agree to $3 \cdot 10^{-16}$, and a derivative check has error below $2 \cdot 10^{-11}$.

References

[1] E. Alpaydin and C. Kaynak. Optical recognition of handwritten digits. UCI Machine Learning Repository, 1998. URL `https://doi.org/10.24432/C50P49`.

[2] Lawrence D. Brown, T. Tony Cai, and Anirban DasGupta. Interval estimation for a binomial proportion. *Statistical Science*, 16(2):101–133, 2001. doi: 10.1214/ss/1009213286. With discussion.

[3] Roberto Fernández and Aldo Procacci. Cluster expansion for abstract polymer models: New bounds from an old approach. *Communications in Mathematical Physics*, 274:123–140, 2007. doi: 10.1007/s00220-007-0279-2.

[4] Stasys Jukna and Hannes Seiwert. Tropical Kirchhoff's formula and postoptimality in matroid optimization. *Discrete Applied Mathematics*, 289:12–21, 2021. doi: 10.1016/j.dam.2020.09.018.

[5] Zhifei Li and Jason Eisner. First- and second-order expectation semirings with applications to minimum-risk training on translation forests. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 40–51. Association for Computational Linguistics, 2009. URL `https://aclanthology.org/D09-1005/`.

[6] scikit-learn developers. `load_digits`. scikit-learn 1.8.0 documentation, n.d. URL `https://scikit-learn.org/1.8/modules/generated/sklearn.datasets.load_digits.html`. Accessed 7 September 2026.

[7] Kevin Swersky, Brendan J. Frey, Daniel Tarlow, Richard S. Zemel, and Ryan P. Adams. Probabilistic n-choose-k models for classification and ranking. In *Advances in Neural Information Processing Systems*, volume 25, 2012. URL `https://papers.neurips.cc/paper/4702-probabilistic-n-choose-k-models-for-classification-and-ranking`.

[8] Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12:389–434, 2012. doi: 10.1007/s10208-011-9099-z.

[9] Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008. doi: 10.1561/2200000001.