THE LAW OF THE PERTURBED OPTIMIZER

Anonymous authors
Paper under double-blind review

ABSTRACT

Perturb-and-MAP samples a combinatorial structure by adding noise to the scores of its elements and returning the optimal structure, as the Gumbel-max trick does for a single choice. Because optimizing can be easy where normalizing the Gibbs distribution is #P-hard, perturb-and-MAP draws are used in place of draws from that distribution, although the law they follow has been regarded as intractable. We prove that for fixed-size subsets, spanning trees and all other matroid bases, the two laws coincide at every score vector only when the structure splits into separate single choices, even with dependent noise entries. For assignments of three or more rows, no noise achieves this, so we instead show that the law under independent Gumbel noise is an integral of a determinant over the assignment problem's dual variables. This law gives the noise scale at which it agrees with the Gibbs distribution to first order in the scores, $n/(n+1)$ for $n \times n$ assignments. Importance weights built from the law turn the draws into consistent estimates of that distribution's marginals, accurate to 0.001 on multi-target tracking scenes where belief propagation and Sinkhorn scaling err by 0.06 to 0.32.

1 INTRODUCTION

The Gumbel-max trick samples by optimizing: when independent standard Gumbel noise is added to n scores, the index of the largest perturbed score follows the softmax of the scores (Gumbel, 1954; Luce, 1959; Yellott, 1977). Perturb-and-MAP (Papandreou & Yuille, 2011; 2014) extends the trick from indices to combinatorial structures. A structure is a set of elements, such as k of n items, the edges of a spanning tree or the entries of an assignment of rows to columns, and its score is the sum of the scores of its elements. Noise is added to the score of every element and the structure of largest perturbed score is returned, so a draw costs one run of a combinatorial algorithm: a sort for subsets, the greedy algorithm for spanning trees, the Hungarian algorithm for assignments (Kuhn, 1955; Munkres, 1957). In machine learning the procedure draws subsets (Kool et al., 2019), permutations (Mena et al., 2018) and other structures (Tarlow et al., 2012; Paulus et al., 2020; Berthet et al., 2020; Niepert et al., 2021) as latent variables and as structured predictions.

The counterpart of the softmax for structures is the Gibbs distribution, $p_\theta(x) \propto e^{\langle \theta, x \rangle}$ for the score vector θ and the indicator vector x of the structure. Its partition function is a sum over all structures, and its marginals, the probabilities that the structure contains a given element, are ratios of such sums. For assignments the partition function is a permanent, a #P-hard quantity (Valiant, 1979a), and the marginals are ratios of permanents. In multi-target tracking these marginals are the association probabilities of tracks and measurements (Fortmann et al., 1983), approximated for instance by belief propagation (Williams & Lau, 2014). The assignment of largest score, unlike the permanent, is found in polynomial time, and perturb-and-MAP is relied on because of this gap between summing and optimizing (Tarlow et al., 2012). Its draws stand in for draws from p_θ and their averages for the marginals (Papandreou & Yuille, 2011; Niepert et al., 2021). The temperature, the scale of the noise relative to the scores, is tuned as a hyperparameter (Berthet et al., 2020; Niepert et al., 2021).

Perturb-and-MAP would sample p_θ if every structure received an independent Gumbel variable of its own (Papandreou & Yuille, 2011, Proposition 3), but there are exponentially many structures, and their perturbed scores would no longer be sums over elements, as the algorithms above require. With noise on the elements the output does not follow p_θ in general (Lázaro-Gredilla et al., 2021), and its distribution is regarded as intractable. Papandreou & Yuille (2011, eq. (4)) wrote the probability of an output as the volume of a shifted polyhedral cone, weighted by the noise density, and remarked: "It is intractable (NP-hard) to compute the volume of general polyhedra in a high-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

dimensional space." For assignments under Gumbel noise, Mena et al. (2018) state that "neither the Gumbel-Matching nor Gumbel-Sinkhorn distributions have tractable densities," the first being the distribution of the output. That distribution would tell how far the draws are from p_θ, whether another noise brings them closer and which temperature brings them closest, and it would give every draw a probability for use in a likelihood or an importance weight.

> *When does adding noise to the scores and returning the optimal structure sample the Gibbs distribution? What does it sample when it does not?*

We call the procedure the perturbed optimizer. At fixed scores, the distribution of its output is its law. Its noise law is the joint law of the noise entries, which may be independent or dependent. As in perturb-and-MAP (Papandreou & Yuille, 2011), it does not depend on the scores, for otherwise the noise could encode a draw from p_θ. A noise law is exact if it makes the perturbed optimizer sample the Gibbs distribution at every score vector.

The answer to the first question is almost never (Theorem 1): over the bases of a matroid, which include the k-subsets and the spanning trees of a graph, an exact noise law exists only when the structure splits into separate single choices, and over assignments of three or more rows none exists. Under independent Gumbel noise, the answer to the second is an explicit integral, and for assignments an integral of a determinant over the dual variables of the assignment problem. The classification and the law over assignments are the first results of their kind: no earlier classification holds for every noise law, and no earlier work computes the law of a perturbed optimizer over assignments. The law then has two uses. It gives the temperature at which the perturbed optimizer agrees with p_θ to first order, in closed form, and an importance weight for every draw.

Over assignments, where the perturbed optimizer is known as Gumbel matching (Mena et al., 2018), the law is an integral of a determinant times a log-concave function over the column potentials, the dual variables of the assignment problem attached to its columns (Theorem 2). For n rows and n columns the integral is over $n-1$ variables, whereas the cone of Papandreou & Yuille (2011) is a region in the space of the n^2 noise entries. The determinant comes from linear programming duality and the matrix-tree theorem (Kirchhoff, 1847; Tutte, 1948): the canonical row and column potentials of an optimal assignment bound the perturbed score of every entry, the non-assigned entries at which the bound is attained link the columns into a tree, each entry joining its column to the column assigned to its row, and each tree contributes a product of softmax probabilities (Section 3.3).

Expanded at zero scores, the law shows what a temperature can correct. Under standard Gumbel noise, for k of n elements with $1 \le k < n$ and for $n \times n$ assignments with $n \ge 2$, exactly one temperature makes the perturbed optimizer agree with p_θ to first order in the scores (Theorem 3). That temperature is $(n/k)(H_n - H_{n-k})$ for the subsets, with H_n the nth harmonic number, and $n/(n+1)$ for the assignments. Standard Gumbel noise, exact for a single choice, is therefore too small for larger subsets and too large for assignments. Under the same noise, for k-subsets with $2 \le k \le n-2$ and square assignments with $n \ge 3$, no deterministic reparametrization of the scores makes the perturbed optimizer agree with p_θ to second order, whereas a reparametrization drawn at random before each optimization does (Theorem C.1 and Proposition C.3).

The law also turns the draws of Gumbel matching into estimates under p_θ. The weight of a draw is the exponential of its score divided by its probability under Gumbel matching, which the law provides. The permanent cancels from the self-normalized average, and the determinant integral is estimated by Monte Carlo, with draws of the column potentials from a log-concave density. The estimator covers square assignments and the rectangular assignments of data association, in which every row has a private column that no other row may use, standing for a missed detection (Section 5). The estimator is consistent at every temperature (Lemma D.1) and runs at $n/(n+1)$, the first-order temperature of square assignments (Theorem 3). It is accurate to 0.001 on multi-target tracking scenes where belief propagation and Sinkhorn scaling err by 0.06 to 0.32 (Section 6).

2 Preliminaries

Notation. $[n] = \{1, \dots, n\}$, and n is the size of the instance throughout: the number of elements for subsets, of rows for assignments. $\langle \cdot, \cdot \rangle$ is the Euclidean inner product, $\|\cdot\|$ the Euclidean norm of a vector or of the entries of a matrix, the latter also written $\|\cdot\|_F$, I the identity matrix, $\mathbf{1}$ the all-ones vector and $\mathbf{1}\{\cdot\}$ the indicator of an event.

2.1 Structures and the Gibbs Distribution

Let E be a finite set, the ground set. A structure is a vector $x \in \{0,1\}^E$, the indicator of a subset of E, and for $B \subseteq E$, x_B is the indicator of B. A structure set is a nonempty $\mathcal{X} \subseteq \{0,1\}^E$, and the scores form a vector $\theta \in \mathbb{R}^E$. The Gibbs distribution on $\mathcal{X}$ with scores θ is $p_\theta(x) = e^{\langle \theta, x\rangle}/Z(\theta)$, with partition function $Z(\theta) = \sum_{x\in\mathcal{X}} e^{\langle\theta,x\rangle}$, and its marginals are $\mathbb{E}_{p_\theta}[x] \in [0,1]^E$, the probabilities that the coordinates equal one. Three structure sets are used throughout. The k-subsets of $[n]$ are $\{x_B : B \subseteq [n],\ |B| = k\}$ on $E = [n]$. A matroid $\mathcal{M}$ on E is given by a nonempty family $\mathcal{B}(\mathcal{M})$ of subsets of E of equal size, its bases, that satisfies the basis exchange axiom (Oxley, 2011): for bases $B \neq B'$ and $b \in B \setminus B'$, some $b' \in B' \setminus B$ makes $(B \setminus \{b\}) \cup \{b'\}$ a basis. Its structure set is $\{x_B : B \in \mathcal{B}(\mathcal{M})\}$. For $N \geq n$, an assignment is an injective map $\sigma : [n] \to [N]$, with assignment matrix $P_\sigma \in \{0,1\}^{n\times N}$, $(P_\sigma)_{ij} = \mathbf{1}\{j = \sigma(i)\}$, and a bijective assignment ($N = n$) is a permutation. The ground set is the set $[n] \times [N]$ of entries, so the scores form a matrix $\theta \in \mathbb{R}^{n\times N}$ with $\langle \theta, P_\sigma\rangle = \sum_i \theta_{i\sigma(i)}$, the marginals form the marginal matrix $M(\theta) = \mathbb{E}_{p_\theta}[P_\sigma]$, and for $N = n$ the partition function is the permanent $Z(\theta) = \operatorname{perm}(\exp\theta)$ of the entrywise exponential of θ.

2.2 The Perturbed Optimizer and Its Law

The noise is a random vector $\xi \in \mathbb{R}^E$ whose law ν, the noise law, does not depend on the scores. Standard Gumbel noise has independent entries with distribution function $\mathbb{P}(\xi_e \leq q) = \exp(-e^{-q})$, and negative Gumbel noise is $-\xi$ for such a ξ.

Definition 1 (perturbed optimizer, its law and temperature)**.** Let $\mathcal{X} \subseteq \{0,1\}^E$ be a structure set and let ξ be noise with law ν. The perturbed optimizer with scores $\theta \in \mathbb{R}^E$ is the random structure $x^*(\theta) = \arg\max_{x\in\mathcal{X}} \langle \theta + \xi, x\rangle$, ties broken by a fixed rule that does not depend on θ. Its law is $Q^\nu_\theta(x) = \mathbb{P}(x^*(\theta) = x)$, $x \in \mathcal{X}$. For standard Gumbel noise we write Q_θ. The perturbed optimizer at temperature $c > 0$ is $\arg\max_{x\in\mathcal{X}} \langle \theta + c\,\xi, x\rangle$, with law $Q^\nu_{\theta/c}$.

An assignment is identified with its matrix, so that $x^* = \sigma$ means $x^* = P_\sigma$. Gumbel top-k (Kool et al., 2019) is the perturbed optimizer over the k-subsets of $[n]$ with standard Gumbel noise, and Gumbel matching the perturbed optimizer over assignments with standard Gumbel noise.

3 The Law of the Perturbed Optimizer

From Section 3.2 on, the scores fed to the optimizer are written η, because Sections 4 and 5 calibrate them while θ stays the score vector of p_θ. The proofs are in Appendices A and B.

3.1 Exact Noise Laws

A noise law ν is exact for $\mathcal{X}$ if $Q^\nu_\theta = p_\theta$ for every $\theta \in \mathbb{R}^E$. The uniform matroid $U_{k,\ell}$ has the k-subsets of an ℓ-element set as bases, and the direct sum of matroids on disjoint ground sets has as bases the unions of one basis of each (Oxley, 2011). A cactus is a connected graph in which every edge lies on at most one cycle. The total variation distance between two distributions on a finite set is $\mathrm{TV}(p,p') = \frac{1}{2}\sum_x |p(x) - p'(x)|$.

Theorem 1 (classification of exact noise laws)**.** *(a) Let $\mathcal{M}$ be a matroid on E and $\mathcal{X} = \{x_B : B \in \mathcal{B}(\mathcal{M})\}$. An exact noise law for $\mathcal{X}$ exists if and only if $\mathcal{M}$ is a direct sum of uniform matroids of the forms $U_{1,\ell}$ and $U_{\ell-1,\ell}$, $\ell \geq 1$. For the k-subsets of $[n]$ with $2 \leq k \leq n-2$, every noise law satisfies $\sup_{\theta\in\mathbb{R}^n} \mathrm{TV}(p_\theta, Q^\nu_\theta) \geq \delta_0 = (8\sqrt{2} - 5\sqrt{5})/12 \approx 0.0111$. For the spanning trees of a connected simple graph, an exact noise law exists if and only if the graph is a cactus.*

(b) For assignments of $n \geq 3$ rows to $N \geq n$ columns no noise law is exact.

For matroid bases an exact noise law thus exists if and only if the structure set is a product of single choices $U_{1,\ell}$, on which standard Gumbel noise is exact by the Gumbel-max trick, and of single exclusions $U_{\ell-1,\ell}$, choices of the element left out, on which negative Gumbel noise is exact by the same trick on the negated scores (Theorem A.1). A spanning tree of a cactus keeps every edge that lies on no cycle and leaves out one edge of each cycle, so the spanning trees of a cactus are such a product. Three elements show why no noise law can be exact for the k-subsets of $[n]$ with

$2 \le k \le n-2$: sending the other scores to $\pm\infty$ makes three elements compete for one slot or for two. Exactness for one slot fixes the joint distribution function of the two differences of their noise entries and hence, by inclusion–exclusion, their joint survival function, while exactness for two slots fixes that survival function differently, by as much as $4\delta_0$. Sending scores to $\pm\infty$ likewise reduces every matroid outside the class of Theorem 1(a) to $U_{2,4}$, where the three-element argument applies, or to a matroid on four elements with five two-element bases, where exactness would prescribe a distribution function with a negative mixed derivative (Appendix A.3). For assignments of three rows or more the same reduction leaves three rows and three columns, where exactness would make the difference of the noise sums $\langle \xi, P_\sigma \rangle$ of any two permutations standard logistic. The even permutation matrices and the odd ones both sum to the all-ones matrix, so a linear combination of the six noise sums vanishes, while the logistic differences give it a positive variance (Appendix A.5). For two rows and two columns standard Gumbel noise on the first row alone is exact (Appendix B.3).

3.2 Matroid Bases

Under standard Gumbel noise the laws below come from exponential clocks. The exponential clock of an element e is $s_e = \exp(-(\eta_e + \xi_e))$. The clocks are independent exponential variables with rates $\exp(\eta_e)$, and a larger perturbed score is a smaller clock. A structure is optimal exactly when every clock outside it is at least a bound set by the clocks of the chosen elements for matroid bases and by the clocks of the assigned entries and the column potentials for assignments. The excess of a clock over its bound is its slack. Integrating out the slack leaves the survival factor at the bound, the probability that the clock exceeds it. For a k-subset B of $[n]$ with $1 \le k < n$, let $s = \min_{e \notin B} s_e$.

Proposition 1 (law of Gumbel top-k, Kool et al., 2020, eq. (31))**.** *For the k-subsets of $[n]$ with $1 \le k < n$, every $\eta \in \mathbb{R}^n$ and every k-subset B,*

$$Q_\eta(x_B) = \int_0^\infty \Big(\sum_{e \notin B} \exp(\eta_e) \Big) e^{-s \sum_{e \notin B} \exp(\eta_e)} \prod_{b \in B} \big(1 - e^{-s \exp(\eta_b)}\big)\, ds.$$

B wins exactly when every clock in B is below s, an exponential clock of rate $\sum_{e \notin B} \exp(\eta_e)$ independent of them, and conditioning on s gives the formula. Conditioning on the clocks in B instead gives the law over the bases of any matroid, and for independent identically distributed noise with a density, conditioning on the perturbed scores of the chosen elements gives it as an integral over them, the form Section 4 uses (Proposition B.1).

3.3 Assignments

For square assignments with finite scores $\eta \in \mathbb{R}^{n \times n}$, the assignment linear program relaxes the assignment problem to the doubly stochastic matrices. Its dual variables are the row potentials u_i and the column potentials $y \in \mathbb{R}^n$, with constraints $u_i - y_j \ge \eta_{ij} + \xi_{ij}$. The row-softmax matrix $S_\eta(y)$ has entries $S_\eta(y)_{ij} = \exp(\eta_{ij} + y_j) / \sum_{j'} \exp(\eta_{ij'} + y_{j'})$. The function $\psi_\eta(y) = \exp(\sum_j y_j) / \prod_{i=1}^n \sum_j \exp(\eta_{ij} + y_j)$ is log-concave in y. Both S_η and ψ_η are unchanged when a constant is added to every y_j, so Theorem 2 fixes $y_1 = 0$.

Theorem 2 (law of Gumbel matching)**.** *For finite $\eta \in \mathbb{R}^{n \times n}$ and every permutation σ of $[n]$, with $y_1 = 0$,*

$$Q_\eta(\sigma) = \frac{e^{\langle \eta, P_\sigma \rangle}}{n} \int_{\mathbb{R}^{n-1}} \psi_\eta(y)\, \det\Big(I - P_\sigma^\top S_\eta(y) + \tfrac{1}{n} \mathbf{1}\mathbf{1}^\top \Big)\, dy,$$

and the determinant is nonnegative. The assignment enters the integral only through the matrix P_σ in the determinant, so the ratio of $Q_\eta(\sigma)$ to the Gibbs probability $p_\eta(\sigma)$ is proportional to that integral.

The determinant of Theorem 2 is a sum over the spanning trees that the tight entries, the non-assigned entries at which the dual bound is attained, can form. On the event $x^* = \sigma$ the canonical potentials are the negatives of the shortest-path distances from each column to a root column, where moving a row to another column is an edge whose length is the perturbed score of the assigned entry minus that of the new entry, and the tight entries form a spanning tree of the columns. After the change of variables to the row potentials, the column potentials, this tree and the slacks of the other entries, the slacks integrate to the survival factors of their clocks, the row potentials integrate in closed form

because the clocks are exponential, and each tree contributes the product of the softmax entries $S_\eta(y)_{ij}$ of its edges. The matrix-tree theorem (Chaiken, 1982) sums these products into a principal cofactor of $I - P_\sigma^\top S_\eta(y)$. The n cofactors sum to the displayed determinant, and averaging over the roots gives the $1/n$ (Appendix B.2). The same change of variables gives the law of rectangular assignments with forbidden entries of score $-\infty$ whenever every row has a private column, which no other row may use, and a finite entry in another column, the setting of data association. It is then an integral over the nonpositive potentials of the used columns, with S_η and ψ_η formed over the N columns (Theorem B.1). Over the perfect matchings of the complete graph K_{2n} with $n \geq 2$, odd cycles of tight edges occur, along which the potentials do not cancel as they do along the even cycles of a bipartite graph, so the tight edges do not form a tree. The law of a perfect matching π is then an integral over vertex potentials of the softmax probability that a fixed vertex is matched to its partner in π times the density g_π of these potentials on the event that the rest of π is optimal (Proposition B.3), and g_π is #P-hard to evaluate to a prescribed absolute accuracy.

4 Calibration

This section expands the laws of Section 3 at zero scores, and its proofs are in Appendix C.

Definition 2 (calibration and order of exactness)**.** A calibration is a smooth map $\eta : \mathbb{R}^E \to \mathbb{R}^E$ with $\eta(0) = 0$, a reparametrization of the scores before the noise is added, under which the perturbed optimizer has law $Q^\nu_{\eta(\theta)}$. A randomized calibration draws η from a fixed law on calibrations, independently of ξ, before each optimization, and the law of the output is the average of $Q^\nu_{\eta(\theta)}$ over the draw. A calibrated perturbed optimizer, deterministic or randomized, is exact to order d if the law of its output is $p_\theta + o(\|\theta\|^d)$ as $\theta \to 0$, first-order exact for $d = 1$ and second-order exact for $d = 2$. The temperature c is the calibration $\eta(\theta) = \theta/c$, and if exactly one temperature is first-order exact, it is the first-order temperature $c_1 = c_1(\mathcal{X}, \nu)$.

The section assumes that the law at zero scores is p_0, the uniform distribution on $\mathcal{X}$. Exactness to any order requires this, and the symmetry below gives it whenever the noise has a density. A constant-score direction is a score vector ϑ with $\langle \vartheta, x \rangle$ the same for every $x \in \mathcal{X}$, and adding it to the scores leaves both Q^ν_θ and p_θ unchanged. The centered scores are the score vectors orthogonal to every constant-score direction: for the k-subsets of $[n]$ with $1 \leq k < n$ those with $\sum_i \theta_i = 0$, and for square assignments those with zero row and column sums. The two Jacobians at zero scores are $J = \partial_\theta Q^\nu_\theta|_{\theta=0}$ and $J_p = \partial_\theta p_\theta|_{\theta=0}$.

The permutations of $[n]$ act on the k-subsets of $[n]$ with $1 \leq k < n$, and the permutations of the rows and of the columns act on square assignments with $n \geq 2$. Each group is transitive on the structures, and the permutations that fix a structure are transitive on its elements and on the other elements. Under every noise law that such a group leaves invariant, with $\theta \mapsto Q^\nu_\theta$ differentiable at $\theta = 0$, the two Jacobians are proportional, $J = c\, J_p$ for a scalar $c \geq 0$ (Proposition C.1). At zero scores the law then moves along the same directions as the Gibbs distribution, at c times the rate. For $c > 0$ rescaling the noise by c removes the whole first-order error and no other factor does, so c is the first-order temperature c_1. The examples of this section are Gumbel, logistic and exponential noise, with independent identically distributed entries of distribution functions $\exp(-e^{-q})$, $1/(1+e^{-q})$ and $(1-e^{-q})\mathbf{1}\{q \geq 0\}$.

Theorem 3 (first-order temperature)**.** *For the k-subsets of $[n]$ with $1 \leq k < n$ and independent identically distributed noise entries whose density f is square-integrable, with distribution function F, the law is continuously differentiable and $c_1 = n\binom{n}{k} \int f(q)^2 F(q)^{n-k-1}(1-F(q))^{k-1}\, dq$, which is $(n/k)(H_n - H_{n-k})$ for Gumbel noise, $n/(n+1)$ for logistic noise and $n/(n-k)$ for exponential noise. For square Gumbel matching with $n \geq 2$, $c_1 = n/(n+1)$.*

A deterministic calibration controls only part of a second-order error: a distribution on $\mathcal{X}$ has $|\mathcal{X}|-1$ degrees of freedom, while the laws that calibrations produce are parametrized by the centered scores alone, $\binom{n}{k} - 1$ against $n-1$ for the k-subsets of $[n]$ with $2 \leq k \leq n-2$ and $n! - 1$ against $(n-1)^2$ for the permutations of $[n]$ with $n \geq 3$. The second-order discrepancy of a calibration η is the quadratic term of the Taylor expansion of $Q^\nu_{\eta(\theta)} - p_\theta$ at $\theta = 0$. The linear part of a first-order exact calibration is division by c_1 up to constant-score directions. Its quadratic part changes the discrepancy only inside the image $\operatorname{im} J$ of J. The component orthogonal to $\operatorname{im} J$ is therefore the same for every such calibration. For the k-subsets of $[n]$ with $2 \leq k \leq n-2$ and independent

identically distributed noise, with an absolutely continuous density of integrable derivative or with the exponential distribution, the component orthogonal to $\operatorname{im} J$ of the second-order term of Q^{ν}_{θ/c_1} is a positive constant multiple of that of p_θ, which is a nonzero quadratic map of the centered scores. A second-order exact calibration therefore exists if and only if the multiple is one, which fails for Gumbel, logistic and exponential noise (Theorem C.1). The multiple is $1 - 1/n^2$ for logistic noise, below one for Gumbel noise at $k = 2$ and above one for exponential noise and for Gumbel noise with $3 \le k \le n - 2$ (Appendix C.5). Randomization removes the rest, since the law averaged over random scores is not the law at their average. Two randomized calibrations are first-order exact on average: a draw of one of two temperatures multiplies the second-order term of the law by any factor above one and a random mixing of two coordinates by any factor below one, so one of them brings the multiple to one, and a quadratic term removes what is left inside $\operatorname{im} J$ (Theorem C.1).

For square Gumbel matching with $n \ge 3$, expanding the law of Theorem 2 at zero scores gives the first-order temperature of Theorem 3 and the whole second-order term (Corollary C.1). The calibration $\eta(\theta) = \frac{n+1}{n}\theta + \frac{n+1}{2(n-2)(n+2)}\Pi(\theta \circ \theta)$ of the centered scores, with Π the double centering of a score matrix and $\circ$ the entrywise product, is first-order exact and removes the whole second-order discrepancy inside $\operatorname{im} J$. Outside $\operatorname{im} J$ the second-order term of p_θ has two components, proportional to the parts of $\sigma \mapsto \langle \theta, P_\sigma \rangle^2 \pm \sum_{i,j} \theta_{i\sigma(j)}\theta_{j\sigma(i)}$ orthogonal to the constants and to $\operatorname{im} J$, the one with the plus sign absent at $n = 3$ (Appendix C.8). On the plus and the minus component the second-order term of the calibrated law is $1 - 1/n^2$ and $1 + 1/(n(n+2))$ times that of p_θ, in that order. Both factors differ from one, by at most n^{-2}, so no deterministic calibration of Gumbel matching is second-order exact. The randomized calibration of Appendix C.8, which draws a pair of columns and a sign, then raises or lowers the temperature according to the sign while mixing the two columns, is second-order exact (Proposition C.3).

5 The Reweighted Estimator

The estimator of $M(\theta)$ is importance sampling with Gumbel matching as the proposal. Let θ be a score matrix and $c > 0$ a temperature, so that Gumbel matching at temperature c has law Q_η with $\eta = \theta/c$. The estimator draws T independent assignments $\sigma_1, \dots, \sigma_T$ from Q_η and weights each by the ratio of its unnormalized Gibbs probability to an estimate $\hat{Q}_\eta(\sigma_t)$ of its probability of being proposed, $w_t = e^{\langle \theta, P_{\sigma_t} \rangle} / \hat{Q}_\eta(\sigma_t)$:

$$\hat{M} = \frac{\sum_{t=1}^{T} w_t P_{\sigma_t}}{\sum_{t=1}^{T} w_t}.$$

The estimate $\hat{Q}_\eta(\sigma)$ is the frequency of σ among the draws if σ is frequent, that is, drawn at least 20 times, and otherwise the law of Section 3.3 evaluated with T_y samples of the column potentials from the density proportional to ψ_η (Algorithms 1 and 2, Appendices D.1 and D.2). The effective sample size of the weights, $\mathrm{ESS} = (\sum_t w_t)^2 / \sum_t w_t^2$ (Kong et al., 1994), is reported as ESS/T.

For the frequent assignments the frequencies suffice: the weights of the draws equal to a frequent assignment σ add up to $Te^{\langle \theta, P_\sigma \rangle}$ whatever its frequency, so $\hat{M}$ weighs the frequent assignments against each other by their unnormalized Gibbs probabilities, as an enumeration of them would. For the rarely drawn assignments the frequencies carry little information, and the law-based estimates take their place. Both algorithms run Gumbel matching at $c = n/(n+1)$, which for square assignments is the first-order temperature of Theorem 3. As T grows and the estimated probabilities converge, the fraction ESS/T tends to $1/(1+\chi^2)$ for the χ^2 divergence of p_θ from the proposal Q_η. For square assignments with $n \ge 3$ the temperature that maximizes this limit over a fixed compact interval around $n/(n+1)$ tends to $n/(n+1)$ as centered scores tend to zero (Proposition D.1).

5.1 Square Assignments

For square assignments Theorem 2 writes $Q_\eta(\sigma)$ as $(\Psi/n)\, e^{\langle \eta, P_\sigma \rangle}$ times the expectation of its determinant under the density ψ_η/Ψ of the column potentials on $\{y_1 = 0\}$, with the normalizer Ψ finite (Appendix B.5). Chains of the Metropolis-adjusted Langevin algorithm (MALA) of Roberts & Tweedie (1996) draw one set of samples of the column potentials, which serves every sampled assignment at one determinant per assignment and potential sample. The assignments are independent draws, and the chains sample the potentials under a density that is log-concave for every score

matrix, so no chain has to move between the modes of p_θ. The law-based estimates are known up to the constant Ψ/n, common to all assignments. An alignment constant puts both kinds of estimates on one scale. It is fitted to the frequencies of the frequent assignments, or of all sampled assignments if none is frequent (Algorithm 1). The estimator is consistent as T grows, and its law-based estimates converge to the probabilities as T and the length of the chains grow (Lemma D.1).

5.2 Private Columns

In the setting of data association (Section 3.3) the shared columns are $1, \dots, m$ and the private column of row i is $m + i$. The shared-column set $\mathcal{G}$ of an assignment is the set of shared columns it uses, and integrating the potentials of the used private columns out of the law of Theorem B.1 restricts its integral and its determinant to $\mathcal{G}$ (Lemma B.1). Write $\Psi(\mathcal{G})$ for the normalizer of ψ_η on $\{y_j \le 0,\ j \in \mathcal{G}\}$, with $y_j = 0$ elsewhere. The probability of an assignment σ is then $e^{\langle \eta, P_\sigma \rangle} \Psi(\mathcal{G})$ times the expectation, under the normalized density, of the determinant of $I - P_\sigma^\top S_\eta(y)$ restricted to $\mathcal{G}$. Assignments with the same shared-column set share one density and one normalizer, so the chains run once per nonempty shared-column set of a sampled assignment (Algorithm 2), and the normalizers of these sets, which no single constant absorbs, are estimated by annealed importance sampling run in both directions (Neal, 2001; Grosse et al., 2015). In Algorithm 2 the law-based estimates converge as well if the normalizer estimates converge (Lemma D.1).

6 Association Probabilities

On the ambiguous entries, those of the true marginals strictly between 0.05 and 0.95, the reweighted estimator is more accurate than Sinkhorn scaling, sequential importance sampling and the unweighted estimators in every row of Table E.1, and on the scenes of Figure 1(a) its error is more than two orders of magnitude below belief propagation's and Sinkhorn scaling's.

A tracker maintains n tracks, and a scan of the sensor reports m measurements $z_1, \dots, z_m$, each track producing at most one measurement and each measurement coming from at most one track (Bar-Shalom & Tse, 1975; Fortmann et al., 1983). The tracks are the rows, the measurements the shared columns, and the private column of track i is its miss column. With the density ϕ_i of the measurement predicted by track i, the detection probability P_D and the Poisson intensity λ of clutter, the measurements from no target, the scores are $\theta_{ij} = \log P_\mathrm{D} + \log \phi_i(z_j) - \log \lambda$ for $j \le m$ and $\theta_{i,m+i} = \log(1 - P_\mathrm{D})$, and every other private entry is forbidden. An assignment is an association hypothesis, and the association probabilities are the marginals $M(\theta)$ of the posterior p_θ (Fortmann et al., 1983), ratios of permanents (Uhlmann, 2004). With every target detected and no clutter a hypothesis is a permutation, the scores are Gaussian log-likelihoods (Appendix E.1), and Algorithm 1 runs in place of Algorithm 2.

The scenes are two-dimensional and come in three families (Appendix E.1). In a formation the predicted measurements sit at equal spacing on a circle and the targets have moved τ spacings along it, τ being the prediction lag. In a random scene the targets are uniform in a square, and in a crossing scene they lie on two perpendicular lines. Figure 1 shows the formations against the lag at $n = 8$ and $n = 12$, Table E.1 the three families at $n \in \{10, 15, 20\}$, and Appendix E.6 the formations at $n = 30, 40, 60$, too large for the subset dynamic program that computes the true marginals (Appendix D.5).

The baselines (Appendix E.2) are belief propagation (BP) on the association model (Williams & Lau, 2014), Sinkhorn scaling of $\exp\theta$ (augmented when detections are missed) to a doubly stochastic matrix, sequential importance sampling (IS) from that scaling (Beichl & Sullivan, 1999) and the unweighted average of T draws of Gumbel matching, at temperature one and at $c = n/(n+1)$. The mean absolute error (MAE) is taken over the ambiguous entries, and the sampling error is the MAE of the empirical marginals of T independent draws from p_θ by the dynamic program. Every sampling method draws $T = 2 \cdot 10^4$ assignments, and the numbers of Figure 1, Table E.1 and Appendix E.6 are means over three draws of each scene (ranges in Figure 1 and Appendix E.4).

On the formations with every target detected, belief propagation and Sinkhorn scaling are off by 0.06 to 0.32 at every lag, while the reweighted estimator's error stays below 0.0002, between 0.03 and 0.16 times the sampling error (Figure 1(a)). On every draw of these scenes the converged beliefs of belief propagation agree with the indicator of the most probable hypothesis to within

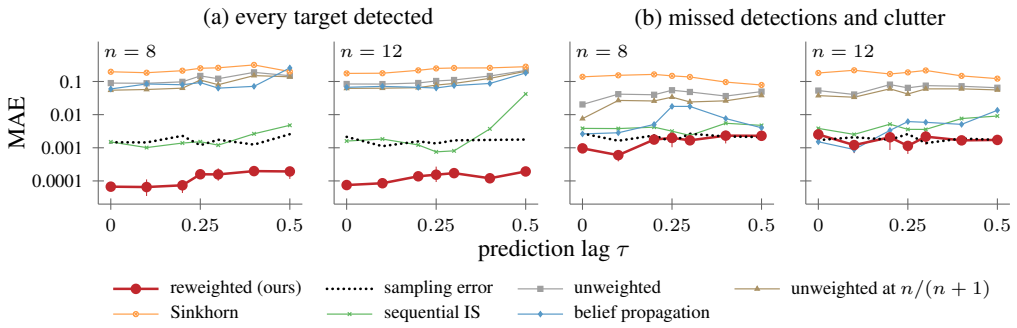


Figure 1: Formation scenes at $n = 8$ and $n = 12$ against the prediction lag: MAE over the ambiguous entries, the reweighted estimator's range over the three draws of each scene, and the sampling error (dotted). The reweighted estimator's mean error stays below the sampling error at every lag in (a) and below twice it in (b). Figure E.1 gives $n = 16$ and $n = 20$.

10^{-14} in every entry (Appendix E.3). The beliefs converge to the minimizer of the Bethe free energy over the doubly stochastic matrices, and when that minimizer is an assignment matrix they collapse onto it (Vontobel, 2013, Theorem 32). The posterior of these scenes is concentrated on frequent hypotheses, which enter the reweighted estimate in the proportions an enumeration of them would give (Section 5), so its error falls below that of T draws from the posterior itself. When no hypothesis is frequent the weights come from the law-based estimates alone, as on one draw of the random scene at $n = 12$ with every target detected (Appendix D.6), where three runs have errors of 1.01 to 1.18 times the sampling error. With missed detections and clutter (Figure 1(b)) the reweighted estimator's error is 0.36 to 1.6 times the sampling error at every lag, the beliefs no longer collapse, and belief propagation's error falls to between 0.4 and 12 times the sampling error.

In Table E.1 the sampling error puts every row on a common scale: the reweighted estimator's error is at most 1.1 times it with every target detected and between 0.7 and 4.5 times it with missed detections and clutter. The unweighted estimator's error at $c = n/(n+1)$ is 1.3 to 870 times the reweighted estimator's, with the largest ratios on the formations with every target detected, the rows where the marginals of Gumbel matching are furthest from those of p_θ.

On the formations at n from 30 to 60, too large for the dynamic program, two independently randomized runs of the reweighted estimator differ on average by 0.0007 to 0.017 on the entries that their average places between 0.05 and 0.95. Sinkhorn scaling, sequential importance sampling and the two unweighted estimators lie 8 to 775 times that far from the runs' average (Appendix E.6). The demonstration ends where the marginals can no longer be computed exactly, the range the estimator is built for: at $n = 60$ it draws its assignments, weights them by the law and reports the effective sample size of the weights, with no ground truth anywhere in the procedure.

7 Related Work

Much of the theory of perturb-and-MAP (Papandreou & Yuille, 2011; 2014) concerns the value of the perturbed maximum, whose expectation under independent zero-mean Gumbel noise on the values of each variable bounds the log-partition function from above (Hazan & Jaakkola, 2012). With one noise variable for every element the maximizer's law has been taken as a model (Tarlow et al., 2012; Gane et al., 2014; Berthet et al., 2020). Under independent Gumbel noise on the values of each variable the maximizer's law is the Gibbs distribution for a sum of functions of single variables and in general only approximates it otherwise (Lázaro-Gredilla et al., 2021). For matroid bases and assignments of three rows or more, whose constraints act as interactions between the coordinates, Theorem 1 decides over all noise laws whether an exact one exists, and Theorem 2 computes the mass of the cone of Papandreou & Yuille (2011, eq. (4)) for Gumbel matching.

In machine learning the law is known when the output is a sequence of softmax choices (Struminsky et al., 2021): the ordered elements of Gumbel top-k follow the Plackett–Luce distribution (Kool et al., 2019), and the probabilities of the orders of a set add up to the integral of Proposition 1

(Kool et al., 2020). Niepert et al. (2021) construct noise whose sum over every k-subset is Gumbel and observe that they still "cannot sample faithfully," as no noise law can for $2 \leq k \leq n-2$ (Theorem 1), and Ahmed et al. (2023) turn to conditional Poisson sampling, which draws from the Gibbs distribution on k-subsets exactly. For assignments the law has been taken as the model (Paulus et al., 2020; Mena et al., 2018) or set aside (Niculae et al., 2025; Charpentier et al., 2022; Cundy et al., 2021). Kool et al. (2021) reweight draws of Gumbel matching row by row and recover expectations under independent softmax choices in the rows, whereas Algorithms 1 and 2 weight each assignment by its unnormalized Gibbs probability over its probability under Gumbel matching and recover the marginals of the Gibbs distribution itself.

Survey sampling uses the perturbed optimizer under other names. Order sampling keeps the k units whose independent ranking variables are smallest (Rosén, 1997a), and when these are rescaled copies of one positive random variable it is a perturbed optimizer over k-subsets: successive sampling (Rosén, 1997a) is Gumbel top-k, Pareto sampling (Rosén, 1997b) has logistic noise and sequential Poisson sampling (Ohlsson, 1998) has exponential noise, while the Gibbs distribution on k-subsets is the rejective sampling of Hájek (1964), also called conditional Poisson sampling. Prescribed feasible marginals are attained by adjusting, for each marginal vector separately, the scores of an order sampling scheme (Aires et al., 2002) or independent noise laws on the elements of a matroid (Dughmi et al., 2019). Theorem 3 gives the temperatures at which the three schemes agree with the Gibbs distribution to first order at equal scores, and for $2 \leq k \leq n-2$ only a randomized reparametrization of the scores makes them agree with it to second order (Theorem C.1).

Under edge weights that are shifts plus independent uniform noise, Babson et al. (2026) show that suitable shifts make the minimum spanning tree uniform on a theta graph, two vertices joined by three internally disjoint paths, and that none do on a complete graph with four or more vertices. The uniform distribution is the Gibbs distribution at zero scores, and Theorem 1 shows that a noise law exact at every score vector exists for the spanning trees of a connected simple graph if and only if the graph is a cactus, which no theta graph is.

A fully polynomial randomized approximation scheme exists for the permanent of a nonnegative matrix (Jerrum et al., 2004), and adaptive partitions sample the Gibbs distribution on assignments (Kuck et al., 2019), while the baselines of Section 6 include sequential importance sampling row by row from a Sinkhorn-scaled proposal (Beichl & Sullivan, 1999) and the deterministic approximations Sinkhorn scaling (Kosowsky & Yuille, 1994; Cuturi, 2013) and sum-product belief propagation (Huang & Jebara, 2009; Vontobel, 2013).

Multi-target tracking uses the association probabilities of Section 6 in probabilistic data association (Bar-Shalom & Tse, 1975; Fortmann et al., 1983). Over one scan the association probabilities are computed exactly by Zhou & Bose (1993) and approximated from the hypotheses that Murty's ranking algorithm enumerates (Murty, 1968; Rezatofighi et al., 2016), by Markov chain Monte Carlo (Oh et al., 2009) and by belief propagation between tracks and measurements (Williams & Lau, 2014; Meyer et al., 2018). Trackers using belief propagation show less track coalescence, the merging of the tracks of nearby targets, than joint probabilistic data association does, a difference that Kropfreiter et al. (2024) trace mainly to overconfident association beliefs that favor the most likely association. On the formations of Figure 1(a) the beliefs collapse to the indicator of the most probable hypothesis, and Algorithms 1 and 2 estimate the association probabilities consistently from independent hypotheses drawn by Gumbel matching (Lemma D.1).

8 Conclusion

On matroid bases, noise on the elements makes the perturbed optimizer sample the Gibbs distribution at every score vector only when the structure splits into single choices, and on assignments of three rows or more it never does (Theorem 1). Under standard Gumbel noise the law is an integral of a product on matroid bases, where every clock outside the basis exceeds its bound, of a determinant on assignments, where the tight entries form no cycle, and of a density, #P-hard to evaluate, on the perfect matchings of the complete graph, where the tight edges close odd cycles (Proposition B.1, Theorem 2, Proposition B.3). Expanded at zero scores, the laws give the first-order temperatures, $n/(n+1)$ for square assignments (Theorem 3). The partition function of the Gibbs distribution on assignments is a permanent, and reweighting the draws of Gumbel matching with these determinant integrals estimates the marginals of that distribution without computing the permanent.

AI USE STATEMENT

The authors used OpenAI's ChatGPT and Anthropic's Claude to help organize and polish the exposition of this paper, based on the authors' own drafts. The same tools assisted in surveying the related literature, in understanding the current state of research in this area and its gaps, and in flagging possible errors in draft proofs. Within an experimental framework of the authors' own design, the tools were further used to implement the methods of this paper in code and to help produce the figures. The authors formulated and proved all mathematical claims in this paper, verified and tested all experimental code, and checked, revised, and finalized every statement, proof, figure, and reference. They take full responsibility for the content of the submission.

REPRODUCIBILITY STATEMENT

The proofs are in Appendices A and B for Section 3 and in Appendices C and D for Sections 4 and 5. The estimator is specified in Algorithms 1 and 2, with the settings of the chains and of the annealing in Appendices D.1 and D.2, and Appendix D.5 gives the subset dynamic program that computes the true marginals. Appendix E specifies the scenes, the budgets, the baselines and the hardware of the experiments in Section 6. The supplementary code, also at `https://anonymous.4open.science/r/law-of-the-perturbed-optimizer`, implements both algorithms, the baselines, the scene generator and the dynamic program, with tests against enumeration on small instances. It contains the scripts that produced every number in Section 6, Appendix D.6 and Appendix E from a fixed seed, together with the outputs of the experiments.

REFERENCES

Kareem Ahmed, Zhe Zeng, Mathias Niepert, and Guy Van den Broeck. SIMPLE: A gradient estimator for k-subset sampling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=GPJVuyX4p_h. arXiv:2210.01941.

Nibia Aires, Johan Jonasson, and Olle Nerman. Order sampling design with prescribed inclusion probabilities. *Scandinavian Journal of Statistics*, 29(1):183–187, 2002. doi: 10.1111/1467-9469.00120.

Eric Babson, Moon Duchin, Annina Iseli, Pietro Poggi-Corradini, Dylan Thurston, and Jamie Tucker-Foltz. Models of random spanning trees. *Random Structures & Algorithms*, 68(3):e70063, 2026. doi: 10.1002/rsa.70063. arXiv:2407.20226.

Yaakov Bar-Shalom and Edison Tse. Tracking in a cluttered environment with probabilistic data association. *Automatica*, 11(5):451–460, 1975. doi: 10.1016/0005-1098(75)90021-7.

Douglas M. Bates and Donald G. Watts. Relative curvature measures of nonlinearity. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(1):1–16, 1980. doi: 10.1111/j.2517-6161.1980.tb01094.x.

Isabel Beichl and Francis Sullivan. Approximating the permanent via importance sampling with application to the dimer covering problem. *Journal of Computational Physics*, 149(1):128–147, 1999. doi: 10.1006/jcph.1998.6149.

Quentin Berthet, Mathieu Blondel, Olivier Teboul, Marco Cuturi, Jean-Philippe Vert, and Francis Bach. Learning with differentiable perturbed optimizers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9508–9519. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/6bb56208f672af0dd65451f869fedfd9-Abstract.html. arXiv:2002.08676.

Seth Chaiken. A combinatorial proof of the all minors matrix tree theorem. *SIAM Journal on Algebraic and Discrete Methods*, 3(3):319–329, 1982. doi: 10.1137/0603033.

Bertrand Charpentier, Simon Kibler, and Stephan Günnemann. Differentiable DAG sampling. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=9wOQOgNe-w. arXiv:2203.08509.

Chris Cundy, Aditya Grover, and Stefano Ermon. BCD Nets: Scalable variational approaches for Bayesian causal discovery. In *Advances in Neural Information Processing Systems*, volume 34, pp. 7095–7110. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/39799c18791e8d7eb29704fc5bc04ac8-Abstract.html.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, pp. 2292–2300. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html. arXiv:1306.0895.

Shaddin Dughmi, David Kempe, and Ruixin Qiang. Alea Iacta Est: Auctions, persuasion, interim rules, and dice. In *10th Innovations in Theoretical Computer Science Conference (ITCS 2019)*, volume 124 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 31:1–31:20. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2019. doi: 10.4230/LIPIcs.ITCS.2019.31. arXiv:1811.11417.

Jack Edmonds. Paths, trees, and flowers. *Canadian Journal of Mathematics*, 17:449–467, 1965. doi: 10.4153/CJM-1965-045-4.

Jack Edmonds. Matroids and the greedy algorithm. *Mathematical Programming*, 1(1):127–136, 1971. doi: 10.1007/BF01584082.

Thomas E. Fortmann, Yaakov Bar-Shalom, and Molly Scheffe. Sonar tracking of multiple targets using joint probabilistic data association. *IEEE Journal of Oceanic Engineering*, 8(3):173–184, 1983. doi: 10.1109/JOE.1983.1145560.

Harold N. Gabow. A combinatoric interpretation of dual variables for weighted matching and f-factors. *Theoretical Computer Science*, 454:136–163, 2012. doi: 10.1016/j.tcs.2012.06.024.

Andreea Gane, Tamir Hazan, and Tommi Jaakkola. Learning with maximum a-posteriori perturbation models. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pp. 247–256. PMLR, 2014. URL https://proceedings.mlr.press/v33/gane14.html.

Roger B. Grosse, Zoubin Ghahramani, and Ryan P. Adams. Sandwiching the marginal likelihood using bidirectional Monte Carlo. *arXiv preprint arXiv:1511.02543*, 2015.

Emil Julius Gumbel. *Statistical Theory of Extreme Values and Some Practical Applications: A Series of Lectures*. Number 33 in National Bureau of Standards Applied Mathematics Series. U.S. Government Printing Office, Washington, D.C., 1954.

Jaroslav Hájek. Asymptotic theory of rejective sampling with varying probabilities from a finite population. *The Annals of Mathematical Statistics*, 35(4):1491–1523, 1964. doi: 10.1214/aoms/1177700375.

Jaroslav Hájek and Zbyněk Šidák. *Theory of Rank Tests*. Academic Press, New York, 1967.

Tamir Hazan and Tommi Jaakkola. On the partition function and random maximum a-posteriori perturbations. In *Proceedings of the 29th International Conference on Machine Learning*, 2012. URL https://icml.cc/2012/papers/528.pdf.

Bert Huang and Tony Jebara. Approximating the permanent with belief propagation. *arXiv preprint arXiv:0908.1769*, 2009.

Mark Jerrum, Alistair Sinclair, and Eric Vigoda. A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries. *Journal of the ACM*, 51(4):671–697, 2004. doi: 10.1145/1008731.1008738.

Han Mao Kiah, Alexander Vardy, and Hanwen Yao. Computing permanents on a trellis. *arXiv preprint arXiv:2107.07377*, 2021.

G. Kirchhoff. Ueber die Auflösung der Gleichungen, auf welche man bei der Untersuchung der linearen Vertheilung galvanischer Ströme geführt wird. *Annalen der Physik und Chemie*, 72(12):497–508, 1847. doi: 10.1002/andp.18471481202.

Augustine Kong, Jun S. Liu, and Wing Hung Wong. Sequential imputations and Bayesian missing data problems. *Journal of the American Statistical Association*, 89(425):278–288, 1994. doi: 10.1080/01621459.1994.10476469.

Wouter Kool, Herke van Hoof, and Max Welling. Stochastic beams and where to find them: The Gumbel-top-k trick for sampling sequences without replacement. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3499–3508. PMLR, 2019. URL https://proceedings.mlr.press/v97/kool19a.html.

Wouter Kool, Herke van Hoof, and Max Welling. Estimating gradients for discrete random variables by sampling without replacement. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rklEj2EFvB. arXiv:2002.06043.

Wouter Kool, Chris J. Maddison, and Andriy Mnih. Unbiased gradient estimation with balanced assignments for mixtures of experts. In *I (Still) Can't Believe It's Not Better Workshop at NeurIPS 2021*, 2021. arXiv:2109.11817.

J. J. Kosowsky and A. L. Yuille. The invisible hand algorithm: Solving the assignment problem with statistical physics. *Neural Networks*, 7(3):477–490, 1994. doi: 10.1016/0893-6080(94)90081-7.

Thomas Kropfreiter, Florian Meyer, David F. Crouse, Stefano Coraluppi, Franz Hlawatsch, and Peter Willett. Track coalescence and repulsion in multitarget tracking: An analysis of MHT, JPDA, and belief propagation methods. *IEEE Open Journal of Signal Processing*, 5:1089–1106, 2024. doi: 10.1109/OJSP.2024.3451167.

Jonathan Kuck, Tri Dao, Hamid Rezatofighi, Ashish Sabharwal, and Stefano Ermon. Approximating the permanent by sampling from adaptive partitions. In *Advances in Neural Information Processing Systems*, volume 32, pp. 8860–8871. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/d914a6c6d93c8df063b9099a076a488c-Abstract.html.

H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955. doi: 10.1002/nav.3800020109.

Miguel Lázaro-Gredilla, Antoine Dedieu, and Dileep George. Perturb-and-max-product: Sampling and learning in discrete energy-based models. In *Advances in Neural Information Processing Systems*, volume 34, pp. 928–940. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/07b1c04a30f798b5506c1ec5acfb9031-Abstract.html.

R. Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, New York, 1959.

Gonzalo E. Mena, David Belanger, Scott Linderman, and Jasper Snoek. Learning latent permutations with Gumbel-Sinkhorn networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Byt3oJ-0W. arXiv:1802.08665.

Florian Meyer, Thomas Kropfreiter, Jason L. Williams, Roslyn A. Lau, Franz Hlawatsch, Paolo Braca, and Moe Z. Win. Message passing algorithms for scalable multitarget tracking. *Proceedings of the IEEE*, 106(2):221–259, 2018. doi: 10.1109/JPROC.2018.2789427.

James Munkres. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 5(1):32–38, 1957. doi: 10.1137/0105003.

Katta G. Murty. Letter to the editor—An algorithm for ranking all the assignments in order of increasing cost. *Operations Research*, 16(3):682–687, 1968. doi: 10.1287/opre.16.3.682.

Radford M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001. doi: 10.1023/A:1008923215028.

Vlad Niculae, Caio Corro, Nikita Nangia, Tsvetomila Mihaylova, and André F. T. Martins. Discrete latent structure in neural networks. *Foundations and Trends in Signal Processing*, 19(2):99–211, 2025. doi: 10.1561/2000000134. arXiv:2301.07473.

Mathias Niepert, Pasquale Minervini, and Luca Franceschi. Implicit MLE: Backpropagating through discrete exponential family distributions. In *Advances in Neural Information Processing Systems*, volume 34, pp. 14567–14579. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/7a430339c10c642c4b2251756fd1b484-Abstract.html.

Songhwai Oh, Stuart Russell, and Shankar Sastry. Markov chain Monte Carlo data association for multi-target tracking. *IEEE Transactions on Automatic Control*, 54(3):481–497, 2009. doi: 10.1109/TAC.2009.2012975.

Esbjörn Ohlsson. Sequential Poisson sampling. *Journal of Official Statistics*, 14(2):149–162, 1998. URL https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/sequential-poisson-sampling.pdf.

Art B. Owen. Monte Carlo theory, methods and examples, 2013. URL https://artowen.su.domains/mc/. Book in progress; copyright 2009–2013, 2018–2019, 2023.

James Oxley. *Matroid Theory*. Oxford University Press, New York, second edition, 2011.

George Papandreou and Alan L. Yuille. Perturb-and-MAP random fields: Using discrete optimization to learn and sample from energy models. In *2011 International Conference on Computer Vision*, pp. 193–200. IEEE, 2011. doi: 10.1109/ICCV.2011.6126242.

George Papandreou and Alan L. Yuille. Perturb-and-MAP random fields: Reducing random sampling to optimization, with applications in computer vision. In Sebastian Nowozin, Peter V. Gehler, Jeremy Jancsary, and Christoph H. Lampert (eds.), *Advanced Structured Prediction*, pp. 159–186. The MIT Press, Cambridge, MA, 2014. doi: 10.7551/mitpress/9969.003.0009.

Max B. Paulus, Dami Choi, Daniel Tarlow, Andreas Krause, and Chris J. Maddison. Gradient estimation with stochastic softmax tricks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5691–5704. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/3df80af53dce8435cf9ad6c3e7a403fd-Abstract.html.

Seyed Hamid Rezatofighi, Anton Milan, Zhen Zhang, Qinfeng Shi, Anthony Dick, and Ian Reid. Joint probabilistic matching using m-best solutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 136–145. IEEE, 2016. doi: 10.1109/CVPR.2016.22.

Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996. doi: 10.2307/3318418.

Bengt Rosén. Asymptotic theory for order sampling. *Journal of Statistical Planning and Inference*, 62(2):135–158, 1997a. doi: 10.1016/S0378-3758(96)00185-1.

Bengt Rosén. On sampling with probability proportional to size. *Journal of Statistical Planning and Inference*, 62(2):159–191, 1997b. doi: 10.1016/S0378-3758(96)00186-3.

Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The Annals of Mathematical Statistics*, 35(2):876–879, 1964. doi: 10.1214/aoms/1177703591.

Kirill Struminsky, Artyom Gadetsky, Denis Rakitin, Danil Karpushkin, and Dmitry Vetrov. Leveraging recursive Gumbel-Max trick for approximate inference in combinatorial spaces. In *Advances in Neural Information Processing Systems*, volume 34, pp. 10999–11011. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/5b658d2a925565f0755e035597f8d22f-Abstract.html.

Daniel Tarlow, Ryan Adams, and Richard Zemel. Randomized optimum models for structured prediction. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pp. 1221–1229. PMLR, 2012. URL https://proceedings.mlr.press/v22/tarlow12b.html.

Luke Tierney. Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22(4):1701–1728, 1994. doi: 10.1214/aos/1176325750.

W. T. Tutte. The dissection of equilateral triangles into equilateral triangles. *Proceedings of the Cambridge Philosophical Society*, 44(4):463–482, 1948. doi: 10.1017/S030500410002449X.

Jeffrey K. Uhlmann. Matrix permanent inequalities for approximating joint assignment matrices in tracking systems. *Journal of the Franklin Institute*, 341(7):569–593, 2004. doi: 10.1016/j.jfranklin.2004.07.003.

L. G. Valiant. The complexity of computing the permanent. *Theoretical Computer Science*, 8(2):189–201, 1979a. doi: 10.1016/0304-3975(79)90044-6.

L. G. Valiant. The complexity of enumeration and reliability problems. *SIAM Journal on Computing*, 8(3):410–421, 1979b. doi: 10.1137/0208032.

Pascal O. Vontobel. The Bethe permanent of a nonnegative matrix. *IEEE Transactions on Information Theory*, 59(3):1866–1901, 2013. doi: 10.1109/TIT.2012.2227109. arXiv:1107.4196.

Jason L. Williams and Roslyn A. Lau. Approximate evaluation of marginal association probabilities with belief propagation. *IEEE Transactions on Aerospace and Electronic Systems*, 50(4):2942–2959, 2014. doi: 10.1109/TAES.2014.120568. arXiv:1209.6299.

John I. Yellott, Jr. The relationship between Luce's Choice Axiom, Thurstone's Theory of Comparative Judgment, and the double exponential distribution. *Journal of Mathematical Psychology*, 15(2):109–144, 1977. doi: 10.1016/0022-2496(77)90026-8.

B. Zhou and N. K. Bose. Multitarget tracking in clutter: Fast algorithms for data association. *IEEE Transactions on Aerospace and Electronic Systems*, 29(2):352–363, 1993. doi: 10.1109/7.210074.

A Exact Noise Laws

Nothing is assumed about the noise law, so the impossibility proofs vary the scores alone. Lemma A.1 uses extreme scores to restrict the problem to a subfamily of structures with the same noise. Every obstruction is then a family of at most six structures on which exactness prescribes a joint distribution function that cannot exist (Lemmas A.2 and A.3) or second moments that contradict a linear identity (Appendix A.5).

The dual matroid $\mathcal{M}^*$ has bases $\{E \setminus B : B \in \mathcal{B}(\mathcal{M})\}$. A set is independent if it lies in a basis, a circuit is a minimal set contained in no basis, a loop lies in no basis and a coloop in every basis, and two elements lie in the same connected component if they are equal or some circuit contains both. A matroid is the direct sum of the restrictions to its connected components, and the connected components of a direct sum of connected matroids are its summands (Oxley, 2011, Chapter 4). A loop is $U_{0,1}$ and a coloop $U_{1,1}$. The matroids $U_{1,\ell}$ and $U_{\ell-1,\ell}$ with $\ell \geq 2$ are connected, so the condition of Theorem 1(a) says that every connected component of $\mathcal{M}$, on its ground set E', is a loop, a coloop, $U_{1,|E'|}$ or $U_{|E'|-1,|E'|}$, the form used below. The following theorem, proved in the sufficiency part of Appendix A.3, gives an exact noise law for each such matroid.

Theorem A.1 (exact noise laws)**.** *When an exact noise law exists for the bases of a matroid $\mathcal{M}$, one is given by independent standard Gumbel entries on each component of the kind $U_{1,|E'|}$, independent negative Gumbel entries on each component of the kind $U_{|E'|-1,|E'|}$ and arbitrary entries on the loops and coloops, the noise blocks of the components being independent of one another. A two-element component is of both kinds, and either choice serves.*

The two uniform kinds are dual to each other (Remark A.1, which also treats a mixture of the two noise laws on a matroid isomorphic to its dual).

A.1 Forcing

Let $\mathsf{X}, \mathsf{Y} \subseteq E$ be disjoint, let $E_0 = E \setminus (\mathsf{X} \cup \mathsf{Y})$, and suppose that the fiber $\mathcal{X}_{\mathsf{X},\mathsf{Y}} = \{x \in \mathcal{X} : x_e = 1 \text{ for } e \in \mathsf{X},\ x_e = 0 \text{ for } e \in \mathsf{Y}\}$ is nonempty. Restriction to E_0, written $x \mapsto x|_{E_0}$, is injective on the fiber, and $\mathcal{X}_0 = \{x|_{E_0} : x \in \mathcal{X}_{\mathsf{X},\mathsf{Y}}\}$ is the fiber structure set. For $\theta \in \mathbb{R}^{E_0}$ and $\Lambda > 0$ let $\theta^\Lambda \in \mathbb{R}^E$ agree with θ on E_0 and equal $+\Lambda$ on X and $-\Lambda$ on Y.

Lemma A.1 (forcing)**.** *As $\Lambda \to \infty$, for every x in the fiber, $p_{\theta^\Lambda}(x) \to p_\theta^{\mathcal{X}_0}(x|_{E_0})$ and $Q^\nu_{\theta^\Lambda}(x) \to Q^{\nu_0}_\theta(x|_{E_0})$, where $p^{\mathcal{X}_0}$ is the Gibbs distribution on $\mathcal{X}_0$ and Q^{ν_0} the law of the perturbed optimizer on $\mathcal{X}_0$ with the noise ξ_{E_0}, of law ν_0, and the tie rule restricted to the fiber. The mass outside the fiber tends to zero under both. Hence if ν is exact for $\mathcal{X}$, then ν_0 is exact for $\mathcal{X}_0$, and if $\sup_\theta \mathrm{TV}(p_\theta, Q^\nu_\theta) \leq \varepsilon$ on $\mathcal{X}$, the same bound holds on $\mathcal{X}_0$ with ν_0.*

Proof. Extreme scores drive the Gibbs mass outside the fiber to zero and make every structure outside the fiber lose to every structure inside it in the perturbed optimization. Let x be in the fiber and $x' \in \mathcal{X}$ outside it. Then x' misses an element of X or contains an element of Y, and $x - x'$ has entries in $\{-1, 0, 1\}$, so $\langle \theta^\Lambda, x - x' \rangle \geq \Lambda - \|\theta\|_1$. Under the Gibbs distribution the mass outside the fiber is therefore at most $|\mathcal{X}|\, e^{\|\theta\|_1 - \Lambda}$, and inside the fiber $\langle \theta^\Lambda, x \rangle = \Lambda|\mathsf{X}| + \langle \theta, x|_{E_0} \rangle$ is the fiber score plus a constant, so the conditional distribution on the fiber is $p_\theta^{\mathcal{X}_0}$ for every Λ. For a realization of the noise, $\langle \theta^\Lambda + \xi, x - x' \rangle \geq \Lambda - \|\theta\|_1 - \|\xi\|_1$, so once Λ exceeds $\|\theta\|_1 + \|\xi\|_1$, every structure outside the fiber loses to every structure inside it, and inside the fiber the perturbed score is $\langle \theta + \xi_{E_0}, x|_{E_0} \rangle$ plus a constant. Hence $x^*(\theta^\Lambda)$ is the fiber optimizer with noise ξ_{E_0} for all large Λ, pointwise in ξ, and bounded convergence gives the limit of $Q^\nu_{\theta^\Lambda}$. Exactness holds at every θ^Λ and passes to the limit on the finite set $\mathcal{X}_0$, and so does the total variation bound, the sum over the fiber being at most the full sum. □

If $\mathcal{X} = \{x_B : B \in \mathcal{B}(\mathcal{M})\}$, X is independent and Y is coindependent, that is, independent in $\mathcal{M}^*$, then $\{B \setminus \mathsf{X} : x_B \in \mathcal{X}_{\mathsf{X},\mathsf{Y}}\}$ is the set of bases of the minor $\mathcal{M}/\mathsf{X} \setminus \mathsf{Y}$, the contraction of X followed by the deletion of Y (Oxley, 2011, Section 3.1), so exactness passes from $\mathcal{M}$ to every such minor, with the restricted noise. If $\mathcal{M} = \mathcal{M}_1 \oplus \mathcal{M}_2$, taking X to be a basis of $\mathcal{M}_2$ and Y the rest of its ground set gives $\mathcal{X}_0 = \{x_B : B \in \mathcal{B}(\mathcal{M}_1)\}$, so exactness passes to every component.

A.2 THREE ELEMENTS

Fix three elements $\mathsf{T} = \{1, 2, 3\} \subseteq E$. Say that T admits singleton fibers if some fiber structure set on T consists of the three singletons of T, and pair fibers if some fiber structure set on T consists of its three pairs. For the k-subsets of $[n]$ with $2 \le k \le n-2$ every triple admits both: force $k-1$ other elements in and the remaining $n-k-2$ out for singleton fibers, and $k-2$ in and $n-k-1$ out for pair fibers.

Lemma A.2 (three elements)**.** *If some triple admits singleton fibers and pair fibers, then* $\sup_\theta \mathrm{TV}(p_\theta, Q_\theta^\nu) \ge \delta_0 = (8\sqrt{2} - 5\sqrt{5})/12$ *for every noise law* ν*.*

Proof. The proof compares the singleton fiber and the pair fiber through inclusion–exclusion. Let $\varepsilon = \sup_\theta \mathrm{TV}(p_\theta, Q_\theta^\nu)$. By Lemma A.1 both fiber problems on T carry the noise (ξ_1, ξ_2, ξ_3) with total variation error at most ε. Let $\mathsf{V} = (\xi_2 - \xi_1, \xi_3 - \xi_1)$, $\mathsf{K}^-(q_1, q_2) = \mathbb{P}(\mathsf{V}_1 < q_1, \mathsf{V}_2 < q_2)$ and $\bar{\mathsf{K}}(q_1, q_2) = \mathbb{P}(\mathsf{V}_1 \ge q_1, \mathsf{V}_2 \ge q_2)$, and write $Q^{\nu_0}_{(q_1,q_2)}$ for the law of either fiber optimizer at any scores with $\theta_1 - \theta_2 = q_1$ and $\theta_1 - \theta_3 = q_2$, since the law depends on the scores only through these differences.

One slot. Element 1 is the unique maximizer of $\theta_j + \xi_j$ on $\{\mathsf{V}_1 < q_1, \mathsf{V}_2 < q_2\}$ and a maximizer only on $\{\mathsf{V}_1 \le q_1, \mathsf{V}_2 \le q_2\}$, while $p_\theta(\{1\}) = \mathsf{f}_1(q_1, q_2) = 1/(1 + e^{-q_1} + e^{-q_2})$. Hence $\mathsf{K}^-(q_1, q_2) \le Q^{\nu_0}_{(q_1,q_2)}(\{1\}) \le \mathsf{f}_1(q_1, q_2) + \varepsilon$, and for every $\mathsf{s} > 0$, $\mathsf{K}^-(q_1, q_2) \ge Q^{\nu_0}_{(q_1-\mathsf{s}, q_2-\mathsf{s})}(\{1\}) \ge \mathsf{f}_1(q_1 - \mathsf{s}, q_2 - \mathsf{s}) - \varepsilon$. Since f_1 is continuous, $|\mathsf{K}^- - \mathsf{f}_1| \le \varepsilon$ on $\mathbb{R}^2$.

Two slots. The excluded element is the one with the smallest perturbed score, so element 1 is excluded on $\{\mathsf{V}_1 > q_1, \mathsf{V}_2 > q_2\}$ and only on $\{\mathsf{V}_1 \ge q_1, \mathsf{V}_2 \ge q_2\}$, while $p_\theta(\{2, 3\}) = \mathsf{f}_2(q_1, q_2) = 1/(1 + e^{q_1} + e^{q_2})$. The same sandwich gives $|\bar{\mathsf{K}} - \mathsf{f}_2| \le \varepsilon$.

Inclusion–exclusion. Since $\{\mathsf{V}_1 \ge q_1, \mathsf{V}_2 \ge q_2\}$ is the complement of $\{\mathsf{V}_1 < q_1\} \cup \{\mathsf{V}_2 < q_2\}$,

$$\bar{\mathsf{K}}(q_1, q_2) = 1 - \mathsf{K}^-(q_1, \infty) - \mathsf{K}^-(\infty, q_2) + \mathsf{K}^-(q_1, q_2),$$

and $\mathsf{K}^-(q_1, \infty) = \lim_{q_2 \to \infty} \mathsf{K}^-(q_1, q_2)$ is within ε of $\lim_{q_2 \to \infty} \mathsf{f}_1(q_1, q_2) = 1/(1 + e^{-q_1})$. With $\mathsf{a} = e^{-q_1}$ and $\mathsf{b} = e^{-q_2}$ this gives, for all q_1, q_2,

$$\left| \mathsf{f}_2(q_1, q_2) - \left(1 - \frac{1}{1+\mathsf{a}} - \frac{1}{1+\mathsf{b}} + \frac{1}{1+\mathsf{a}+\mathsf{b}}\right) \right| \le 4\varepsilon.$$

The expression in parentheses equals $\mathsf{ab}(2 + \mathsf{a} + \mathsf{b})/((1 + \mathsf{a})(1 + \mathsf{b})(1 + \mathsf{a} + \mathsf{b}))$ and $\mathsf{f}_2 = \mathsf{ab}/(\mathsf{a} + \mathsf{b} + \mathsf{ab})$, so the left side is $|\mathsf{h}(\mathsf{a}, \mathsf{b})|$ with

$$\mathsf{h}(\mathsf{a}, \mathsf{b}) = \frac{\mathsf{ab}\,(1 - \mathsf{ab})}{(\mathsf{a} + \mathsf{b} + \mathsf{ab})(1 + \mathsf{a})(1 + \mathsf{b})(1 + \mathsf{a} + \mathsf{b})}.$$

The supremum. Since $\mathsf{h}(1/\mathsf{a}, 1/\mathsf{b}) = -\mathsf{h}(\mathsf{a}, \mathsf{b})$, it suffices to maximize h over $\mathsf{ab} < 1$. With $\mathsf{p} = \mathsf{ab}$ and $\mathsf{t} = \mathsf{a} + \mathsf{b}$, $\mathsf{h} = \mathsf{p}(1 - \mathsf{p})/((\mathsf{t} + \mathsf{p})(1 + \mathsf{t} + \mathsf{p})(1 + \mathsf{t}))$ decreases in t at fixed p, and $\mathsf{t} \ge 2\sqrt{\mathsf{p}}$ with equality if and only if $\mathsf{a} = \mathsf{b}$. So the supremum is on the diagonal $\mathsf{a} = \mathsf{b} < 1$, where $\mathsf{h} = \mathsf{a}(1 - \mathsf{a})/((2 + \mathsf{a})(1 + \mathsf{a})(1 + 2\mathsf{a}))$. The derivative vanishes if and only if $\mathsf{a}^4 - 2\mathsf{a}^3 - 7\mathsf{a}^2 - 2\mathsf{a} + 1 = 0$, that is, $(\mathsf{a} + \mathsf{a}^{-1})^2 - 2(\mathsf{a} + \mathsf{a}^{-1}) - 9 = 0$, so $\mathsf{a} + \mathsf{a}^{-1} = 1 + \sqrt{10}$. At this point $(1 - \mathsf{a})/(1 + \mathsf{a}) = 2\sqrt{2} - \sqrt{5}$ and $(2 + \mathsf{a})(1 + 2\mathsf{a}) = \mathsf{a}(7 + 2\sqrt{10}) = \mathsf{a}(\sqrt{5} + \sqrt{2})^2$, so the supremum of h is $(2\sqrt{2} - \sqrt{5})/(\sqrt{5} + \sqrt{2})^2 = (8\sqrt{2} - 5\sqrt{5})/3 = 4\delta_0$. Hence $4\varepsilon \ge 4\delta_0$. □

The one-slot and two-slot prescriptions are those of standard and of negative Gumbel noise. Were ν exact, K^- would be the joint distribution function of the two differences of three independent standard Gumbel variables and $\bar{\mathsf{K}}$ the joint survival function of the two differences of three independent negative Gumbel variables. Each difference is standard logistic under both, and the joint survival functions of the two differences under the two prescriptions differ by h, which vanishes if and only if $q_1 + q_2 = 0$.

A.3 CLASSIFICATION

The rank of a matroid is the common size of its bases and its corank the size of their complements. Two elements are parallel if they form a circuit. The pair $\{a, b\}$ is written ab and the set $(B \setminus \{b\}) \cup \{e\}$ is written $B - b + e$.

Sufficiency. Let every component of $\mathcal{M}$ be a loop, a coloop, $U_{1,\ell}$ or $U_{\ell-1,\ell}$. A basis of $\mathcal{M}$ is a union of bases of the components, its score is the sum of their scores, and $\langle \theta + \xi, x_B \rangle$ is maximized component by component. With the noise blocks of the components independent, Q_θ^ν is the product of the component laws, and p_θ factors in the same way. So it suffices to treat one component. A loop is in no basis and a coloop in every basis, so every noise law is exact for such a component. On $U_{1,\ell}$ the bases are the singletons and, with independent standard Gumbel entries, $\mathbb{P}(\arg\max_j(\theta_j + \xi_j) = e) = \exp(\theta_e)/\sum_j \exp(\theta_j) = p_\theta(\{e\})$, the Gumbel-max trick (Gumbel, 1954; Luce, 1959; Yellott, 1977). On $U_{\ell-1,\ell}$ the bases are the complements of singletons and the perturbed optimizer drops the element with the smallest perturbed score. With $\xi = -\xi'$ for standard Gumbel ξ', $\mathbb{P}(\arg\min_j(\theta_j - \xi'_j) = e) = \mathbb{P}(\arg\max_j(-\theta_j + \xi'_j) = e) = \exp(-\theta_e)/\sum_j \exp(-\theta_j)$, which is p_θ of the complement of e because $\langle \theta, x_{E'\setminus\{e\}} \rangle = \langle \theta, \mathbf{1} \rangle - \theta_e$. Ties have probability zero. A two-element component is $U_{1,2} = U_{2-1,2}$, and either construction serves.

Necessity. The proof reduces every remaining component to a rank-two minor on four elements. Let ν be exact for the bases of $\mathcal{M}$ and let $\mathcal{K}$ be a component on E' that is none of the four kinds. By Lemma A.1 the restricted noise law is exact for $\mathcal{K}$. A connected matroid with $|E'| \geq 2$ has no loop and no coloop. If its rank is one, all its elements are parallel and it is $U_{1,|E'|}$. If its corank is one, its dual is $U_{1,|E'|}$ and it is $U_{|E'|-1,|E'|}$. So $\mathcal{K}$ has rank $\mathsf{r} \geq 2$ and corank $|E'| - \mathsf{r} \geq 2$. Fix a basis B of $\mathcal{K}$ and let G_B be its fundamental graph, the bipartite graph on $B \sqcup (E' \setminus B)$ in which $b \sim e$ if $B - b + e$ is a basis. It is connected because $\mathcal{K}$ is (Oxley, 2011, Proposition 4.3.2), and both sides have at least two vertices, so a spanning tree of G_B is not a star and contains a path on four vertices, alternating between the sides, $b_1e_1b_2e_2$, say. Put $\mathsf{X} = B \setminus \{b_1, b_2\}$, which is independent, and $\mathsf{Y} = (E' \setminus B) \setminus \{e_1, e_2\}$, which is coindependent because B avoids it. The minor $\mathcal{N} = \mathcal{K}/\mathsf{X}\setminus\mathsf{Y}$ on $\{b_1, b_2, e_1, e_2\}$ has rank two, and the bases B, $B - b_1 + e_1$, $B - b_2 + e_1$ and $B - b_2 + e_2$ of $\mathcal{K}$ contain X and avoid Y, so b_1b_2, b_2e_1, b_1e_1 and b_1e_2 are bases of $\mathcal{N}$. Every element of $\mathcal{N}$ lies in a basis, so $\mathcal{N}$ has no loop, and in a loopless matroid of rank two a pair that is not a basis is a parallel pair. Parallelism is transitive, so b_2e_2 and e_1e_2 cannot both be non-bases, since then b_2e_1 would be one. Hence $\mathcal{N}$ has six bases and is $U_{2,4}$, or five bases with one parallel pair and is, up to relabeling, the matroid $\mathcal{M}_5$ of Lemma A.3. By Lemma A.1 the restricted noise law is exact for $\mathcal{N}$. Lemma A.2 excludes $U_{2,4}$, any three of its elements admitting singleton fibers and pair fibers, and the next lemma excludes $\mathcal{M}_5$.

Lemma A.3 ($\mathcal{M}_5$)**.** *Label the elements of $\mathcal{M}_5$ by $1, 2, 3, 4$ with 12 the parallel pair, so that the bases are $13, 14, 23, 24, 34$. No noise law is exact for $\mathcal{M}_5$.*

Proof. Exactness would make one explicit function a distribution function on $\mathbb{R}^3$, and its mixed third derivative is negative somewhere. Let $q_1 = \theta_3 - \theta_4$, $q_2 = \theta_1 - \theta_2$ and $q_3 = \theta_1 - \theta_4$. The map $\theta \mapsto (q_1, q_2, q_3)$ is onto $\mathbb{R}^3$. The basis 13 beats 14, 23 and 34 if and only if, respectively, $\xi_4 - \xi_3 < q_1$, $\xi_2 - \xi_1 < q_2$ and $\xi_4 - \xi_1 < q_3$, and then beats 24 as well, by adding the first two inequalities. It is a maximizer only if the three non-strict inequalities hold. With $q = (q_1, q_2, q_3)$, $\mathsf{V} = (\xi_4 - \xi_3, \xi_2 - \xi_1, \xi_4 - \xi_1)$, $\mathsf{C}(q) = \mathbb{P}(\mathsf{V} \leq q)$ and $\mathsf{a} = e^{-q_1}$, $\mathsf{b} = e^{-q_2}$, $\mathsf{c} = e^{-q_3}$, exactness gives $\mathsf{f}(q) \leq \mathsf{C}(q) \leq \mathsf{f}(q + \mathsf{s}\mathbf{1})$ for every $\mathsf{s} > 0$ as in Lemma A.2, where

$$\mathsf{f}(q) = p_\theta(13) = \frac{1}{1 + e^{-q_1} + e^{-q_2} + e^{-q_1-q_2} + e^{-q_3}} = \frac{1}{(1+\mathsf{a})(1+\mathsf{b}) + \mathsf{c}}.$$

Since f is continuous, $\mathsf{C} = \mathsf{f}$. But C gives nonnegative mass to every box, and the mass of a small box under a smooth distribution function has the sign of its mixed derivative, which here, with $\mathsf{A} = (1+\mathsf{a})(1+\mathsf{b})$, is

$$\frac{\partial^3 \mathsf{f}}{\partial q_1\, \partial q_2\, \partial q_3} = \frac{2\mathsf{abc}\,(2\mathsf{A} - \mathsf{c})}{(\mathsf{A} + \mathsf{c})^4},$$

negative wherever $\mathsf{c} > 2\mathsf{A}$, for instance as $q_3 \to -\infty$ at fixed q_1, q_2. □

This proves the matroid classification of Theorem 1(a). The k-subset bound is Lemma A.2, every triple of $[n]$ admitting singleton fibers and pair fibers.

A.4 Spanning Trees

Let G be a connected simple graph and $\mathcal{M}(G)$ its cycle matroid. The components of $\mathcal{M}(G)$ are the cycle matroids of the blocks of G that contain an edge, a block being a maximal connected subgraph

without a cut vertex and a bridge being a block by itself (Oxley, 2011, Section 4.1). A bridge is a coloop. A block that is a cycle of length $\ell \geq 3$ has cycle matroid $U_{\ell-1,\ell}$. A block with at least two edges that is not a cycle is 2-connected, has at least three vertices, hence rank at least two, and more edges than vertices, hence corank at least two. No block of a simple graph has cycle matroid $U_{1,\ell}$ with $\ell \geq 2$, which would require ℓ parallel edges. So by Theorem 1(a) an exact noise law exists if and only if every block of G with an edge is an edge or a cycle. This condition says that G is a cactus: a cycle lies inside one block, and an edge of a cycle block lies on that one cycle. Conversely, a 2-connected block that is not a cycle has a vertex v of degree at least three, and any two edges at v lie on a common cycle because the cycle matroid of the block is connected. Since a cycle uses only two edges at v, an edge at v then lies on two cycles.

A.5 Assignments

Assignments of three rows or more are not matroid bases, and the obstruction is a second-moment identity. Let $\mathcal{X}$ be the six permutation matrices of size three and $\mathsf{s}_\sigma = \langle \xi, P_\sigma \rangle = \sum_i \xi_{i\sigma(i)}$. Two distinct permutations σ, σ' of $[3]$ differ on the rows moved by $\sigma^{-1}\sigma'$, a transposition or a 3-cycle, and the entries on which they differ form one alternating cycle. Forcing the common entries in and every entry where $P_\sigma + P_{\sigma'}$ vanishes out leaves the fiber $\{\sigma, \sigma'\}$, since an even cycle has exactly two perfect matchings. On this fiber the score difference $\langle \theta, P_\sigma - P_{\sigma'} \rangle$ ranges over $\mathbb{R}$. With $q = \langle \theta, P_\sigma - P_{\sigma'} \rangle$, the probability that the fiber optimizer returns σ lies between $\mathbb{P}(\mathsf{s}_\sigma - \mathsf{s}_{\sigma'} > -q)$ and $\mathbb{P}(\mathsf{s}_\sigma - \mathsf{s}_{\sigma'} \geq -q)$, and exactness makes it $1/(1+e^{-q})$ for every q, so $\mathsf{s}_\sigma - \mathsf{s}_{\sigma'}$ is standard logistic by Lemma A.1 and the sandwich of Lemma A.2 in one dimension. Hence every difference $\mathsf{s}_\sigma - \mathsf{s}_{\sigma'}$ with $\sigma \neq \sigma'$ has mean zero and a common variance $\mathsf{v} > 0$. Now the three even permutation matrices sum to the all-ones matrix, and so do the three odd ones, so $\sum_{\sigma \text{ even}} \mathsf{s}_\sigma - \sum_{\sigma \text{ odd}} \mathsf{s}_\sigma = 0$ identically. Fix σ_0 and put $\mathsf{c}_\sigma = \mathsf{s}_\sigma - \mathsf{s}_{\sigma_0}$, which has finite variance. With $\mathsf{p}_\sigma = \pm 1$ for even and odd σ, $\sum_\sigma \mathsf{p}_\sigma = 0$ gives $\sum_\sigma \mathsf{p}_\sigma \mathsf{s}_\sigma = \sum_\sigma \mathsf{p}_\sigma \mathsf{c}_\sigma$, and expanding $\operatorname{Var}(\mathsf{c}_\sigma - \mathsf{c}_{\sigma'})$ with $\sum_\sigma \mathsf{p}_\sigma = 0$ gives

$$\operatorname{Var}\Big(\sum_\sigma \mathsf{p}_\sigma \mathsf{c}_\sigma\Big) = -\frac{1}{2}\sum_{\sigma,\sigma'} \mathsf{p}_\sigma \mathsf{p}_{\sigma'} \operatorname{Var}(\mathsf{c}_\sigma - \mathsf{c}_{\sigma'}) = -\frac{\mathsf{v}}{2}\sum_{\sigma \neq \sigma'} \mathsf{p}_\sigma \mathsf{p}_{\sigma'} = \frac{\mathsf{v}}{2}\sum_\sigma \mathsf{p}_\sigma^2 = 3\mathsf{v} > 0,$$

which contradicts the vanishing of $\sum_\sigma \mathsf{p}_\sigma \mathsf{c}_\sigma$. For $n \geq 3$ rows and $N \geq n$ columns, force the entries (i,i) with $4 \leq i \leq n$ in and every other entry outside the block $[3] \times [3]$ out. The fiber is nonempty, its structure set is the six permutations of the block, and by Lemma A.1 the restricted noise law would be exact for it.

A.6 Duality

Remark A.1 (duality). For scores without ties, the maximum-weight basis of $\mathcal{M}$ is the complement of the maximum-weight basis of $\mathcal{M}^*$ for the negated scores. Applied to the perturbed scores, this makes the perturbed optimizer with negative Gumbel noise on $\mathcal{M}$ at θ the complement of the perturbed optimizer with standard Gumbel noise on $\mathcal{M}^*$ at $-\theta$, and the Gibbs distributions correspond in the same way, so the two uniform kinds of Theorem 1(a) are dual to each other. A matroid is self-dual if it is isomorphic to its dual, as the cycle matroid of K_4 is (Oxley, 2011). The f-divergence of a distribution p from a distribution p' on $\mathcal{X}$ is $\sum_x p'(x) f(p(x)/p'(x))$ for a convex f with $f(1) = 0$, and the total variation distance and the Kullback–Leibler divergence are instances. On a self-dual matroid, use standard Gumbel noise on $\mathcal{M}$ with a fixed probability and negative Gumbel noise otherwise. For every f-divergence, the divergence of the output law from p_θ, averaged over a score prior invariant under the sign change composed with a self-duality isomorphism, is a convex function of the probability, symmetric about one half, and so is smallest there.

The identity. Let $\mathcal{M}$ have bases $\mathcal{B}$, so that $\mathcal{M}^*$ has bases $\mathcal{B}^* = \{E \setminus B : B \in \mathcal{B}\}$. For $\mathsf{w} \in \mathbb{R}^E$ whose maximum over $\mathcal{B}$ is attained once, $\langle \mathsf{w}, x_B \rangle = \langle \mathsf{w}, \mathbf{1} \rangle + \langle -\mathsf{w}, x_{E \setminus B} \rangle$ and complementation is a bijection $\mathcal{B} \to \mathcal{B}^*$, so $\arg\max_{B \in \mathcal{B}} \langle \mathsf{w}, x_B \rangle = E \setminus \arg\max_{B' \in \mathcal{B}^*} \langle -\mathsf{w}, x_{B'} \rangle$. With $\mathsf{w} = \theta - \xi$ for standard Gumbel ξ, ties having probability zero, the perturbed optimizer with negative Gumbel noise on $\mathcal{M}$ at scores θ is the complement of the perturbed optimizer with standard Gumbel noise on $\mathcal{M}^*$ at scores $-\theta$, for the same draw of ξ. The Gibbs distributions correspond in the same way, $p_\theta^{\mathcal{M}}(B) = p_{-\theta}^{\mathcal{M}^*}(E \setminus B)$, because $\langle \theta, x_B \rangle = \langle \theta, \mathbf{1} \rangle - \langle \theta, x_{E \setminus B} \rangle$ and the factor $e^{\langle \theta, \mathbf{1} \rangle}$ cancels in the normalization. So the negative Gumbel noise law is exact on $U_{\ell-1,\ell}$ because the standard Gumbel noise law is exact on its dual $U_{1,\ell}$.

The mixture. Let $\mathcal{M}$ be self-dual: a bijection $\mathsf{p} : E \to E$ maps $\mathcal{B}$ onto $\mathcal{B}^*$, so that $B \in \mathcal{B}$ if and only if $E \setminus \mathsf{p}(B) \in \mathcal{B}$. Let $\mathsf{Q}_1(\theta)$ be the law of the perturbed optimizer with standard Gumbel noise on $\mathcal{M}$ at θ and $\mathsf{Q}_2(\theta)$ that with negative Gumbel noise, both compared with $p_\theta = p_\theta^{\mathcal{M}}$. Define $\mathsf{T}\theta$ by $(\mathsf{T}\theta)_e = -\theta_{\mathsf{p}(e)}$ and the bijection t of $\mathcal{B}$ by $\mathsf{t}B = \mathsf{p}^{-1}(E \setminus B)$. Then for every θ and B,

$$p_\theta(B) = p_{\mathsf{T}\theta}(\mathsf{t}B), \qquad \mathsf{Q}_2(\theta)(B) = \mathsf{Q}_1(\mathsf{T}\theta)(\mathsf{t}B), \qquad \mathsf{Q}_1(\theta)(B) = \mathsf{Q}_2(\mathsf{T}\theta)(\mathsf{t}B).$$

The first holds because $\langle \mathsf{T}\theta, x_{\mathsf{t}B} \rangle = -\sum_{e \notin B} \theta_e = \langle \theta, x_B \rangle - \langle \theta, \mathbf{1} \rangle$ and t permutes $\mathcal{B}$. For the second, the identity gives $\mathsf{Q}_2(\theta)(B) = \mathbb{P}(\arg\max_{B' \in \mathcal{B}^*} \langle -\theta + \xi, x_{B'} \rangle = E \setminus B)$. Writing $B' = \mathsf{p}(B'')$ with $B'' \in \mathcal{B}$, using $\langle \mathsf{v}, x_{\mathsf{p}(B'')} \rangle = \sum_{e \in B''} \mathsf{v}_{\mathsf{p}(e)}$ for $\mathsf{v} \in \mathbb{R}^E$ and the fact that the reindexed noise $(\xi_{\mathsf{p}(e)})_{e \in E}$ has the law of ξ, this is $\mathbb{P}(\arg\max_{B'' \in \mathcal{B}} \langle \mathsf{T}\theta + \xi, x_{B''} \rangle = \mathsf{p}^{-1}(E \setminus B)) = \mathsf{Q}_1(\mathsf{T}\theta)(\mathsf{t}B)$. The third follows in the same way from the identity with the sign of the noise reversed, which writes the perturbed optimizer with standard Gumbel noise on $\mathcal{M}$ at θ as the complement of the one with negative Gumbel noise on $\mathcal{M}^*$ at $-\theta$. Hence the mixture $\mathsf{Q}_\mathsf{a} = \mathsf{a}\mathsf{Q}_1 + (1 - \mathsf{a})\mathsf{Q}_2$ satisfies $\mathsf{Q}_\mathsf{a}(\theta)(B) = \mathsf{Q}_{1-\mathsf{a}}(\mathsf{T}\theta)(\mathsf{t}B)$. For every f-divergence, the risk $\mathsf{r}_\theta(\mathsf{a})$, the divergence of $\mathsf{Q}_\mathsf{a}(\theta)$ from p_θ, therefore satisfies $\mathsf{r}_\theta(\mathsf{a}) = \mathsf{r}_{\mathsf{T}\theta}(1 - \mathsf{a})$, an f-divergence being invariant under a relabeling of the outcomes. The risk is convex in a, because f is convex and Q_a is affine in a. Let q be a prior on scores invariant under T, such as a centered isotropic Gaussian, T being a signed permutation of the coordinates. The averaged risk $\bar{\mathsf{r}}(\mathsf{a}) = \int \mathsf{r}_\theta(\mathsf{a})\, \mathsf{q}(d\theta)$ is then convex and symmetric about $\frac{1}{2}$, hence minimal at $\mathsf{a} = \frac{1}{2}$, and the same holds for the supremum over any T-invariant set of scores. The conclusion does not hold pointwise in θ: on $U_{2,4}$ at $\theta = (\log 2, 0, 0, 0)$ the probability that the basis contains element 1 is $\frac{2}{3}$ under p_θ, $\frac{7}{10}$ under $\mathsf{Q}_1(\theta)$ and $\frac{24}{35}$ under $\mathsf{Q}_2(\theta)$. All three are uniform on the bases containing 1 and on those avoiding it, so at this θ every risk is smallest at $\mathsf{a} = 0$.

B Matroid Bases and Assignments

Standard Gumbel noise turns every perturbed score into an exponential clock, and in the clocks the optimality of a structure is a set of inequalities: one per exchange for matroid bases (Appendix B.1), one per cycle and path of reassignments for assignments (Appendix B.2). Appendix B.1 states and proves the law over the bases of any matroid (Proposition B.1), and Appendix B.2 states the law of Gumbel matching for rectangular assignments with forbidden entries (Theorem B.1), the form that data association uses. The proof of Theorems 2 and B.1 is one change of variables on the event that a given assignment is optimal, from the clocks of the entries with finite score to the clocks of the assigned entries, the canonical potentials of the columns, the forest of tight non-assigned entries that attains them, and the slacks of the other entries. The slacks and the clocks of the assigned entries integrate in closed form, and the sum over forests is a determinant. Appendix B.3 computes the law for two rows and two columns and checks Theorem 2 against it, Appendix B.4 integrates out of Theorem B.1 the potentials of the columns that only one row may use, Appendix B.5 proves Proposition B.2, and Appendix B.6 carries the change of variables to the perfect matchings of the complete graph K_{2n}, where nested blossom structures take the place of the forests.

B.1 Matroid Bases

For a basis B and an element $e \notin B$, the fundamental circuit $C(e, B)$ is the unique circuit contained in $B \cup \{e\}$, and it contains e (Oxley, 2011, Section 1.2).

The exchange criterion. Let $\mathsf{w} \in \mathbb{R}^E$ have distinct entries. A basis B of $\mathcal{M}$ maximizes $\langle \mathsf{w}, x_{B'} \rangle$ over the bases B' if and only if $\mathsf{w}_e \leq \mathsf{w}_b$ for every $e \notin B$ and every $b \in C(e, B) \setminus \{e\}$. The condition is necessary because $(B \setminus \{b\}) \cup \{e\}$ is a basis exactly when b lies on $C(e, B)$, the only circuit in $B \cup \{e\}$, and the weight $\langle \mathsf{w}, x_B \rangle - \mathsf{w}_b + \mathsf{w}_e$ of this basis exceeds that of B when $\mathsf{w}_e > \mathsf{w}_b$. It is sufficient by the greedy algorithm, which processes the elements in order of decreasing weight, accepts each element whose addition keeps the accepted set independent, and returns a maximum-weight basis (Edmonds, 1971). By induction the accepted set stays inside B: an element $e \notin B$ is reached after all of $C(e, B) \setminus \{e\}$, which have larger weight and, lying in B, were accepted, so adding e would complete the circuit $C(e, B)$ and e is rejected, while an element of B keeps the accepted subset of B independent and is accepted. So the greedy basis is B.

Proposition B.1 (law of the perturbed optimizer over matroid bases). *Let $\mathcal{X} = \{x_B : B \in \mathcal{B}(\mathcal{M})\}$ for a matroid $\mathcal{M}$ on E, let $\eta \in \mathbb{R}^E$, and let B be a basis. (a) For standard Gumbel noise,*

$$Q_\eta(x_B) = \int_{[0,\infty)^B} \prod_{b\in B} \exp(\eta_b)\, e^{-\exp(\eta_b)\, s_b} \prod_{e\notin B} \exp\Big(-\exp(\eta_e) \max_{b\in C(e,B)\setminus\{e\}} s_b\Big) \prod_{b\in B} ds_b,$$

with $\max \varnothing = 0$ (loops). (b) For a noise law ν with independent entries of distribution function F and density f,

$$Q^\nu_\eta(x_B) = \int_{\mathbb{R}^B} \prod_{b\in B} f(q_b - \eta_b) \prod_{e\notin B} F\Big(\min_{b\in C(e,B)\setminus\{e\}} q_b - \eta_e\Big)\, dq,$$

with $\min \varnothing = +\infty$ and $F(+\infty) = 1$ (loops).

For the k-subsets of $[n]$ with $1 \le k < n$, $C(e,B)\setminus\{e\} = B$ for every $e \notin B$, and part (a) reduces to Proposition 1.

Proof of Proposition B.1. (a) For perturbed scores without ties, B is the maximum-weight basis if and only if $s_e \ge \max_{b\in C(e,B)\setminus\{e\}} s_b$ for every $e \notin B$, by the exchange criterion. Conditionally on the basis clocks these events are independent, with probabilities $\exp(-\exp(\eta_e)\max_b s_b)$, the survival factors of the clocks s_e at their bounds, and integrating against the density of the basis clocks gives the formula. A loop e has $C(e,B) = \{e\}$ and imposes no condition, in keeping with the convention $\max\varnothing = 0$. For k-subsets, B wins if and only if every basis clock is below $s = \min_{e\notin B} s_e$, an exponential clock of rate $\sum_{e\notin B}\exp(\eta_e)$ independent of them, and conditioning on s gives the one-dimensional integral of Proposition 1. (b) Ties have probability zero. With $q_b = \eta_b + \xi_b$ the perturbed scores of the basis elements, of density $\prod_{b\in B} f(q_b - \eta_b)$, the exchange criterion says that B wins if and only if $\eta_e + \xi_e \le \min_{b\in C(e,B)\setminus\{e\}} q_b$ for every $e \notin B$. Conditionally on q these events are independent, with probabilities $F(\min_b q_b - \eta_e)$, a loop contributing $F(+\infty) = 1$, and integrating against the density of q gives the formula. □

Three remarks relate Propositions 1 and B.1 to the dual of the k-subset problem, to the spanning-tree formula of Babson et al. (2026) and to Laplace's method, and Appendix B.5 uses the first.

The expected dual objective and Rosén's threshold. The expected dual objective of a perturbed linear program is a function of some of its dual variables: the dual objective minimized over the others for each realization of the noise and then averaged. Write $(q)_+ = \max(q, 0)$ for the positive part. The k-subset problem is the linear program $\max\langle \eta + \xi, x\rangle$ over $0 \le x \le \mathbf{1}$ with $\sum_i x_i = k$, whose dual is $\min\, k\mathsf{a} + \sum_i \mathsf{b}_i$ subject to $\mathsf{a} + \mathsf{b}_i \ge \eta_i + \xi_i$ and $\mathsf{b}_i \ge 0$. At a fixed multiplier $\mathsf{a} = -\log s$ the optimal b_i is $(\eta_i + \xi_i + \log s)_+$, so the expected dual objective at the clock time s is

$$\mathsf{J}(s) = -k\log s + \sum_i \mathbb{E}(\eta_i + \xi_i + \log s)_+ = -k\log s + \sum_i \mathrm{Ein}\big(s\exp(\eta_i)\big),$$

where $\mathrm{Ein}(q) = \int_0^q (1 - e^{-q'})\, dq'/q'$. The second form follows by integrating $\log(s/s_i)$ against the density of the clock s_i over $\{s_i < s\}$. It is convex in $\log s$, because $k\mathsf{a} + \sum_i \mathbb{E}(\eta_i + \xi_i - \mathsf{a})_+$ is convex in a, and

$$\frac{d\mathsf{J}}{d\log s} = -k + \sum_i \mathbb{P}(\eta_i + \xi_i > -\log s) = -k + \sum_i \mathbb{P}(s_i < s).$$

The derivative vanishes when the expected number of clocks that have rung by the clock time s is k. For standard Gumbel noise $\mathbb{P}(s_i < s) = 1 - e^{-s\exp(\eta_i)}$, and the minimizer is Rosén's threshold s^* (Rosén, 1997a), the solution of $\sum_i (1 - e^{-s^*\exp(\eta_i)}) = k$.

The greedy-order sum. Processing the clocks in increasing order, the greedy algorithm accepts an element if and only if its addition keeps the accepted set A independent, and by the memorylessness of the clocks the next accepted element is e with probability $\exp(\eta_e)/\sum_{e'}\exp(\eta_{e'})$, the sum over the elements $e' \notin \mathsf{A}$ for which $\mathsf{A}\cup\{e'\}$ is independent, whose clocks have not rung yet. Hence

$$Q_\eta(x_B) = \sum_{(b_1,\dots,b_r)} \prod_{i=1}^{r} \frac{\exp(\eta_{b_i})}{\sum_{e'}\exp(\eta_{e'})},$$

the outer sum over the orderings of B, $\mathsf{r} = |B|$, and the inner sum over the $e' \notin \{b_1, \dots, b_{i-1}\}$ for which $\{b_1, \dots, b_{i-1}, e'\}$ is independent: a rational function of the rates. At $\eta = 0$ on the spanning trees of a graph this is the formula of Babson et al. (2026, Theorem 3.5) for the law of the minimum spanning tree under independent identically distributed continuous edge weights, which give the same tree law as unit-rate clocks because only the order of the weights matters. On K_4 a star has probability $\frac{1}{15}$ (Babson et al., 2026, Corollary 3.7) and a path $\frac{11}{180}$ (Babson et al., 2026, Table 1), against the Gibbs value $\frac{1}{16}$.

Laplace's method in the clock variable. For k-subsets, write the integral of Proposition 1 as $Q_\eta(x_B) = \int_0^\infty \Sigma_B e^{\mathsf{h}_B(s)}\, ds$ with $\Sigma_B = \sum_{e \notin B} \exp(\eta_e)$ and $\mathsf{h}_B(s) = -s\Sigma_B + \sum_{b \in B} \log(1 - e^{-s\exp(\eta_b)})$, which is concave in s. Its Laplace approximation $Q^{\mathrm{L}}(x_B) = \Sigma_B e^{\mathsf{h}_B(\mathsf{s}_B)} \big(2\pi/|\mathsf{h}_B''(\mathsf{s}_B)|\big)^{1/2}$ at the maximizer s_B of h_B defines a temperature c^{L} through its first-order term. At $\eta = 0$ the maximizer is $\mathsf{s}_0 = \log \frac{n}{n-k}$, Rosén's threshold at equal rates, and in general it depends on B. For η with $\sum_i \eta_i = 0$ the first-order term of $\log Q^{\mathrm{L}}(x_B)$ is $c^{\mathrm{L}}\eta_B$ with $\eta_B = \sum_{b \in B} \eta_b$, and normalizing over B does not change c^{L}, because $\sum_B \eta_B = 0$. With $\bar{k} = n - k$ and $e^{-\mathsf{s}_0} = \bar{k}/n$, three terms make up c^{L}. The envelope term of h_B is $\frac{n}{k}\mathsf{s}_0$: at the maximizer only the explicit dependence on η counts, $\partial \mathsf{h}_B/\partial \eta_e = -\mathsf{s}_0$ for $e \notin B$ and $\partial \mathsf{h}_B/\partial \eta_b = \mathsf{s}_0 \bar{k}/k$ for $b \in B$, and $\sum_{e \notin B} \eta_e = -\eta_B$. The term of $\log \Sigma_B$ is $-\frac{1}{\bar{k}}$. The term of $-\frac{1}{2}\log|\mathsf{h}_B''|$ is $\frac{1}{2\bar{k}}$: $|\mathsf{h}_B''(s)| = \sum_{b \in B} \mathsf{p}(\exp(\eta_b), s)$ with $\mathsf{p}(\mathsf{v}, s) = \mathsf{v}^2 e^{-s\mathsf{v}}/(1 - e^{-s\mathsf{v}})^2$. At $\mathsf{v} = 1$ and $s = \mathsf{s}_0$ one has $\partial_{\mathsf{v}} \log \mathsf{p} = 2 - \mathsf{s}_0 \frac{k+2\bar{k}}{k}$ and $\partial_s \log \mathsf{p} = -\frac{k+2\bar{k}}{k}$, and the maximizer shifts by $d\mathsf{s}_B = \eta_B(\frac{1}{\bar{k}} - \frac{\mathsf{s}_0}{k})$, so

$$-\tfrac{1}{2}\, d \log |\mathsf{h}_B''| = -\tfrac{1}{2}\Big[\tfrac{1}{k}\,\partial_{\mathsf{v}} \log \mathsf{p} + \Big(\tfrac{1}{\bar{k}} - \tfrac{\mathsf{s}_0}{k}\Big)\partial_s \log \mathsf{p}\Big]\eta_B = \frac{\eta_B}{2\bar{k}}.$$

Hence

$$c^{\mathrm{L}} = \frac{n}{k} \log \frac{n}{n-k} - \frac{1}{2(n-k)},$$

and the first-order temperature $\frac{n}{k}(H_n - H_{n-k})$ of Theorem 3 exceeds it by $\frac{2n-k}{12n(n-k)^2} + O((n-k)^{-4})$ as $n - k \to \infty$, uniformly in k, by the Euler–Maclaurin expansion of $H_n - H_{n-k}$, just as $\frac{n}{n+1}$ exceeds the Laplace value $\frac{n-1}{n}$ of Proposition B.2.

B.2 Gumbel Matching

In the logarithms of the clocks of the entries with finite score, the event $\{x^* = \sigma\}$ is a polyhedral cone whose facets depend on σ. The canonical potentials and the forest of tight non-assigned entries are the coordinates in which it becomes, forest by forest, a product of half-lines, orthants and, in the square case, a hyperplane.

Setting for rectangular assignments. An assignment σ uses the columns $\sigma([n])$ and leaves the other columns unused. A forbidden entry has score $-\infty$ and is never used. The other entries are allowed. For a score matrix with forbidden entries and at least one allowed assignment, $x^*(\eta)$, Q_η and p_η are the objects of Section 2 for the structure set of assignments that avoid the forbidden entries. The assignment linear program relaxes the assignment problem to the nonnegative matrices that vanish on the forbidden entries and have unit row sums and column sums at most one, equal to one when $N = n$. Its dual variables are the row potentials u_i and the column potentials $y \in \mathbb{R}^N$, with the constraints $u_i - y_j \geq \eta_{ij} + \xi_{ij}$ and, for $N > n$, $y_j \leq 0$. The row-softmax matrix $S_\eta(y)$, now $n \times N$, and the function $\psi_\eta(y)$ are given by the formulas of Section 3.3 with the sums over the N columns and $y_j = 0$ on the unused columns, and ψ_η is log-concave in y. For an $N \times N$ matrix A, A_σ is its principal submatrix on the used columns $\sigma([n])$. The reachability condition for σ asks that every used column reach an unused column through a chain of columns in which each step passes through an entry of finite score in the row assigned to the current column. It holds whenever all scores are finite and $N > n$, and in the private-column setting of Appendix B.4.

For $N > n$, the random potentials Y of the perturbed problem are indexed by the columns. For a used column j of $x^*(\eta)$, Y_j is the optimal value of the perturbed problem with column j forbidden minus the optimal value of the perturbed problem itself, so $Y_j \leq 0$, with $Y_j = -\infty$ when forbidding column j leaves no allowed assignment. On the unused columns $Y_j = 0$. Step 2 of the proof identifies Y with the canonical potentials.

Theorem B.1 (law of Gumbel matching for rectangular assignments). *Let $\eta \in (\mathbb{R} \cup \{-\infty\})^{n \times N}$ with $N > n$, and let σ be an assignment that uses only entries with finite score and satisfies the reachability condition. Then*

$$Q_\eta(\sigma) = e^{\langle \eta, P_\sigma \rangle} \int_{\{y_j \le 0,\ j \in \sigma([n])\}} \psi_\eta(y)\, \det\Big(\big(I - P_\sigma^\top S_\eta(y)\big)_\sigma \Big)\, dy, \qquad y_j = 0 \text{ for } j \notin \sigma([n]),$$

and the determinant is nonnegative. The product of $e^{\langle \eta, P_\sigma \rangle}$ and the integrand is the joint density of $(x^*(\eta), Y)$ at (σ, y) with respect to Lebesgue measure on the used coordinates. The assignment enters through P_σ and through the set of used columns, which fixes the integration domain.

Proof of Theorems 2 and B.1. Setting. Only allowed entries carry noise, their clocks $s_{ij} = \exp(-(\eta_{ij} + \xi_{ij}))$ are independent exponential variables with rates $\exp(\eta_{ij})$, and x^* minimizes $\sum_i \log s_{i\sigma(i)}$ over the allowed assignments. Fix an allowed σ and relabel the used columns so that $\sigma(i) = i$, which makes σ the identity id. The used columns are then $\mathsf{U} = [n]$ and the unused ones $\bar{\mathsf{U}} = [N] \setminus [n]$. Write $\mathsf{c}_{ij} = \log s_{ij}$. The column graph has the columns as vertices and an edge $i \to j$ from each used column i to every column $j \neq i$ with (i, j) allowed, and the reduced cost of this edge is $\mathsf{d}_{ij} = \mathsf{c}_{ij} - \mathsf{c}_{ii}$. Almost surely no directed cycle has reduced cost zero and no two distinct directed paths from the same column have equal cost, each such equality being a nontrivial linear relation among independent continuous variables. So on the optimality event every cycle has positive cost, and shortest paths, when they exist, are unique.

Step 1 (optimality is a condition on cycles and paths). Let $\sigma' \neq \mathrm{id}$ be an allowed assignment. The edges $i \to \sigma'(i)$ over the rows with $\sigma'(i) \neq i$ have in-degree and out-degree at most one at every column, so they form vertex-disjoint directed cycles among used columns and directed paths from a used column to an unused one (a path cannot end at a used column j, since then $\sigma'(j) = j$ would use j twice), and $\sum_i \mathsf{c}_{i\sigma'(i)} - \sum_i \mathsf{c}_{ii}$ is the sum of the reduced costs along these cycles and paths. Reassigning along a single cycle or path is itself an allowed assignment. Hence the identity is optimal if and only if every directed cycle among used columns and every directed path from a used column to $\bar{\mathsf{U}}$ has nonnegative reduced cost.

Step 2 (the canonical potentials). In the rectangular case, $N > n$, the reachability condition says that every used column reaches $\bar{\mathsf{U}}$ in the column graph. On the optimality event let h_i be the shortest-path distance from the used column i to $\bar{\mathsf{U}}$, finite by reachability and well defined because no cycle is negative, let $\mathsf{h}_j = 0$ for $j \in \bar{\mathsf{U}}$, and put $y = -\mathsf{h}$, so that $y_{\mathsf{U}} \le 0$ and $y_{\bar{\mathsf{U}}} = 0$. The shortest-path inequality $\mathsf{h}_i \le \mathsf{d}_{ij} + \mathsf{h}_j$ along every edge reads $\mathsf{c}_{ij} \ge \mathsf{c}_{ii} + y_j - y_i$. With $u_i = \eta_{ii} + \xi_{ii} + y_i$, it reads $u_i - y_j \ge \eta_{ij} + \xi_{ij}$ for every allowed (i, j), with equality on the assigned entries, so by complementary slackness the row potentials u_i and the column potentials y are optimal dual variables of the assignment problem. These are the canonical potentials. In terms of assignments, h_i is the least cost increase of an allowed assignment that leaves column i unused, because such an assignment decomposes as in Step 1 into components of nonnegative cost, one of which is a path from i to $\bar{\mathsf{U}}$ (row i moves and no row moves into column i), of cost at least h_i, and reassigning along a shortest path attains h_i. So $-\mathsf{h}_i$ is the optimal value of the perturbed problem with column i forbidden minus the optimal value, which is the definition of Y_i above. In the square case, $N = n$ with all scores finite, fix a root column r, let h_i be the shortest-path distance from i to r, with $\mathsf{h}_\mathsf{r} = 0$, and put $y = -\mathsf{h}$, with no sign constraint. The same inequalities hold, an edge out of r satisfying $0 \le \mathsf{d}_{\mathsf{r}j} + \mathsf{h}_j$ because the closed walk $\mathsf{r} \to j \to \cdots \to \mathsf{r}$ decomposes into cycles of nonnegative cost.

Step 3 (the forest of tight non-assigned entries). For every used column i (every $i \neq \mathsf{r}$ in the square case) let the parent p_i be the next column on the unique shortest path from i to $\bar{\mathsf{U}}$ (to r). The edges $i \to \mathsf{p}_i$ form a forest $\mathcal{T}$ rooted in $\bar{\mathsf{U}}$ (a tree rooted at r), since following parents traces a shortest path, and $\mathsf{d}_{i\mathsf{p}_i} = \mathsf{h}_i - \mathsf{h}_{\mathsf{p}_i} = y_{\mathsf{p}_i} - y_i$: the parent entry is tight. Every other allowed off-diagonal entry is slack, $\mathsf{d}_{ij} > y_j - y_i$: equality would make the edge $i \to j$ followed by the shortest path from j a walk from i of cost h_i, a second shortest path if it is simple, and otherwise, once its positive-cost cycle is removed, a path of cost below h_i.

Step 4 (the change of variables). Given the forest $\mathcal{T}$, introduce three kinds of variables: the diagonal clocks $\mathsf{x}_i = s_{ii} > 0$, which at fixed y parametrize the row potentials through $\mathsf{x}_i = e^{y_i - u_i}$; the potentials y; and, for every allowed off-diagonal entry that is not a parent entry, the slack $\mathsf{z}_{ij} > 0$

(in the square case the root row has no parent entry, and all its off-diagonal entries carry slacks):

$$s_{ii} = \mathsf{x}_i, \qquad s_{i\mathsf{p}_i} = \mathsf{x}_i e^{y_{\mathsf{p}_i} - y_i}, \qquad s_{ij} = \mathsf{x}_i e^{y_j - y_i} + \mathsf{z}_{ij} \quad (j \neq i,\ j \neq \mathsf{p}_i).$$

Conversely, for any forest $\mathcal{T}$ rooted in $\bar{\mathsf{U}}$ (tree rooted at r), any y with $y_{\mathsf{U}} \leq 0$ (with $y_{\mathsf{r}} = 0$) and any $\mathsf{x} > 0$, $\mathsf{z} > 0$, the clocks so defined satisfy $\mathsf{d}_{ij} \geq y_j - y_i$ on every edge, with equality exactly on the forest edges. So every cycle has nonnegative cost, the potential differences telescoping to zero; every path from i to $\bar{\mathsf{U}}$ (to r) has cost at least $-y_i$; and the forest path has cost exactly $-y_i$ and is the unique shortest one. Hence the identity is optimal, $\mathsf{h}_i = -y_i$, and the forest of tight non-assigned entries is $\mathcal{T}$. The map is therefore a bijection, up to null sets, between the optimality event with forest $\mathcal{T}$ and $\{\mathsf{x} > 0\} \times \{y_{\mathsf{U}} \leq 0\} \times \{\mathsf{z} > 0\}$, the middle factor being $\{y_{\mathsf{r}} = 0\}$ in the square case. For its Jacobian, order the clocks as (diagonal entries, parent entries, slack entries) and the variables as $(\mathsf{x}, y, \mathsf{z})$: the matrix is block lower-triangular with identity blocks in the first and third diagonal positions, and the middle block is $\operatorname{diag}(\mathsf{x}_i e^{y_{\mathsf{p}_i} - y_i})\, \mathsf{B}_{\mathcal{T}}$, where the row of $\mathsf{B}_{\mathcal{T}}$ for i is $e_{\mathsf{p}_i}^\top - e_i^\top$ in the standard basis vectors e_j indexed by the used columns, the term $e_{\mathsf{p}_i}^\top$ being absent when $\mathsf{p}_i \in \bar{\mathsf{U}}$ (rows $i \neq \mathsf{r}$ and columns $j \neq \mathsf{r}$ in the square case). Listing the used columns in an order in which every parent precedes its children makes $\mathsf{B}_{\mathcal{T}}$ triangular with diagonal -1, so

$$|\det \mathsf{J}| = \prod_i \mathsf{x}_i e^{y_{\mathsf{p}_i} - y_i},$$

the product over the used columns (over $i \neq \mathsf{r}$ in the square case).

Step 5 (integrating the slacks and the diagonal clocks). The density of the clocks is $\prod_{(i,j)\text{ allowed}} \exp(\eta_{ij}) e^{-\exp(\eta_{ij}) s_{ij}}$, and in the new coordinates $\sum_j \exp(\eta_{ij}) s_{ij} = \mathsf{x}_i \mathsf{A}_i(y) + \sum_j \exp(\eta_{ij}) \mathsf{z}_{ij}$, the second sum over the slack entries of row i, with $\mathsf{A}_i(y) = \sum_j \exp(\eta_{ij}) e^{y_j - y_i} = e^{-y_i} \mathsf{Z}_i(y)$ and $\mathsf{Z}_i(y) = \sum_j \exp(\eta_{ij}) e^{y_j}$. The survival factors at the bounds are collected in $e^{-\mathsf{x}_i \mathsf{A}_i(y)}$, and each slack integrates to $\int_0^\infty \exp(\eta_{ij}) e^{-\exp(\eta_{ij})\mathsf{z}}\, d\mathsf{z} = 1$. Each diagonal clock that carries the Jacobian factor x_i integrates to $\int_0^\infty \mathsf{x}_i e^{-\mathsf{A}_i \mathsf{x}_i}\, d\mathsf{x}_i = \mathsf{A}_i^{-2}$, and in the square case the clock of the root row, which carries no factor, to $\mathsf{A}_{\mathsf{r}}^{-1}$. What remains is the joint density of (the identity, $\mathcal{T}$, y):

$$\prod_i \frac{\exp(\eta_{ii})}{\mathsf{A}_i(y)} \prod_i \frac{\exp(\eta_{i\mathsf{p}_i}) e^{y_{\mathsf{p}_i} - y_i}}{\mathsf{A}_i(y)} = \prod_i S_\eta(y)_{ii} \prod_i S_\eta(y)_{i\mathsf{p}_i},$$

the second product over the used columns (over $i \neq \mathsf{r}$), because $S_\eta(y)_{ij} = \exp(\eta_{ij}) e^{y_j} / \mathsf{Z}_i(y)$. The first product is $\prod_i S_\eta(y)_{ii} = e^{\langle \eta, P_{\mathrm{id}} \rangle} e^{\sum_{i \in \mathsf{U}} y_i} / \prod_i \mathsf{Z}_i(y) = e^{\langle \eta, P_{\mathrm{id}} \rangle} \psi_\eta(y)$, with $y_j = 0$ off the used columns.

Step 6 (the forest sum is a determinant). For a matrix P with nonnegative entries indexed by U and row sums at most one, put $\mathsf{b}_i = 1 - \sum_{j \in \mathsf{U}} \mathsf{P}_{ij}$. Then

$$\det(I - \mathsf{P}) = \sum_{\mathcal{F}} \prod_{i \text{ with a parent}} \mathsf{P}_{i\mathsf{p}_i} \prod_{i \text{ a root}} \mathsf{b}_i,$$

the sum over the forests $\mathcal{F}$ on U in which every vertex either has one parent in U or is a root. Row i of $I - \mathsf{P}$ is $\mathsf{b}_i e_i^\top + \sum_{j \neq i} \mathsf{P}_{ij} (e_i - e_j)^\top$. Expanding the determinant multilinearly in its rows chooses for every i either the root term or an edge $i \to j$. A choice containing a directed cycle has rows summing to zero along the cycle and contributes nothing. An acyclic choice is a forest, and its matrix, with rows and columns listed in an order in which every parent precedes its children, is unitriangular. In the rectangular case take $\mathsf{P} = S_\eta(y)_{\mathsf{U}\mathsf{U}}$: then $\mathsf{b}_i = \sum_{j \in \bar{\mathsf{U}}} S_\eta(y)_{ij}$ is the sum of $S_\eta(y)_{i\mathsf{p}_i}$ over the possible unused parents, so summing the density of Step 5 over the forests $\mathcal{T}$ rooted in $\bar{\mathsf{U}}$ gives $e^{\langle \eta, P_{\mathrm{id}} \rangle} \psi_\eta(y) \det(I - S_\eta(y)_{\mathsf{U}\mathsf{U}})$. In the square case take P to be the block of $S_\eta(y)$ on the rows and columns other than r: since the rows of $S_\eta(y)$ sum to one, $\mathsf{b}_i = S_\eta(y)_{i\mathsf{r}}$, and summing over the trees rooted at r gives $e^{\langle \eta, P_{\mathrm{id}} \rangle} \psi_\eta(y) \det\left((I - S_\eta(y))^{(\mathsf{r})}\right)$, where $\mathsf{L}^{(\mathsf{r})}$ denotes the principal submatrix of a matrix L without row and column r. Both determinants are sums of nonnegative terms.

Step 7 (the two cases). Rectangular: integrating over $y_{\mathsf{U}} \leq 0$ gives $Q_\eta(\mathrm{id})$. When the relabeling is undone, row j of $P_\sigma^\top S_\eta(y)$, for a used column j, is row $\sigma^{-1}(j)$ of $S_\eta(y)$, so $S_\eta(y)_{\mathsf{U}\mathsf{U}}$ is the principal submatrix $(P_\sigma^\top S_\eta(y))_\sigma$ on the used columns. This is Theorem B.1, with the integrand

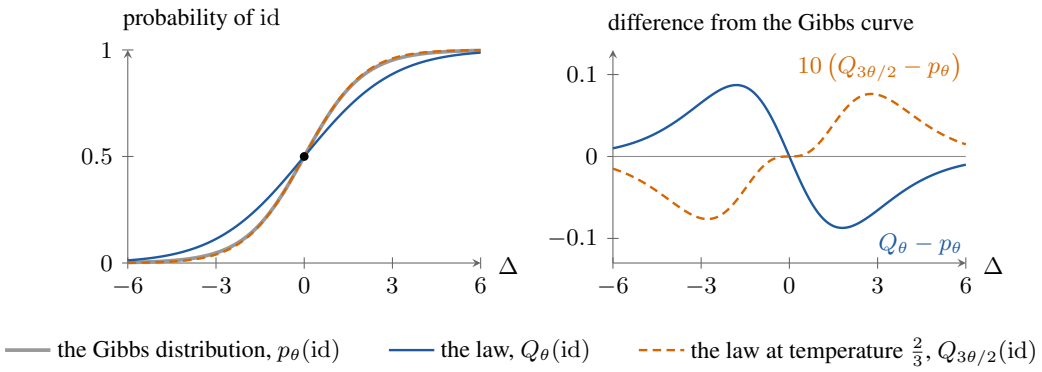


Figure B.1: The two-by-two example. Left: the probability of the identity permutation against the score difference Δ under the Gibbs distribution, under the law of Gumbel matching and under that law at temperature $\frac{2}{3}$, with slopes $\frac{1}{4}$, $\frac{1}{6}$ and $\frac{1}{4}$ at $\Delta = 0$. Right: the law and the calibrated law minus the Gibbs curve, the calibrated difference multiplied by ten. The temperature matches the slope at zero and leaves a residual of third order.

identified in Steps 5 and 6 as the joint density of (x^*, Y) on the used coordinates. Square: for each root r, $Q_\eta(\mathrm{id}) = e^{\langle \eta, P_{\mathrm{id}} \rangle} \int_{\{y_\mathsf{r}=0\}} \psi_\eta(y) \det\left((I - S_\eta(y))^{(\mathsf{r})}\right) dy$, and the integrand is invariant under $y \mapsto y + \mathsf{t}\mathbf{1}$, so the hyperplane $\{y_\mathsf{r} = 0\}$ may be replaced by $\{y_1 = 0\}$ by a shift of unit Jacobian. The n integrands need not coincide, but each integrates to $Q_\eta(\mathrm{id})$. Averaging them and using, for a row-stochastic matrix S with $\mathsf{L} = I - \mathsf{S}$, the identity $\sum_\mathsf{r} \det(\mathsf{L}^{(\mathsf{r})}) = \det(\mathsf{L} + \frac{1}{n}\mathbf{1}\mathbf{1}^\top)$ gives Theorem 2. To prove the identity, observe that $\mathsf{L}\mathbf{1} = 0$, so $\mathsf{L}\,\mathrm{adj}(\mathsf{L}) = \det(\mathsf{L})I = 0$ puts every column of $\mathrm{adj}(\mathsf{L})$ in $\ker \mathsf{L}$. If $\mathrm{rank}\,\mathsf{L} = n - 1$, then $\ker \mathsf{L} = \mathbb{R}\mathbf{1}$ and $\mathrm{adj}(\mathsf{L}) = \mathbf{1}\mathsf{a}^\top$ with $\mathsf{a}_\mathsf{r} = \det(\mathsf{L}^{(\mathsf{r})})$ the principal cofactors, and $\det(\mathsf{L} + \mathsf{u}\mathsf{v}^\top) = \det \mathsf{L} + \mathsf{v}^\top \mathrm{adj}(\mathsf{L})\mathsf{u}$ with $\mathsf{u} = \mathsf{v} = \mathbf{1}/\sqrt{n}$ gives $\det(\mathsf{L} + \frac{1}{n}\mathbf{1}\mathbf{1}^\top) = \frac{1}{n}\mathbf{1}^\top \mathbf{1}\, \mathsf{a}^\top \mathbf{1} = \sum_\mathsf{r} \mathsf{a}_\mathsf{r}$. If $\mathrm{rank}\,\mathsf{L} < n - 1$, then $\mathrm{adj}(\mathsf{L}) = 0$ and both sides vanish. □

Without reachability the formula fails: for $n = 1$, $N = 2$ and the second entry forbidden, the only assignment has probability one, the integrand vanishes identically because $S_\eta(y)_{11} = 1$, and $Y_1 = -\infty$. The square case is not the rectangular case at $N = n$: with no unused column there is no root at potential zero, the determinant of I minus the full row-stochastic block vanishes, and a root has to be chosen.

B.3 Two Rows and Two Columns

For two rows and two columns the law can be computed by hand and Theorem 2 checked against it. Let $n = N = 2$, let id be the identity permutation, and let $\Delta = \theta_{11} + \theta_{22} - \theta_{12} - \theta_{21}$ be the score difference between the two permutations, so that $p_\theta(\mathrm{id}) = 1/(1 + e^{-\Delta})$. Let ζ_1, ζ_2 be independent standard logistic variables. With standard Gumbel noise the perturbed optimizer returns id when $(\xi_{12} - \xi_{11}) + (\xi_{21} - \xi_{22}) < \Delta$. The difference of two independent standard Gumbel variables is standard logistic, so the two differences are independent standard logistic variables and

$$Q_\theta(\mathrm{id}) = \mathbb{P}(\zeta_1 + \zeta_2 \le \Delta). \tag{B.1}$$

As functions of Δ, both probabilities pass through $\frac{1}{2}$ at $\Delta = 0$, with slopes $\frac{1}{4}$ and $\frac{1}{6}$: the density of $\zeta_1 + \zeta_2$ at zero is the integral of the squared standard logistic density, $\frac{1}{6}$. At temperature $\frac{2}{3}$ the law is $Q_{3\theta/2}$, whose curve $\mathbb{P}(\zeta_1 + \zeta_2 \le \frac{3}{2}\Delta)$ has slope $\frac{1}{4}$ at zero. At this temperature, and at no other, the curve of the law is tangent to the Gibbs curve at $\Delta = 0$ (Figure B.1).

The law Q_θ differs from p_θ away from $\Delta = 0$, and $c_1 = \frac{2}{3}$ is the value of $\frac{n}{n+1}$ at $n = 2$. Because $Q_\theta(\mathrm{id})$, as a function of Δ, is a smooth increasing bijection of $\mathbb{R}$ onto $(0, 1)$, replacing Δ by the point at which it reaches $p_\theta(\mathrm{id})$ is a calibration under which the two laws coincide for every θ. There is one degree of freedom in the scores and one in the distribution, and Proposition C.3 shows that no

deterministic calibration does the same from $n = 3$ on. An exact noise law also exists for two rows and two columns: with standard Gumbel noise on the first row only and zero noise elsewhere, the perturbed optimizer returns id when the standard logistic variable $\xi_{12} - \xi_{11}$ is below Δ, so $Q^{\nu}_{\theta} = p_{\theta}$. This is why Theorem 1 starts at three rows.

At $n = 2$ the formula of Theorem 2 can be evaluated by hand. With $y = (0, y_2)$, $\mathsf{w} = e^{y_2}$, $\mathsf{a} = e^{\eta_{11}}$, $\mathsf{b} = e^{\eta_{12}}$, $\mathsf{c} = e^{\eta_{21}}$, $\mathsf{d} = e^{\eta_{22}}$, $\mathsf{Z}_1 = \mathsf{a} + \mathsf{bw}$ and $\mathsf{Z}_2 = \mathsf{c} + \mathsf{dw}$, the determinant for $\sigma = \mathrm{id}$ is the sum of the two principal cofactors, $(1 - \mathsf{a}/\mathsf{Z}_1) + (1 - \mathsf{dw}/\mathsf{Z}_2) = \mathsf{bw}/\mathsf{Z}_1 + \mathsf{c}/\mathsf{Z}_2$, and with $dy_2 = d\mathsf{w}/\mathsf{w}$,

$$Q_{\eta}(\mathrm{id}) = \frac{\mathsf{ad}}{2} \int_0^{\infty} \Big(\frac{\mathsf{bw}}{\mathsf{Z}_1^2 \mathsf{Z}_2} + \frac{\mathsf{c}}{\mathsf{Z}_1 \mathsf{Z}_2^2} \Big) d\mathsf{w} = \frac{e^{\Delta}\big(e^{\Delta} - 1 - \Delta\big)}{(e^{\Delta} - 1)^2},$$

where $\Delta = \log \frac{\mathsf{ad}}{\mathsf{bc}} = \eta_{11} + \eta_{22} - \eta_{12} - \eta_{21}$ is the score difference above at the scores η: with $\mathsf{p} = \mathsf{a}/\mathsf{b}$ and $\mathsf{q} = \mathsf{c}/\mathsf{d}$, partial fractions in w give both integrals the value $\frac{1}{\mathsf{bd}}\big(\frac{\mathsf{q}\log(\mathsf{q}/\mathsf{p})}{(\mathsf{p}-\mathsf{q})^2} + \frac{1}{\mathsf{p}-\mathsf{q}}\big)$, and $e^{\Delta} = \mathsf{p}/\mathsf{q}$. This is the distribution function at Δ of the sum of two independent standard logistic variables, Equation (B.1), with value $\frac{1}{2}$ and slope $\frac{1}{6}$ at $\Delta = 0$ by continuity.

B.4 Private Columns

Let $N = m + n$, let the columns $1, \dots, m$ be shared, and let column $m + i$ be private to row i: finite in row i and forbidden in every other row. Every row also has a finite shared entry. Reachability then holds for every assignment σ: a used shared column $\sigma(i) \le m$ reaches the unused private column $m + i$ in one step, and a used private column $m + i$ reaches in one step a shared column j with a finite entry in row i, which is unused or else used by a row i' whose private column $m + i'$ is unused and reachable from j in one further step. The shared-column set of σ is $\mathcal{G} = \sigma([n]) \cap [m]$, and $\sigma_{\varnothing}$ is the assignment that uses only private columns. For a set $\mathcal{G}$ of columns, $(\cdot)_{\mathcal{G}}$ is the principal submatrix on $\mathcal{G}$, and $\Psi(\mathcal{G}) = \int_{\{y_j \le 0,\ j \in \mathcal{G}\}} \psi_{\eta}(y)\, dy$, with $y_j = 0$ for $j \notin \mathcal{G}$, is the normalizer of ψ_{η} on that domain. The next lemma reduces the integral of Theorem B.1 from the used columns to the used shared columns, so that the assignments with the same shared-column set share the density proportional to ψ_{η} and its normalizer, which Section 5 uses.

Lemma B.1 (reduced law for rectangular assignments with private columns)**.** *For every assignment σ with shared-column set $\mathcal{G}$,*

$$Q_{\eta}(\sigma) = e^{\langle \eta, P_{\sigma} \rangle} \int_{\{y_j \le 0,\ j \in \mathcal{G}\}} \psi_{\eta}(y)\, \det\Big(\big(I - P_{\sigma}^{\top} S_{\eta}(y)\big)_{\mathcal{G}} \Big)\, dy, \qquad y_j = 0 \text{ for } j \notin \mathcal{G}.$$

In particular $Q_{\eta}(\sigma_{\varnothing}) = \prod_i \frac{e^{\eta_{i,m+i}}}{e^{\eta_{i,m+i}} + \sum_{j \le m} e^{\eta_{ij}}}$, and the restriction of ψ_{η} to $\{y_j = 0,\ j \notin \mathcal{G}\}$ and its normalizer $\Psi(\mathcal{G})$ depend on σ only through $\mathcal{G}$.

Proof. The potentials of the used private columns integrate out of Theorem B.1 one at a time. Let $\mathcal{R}$ be the set of rows using their private column, so that the used columns are $\mathcal{G} \cup \{m + i : i \in \mathcal{R}\}$, listed with $\mathcal{G}$ first. A private column $m + i$ has a single allowed entry, in row i, so the column of $P_{\sigma}^{\top} S_{\eta}(y)$ indexed by $m + i$ has its only nonzero entry, $S_{\eta}(y)_{i,m+i}$, in the row indexed by $m + i$ itself. Hence the block of $I - P_{\sigma}^{\top} S_{\eta}(y)$ on the used columns is block lower-triangular with diagonal blocks $(I - P_{\sigma}^{\top} S_{\eta}(y))_{\mathcal{G}}$ and $\mathrm{diag}(1 - S_{\eta}(y)_{i,m+i})_{i \in \mathcal{R}}$, and its determinant is the product of the determinants of the two diagonal blocks. For $i \in \mathcal{R}$ write $\mathsf{Z}_i(y) = \mathsf{a}_i e^{y_{m+i}} + \mathsf{b}_i$ with $\mathsf{a}_i = e^{\eta_{i,m+i}}$ and $\mathsf{b}_i = \sum_{j \le m} e^{\eta_{ij} + y_j} > 0$, which does not involve y_{m+i}. The rows $i \notin \mathcal{R}$ have their private column unused, so their Z_i involves no private potential either, and neither does $(I - P_{\sigma}^{\top} S_{\eta}(y))_{\mathcal{G}}$, whose rows correspond to rows outside $\mathcal{R}$. The factors of the integrand that contain y_{m+i} are $e^{y_{m+i}}/\mathsf{Z}_i(y)$ from ψ_{η} and $1 - S_{\eta}(y)_{i,m+i} = \mathsf{b}_i/\mathsf{Z}_i(y)$ from the determinant, and, with $\mathsf{w} = e^{y_{m+i}}$,

$$\int_{-\infty}^{0} \frac{\mathsf{b}_i e^{y_{m+i}}}{(\mathsf{a}_i e^{y_{m+i}} + \mathsf{b}_i)^2}\, dy_{m+i} = \int_0^1 \frac{\mathsf{b}_i\, d\mathsf{w}}{(\mathsf{a}_i \mathsf{w} + \mathsf{b}_i)^2} = \frac{1}{\mathsf{a}_i + \mathsf{b}_i},$$

which is the factor $e^{y_{m+i}}/\mathsf{Z}_i(y)$ of ψ_{η} evaluated at $y_{m+i} = 0$. So integrating the used private potentials out of Theorem B.1 gives the display. With $\mathcal{G} = \varnothing$ there is no integration and the determinant is empty, which gives the closed form. Finally, on $\{y_j = 0,\ j \notin \mathcal{G}\}$, $\psi_{\eta}(y) = e^{\sum_{j \in \mathcal{G}} y_j} / \prod_i \mathsf{Z}_i(y)$ with $\mathsf{Z}_i(y) = e^{\eta_{i,m+i}} + \sum_{j \le m} e^{\eta_{ij} + y_j}$, so this function depends on σ only through $\mathcal{G}$. □

B.5 SINKHORN SCALING

The Sinkhorn scaling of a positive square matrix is its unique doubly stochastic scaling by positive diagonal matrices (Sinkhorn, 1964).

Proposition B.2 (Sinkhorn scaling as the maximizer of ψ_η)**.** *For $N = n$ and finite η, $\log \psi_\eta$ is strictly concave on $\{y_1 = 0\}$. Its maximizer y^* is the unique point of $\{y_1 = 0\}$ at which $S_\eta(y)$ is doubly stochastic, that is, $S_\eta(y^*)$ is the Sinkhorn scaling of $\exp(\eta)$. The law of Theorem 2 with its integral replaced by the Laplace approximation at y^*, taken at $\eta = \theta$ and normalized over σ, has the same first-order term at $\theta = 0$ as $p_{\frac{n-1}{n}\theta}$, whereas $Q_\theta = p_{\frac{n}{n+1}\theta} + O(\|\theta\|_F^2)$.*

The Sinkhorn column potentials are the potentials y^* of the proposition. The exact law integrates the determinant against $\psi_\theta(y)$, whereas the Laplace approximation evaluates it at y^*, and the first-order coefficients differ by $\frac{1}{n(n+1)}$.

Proof of Proposition B.2. Strict concavity and the maximizer. For $N = n$ and finite η, $\log \psi_\eta(y) = \sum_j y_j - \sum_i \log \mathsf{Z}_i(y)$ with $\mathsf{Z}_i(y) = \sum_j \exp(\eta_{ij}) e^{y_j}$, and for $\mathsf{v} \in \mathbb{R}^n$, $-\mathsf{v}^\top \nabla^2 \log \psi_\eta(y)\, \mathsf{v} = \sum_i \mathrm{Var}_{S_\eta(y)_{i\cdot}}(\mathsf{v})$, each summand being the variance of a coordinate of v drawn from the softmax row i. The sum vanishes if and only if v is constant, so $\log \psi_\eta$ is strictly concave on $\{y_1 = 0\}$. It tends to $-\infty$ as $\|y\| \to \infty$ with $y_1 = 0$, because $\log \mathsf{Z}_i(y) \geq \max_j(\eta_{ij} + y_j)$ gives $\log \psi_\eta(y) \leq \sum_j (y_j - \max_{j'} y_{j'}) - n \min \eta \leq \min_j y_j - \max_j y_j - n \min \eta$. So the maximizer y^* exists and is unique, and ψ_η is integrable on $\{y_1 = 0\}$. Its gradient condition is $\partial_j \log \psi_\eta = 1 - \sum_i S_\eta(y)_{ij} = 0$ for $j \geq 2$, and $\sum_j \partial_j \log \psi_\eta = n - \sum_{ij} S_\eta(y)_{ij} = 0$ identically, so every column sum of $S_\eta(y^*)$ is one: $S_\eta(y^*) = \mathrm{diag}(\mathsf{Z}_i(y^*)^{-1}) \exp(\eta)\, \mathrm{diag}(e^{y_j^*})$ is doubly stochastic, hence the Sinkhorn scaling of $\exp(\eta)$, which is unique (Sinkhorn, 1964).

The Laplace approximation. Let H be the Hessian of $-\log \psi_\theta$ at y^* in the free coordinates $y_2, \ldots, y_n$. Laplace's method at y^* gives

$$Q_\theta^{\mathrm{L}}(\sigma) \propto e^{\langle \theta, P_\sigma \rangle}\, \psi_\theta(y^*)\, \det\big(I - P_\sigma^\top S_\theta(y^*) + \tfrac{1}{n}\mathbf{1}\mathbf{1}^\top\big)\, (\det \mathsf{H})^{-1/2},$$

and the factors $\psi_\theta(y^*)$ and $\det \mathsf{H}$ do not depend on σ and cancel upon normalization over σ. Take θ with zero row and column sums, without loss of generality: a row shift of the scores multiplies the prefactor of every σ by the same constant and leaves y^* and $S_\theta(y^*)$ unchanged, and a column shift translates y^*. The maximizer solves the column-sum equations $\sum_i S_\theta(y)_{ij} = 1$, $j \geq 2$, whose Jacobian in $(y_2, \ldots, y_n)$ at $\theta = 0$, $y = 0$ is the $(n-1) \times (n-1)$ matrix $I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$, invertible, and whose derivative in a score direction with zero row and column sums vanishes there. By the implicit function theorem $y^*(\theta) = O(\|\theta\|_F^2)$, so $S_\theta(y^*) = S_\theta(0) + O(\|\theta\|_F^2) = \frac{1}{n}(\mathbf{1}\mathbf{1}^\top + \theta) + O(\|\theta\|_F^2)$, the row sums of θ vanishing. With $P_\sigma^\top \mathbf{1}\mathbf{1}^\top = \mathbf{1}\mathbf{1}^\top$ the determinant is $\det(I - \frac{1}{n}P_\sigma^\top \theta) + O(\|\theta\|_F^2) = 1 - \frac{1}{n}\langle \theta, P_\sigma \rangle + O(\|\theta\|_F^2)$. Hence $Q_\theta^{\mathrm{L}}(\sigma) \propto 1 + (1 - \frac{1}{n})\langle \theta, P_\sigma \rangle + O(\|\theta\|_F^2)$, and since $\sum_\sigma \langle \theta, P_\sigma \rangle = 0$, $n!\, Q_\theta^{\mathrm{L}}(\sigma) = 1 + \frac{n-1}{n}\langle \theta, P_\sigma \rangle + O(\|\theta\|_F^2)$, the first-order term of $p_{\frac{n-1}{n}\theta}$. The law itself has $n!\, Q_\theta(\sigma) = 1 + \frac{n}{n+1}\langle \theta, P_\sigma \rangle + O(\|\theta\|_F^2)$, by Corollary C.1. □

The expected dual objective. For the square problem $\max\langle \eta + \xi, W \rangle$ over doubly stochastic W, the dual is $\min \sum_i u_i - \sum_j y_j$ subject to $u_i - y_j \geq \eta_{ij} + \xi_{ij}$. At fixed y the optimal u_i is $\max_j(\eta_{ij} + y_j + \xi_{ij})$, whose expectation is $\log \mathsf{Z}_i(y) + \mathsf{c}_{\mathrm{EM}}$, because the maximum of $\mathsf{a}_j + \xi_j$ over independent standard Gumbel ξ_j is Gumbel with location $\log \sum_j e^{\mathsf{a}_j}$, and its mean is that location plus the Euler–Mascheroni constant c_{EM}. So the expected dual objective at column potentials y is $\sum_i \log \mathsf{Z}_i(y) - \sum_j y_j + n\mathsf{c}_{\mathrm{EM}} = -\log \psi_\eta(y) + n\mathsf{c}_{\mathrm{EM}}$, and the Sinkhorn column potentials minimize it, as Rosén's threshold minimizes the expected dual objective of the perturbed k-subset problem (Appendix B.1).

B.6 GENERAL GRAPHS

Over the perfect matchings of a general graph, odd cycles of tight edges occur, and the nested blossom structures defined below take the place of the forests of Appendix B.2.

The perfect matchings of K_{2n}, $n \geq 2$, form a structure set on its edges once a perfect matching π is identified with the indicator of its edges, so that $x^* = \pi$ means $x^* = x_\pi$. Fix π, a vertex i_0 with

partner j_0 in π, and $\pi' = \pi \setminus \{i_0 j_0\}$. A perfect matching pairs i_0 with some vertex u and matches the other vertices among themselves, so for $u \neq i_0$ let val_u be the optimal value of the perturbed problem on the vertices other than i_0 and u. Let $\mathcal{H}$ be the union of the optimal perfect matchings of these $2n-1$ problems. The potentials of the vertices $u \neq i_0$ are

$$\tilde{Y}_u = \mathrm{val}_u - \min_{vw \in \mathcal{H}} \tfrac{1}{2}\big(\eta_{vw} + \xi_{vw} + \mathrm{val}_v + \mathrm{val}_w\big),$$

and $\tilde{y} \in \mathbb{R}^{2n-1}$ denotes a value of $\tilde{Y}$. Their negatives are canonical vertex duals of Gabow (2012) for these problems, and the minimum fixes the additive constant that such duals leave free (Lemma B.2). On the event that π' is the optimal perfect matching of the vertices other than i_0 and j_0, $\tilde{Y}$ has a continuous density g_π with respect to Lebesgue measure, whose integral is the probability of that event.

The density g_π is a sum over the structures that Edmonds' algorithm builds (Edmonds, 1965), and the vertex sets it contracts are the blossoms. A family of sets is laminar if any two of its sets are disjoint or nested. The children of a set in a laminar family are the maximal sets of the family properly contained in it, together with the singletons of its elements that lie in none of them. A nested blossom structure is a laminar family of odd sets of at least three vertices other than i_0, among them the set of all these vertices, together with a cycle through the children of each of its sets, formalized below. When the family consists of the set of all these vertices alone, its cycle is a Hamiltonian cycle of the vertices other than i_0, and Γ denotes such a cycle that contains π' and alternates between edges of π' and other edges, except at j_0, where two other edges meet. FP is the class of functions computable in polynomial time, and #P the class of functions that count the accepting computations of nondeterministic polynomial-time Turing machines (Valiant, 1979a).

Proposition B.3 (perfect matchings of the complete graph)**.** *For the perfect matchings of K_{2n}, $n \geq 2$, with standard Gumbel noise on the edges and scores η,*

$$Q_\eta(\pi) = \int_{\mathbb{R}^{2n-1}} \frac{\exp(\eta_{i_0 j_0} + \tilde{y}_{j_0})}{\sum_{v \neq i_0} \exp(\eta_{i_0 v} + \tilde{y}_v)}\, g_\pi(\tilde{y})\, d\tilde{y}.$$

The integrand is the joint density of $(x^*(\eta), \tilde{Y})$ at $(\pi, \tilde{y})$, and g_π is a finite sum, over the nested blossom structures, of explicit integrals. If every rate $\exp(\eta_e)$ is multiplied by $\varsigma > 0$, then, for every $\tilde{y}$, as $\varsigma \to 0$,

$$g_\pi(\tilde{y}) = 2\,\varsigma^{2n-1} e^{2\sum_{v \neq i_0} \tilde{y}_v} \sum_{\Gamma} \prod_{e \in \Gamma} \exp(\eta_e) + O(\varsigma^{2n}).$$

Consequently, unless $\mathrm{FP} = \#\mathrm{P}$, no algorithm computes $g_\pi(\tilde{y})$ to a prescribed absolute accuracy from the rates $\exp(\eta_e)$ and the values $e^{\tilde{y}_v}$, given as rationals, in time polynomial in the input length and the number of bits of accuracy.

The proof of Proposition B.3 follows Appendix B.2, with vertex potentials and multipliers on odd sets of vertices in place of the potentials of the assignment problem. Given the clocks of the edges not at i_0, the edges at i_0 contribute the fraction of the first display, which leaves the density g_π of the potentials to compute. Almost surely the potentials $\tilde{Y}$ extend uniquely to a dual solution with positive multipliers on nested odd sets of vertices (Lemma B.2). The edges whose clocks are at their bounds form one odd cycle through the children of each of these sets and of the set of all vertices other than i_0, and this nested blossom structure replaces the forest of tight non-assigned entries of Appendix B.2 as the discrete coordinate of the change of variables. Each blossom contributes a factor 2 to the Jacobian (Lemma B.3), the slacks integrate out as before, and at small rates the structures with a single blossom, the cycles Γ, dominate, so that the leading term of g_π is a weighted count of Hamiltonian cycles.

Setting. V is the vertex set of K_{2n} and $V' = V \setminus \{i_0\}$. For $B \subseteq V'$, $E(B)$ is the set of edges with both endpoints in B. The clocks $s_e = \exp(-(\eta_e + \xi_e))$ are independent exponential variables with rates $\exp(\eta_e)$. The cost of an edge is $\mathsf{c}_e = \log s_e$, the cost of a matching is the sum of the costs of its edges, and x^* is the perfect matching of least cost. For $u \in V'$, M_u is the perfect matching of $V' \setminus \{u\}$ of least cost, and Y_u is its cost, so that $\mathrm{val}_u = -\mathsf{Y}_u$ and $\mathcal{H} = \bigcup_{u \in V'} \mathsf{M}_u$. Let $\mathsf{C} = \max_{vw \in \mathcal{H}} \frac{1}{2}(\mathsf{c}_{vw} + \mathsf{Y}_v + \mathsf{Y}_w)$. In this notation the potentials are $\tilde{Y}_u = \mathsf{C} - \mathsf{Y}_u$. The event $\mathcal{E}$ is $\{\mathsf{M}_{j_0} = \pi'\}$, and g_π is the density of $\tilde{Y}$ on $\mathcal{E}$: the probability that $\mathcal{E}$ holds and $\tilde{Y}$ lies in a Borel set is the integral of g_π over that set. Both $\mathcal{E}$ and $\tilde{Y}$ are functions of the clocks of $E(V')$.

The edges at i_0. The fraction in Proposition B.3 is the conditional probability that x^* matches i_0 to j_0. A perfect matching of V consists of an edge i_0u and a perfect matching of $V' \setminus \{u\}$, so $x^* = \pi$ exactly when $\mathcal{E}$ holds and $u = j_0$ minimizes $\mathsf{c}_{i_0u} + \mathsf{Y}_u$ over $u \in V'$. Given the clocks of $E(V')$, the variables $s_{i_0u}e^{\mathsf{Y}_u}$ are independent exponential variables with rates $\exp(\eta_{i_0u})e^{-\mathsf{Y}_u} = \exp(\eta_{i_0u})e^{\tilde{Y}_u - \mathsf{C}}$, and the one at j_0 is the smallest with probability proportional to its rate:

$$\mathbb{P}\big(x^* = \pi \;\big|\; (s_e)_{e\in E(V')}\big) = \mathbf{1}\{\mathcal{E}\}\,\frac{\exp(\eta_{i_0j_0} + \tilde{Y}_{j_0})}{\sum_{v\in V'}\exp(\eta_{i_0v} + \tilde{Y}_v)}.$$

The first display of Proposition B.3 and its joint-density clause therefore hold once $\tilde{Y}$ is shown to have a density g_π on $\mathcal{E}$, which the rest of this section computes.

Nested blossom structures. A nested blossom structure S is a laminar family $\mathcal{L}_S$ of odd subsets of V' of size at least three with $V' \in \mathcal{L}_S$, its blossoms, together with a set of edges Γ_B for every $B \in \mathcal{L}_S$. The set $\Gamma_B \subseteq E(B)$ consists of edges that join different children and form a cycle through all the children, each child taken as a single vertex. Let q_B be the number of children of B. A child has odd size and so does B, so q_B is odd, and $\mathsf{q}_B \geq 3$ because Γ_B is a cycle. Write $\mathsf{b}_S = |\mathcal{L}_S|$ and $E_S = \bigcup_B \Gamma_B$. Every singleton of a vertex and every blossom other than V' is a child of exactly one blossom, so $|E_S| = \sum_B \mathsf{q}_B = |V'| + \mathsf{b}_S - 1$. For $u \in V'$ the matching M_u^S is built recursively, starting from V'. The odd cycle $\Gamma_{V'}$ has exactly one matching that covers every child except the one containing u, and every other child A contains exactly one endpoint v_A of the edges of that matching. Inside every child that is a blossom the construction is repeated, with v_A in place of u inside A and with u itself inside the child containing u. So $\mathsf{M}_u^S \subseteq E_S$ is a perfect matching of $V' \setminus \{u\}$ with $(|B| - 1)/2$ edges inside every blossom B.

The dual variables for S are potentials $\tilde{y} \in \mathbb{R}^{V'}$ and multipliers z_B on the blossoms $B \neq V'$. For an edge $e = uw$, $\mathsf{k}(e) = \mathsf{k}(u, w) = \sum_{B\ni u,w,\ B\neq V'} z_B$ is the total multiplier of the blossoms other than V' that contain both u and w. The bound on the cost of an edge $uw \in E(V')$ is

$$\mathrm{bd}_{uw} = \tilde{y}_u + \tilde{y}_w - 2\mathsf{k}(u, w),$$

and $e^{\mathrm{bd}_{uw}}$ is the corresponding bound on the clock s_{uw}. A dual solution for S is a pair $(\tilde{y}, z)$ such that $z_B > 0$ for every blossom $B \neq V'$ and $\mathsf{c}_e \geq \mathrm{bd}_e$ for every $e \in E(V')$, with equality exactly on E_S. The clocks of $E(V')$ are generic if no two perfect matchings of the same $V' \setminus \{u\}$ have equal cost and, for every laminar family $\mathcal{L}$ of odd subsets of V' of size at least three with $V' \in \mathcal{L}$, no $|V'| + |\mathcal{L}|$ edges satisfy $\mathsf{c}_e = \mathrm{bd}_e$ for a common $(\tilde{y}, z) \in \mathbb{R}^{V'} \times \mathbb{R}^{\mathcal{L}\setminus\{V'\}}$. Each exception confines the costs to a proper linear subspace, and there are finitely many, so the clocks are generic almost surely.

Weak duality. The bounds on the edges add up to a bound on the cost of every matching. Let S be a nested blossom structure, let $\tilde{y} \in \mathbb{R}^{V'}$ and $z_B \geq 0$ for the blossoms $B \neq V'$, and suppose that $\mathsf{c}_e \geq \mathrm{bd}_e$ for every $e \in E(V')$. Let $u \in V'$ and let M be a perfect matching of $V' \setminus \{u\}$. In the sum of the bounds over the edges of M every vertex other than u appears once, an edge inside a blossom $B \neq V'$ carries $-2z_B$, and a matching has at most $(|B| - 1)/2$ edges inside the odd set B. So

$$\sum_{e\in\mathsf{M}} \mathsf{c}_e \;\geq\; \sum_{e\in\mathsf{M}} \mathrm{bd}_e = \sum_{v\neq u} \tilde{y}_v - 2\sum_{B\neq V'} z_B\,|\mathsf{M}\cap E(B)| \;\geq\; \sum_{v\neq u}\tilde{y}_v - \sum_{B\neq V'}(|B| - 1)z_B.$$

Equality holds throughout exactly when every edge of M attains its bound and M has $(|B| - 1)/2$ edges inside every B with $z_B > 0$. If $(\tilde{y}, z)$ is a dual solution for S, the matching M_u^S satisfies both conditions and attains the bound.

Lemma B.2 (the canonical dual solution)**.** *For generic clocks exactly one nested blossom structure S has a dual solution, and its dual solution is unique. Its potentials are $\tilde{Y}$, $E_S = \mathcal{H}$, and $\mathsf{M}_u = \mathsf{M}_u^S$ for every $u \in V'$. In particular $\mathcal{E} = \{\mathsf{M}_{j_0}^S = \pi'\}$.*

Proof. Existence. The complete graph on V' has a perfect matching after the deletion of any vertex, so the construction of Gabow (2012, Section 4.2 and Theorem 4.2) applies to it, with the edge weights $-\mathsf{c}$. It yields a laminar family $\mathcal{L}$ of odd subsets of V' of size at least three with $V' \in \mathcal{L}$, in which every set carries a cycle of edges through its children; vertex duals, which are the costs Y_u; and a multiplier for every set of $\mathcal{L}$, nonnegative except possibly for V'. For every edge uw, $-\mathsf{c}_{uw}$ is

at most $\mathsf{Y}_u + \mathsf{Y}_w$ plus the multipliers of the sets that contain u and w, with equality on the edges of the cycles. Let $-2\mathsf{C}'$ be the multiplier of V' and $2z_B$ that of $B \neq V'$, and put $\tilde{y}_u = \mathsf{C}' - \mathsf{Y}_u$. The inequality then reads $\mathsf{c}_e \geq \mathrm{bd}_e$, with equality on the cycles. The cycles have $|V'| + |\mathcal{L}| - 1$ edges, so genericity excludes any further equality. It also gives $z_B > 0$ for $B \neq V'$, since with $z_B = 0$ the same $|V'| + |\mathcal{L}| - 1$ equalities would hold for the family $\mathcal{L} \setminus \{B\}$. So $(\tilde{y}, z)$ is a dual solution for the structure formed by $\mathcal{L}$ and the cycles.

Identification. Every dual solution has the potentials $\tilde{Y}$ and the edge set $\mathcal{H}$. Let $(\tilde{y}, z)$ be a dual solution for a structure S, and put $\mathsf{C}_S = \sum_v \tilde{y}_v - \sum_{B \neq V'} (|B| - 1) z_B$. By weak duality, M_u^S is a perfect matching of $V' \setminus \{u\}$ of least cost, so $\mathsf{M}_u^S = \mathsf{M}_u$ by genericity and $\mathsf{Y}_u = \mathsf{C}_S - \tilde{y}_u$. For a child A of a blossom B and $u \in \mathsf{A}$, M_u^S contains the matching of Γ_B that covers every child but A, and every edge of an odd cycle lies in one of these matchings. So $E_S \subseteq \mathcal{H}$, and $\mathcal{H} \subseteq E_S$ because every $\mathsf{M}_u = \mathsf{M}_u^S$ lies in E_S. On an edge $uw \in \mathcal{H}$ the equality $\mathsf{c}_{uw} = \mathrm{bd}_{uw}$ reads

$$\tfrac{1}{2}(\mathsf{c}_{uw} + \mathsf{Y}_u + \mathsf{Y}_w) = \mathsf{C}_S - \mathsf{k}(u, w),$$

which is at most C_S, with equality on the edges of $\Gamma_{V'}$, which lie in no blossom other than V'. So $\mathsf{C}_S = \mathsf{C}$ and $\tilde{y} = \tilde{Y}$. In particular $-\tilde{Y} = \mathsf{Y} - \mathsf{C}$ are Gabow's vertex duals translated so that V' carries no multiplier.

Uniqueness. The blossoms are recovered by thresholding k on $\mathcal{H}$. Let $(\tilde{y}, z)$ be a dual solution for a structure S. By the identification, $\tilde{y} = \tilde{Y}$, $E_S = \mathcal{H}$, and $\mathsf{k}(u, w) = \mathsf{C} - \frac{1}{2}(\mathsf{c}_{uw} + \mathsf{Y}_u + \mathsf{Y}_w)$ on every edge $uw \in \mathcal{H}$, a quantity fixed by the clocks. For $B \in \mathcal{L}_S$ let $\tilde{z}_B = \sum_{B' \supseteq B,\ B' \neq V'} z_{B'}$, the accumulated multiplier, so that $\tilde{z}_{V'} = 0$. An edge of $\mathcal{H}$ has $\mathsf{k} = \tilde{z}_{B'}$ for the smallest blossom B' containing it. So the edges of $\mathcal{H}$ inside a blossom $B \neq V'$ have $\mathsf{k} \geq \tilde{z}_B$ and connect B, being the cycles of B and of the blossoms inside it, while an edge of $\mathcal{H}$ leaving B has $\mathsf{k} \leq \tilde{z}_{B'}$ for the smallest blossom B' properly containing B. Hence, for $\mathsf{t} > 0$, the components with at least two vertices of the graph on V' with the edges $\{e \in \mathcal{H} : \mathsf{k}(e) \geq \mathsf{t}\}$ are the maximal blossoms $B \neq V'$ with $\tilde{z}_B \geq \mathsf{t}$. As t ranges over $(0, \infty)$ these components determine $\mathcal{L}_S$ and $\tilde{z}$, hence z, and the cycles are the edges of $\mathcal{H}$ inside each blossom that join different children. So any two dual solutions coincide, and so do their structures. □

Lemma B.3 (the Jacobian)**.** *For a nested blossom structure S, the linear map $(\tilde{y}, z) \mapsto (\mathrm{bd}_e)_{e \in E_S}$ from $\mathbb{R}^{V'} \times \mathbb{R}^{\mathcal{L}_S \setminus \{V'\}}$ to $\mathbb{R}^{E_S}$ has determinant $\pm 2^{\mathsf{b}_S}$.*

Proof. The matrix is block triangular in suitable coordinates. Let $\mathsf{y}_v = \tilde{y}_v - \sum_{B \ni v,\ B \neq V'} z_B$. The change from $(\tilde{y}, z)$ to (y, z) has determinant one, and in the new coordinates $\mathrm{bd}_{uw} = \mathsf{y}_u + \mathsf{y}_w + \sum_B z_B$, the sum over the blossoms $B \neq V'$ that contain exactly one of u and w. Give each child A of a blossom one coordinate, y_v if $\mathsf{A} = \{v\}$ and z_{A} if A is a blossom. Every coordinate belongs to exactly one child. Let $uw \in \Gamma_B$ join the children $\mathsf{A} \ni u$ and $\mathsf{A}' \ni w$ of B. A blossom that contains u but not w lies inside A, by laminarity and the maximality of A, and likewise with u and w exchanged. So bd_{uw} is the sum of the coordinates of A and A' and of coordinates of children of blossoms contained in A or A'. List the blossoms so that each precedes the blossoms inside it, and order the equations and the coordinates accordingly, the equations of Γ_B and the coordinates of the children of B forming one block for each B. The matrix is then block triangular, and its diagonal block for B is the $0/1$ incidence matrix of the odd cycle Γ_B on the children of B, whose determinant is ± 2. □

Proof of Proposition B.3. The law. A change of variables to the potentials, the multipliers and the slacks writes g_π as a sum over structures. For a structure S, map $\tilde{y} \in \mathbb{R}^{V'}$, $z \in (0, \infty)^{\mathcal{L}_S \setminus \{V'\}}$ and slacks $\mathsf{s} \in (0, \infty)^{E(V') \setminus E_S}$, the excesses of the clocks over their bounds, to the clocks

$$s_e = e^{\mathrm{bd}_e} \quad (e \in E_S), \qquad s_e = e^{\mathrm{bd}_e} + \mathsf{s}_e \quad (e \in E(V') \setminus E_S).$$

The map is injective, because by Lemma B.3 the clocks of E_S determine $(\tilde{y}, z)$. Its image consists of the clocks with a dual solution for S. So by Lemma B.2, up to a null set, the images of the different structures are disjoint and cover all clocks, and the image of S lies in $\mathcal{E}$ if $\mathsf{M}_{j_0}^S = \pi'$ and is disjoint from $\mathcal{E}$ otherwise. Write $\mathsf{h}_e = \exp(\eta_e) e^{\mathrm{bd}_e}$. By Lemma B.3 the Jacobian of the map is $2^{\mathsf{b}_S} \prod_{e \in E_S} e^{\mathrm{bd}_e}$, so the density $\prod_e \exp(\eta_e) e^{-\exp(\eta_e) s_e}$ of the clocks becomes $2^{\mathsf{b}_S} \prod_{e \in E_S} \mathsf{h}_e e^{-\mathsf{h}_e}$

times $\prod_{e\notin E_S}\exp(\eta_e)e^{-\mathsf{h}_e}e^{-\exp(\eta_e)\mathsf{s}_e}$. Each slack integrates to the survival factor $e^{-\mathsf{h}_e}$ of its clock at the bound, and therefore

$$g_\pi(\tilde y)=\sum_{S:\,\mathsf{M}^S_{j_0}=\pi'}2^{\mathsf{b}_S}\int_{(0,\infty)^{\mathcal{L}_S\setminus\{V'\}}}\prod_{e\in E_S}\mathsf{h}_e\prod_{e\in E(V')}e^{-\mathsf{h}_e}\,dz,$$

with $\mathsf{h}_{uw}=\exp(\eta_{uw})e^{\tilde y_u+\tilde y_w-2\mathsf{k}(u,w)}$, a finite sum of explicit integrals. Each is continuous in $\tilde y$ by dominated convergence, since for $\tilde y$ in a compact set its integrand is at most a constant times $\prod_{e\in E_S}e^{-2\mathsf{k}(e)}$, which is integrable over $z>0$. This version of the density is the g_π of the proposition, and with the fraction contributed by the edges at i_0 it proves the first display and the joint-density clause. For K_4, where V' has three vertices, the only structure is the triangle, so that, with $\mathsf{h}_{uw}=\exp(\eta_{uw})e^{\tilde y_u+\tilde y_w}$,

$$Q_\eta(\pi)=2\int_{\mathbb{R}^3}\frac{\exp(\eta_{i_0j_0}+\tilde y_{j_0})}{\sum_{v\in V'}\exp(\eta_{i_0v}+\tilde y_v)}\prod_{uw\in E(V')}\mathsf{h}_{uw}e^{-\mathsf{h}_{uw}}\,d\tilde y.$$

Leading order. At small rates the structures with one blossom dominate. Multiply every rate by $\varsigma\in(0,1]$, and write $\mathsf{r}_{uw}=\exp(\eta_{uw})e^{\tilde y_u+\tilde y_w}$, with η the scores before the scaling, so that $\mathsf{h}_e=\varsigma\mathsf{r}_e e^{-2\mathsf{k}(e)}$. For a blossom $B\neq V'$ let n_B be the number of edges of E_S inside B, so that $\mathsf{n}_B\geq|B|$ and $\sum_{e\in E_S}\mathsf{k}(e)=\sum_{B\neq V'}\mathsf{n}_B z_B$. Let

$$\mathsf{I}_S=\int_{(0,\infty)^{\mathcal{L}_S\setminus\{V'\}}}\prod_{B\neq V'}e^{-2\mathsf{n}_B z_B}\exp\Big(-\varsigma\sum_{e\in E(V')}\mathsf{r}_e e^{-2\mathsf{k}(e)}\Big)\,dz;$$

then $\mathsf{I}_S\leq\prod_{B\neq V'}(2\mathsf{n}_B)^{-1}$, and, since E_S has $2n-2+\mathsf{b}_S$ edges, the term of S in g_π is $2^{\mathsf{b}_S}\varsigma^{2n-2+\mathsf{b}_S}\prod_{e\in E_S}\mathsf{r}_e\,\mathsf{I}_S$. Let $\mathsf{W}=\sum_{e\in E(V')}\mathsf{r}_e$. A structure with $\mathsf{b}_S=1$ is a Hamiltonian cycle of V', and the condition $\mathsf{M}^S_{j_0}=\pi'$ makes it one of the cycles Γ. For such a structure, $\mathsf{I}_S=e^{-\varsigma\mathsf{W}}$, and $\prod_{e\in\Gamma}\mathsf{r}_e=e^{2\sum_v\tilde y_v}\prod_{e\in\Gamma}\exp(\eta_e)$ because every vertex of V' has degree two in Γ. With $1-e^{-\varsigma\mathsf{W}}\leq\varsigma\mathsf{W}$, and $2^{\mathsf{b}_S}\varsigma^{2n-2+\mathsf{b}_S}\mathsf{I}_S\leq2\varsigma^{2n}$ when $\mathsf{b}_S\geq2$,

$$\Big|g_\pi(\tilde y)-2\varsigma^{2n-1}e^{2\sum_v\tilde y_v}\sum_\Gamma\prod_{e\in\Gamma}\exp(\eta_e)\Big|\;\leq\;2\varsigma^{2n}\Big(\mathsf{W}\sum_\Gamma\prod_{e\in\Gamma}\mathsf{r}_e+\sum_{S:\,\mathsf{b}_S\geq2}\prod_{e\in E_S}\mathsf{r}_e\Big),$$

which proves the second display. Let $\mathsf{N}_0=2^{|V'|^2}$. A structure is determined by $E_S\subseteq E(V')$ and by its blossoms, at most $(|V'|-1)/2$ subsets of V' because $|V'|+\mathsf{b}_S-1=\sum_B\mathsf{q}_B\geq3\mathsf{b}_S$. So there are at most N_0 structures, and if every $\mathsf{r}_e\leq1$ the right side is at most $2\mathsf{N}_0(|E(V')|+1)\varsigma^{2n}$.

Hardness. One evaluation of g_π counts the Hamiltonian cycles of a graph. Let G_0 be a graph on the vertices $0,1,\dots,n-1$ with $n\geq3$, and let p be the number of its Hamiltonian cycles, counted as edge sets. Counting the Hamiltonian cycles of a graph is #P-complete (Valiant, 1979b). Let V' consist of j_0 and $n-1$ pairs $\{v_i,w_i\}$, and let $\pi'=\{v_iw_i:1\leq i\leq n-1\}$. The vertex j_0 stands for the vertex 0 of G_0, and v_i and w_i stand for the vertex i. Let $\mathsf{d}=1/(16\mathsf{N}_0)$ and $\varsigma=1/(16\mathsf{N}_0(|E(V')|+1))$. Take $\tilde y=0$, set $\eta_e=0$ on the edges v_iw_i and on every edge whose endpoints stand for adjacent vertices of G_0, set $\exp(\eta_e)=\mathsf{d}$ on every other edge of $E(V')$, and multiply every rate by ς. The edges at i_0 do not enter g_π. A cycle Γ leaves j_0, passes through the pairs one at a time, entering each pair at one vertex and leaving it at the other, and returns to j_0. If it avoids the edges with $\exp(\eta_e)=\mathsf{d}$, the pairs follow a Hamiltonian cycle of G_0. Each Hamiltonian cycle of G_0 gives 2^{n-1} such Γ: fix one direction of traversal and choose the entry vertex of every pair. The cycles Γ that use an edge with $\exp(\eta_e)=\mathsf{d}$ are at most N_0 in number and contribute at most d each, so $\sum_\Gamma\prod_{e\in\Gamma}\exp(\eta_e)$ lies between $2^{n-1}\mathsf{p}$ and $2^{n-1}\mathsf{p}+\frac{1}{16}$, and the leading-order bound, with $\mathsf{r}_e\in\{1,\mathsf{d}\}$, gives

$$\Big|\frac{g_\pi(0)}{\varsigma^{2n-1}}-2^n\mathsf{p}\Big|\;\leq\;\frac18+2\mathsf{N}_0\big(|E(V')|+1\big)\varsigma\;=\;\frac14.$$

So an evaluation of $g_\pi(0)$ to within $\varsigma^{2n-1}/8$, divided by ς^{2n-1} and rounded to the nearest integer, gives $2^n\mathsf{p}$. The rates are rationals of $O(n^2)$ bits, and the accuracy $\varsigma^{2n-1}/8$ needs only $O(n^3)$ bits. An algorithm as in the proposition would therefore count Hamiltonian cycles in polynomial time, which is impossible unless FP $=$ #P. □

The factor 2 of Lemma B.3 and the forests of Appendix B.2 both come from the $0/1$ incidence matrix of a cycle. The edges of an even cycle satisfy $\mathsf{c}_{uw} = \tilde{y}_u + \tilde{y}_w$ for a common $\tilde{y}$ with probability zero, because the incidence matrix of an even cycle is singular, while that of an odd cycle has determinant ± 2. Over assignments every cycle in the bipartite graph of rows and columns is even, so the tight entries form a forest, and contracting the assigned entries turns it into the forest of tight non-assigned entries of Appendix B.2.

C Calibration

The proofs here expand the law at zero scores and use the symmetry of the structure set. In Appendix C.1, linear algebra and one counting argument characterize the linear maps that respect the symmetry. Appendices C.2 and C.3 evaluate a single derivative twice, once through the order statistics of the noise and once through its ranks. Appendix C.4 computes the second-order coefficients of an independent identically distributed noise law from two conditional probabilities and normalization, and evaluates them for the three families. Appendix C.5 separates the part of the second-order discrepancy that a calibration can remove from the part it cannot, and Appendix C.6 shows how randomization removes the rest. Appendix C.7 expands the law of square Gumbel matching at zero scores, where Appendix B.2 makes the conditional law of the clocks explicit, and Appendix C.8 calibrates that expansion.

C.1 First-Order Calibration

A linear calibration is $\eta(\theta) = L\theta$ for a linear map L of the scores. The stabilizer of a structure in a group of permutations of E is the subgroup of those permutations that leave the structure unchanged. The hypothesis of Proposition C.1 on the group is that the stabilizer of some structure x acts transitively on $\{e : x_e = 1\}$ and on $\{e : x_e = 0\}$.

Proposition C.1 (first-order calibration under symmetry)**.** *Let ν be a noise law for which $\theta \mapsto Q^{\nu}_{\theta}$ is differentiable at $\theta = 0$ and $Q^{\nu}_{0} = p_0$. Suppose that $\mathcal{X}$ has at least two structures, that a group of permutations of E maps $\mathcal{X}$ onto itself, leaves ν invariant and acts transitively on $\mathcal{X}$, and that the permutations in the group fixing some structure act transitively on its elements and on the other elements. Then $J = c\, J_p$ for a scalar $c \geq 0$, and if $c > 0$ the first-order temperature exists and $c_1 = c$. The permutations of $[n]$ are such a group for the k-subsets of $[n]$ with $1 \leq k < n$, and the row and column permutations for square assignments with $n \geq 2$, under every noise law they leave invariant.*

Lemma C.1 (linear calibrations)**.** *Let ν be a noise law for which $\theta \mapsto Q^{\nu}_{\theta}$ is differentiable at $\theta = 0$ and $Q^{\nu}_{0} = p_0$. A first-order exact linear calibration exists if and only if $\operatorname{im} J \supseteq \operatorname{im} J_p$, and any two of them differ by a linear map into the constant-score directions.*

Proof of Lemma C.1 and of Proposition C.1. The lemma. The maps $\theta \mapsto Q^{\nu}_{L\theta}$ and $\theta \mapsto p_\theta$ are differentiable at 0 and agree there, so $Q^{\nu}_{L\theta} - p_\theta = (JL - J_p)\theta + o(\|\theta\|)$, and the linear calibration L is first-order exact if and only if $JL = J_p$. Such an L exists if and only if $\operatorname{im} J \supseteq \operatorname{im} J_p$. The kernel of J_p is the space of constant-score directions, because $J_p\vartheta(x) = p_0(x)\big(\langle \vartheta, x\rangle - \mathbb{E}_{p_0}\langle \vartheta, x\rangle\big)$ vanishes for every x exactly when $\langle \vartheta, \cdot\rangle$ is constant on $\mathcal{X}$. The kernel of J contains that space, because adding a constant-score direction to the scores leaves the law unchanged. Under the image condition $\dim\ker J = |E| - \dim\operatorname{im} J \leq |E| - \dim\operatorname{im} J_p = \dim\ker J_p$, so the two kernels are equal. If two linear calibrations L and L' are first-order exact then $J(L - L') = 0$, so $L - L'$ maps into $\ker J$.

The proposition. Write G for the group and, for $\mathsf{g} \in \mathsf{G}$, $(\mathsf{g}\vartheta)_e = \vartheta_{\mathsf{g}^{-1}(e)}$ and $(\mathsf{g}x)_e = x_{\mathsf{g}^{-1}(e)}$, so that $\langle \mathsf{g}\vartheta, \mathsf{g}x\rangle = \langle \vartheta, x\rangle$.

Equivariance. For $\mathsf{g} \in \mathsf{G}$ the noise $\mathsf{g}\xi$ has law ν and g permutes $\mathcal{X}$, so $Q^{\nu}_{\mathsf{g}\theta}(\mathsf{g}x) = Q^{\nu}_{\theta}(x)$ at every θ where ties have probability zero. For a fixed realization of the noise the score vectors at which two structures tie form finitely many hyperplanes, a Lebesgue null set, so by Fubini ties have probability zero for almost every θ. Two functions that agree almost everywhere and are differentiable at 0 have the same derivative there, whence $(J\mathsf{g}\vartheta)(\mathsf{g}x) = (J\vartheta)(x)$ for all ϑ and x. The same identity holds for J_p, where it needs no exceptional set.

One degree of freedom. Let x be a structure whose stabilizer H is transitive on $\mathsf{S} = \{e : x_e = 1\}$ and on $E \setminus \mathsf{S}$, and let $\mathsf{T} : \mathbb{R}^E \to \mathbb{R}^{\mathcal{X}}$ be linear, equivariant in the sense above, and zero at $\mathbf{1}$. By the transitivity of G on $\mathcal{X}$, T is determined by the linear functional $\vartheta \mapsto (\mathsf{T}\vartheta)(x)$, since $(\mathsf{T}\vartheta)(\mathsf{g}x) = (\mathsf{T}\mathsf{g}^{-1}\vartheta)(x)$. That functional is H-invariant, so its coefficients are constant on the orbits of H and

$$(\mathsf{T}\vartheta)(x) = \mathsf{a}\sum_{e\in\mathsf{S}} \vartheta_e + \mathsf{b}\sum_{e\notin\mathsf{S}} \vartheta_e, \qquad \mathsf{a}|\mathsf{S}| + \mathsf{b}\,(|E| - |\mathsf{S}|) = 0,$$

the second equation being $\mathsf{T}\mathbf{1} = 0$. So these maps form a space of dimension at most one. All structures have $|\mathsf{S}|$ elements, G acting transitively on $\mathcal{X}$ by permutations of E, so $\mathbf{1}$ is a constant-score direction and both J and J_p are maps of this kind. The map J_p is not zero, because $\langle x, x'\rangle < \langle x, x\rangle$ for every structure $x' \neq x$ of the same size, so that $\langle x, \cdot\rangle$ is not constant on $\mathcal{X}$ and $J_p x \neq 0$. Hence $J = c\,J_p$ for a scalar c.

The sign. For every realization of the noise and every $\mathsf{t} > 0$, the optimality of $x^*(\mathsf{t}\vartheta)$ at the scores $\mathsf{t}\vartheta + \xi$ and that of $x^*(0)$ at ξ add up to $\mathsf{t}\langle\vartheta, x^*(\mathsf{t}\vartheta) - x^*(0)\rangle \geq 0$. Taking expectations, dividing by t^2 and letting $\mathsf{t} \to 0$ gives $\sum_x \langle\vartheta, x\rangle (J\vartheta)(x) \geq 0$, while the same sum for J_p is $\mathrm{Var}_{p_0}\langle\vartheta, x\rangle$, positive for every ϑ outside the constant-score directions. Hence $c \geq 0$.

The temperature. If $c > 0$ then $L = I/c$ solves $JL = J_p$, so the temperature c is first-order exact, and any first-order exact temperature c' has $\frac{c}{c'} J_p = J_p$, hence equals c. By Definition 2 it is the first-order temperature.

The two instances. For the k-subsets of $[n]$ with $1 \leq k < n$, the permutations of $[n]$ preserve $\mathcal{X}$ and are transitive on it, and the stabilizer of x_B consists of the permutations preserving B, which are transitive on B and on its complement. For square assignments a pair of permutations of $[n]$ acts on the entries by permuting the rows by the first and the columns by the second, preserving the set of assignment matrices and acting transitively on it. The stabilizer of P_{id} consists of the pairs that permute rows and columns alike. These are transitive on the diagonal entries, and on the entries off the diagonal because the permutations of $[n]$ are transitive on ordered pairs of distinct indices. □

C.2 First-Order Temperatures

Proof of Theorem 3. The proof differentiates the integral of Proposition B.1(b) in one coordinate outside B and compares the derivative with that of p_θ. Let B be a k-subset and let e_i denote the ith standard basis vector of $\mathbb{R}^n$. Since $C(e, B) \setminus \{e\} = B$ for every $e \notin B$, Proposition B.1(b) reads

$$Q^\nu_\theta(x_B) = \int_{\mathbb{R}^B} \prod_{b\in B} f(q_b - \theta_b) \prod_{j\notin B} F\Big(\min_{b\in B} q_b - \theta_j\Big)\, dq.$$

Splitting the integral according to which coordinate attains the minimum and substituting $q \mapsto q + \theta_b$ in the term of b gives

$$Q^\nu_\theta(x_B) = \sum_{b\in B} \int f(q) \prod_{b'\in B\setminus\{b\}} \big(1 - F(q + \theta_b - \theta_{b'})\big) \prod_{j\notin B} F(q + \theta_b - \theta_j)\, dq,$$

in which θ enters only through the arguments of F.

Regularity. Increasing one of those arguments by t changes its factor by $\pm\mathsf{t}$ times the average of f over an interval of length t, so the difference quotient of the integral in t is, up to sign, the average over the shifts in $[0, \mathsf{t}]$ of the same integral with that factor replaced by f at the shifted argument. Each of these integrals is at most $\|f\|^2_{L^2}$ in absolute value, and as the shift tends to zero it converges to the integral with the unshifted factor f, by the continuity of translation in L^2. Hence the partial derivatives of Q^ν exist, are obtained by differentiating under the integral sign, and are continuous in θ, so Q^ν is continuously differentiable. The law at zero scores is p_0, because the noise is exchangeable and has a density.

The derivative. At $\theta = 0$ and for $j \notin B$ only the factors $F(q + \theta_b - \theta_j)$ depend on θ_j, so

$$(Je_j)(x_B) = -k \int f(q)^2 F(q)^{n-k-1} \big(1 - F(q)\big)^{k-1} dq,$$

the factor k counting the terms of the sum, while $(J_p e_j)(x_B) = -\frac{k}{n}\binom{n}{k}^{-1}$. By Proposition C.1 the first-order temperature is the ratio of the two, which is the stated formula and is positive.

The three families. Let $F^{-1}(\mathsf{u}) = \inf\{q : F(q) \geq \mathsf{u}\}$ for $0 < \mathsf{u} < 1$. Since $F(\xi_1)$ is uniform on $(0,1)$ and $F^{-1}(F(\xi_1)) = \xi_1$ almost surely, the substitution $\mathsf{u} = F(q)$ turns $\int f^2 F^{\mathsf{a}}(1-F)^{\mathsf{b}}\,dq$ into $\int_0^1 f\big(F^{-1}(\mathsf{u})\big)\mathsf{u}^{\mathsf{a}}(1-\mathsf{u})^{\mathsf{b}}\,d\mathsf{u}$, and $f(F^{-1}(\mathsf{u}))$ is $\mathsf{u}(-\log \mathsf{u})$ for Gumbel noise, $\mathsf{u}(1-\mathsf{u})$ for logistic noise and $1-\mathsf{u}$ for exponential noise. The three integrals are $\mathrm{B}(n-k+1,k)(H_n - H_{n-k})$, $\mathrm{B}(n-k+1,k+1)$ and $\mathrm{B}(n-k,k+1)$, with B the Beta function, the first because

$$\int_0^1 \mathsf{u}^{\mathsf{a}-1}(1-\mathsf{u})^{\mathsf{b}-1}(-\log \mathsf{u})\,d\mathsf{u} = \mathrm{B}(\mathsf{a},\mathsf{b})\big(\psi_\Gamma(\mathsf{a}+\mathsf{b}) - \psi_\Gamma(\mathsf{a})\big)$$

with ψ_Γ the digamma function and $\psi_\Gamma(n+1) - \psi_\Gamma(n-k+1) = H_n - H_{n-k}$. Multiplying by $n\binom{n}{k}$ gives $\frac{n}{k}(H_n - H_{n-k})$, $\frac{n}{n+1}$ and $\frac{n}{n-k}$. For square Gumbel matching, Corollary C.1 gives $c_1 = \frac{n}{n+1}$. □

Moment matching. The moment matching of Papandreou & Yuille (2011) moves the scores of the perturb-and-MAP model until its expected sufficient statistics, the marginals of the perturbed optimizer, match those of the data. When the data follow p_θ, moment matching returns the scores θ/c_1 to first order up to constant-score directions, because $J = c_1 J_p$ makes the marginals of the perturbed optimizer at θ/c_1 agree with those of p_θ to first order and the marginal map is invertible near zero up to those directions. Theorem 3 gives c_1 in closed form.

C.3 Rank Scores

For independent identically distributed noise entries whose density f is absolutely continuous with $\int |f'| < \infty$, the rank scores of f are

$$a(\ell) = \mathbb{E}\Big[-\frac{f'(\xi_{(\ell)})}{f(\xi_{(\ell)})}\Big], \qquad \ell = 1,\dots,n,$$

where $\xi_{(1)} \leq \cdots \leq \xi_{(n)}$ are the order statistics of n entries. The statistic of the locally most powerful rank test for a location shift in part of the sample is the sum of the rank scores over the ranks in that part (Hájek & Šidák, 1967). Rank scores appear because the perturbed optimizer over k-subsets depends on the perturbed scores only through their order.

Proposition C.2 (the first-order temperature as a rank-score difference)**.** *For the k-subsets of $[n]$ with $1 \leq k < n$ and independent identically distributed noise entries whose density f is absolutely continuous with $\int |f'| < \infty$,*

$$c_1 = \frac{1}{k}\sum_{\ell=n-k+1}^{n} a(\ell) \;-\; \frac{1}{n-k}\sum_{\ell=1}^{n-k} a(\ell).$$

The Wilcoxon rank scores $a(\ell) = \frac{2\ell}{n+1} - 1$ of logistic noise give $\frac{n}{n+1}$, and the Gumbel rank scores $a(\ell) = 1 - H_n + H_{\ell-1}$ give $\frac{n}{k}(H_n - H_{n-k})$.

Proof of Proposition C.2. The proof differentiates through the noise density and conditions on the ranks. A density that is absolutely continuous with $\int |f'| < \infty$ is bounded, hence square-integrable, so Theorem 3 applies. Translation is differentiable in L^1: $\frac{1}{\mathsf{t}}(f(\cdot - \mathsf{t}) - f) \to -f'$ in L^1 as $\mathsf{t} \to 0$. Write $\mathsf{s} = -f'/f$ on $\{f > 0\}$, off which $f' = 0$ almost everywhere. Consequently, for every bounded measurable function h of the noise vector and every index j,

$$\frac{d}{d\mathsf{t}}\,\mathbb{E}\big[\mathsf{h}(\xi + \mathsf{t}e_j)\big]\Big|_{\mathsf{t}=0} = \mathbb{E}\big[\mathsf{h}(\xi)\,\mathsf{s}(\xi_j)\big].$$

Since shifting ξ_j is the same as raising θ_j, applying this to $\mathsf{h} = \mathbf{1}\{x^*(0) = x_B\}$ and $j \notin B$ gives $(Je_j)(x_B) = \mathbb{E}[\mathbf{1}\{x^*(0) = x_B\}\,\mathsf{s}(\xi_j)]$. At zero scores the ranks of the noise entries are uniformly distributed and independent of the order statistics. The event $\{x^*(0) = x_B\}$ says that the entries indexed by B hold the k largest ranks, and on that event the rank of ξ_j is uniform on $\{1,\dots,n-k\}$. Hence

$$(Je_j)(x_B) = \binom{n}{k}^{-1}\frac{1}{n-k}\sum_{\ell=1}^{n-k} a(\ell),$$

and dividing by $(J_p e_j)(x_B) = -\frac{k}{n}\binom{n}{k}^{-1}$ gives $c_1 = -\frac{n}{k(n-k)}\sum_{\ell \le n-k} a(\ell)$, which is the stated difference because $\sum_{\ell=1}^{n} a(\ell) = n\,\mathbb{E}\,\mathsf{s}(\xi_1) = -n\int f' = 0$.

The two families. For logistic noise $\mathsf{s} = 2F - 1$, and $F(\xi_{(\ell)})$ has the $\mathrm{Beta}(\ell, n-\ell+1)$ distribution, so $a(\ell) = \frac{2\ell}{n+1} - 1$ and $\sum_{\ell > n-k} a(\ell) = \frac{k(n-k)}{n+1}$, whence the difference of the two means is $\frac{n}{n+1}$. For Gumbel noise $\mathsf{s}(q) = 1 - e^{-q}$ and $e^{-\xi_1}$ is a unit-rate exponential variable, so $e^{-\xi_{(\ell)}}$ is the $(n-\ell+1)$-st smallest of n such variables, with mean $H_n - H_{\ell-1}$, and $a(\ell) = 1 - H_n + H_{\ell-1}$. With $\sum_{j<\mathsf{n}} H_j = \mathsf{n}H_{\mathsf{n}} - \mathsf{n}$ for every integer $\mathsf{n} \ge 1$ this gives $\sum_{\ell > n-k} a(\ell) = (n-k)(H_n - H_{n-k})$, and the difference of the two means is $\frac{n}{k}(H_n - H_{n-k})$. □

C.4 Second-Order Constants

For the k-subsets of $[n]$ with $2 \le k \le n-2$ and independent identically distributed noise entries whose density f is absolutely continuous with $\int |f'| < \infty$, two coefficients determine the second-order term of the law. Write $\theta_B = \langle \theta, x_B \rangle$ for the score of a subset and $\theta_B^{(2)} = \langle \theta \circ \theta, x_B - \frac{k}{n}\mathbf{1} \rangle$ for its centered square score. The law is then twice continuously differentiable (Lemma C.2) and, for centered θ,

$$\binom{n}{k} Q_\theta^\nu(x_B) = 1 + c_1\theta_B + \alpha\Big(\theta_B^2 - \frac{k(n-k)}{n(n-1)}\|\theta\|^2\Big) + \beta\,\theta_B^{(2)} + o(\|\theta\|^2), \tag{C.1}$$

which defines the second-order coefficients α and β. A quadratic form in the centered scores invariant under the permutations of B and of its complement is a combination of θ_B^2, $\theta_B^{(2)}$ and $\|\theta\|^2$, and normalization fixes the coefficient of $\|\theta\|^2$. The same expansion of $\binom{n}{k} p_\theta(x_B)$ has 1, $\frac{1}{2}$ and 0 in place of c_1, α and β. At temperature c_1 the coefficient of θ_B^2 in the law becomes α/c_1^2, against $\frac{1}{2}$ for the Gibbs distribution, and $r = \alpha/c_1^2 - \frac{1}{2}$ is their difference. Lemma C.2 computes α and β from f, and Table C.1 evaluates them for the three noise families. The exponential density jumps at 0, and the end of this subsection expands its law directly to the same form, with the constants of that table.

Lemma C.2 (second-order constants of independent noise)**.** *Let $2 \le k \le n-2$ and let the noise entries be independent and identically distributed with a density f that is absolutely continuous with $\int |f'| < \infty$ and distribution function F. Write $\bar{k} = n - k$ and*

$$A_1 = \int f^3 F^{\bar{k}-1}(1-F)^{k-2}, \qquad A_2 = \int f^3 F^{\bar{k}-2}(1-F)^{k-1}.$$

Then $\theta \mapsto Q_\theta^\nu$ is twice continuously differentiable, Equation (C.1) holds with

$$\alpha = \binom{n}{k}\frac{n-1}{2}(A_1 + A_2), \qquad \beta = \binom{n}{k}\frac{1}{4}\Big[\big(n\bar{k} + k - \bar{k}\big)A_2 - \big(nk + \bar{k} - k\big)A_1\Big],$$

and $A_1 + A_2 = \int f^3 F^{\bar{k}-2}(1-F)^{k-2}$, so $\alpha > 0$. Moreover $r \ge -\frac{1}{2n^2}$, with equality if and only if f is a logistic density.

Proof. Regularity. The perturbed scores $\mathsf{x} = \theta + \xi$ have density $\prod_j f(\mathsf{x}_j - \theta_j)$ and $x^* = x_B$ on the region $\mathsf{W}_B = \{\mathsf{x} : \min_{b\in B} \mathsf{x}_b > \max_{j\notin B} \mathsf{x}_j\}$, so $Q_\theta^\nu(x_B) = \int_{\mathsf{W}_B} \prod_j f(\mathsf{x}_j - \theta_j)\,d\mathsf{x}$, which is Proposition B.1(b) before the coordinates outside B are integrated out. The difference quotient of $f(\cdot - \theta_j)$ in θ_j is bounded in L^1 by $\int |f'|$ and converges to $-f'$ in L^1, and the remaining factors are bounded in the differentiated variable and integrable in their own variables. Differentiating in two distinct coordinates therefore gives

$$\partial_i \partial_{i'} Q_\theta^\nu(x_B) = \int_{\mathsf{W}_B} f'(\mathsf{x}_i - \theta_i)\, f'(\mathsf{x}_{i'} - \theta_{i'}) \prod_{j \ne i, i'} f(\mathsf{x}_j - \theta_j)\,d\mathsf{x},$$

continuous in θ by the continuity of translation in L^1. A common shift of the scores leaves the law unchanged, so $\sum_j \partial_j Q_\theta^\nu(x_B) = 0$ identically and $\partial_i Q^\nu = -\sum_{j\ne i} \partial_j Q^\nu$ is differentiable in θ_i as well, with continuous derivative. So Q^ν is twice continuously differentiable, and every row of the Hessian of $Q_\theta^\nu(x_B)$ sums to zero.

Two entries of the Hessian. By the symmetry of the law in the coordinates of B and in those outside it, the Hessian at $\theta = 0$ has five entries up to symmetry: h_{bb} and $\mathsf{h}_{bb'}$ for one and for two coordinates in B, h_{jj} and $\mathsf{h}_{jj'}$ for those outside it, and h_{bj} for one of each. Condition on all entries but $\xi_j, \xi_{j'}$ for two indices outside B, and let a be the smallest perturbed score in B and c the largest among the other $\bar{k} - 2$ indices outside B, with $\mathsf{c} = -\infty$ when $\bar{k} = 2$. Given these, B wins if and only if $\mathsf{a} > \mathsf{c}$ and both $\theta_j + \xi_j$ and $\theta_{j'} + \xi_{j'}$ lie below a, an event of conditional probability $\mathbf{1}\{\mathsf{a} > \mathsf{c}\}F(\mathsf{a} - \theta_j)F(\mathsf{a} - \theta_{j'})$. Differentiating in θ_j and $\theta_{j'}$ at zero gives

$$\mathsf{h}_{jj'} = \mathbb{E}\big[f(\mathsf{a})^2 \mathbf{1}\{\mathsf{a} > \mathsf{c}\}\big] = k \int f^3 F^{\bar{k}-2}(1-F)^{k-1} = kA_2,$$

since at $\theta = 0$ the variable a has density $kf(1-F)^{k-1}$ and $\mathbb{P}(\mathsf{c} < \mathsf{a} \mid \mathsf{a}) = F(\mathsf{a})^{\bar{k}-2}$. Symmetrically, conditioning on all entries but $\xi_b, \xi_{b'}$ for two indices in B, with c' the largest perturbed score outside B and a' the smallest among the other $k - 2$ indices in B ($+\infty$ when $k = 2$), the conditional probability is $\mathbf{1}\{\mathsf{a}' > \mathsf{c}'\}\big(1 - F(\mathsf{c}' - \theta_b)\big)\big(1 - F(\mathsf{c}' - \theta_{b'})\big)$ and $\mathsf{h}_{bb'} = \mathbb{E}[f(\mathsf{c}')^2 \mathbf{1}\{\mathsf{a}' > \mathsf{c}'\}] = \bar{k}A_1$.

The third entry. Differentiating $\sum_{B'} Q_\theta^\nu(x_{B'}) = 1$ twice, in two distinct coordinates, gives

$$\binom{n-2}{k-2}\mathsf{h}_{bb'} + 2\binom{n-2}{k-1}\mathsf{h}_{bj} + \binom{n-2}{k}\mathsf{h}_{jj'} = 0,$$

the three coefficients counting the k-subsets that contain both differentiated coordinates, exactly one of them, and neither. Hence $\mathsf{h}_{bj} = -\frac{1}{2}[(k-1)A_1 + (\bar{k}-1)A_2]$.

Assembly. For centered θ the second-order term of $\binom{n}{k}Q_\theta^\nu(x_B)$ is $\frac{1}{2}\binom{n}{k}\sum_{i,i'} \mathsf{h}_{ii'}\theta_i\theta_{i'}$, which with $\sum_{j \notin B} \theta_j = -\theta_B$ and $\mathsf{P}_B = \sum_{b \in B} \theta_b^2$ equals

$$\tfrac{1}{2}\tbinom{n}{k}\Big[(\mathsf{h}_{bb} - \mathsf{h}_{bb'})\mathsf{P}_B + (\mathsf{h}_{jj} - \mathsf{h}_{jj'})\big(\|\theta\|^2 - \mathsf{P}_B\big) + (\mathsf{h}_{bb'} + \mathsf{h}_{jj'} - 2\mathsf{h}_{bj})\theta_B^2\Big].$$

Comparing with Equation (C.1) through $\theta_B^{(2)} = \mathsf{P}_B - \frac{k}{n}\|\theta\|^2$ gives $\alpha = \frac{1}{2}\binom{n}{k}(\mathsf{h}_{bb'} + \mathsf{h}_{jj'} - 2\mathsf{h}_{bj}) = \binom{n}{k}\frac{n-1}{2}(A_1 + A_2)$ and $\beta = \frac{1}{2}\binom{n}{k}(\mathsf{h}_{bb} - \mathsf{h}_{bb'} - \mathsf{h}_{jj} + \mathsf{h}_{jj'})$. The two row sums $\mathsf{h}_{bb} + (k-1)\mathsf{h}_{bb'} + \bar{k}\mathsf{h}_{bj} = 0$ and $\mathsf{h}_{jj} + (\bar{k}-1)\mathsf{h}_{jj'} + k\mathsf{h}_{bj} = 0$ turn β into $\frac{1}{2}\binom{n}{k}\big(\bar{k}\mathsf{h}_{jj'} - k\mathsf{h}_{bb'} + (k-\bar{k})\mathsf{h}_{bj}\big)$, the displayed expression. The term of Equation (C.1) that does not depend on B is then fixed by $\sum_{B'} Q_\theta^\nu(x_{B'}) = 1$.

The bound. Substituting $\mathsf{u} = F(q)$ as in Appendix C.2 and writing $\mathsf{w}(\mathsf{u}) = f(F^{-1}(\mathsf{u}))/(\mathsf{u}(1-\mathsf{u}))$ and $A_0 = \int f^2 F^{\bar{k}-1}(1-F)^{k-1}$,

$$A_0 = \mathrm{B}(\bar{k}+1, k+1)\,\mathbb{E}\,\mathsf{w}(U), \qquad A_1 + A_2 = \mathrm{B}(\bar{k}+1, k+1)\,\mathbb{E}\,\mathsf{w}(U)^2,$$

where U has the $\mathrm{Beta}(\bar{k}+1, k+1)$ distribution. With $\binom{n}{k}\mathrm{B}(\bar{k}+1, k+1) = \frac{1}{n+1}$ and Theorem 3, $c_1 = \frac{n}{n+1}\mathbb{E}\,\mathsf{w}(U)$ and $\alpha = \frac{n-1}{2(n+1)}\mathbb{E}\,\mathsf{w}(U)^2$, so

$$r = \frac{\alpha}{c_1^2} - \frac{1}{2} = \frac{(n^2-1)\,\mathrm{Var}\,\mathsf{w}(U) - \big(\mathbb{E}\,\mathsf{w}(U)\big)^2}{2n^2\big(\mathbb{E}\,\mathsf{w}(U)\big)^2} \;\geq\; -\frac{1}{2n^2},$$

with equality if and only if $\mathsf{w}(U)$ is almost surely constant, say equal to $\mathsf{s} > 0$. In that case $f(F^{-1}(\mathsf{u})) = \mathsf{s}\,\mathsf{u}(1-\mathsf{u})$ for almost every $\mathsf{u} \in (0,1)$, hence $f = \mathsf{s}F(1-F)$ almost everywhere on $\{f > 0\}$ and, both sides being continuous, everywhere on the open set $\mathcal{S} = \{f > 0\} \cap \mathcal{I}$, where $\mathcal{I} = \{0 < F < 1\}$ is an open interval. At a boundary point of $\mathcal{S}$ inside $\mathcal{I}$ continuity would give both $f = 0$ and $f = \mathsf{s}F(1-F) > 0$, so $\mathcal{S}$ is closed in $\mathcal{I}$ as well and $\mathcal{S} = \mathcal{I}$. On this interval $\big(\log\frac{F}{1-F}\big)' = \mathsf{s}$, so $F(q) = 1/(1 + e^{-\mathsf{s}(q-q_0)})$ on $\mathcal{I}$ for some $q_0 \in \mathbb{R}$, which forces $\mathcal{I} = \mathbb{R}$: f is a logistic density of scale $1/\mathsf{s}$.

The Gumbel and logistic rows. Under the same substitution, f^3 against powers of F and $1-F$ becomes $f(F^{-1}(\mathsf{u}))^2$ against the same powers of u and $1-\mathsf{u}$. For Gumbel noise this makes $A_2 = \int_0^1 \mathsf{u}^{\bar{k}}(1-\mathsf{u})^{k-1}(\log\mathsf{u})^2\,d\mathsf{u}$ and $A_1 + A_2 = \int_0^1 \mathsf{u}^{\bar{k}}(1-\mathsf{u})^{k-2}(\log\mathsf{u})^2\,d\mathsf{u}$. Differentiating the Beta integral twice in its first parameter evaluates them as $A_2 = \mathrm{B}(\bar{k}+1, k)(h_1^2 + h_2)$ and $A_1 + A_2 = \mathrm{B}(\bar{k}+1, k-1)\big(h_1^2 + h_2 - \frac{2h_1}{n}\big)$ with $h_1 = H_n - H_{\bar{k}}$ and $h_2 = H_n^{(2)} - H_{\bar{k}}^{(2)}$, where $H_n^{(2)} = \sum_{j \leq n} j^{-2}$.

Table C.1: The first- and second-order constants of the three noise families for the k-subsets of $[n]$ with $2 \le k \le n-2$, from Theorem 3 and Lemma C.2, the exponential row from the direct expansion at the end of Appendix C.4. Here $h_1 = H_n - H_{n-k}$, $h_2 = H_n^{(2)} - H_{n-k}^{(2)}$ with $H_n^{(2)} = \sum_{j \le n} j^{-2}$, $h_3 = h_1^2 + h_2$ and $\mathsf{N} = n(n-k)h_1^2 + kn(n-1)h_2 - 2k(n-1)h_1$. The last row is square Gumbel matching, whose second-order term is not of the k-subset form: Corollary C.1 gives it, and the coefficients r' and r'' of Proposition C.3 take the place of r.

noise	c_1	α	β	$r = \alpha/c_1^2 - \frac{1}{2}$
Gumbel	$\frac{n}{k}\, h_1$	$\frac{(n-1)(nh_3 - 2h_1)}{2k(k-1)}$	$\frac{(k(n-1)+n-k)h_1 - (n-k)nh_3}{2k(k-1)}$	$\frac{\mathsf{N}}{2(k-1)n^2h_1^2}$
logistic	$\frac{n}{n+1}$	$\frac{n-1}{2(n+1)}$	$\frac{2k-n}{2(n+1)(n+2)}$	$-\frac{1}{2n^2}$
exponential	$\frac{n}{n-k}$	$\frac{n(n-1)}{2(n-k)(n-k-1)}$	$\frac{nk}{2(n-k)(n-k-1)}$	$\frac{k}{2n(n-k-1)}$
Gumbel matching	$\frac{n}{n+1}$	—	—	$r' = -\frac{1}{2n^2}$ $r'' = \frac{1}{2n(n+2)}$

With $\binom{n}{k}\mathrm{B}(\bar{k}+1, k) = \frac{1}{k}$ and $\binom{n}{k}\mathrm{B}(\bar{k}+1, k-1) = \frac{n}{k(k-1)}$ the lemma gives the Gumbel row of Table C.1. For logistic noise $A_1 = \mathrm{B}(\bar{k}+2, k+1)$ and $A_2 = \mathrm{B}(\bar{k}+1, k+2)$, whose sum is $\mathrm{B}(\bar{k}+1, k+1)$, and the logistic row follows. □

The exponential row. The exponential density $e^{-q}\mathbf{1}\{q > 0\}$ jumps at 0 and lies outside the lemma. Conditioning on the smallest perturbed score in B, the integrand of $Q_\theta^\nu(x_B)$ equals $ke^{-kq+\theta_B}\prod_{j\notin B}(1-e^{\theta_j - q})$ for every q above all the scores, so the probability may be compared with the surrogate $\tilde{Q}(x_B) = ke^{\theta_B}\int_0^\infty e^{-kq}\prod_{j\notin B}(1-e^{\theta_j-q})\,dq$. The two integrands differ only on an interval of length $O(\|\theta\|)$ around zero, on which each of the $\bar{k} \ge 2$ factors indexed by $j \notin B$ is $O(\|\theta\|)$ in both, so the two integrals differ by $O(\|\theta\|^{\bar{k}+1}) = O(\|\theta\|^3)$. Substituting $\mathsf{v} = e^{-q}$ and expanding to second order for centered θ,

$$\tilde{Q}(x_B) = k\int_0^1 \mathsf{v}^{k-1}\Big[(1-\mathsf{v})^{\bar{k}} + (1-\mathsf{v})^{\bar{k}-1}\theta_B + \tfrac{1}{2}(1-\mathsf{v})^{\bar{k}-2}\Big(\theta_B^2 - \mathsf{v}\big(\|\theta\|^2 - \mathsf{P}_B\big)\Big)\Big]d\mathsf{v} + O(\|\theta\|^3),$$

and the ratios $\mathrm{B}(k,\bar{k})/\mathrm{B}(k,\bar{k}+1) = \frac{n}{\bar{k}}$, $\mathrm{B}(k,\bar{k}-1)/\mathrm{B}(k,\bar{k}+1) = \frac{n(n-1)}{\bar{k}(\bar{k}-1)}$ and $\mathrm{B}(k+1,\bar{k}-1)/\mathrm{B}(k,\bar{k}+1) = \frac{kn}{\bar{k}(\bar{k}-1)}$ give the exponential row of Table C.1, which is what the lemma's formulas return at $A_1 = \mathrm{B}(\bar{k}, k+1)$ and $A_2 = \mathrm{B}(\bar{k}-1, k+2)$.

C.5 Intrinsic Residual

Section 4 states the following theorem in prose. For a first-order exact calibration, the component orthogonal to $\mathrm{im}\, J$ of the second-order discrepancy is the intrinsic residual, the same for every such calibration by the chain rule below. Part (a) is proved in this subsection and part (b) in Appendix C.6.

Theorem C.1 (second-order calibration)**.** *Let $\mathcal{X}$ be the k-subsets of $[n]$ with $2 \le k \le n-2$, and let the noise entries be independent and identically distributed, either with a density f that is absolutely continuous with $\int |f'| < \infty$ or with the exponential distribution.*

(a) Every first-order exact calibration leaves the same intrinsic residual: $2r$ times a nonzero quadratic map of the centered scores that depends only on n and k, the component orthogonal to $\mathrm{im}\, J$ of the second-order term of p_θ, where

$$\alpha = \binom{n}{k}\frac{n-1}{2}\int f(q)^3 F(q)^{n-k-2}\big(1-F(q)\big)^{k-2}\,dq > 0, \qquad r = \frac{\alpha}{c_1^2} - \frac{1}{2}.$$

A second-order exact calibration exists if and only if $r = 0$. Here $r \geq -\frac{1}{2n^2}$, with equality if and only if f is a logistic density. For Gumbel noise $r = -\frac{1}{2(2n-1)^2}$ at $k = 2$ and $r > 0$ for $3 \leq k \leq n-2$; for logistic noise $r = -\frac{1}{2n^2}$; for exponential noise $r = \frac{k}{2n(n-k-1)}$.

(b) If $r \neq 0$, the following randomized calibrations, applied to the centered part of the scores, are second-order exact. If $r > 0$, draw a pair $\{i, j\} \subseteq [n]$ and a sign uniformly at random, set $\eta_i = (\theta_i \pm \varepsilon\theta_j)/c_1$ and $\eta_j = (\theta_j \mp \varepsilon\theta_i)/c_1$, and divide the remaining coordinates by c_1, with $\varepsilon^2 = \frac{2r}{1+2r}\binom{n}{2}$. If $r < 0$, draw a sign uniformly at random and set $\eta = (1 \pm \varepsilon)\theta/c_1$ with $\varepsilon^2 = \frac{-2r}{1+2r}$. In both cases add $\gamma\big(\theta_{i'}^2 - \frac{1}{n}\|\theta\|^2\big)$ to every coordinate $\eta_{i'}$, with $\gamma = -\beta/(2\alpha c_1)$.

The residual is intrinsic in the sense of Bates & Watts (1980): the calibrated laws and the Gibbs distributions pass through p_0 with the same tangent space $\operatorname{im} J$, and by the chain rule a reparametrization of the scores controls only the discrepancy inside that space. The proof below writes the intrinsic residual as $r/\binom{n}{k}$ times the component of θ_B^2 orthogonal to the constants and to $\operatorname{im} J$. Since the second-order term of $\binom{n}{k}p_\theta(x_B)$ is $\frac{1}{2}\theta_B^2$ up to a multiple of $\|\theta\|^2$, which is constant on $\mathcal{X}$, this is $2r$ times the component orthogonal to $\operatorname{im} J$ of the second-order term of p_θ, the form in which Theorem C.1(a) states it. Equivalently, the component orthogonal to $\operatorname{im} J$ of the second-order term of Q^ν_{θ/c_1} is $1 + 2r$ times that of p_θ, the multiple of Section 4.

Proof of Theorem C.1(a). The chain rule. Let η be a calibration, L its differential at 0, which is a linear calibration, and η_2 the quadratic term of its Taylor expansion there. Write Q_2 and p_2 for the quadratic terms of the expansions of Q^ν and of p at 0. Then

$$Q^\nu_{\eta(\theta)} - p_\theta = \big(JL\theta - J_p\theta\big) + \Big(J\eta_2(\theta) + Q_2(L\theta) - p_2(\theta)\Big) + o(\|\theta\|^2),$$

so η is first-order exact if and only if $JL = J_p$, which by Lemma C.1 and Proposition C.1 holds if and only if $L\theta = \theta/c_1$ up to constant-score directions. These leave the law unchanged, hence also its Taylor terms, so $Q_2(L\theta) = Q_2(\theta/c_1)$ and the second-order discrepancy of every first-order exact calibration is $J\eta_2(\theta) + Q_2(\theta/c_1) - p_2(\theta)$. Its component orthogonal to $\operatorname{im} J$ does not depend on η_2, and if that component vanishes then η_2 can be chosen with $J\eta_2 = p_2 - Q_2(\cdot/c_1)$, by composing with a linear right inverse of J on $\operatorname{im} J$. So a second-order exact calibration exists if and only if the component vanishes.

The residual. Both $Q_2(\vartheta)$ and $p_2(\vartheta)$ sum to zero over the k-subsets, so the discrepancy has no component along the constants, and by Equation (C.1),

$$\binom{n}{k}\big(Q_2(\theta/c_1) - p_2(\theta)\big)(x_B) = r\Big(\theta_B^2 - \frac{k(n-k)}{n(n-1)}\|\theta\|^2\Big) + \frac{\beta}{c_1^2}\theta_B^{(2)}$$

for centered θ. Here $\operatorname{im} J = \operatorname{im} J_p$, because $J = c_1 J_p$ with $c_1 > 0$, and $\operatorname{im} J_p$ is the space V_1 of functions $x_B \mapsto \langle\vartheta, x_B\rangle$ with $\sum_i \vartheta_i = 0$. Since $\theta_B^{(2)}$ belongs to V_1, the intrinsic residual is $r/\binom{n}{k}$ times the component of θ_B^2 orthogonal to the constants and to V_1. A function of k-subsets is orthogonal to the constants and to V_1 exactly when its mean and all its conditional expectations $\mathbb{E}_{p_0}[\,\cdot \mid i \in B]$ under the uniform distribution on k-subsets vanish, so the projection is read off from those conditional expectations. A function $\langle\vartheta, x_B\rangle$ with $\sum_i \vartheta_i = 0$ has $\mathbb{E}_{p_0}[\langle\vartheta, x_B\rangle \mid i \in B] = \frac{n-k}{n-1}\vartheta_i$. Expanding $\theta_B^2 = \sum_{j,j'}\theta_j\theta_{j'}\mathbf{1}\{j, j' \in B\}$ and using $\mathbb{P}(j \in B \mid i \in B) = \frac{k-1}{n-1}$ and $\mathbb{P}(j, j' \in B \mid i \in B) = \frac{(k-1)(k-2)}{(n-1)(n-2)}$ for distinct $j, j' \neq i$ gives, for centered θ,

$$\mathbb{E}_{p_0}\big[\theta_B^2 \mid i \in B\big] = \frac{(n-k)(n-2k)}{(n-1)(n-2)}\Big(\theta_i^2 - \frac{\|\theta\|^2}{n}\Big) + \text{a term independent of } i.$$

So the projection of θ_B^2 onto the constants and V_1 is $\frac{n-2k}{n-2}\theta_B^{(2)}$ plus a constant, and the intrinsic residual is

$$\rho_2(\theta)(x_B) = \frac{r}{\binom{n}{k}}\Big(\theta_B^2 - \frac{k(n-k)}{n(n-1)}\|\theta\|^2 - \frac{n-2k}{n-2}\theta_B^{(2)}\Big).$$

The quadratic map in parentheses is not zero: with $\theta = e_1 + e_2 - e_3 - e_4$, a fixed $(k-2)$-subset C of $[n] \setminus \{1, 2, 3, 4\}$ and $\mathsf{B}_{ii'} = \mathsf{C} \cup \{i, i'\}$, the alternating sum $\mathsf{f}(\mathsf{B}_{12}) - \mathsf{f}(\mathsf{B}_{13}) - \mathsf{f}(\mathsf{B}_{42}) + \mathsf{f}(\mathsf{B}_{43})$

vanishes for every constant and every member of V_1 and equals 8 for $\mathsf{f}(x_B) = \theta_B^2$. Hence a second-order exact calibration exists if and only if $r = 0$, and Lemma C.2 supplies α, the bound on r and the logistic case. The exponential values are those of Table C.1.

The sign of r for Gumbel noise. By Table C.1, $r = \mathsf{N}/\big(2(k-1)n^2h_1^2\big)$ with

$$\mathsf{N} = n(n-k)h_1^2 + kn(n-1)h_2 - 2k(n-1)h_1.$$

At $k = 2$, $h_1 = \frac{1}{n} + \frac{1}{n-1}$ and $h_2 = \frac{1}{n^2} + \frac{1}{(n-1)^2}$ give $\mathsf{N} = -\frac{1}{(n-1)^2}$ and $r = -\frac{1}{2(2n-1)^2}$. For $3 \le k \le n-2$, two mean inequalities bound N from below once the summation range is centered: let $\mathsf{a} = \frac{k-1}{2}$, $\mathsf{c} = n - \mathsf{a}$, $\mathsf{z} = 1/\mathsf{c}$ and $\mathsf{d}_i = i - \mathsf{a}$ for $i = 0, \dots, k-1$, so that the summands of h_1 and h_2 are $1/(\mathsf{c} - \mathsf{d}_i)$ and its square and the d_i are symmetric about zero. Pairing $\pm\mathsf{d}_i$ and writing $\mathsf{U}_i = (1 - \mathsf{d}_i^2\mathsf{z}^2)^{-1}$, which is positive because $|\mathsf{d}_i| \le \mathsf{a} < \mathsf{c}$,

$$h_1 = k\mathsf{z}\mathsf{S}, \qquad h_2 = k\mathsf{z}^2\mathsf{T}, \qquad \mathsf{S} = \frac{1}{k}\sum_i \mathsf{U}_i, \qquad \mathsf{T} = \frac{2}{k}\sum_i \mathsf{U}_i^2 - \mathsf{S}.$$

Put $\mathsf{m} = \frac{1}{k}\sum_i \mathsf{d}_i^2 = \frac{k^2-1}{12}$, $\mathsf{b} = \mathsf{a}(\mathsf{a}-1)$ and $\mathsf{E} = 1 - \mathsf{m}\mathsf{z}^2$, which is positive because $\mathsf{m} \le \mathsf{a}^2 < \mathsf{c}^2$. With $n(n-k) = \mathsf{c}^2 - \mathsf{c} - 3\mathsf{m}$ and $n(n-1) = \mathsf{c}^2 + (2\mathsf{a}-1)\mathsf{c} + \mathsf{b}$,

$$\frac{\mathsf{N}}{k^2} = (1 - \mathsf{z} - 3\mathsf{m}\mathsf{z}^2)\mathsf{S}^2 + \big(1 + (2\mathsf{a}-1)\mathsf{z} + \mathsf{b}\mathsf{z}^2\big)\mathsf{T} - 2\big(1 + (\mathsf{a}-1)\mathsf{z}\big)\mathsf{S}.$$

Cauchy–Schwarz gives $\mathsf{T} \ge 2\mathsf{S}^2 - \mathsf{S}$, and Jensen, applied to the convex function $q \mapsto (1 - \mathsf{z}^2 q)^{-1}$ at the mean m of the d_i^2, gives $\mathsf{S} \ge 1/\mathsf{E}$. For $k \ge 3$ one has $\mathsf{a} \ge 1$, so the coefficient of T is positive and, after the first substitution, the bracket below is increasing in S:

$$\frac{\mathsf{N}}{k^2} \;\ge\; \mathsf{S}\Big[\big(3\mathsf{E} + (4\mathsf{a}-3)\mathsf{z} + 2\mathsf{b}\mathsf{z}^2\big)\mathsf{S} - \big(3 + (4\mathsf{a}-3)\mathsf{z} + \mathsf{b}\mathsf{z}^2\big)\Big] \;\ge\; \frac{\mathsf{S}\mathsf{z}^2}{\mathsf{E}}\Big[\mathsf{b}\big(1 + \mathsf{m}\mathsf{z}^2\big) + (4\mathsf{a}-3)\mathsf{m}\mathsf{z}\Big] > 0,$$

since $\mathsf{b} \ge 0$ and $4\mathsf{a} - 3 = 2k - 5 \ge 1$. So $r > 0$ for $3 \le k \le n-2$.

Its size. As $n \to \infty$ with k/n bounded away from 0 and 1, $h_1 = \log\frac{n}{n-k} + O(n^{-1})$ and $h_2 = \frac{k}{n(n-k)} + O(n^{-2})$, so $\mathsf{N} = \frac{n}{n-k}\big((n-k)h_1 - k\big)^2 + O(n)$ and

$$r = \frac{\big((n-k)\log\frac{n}{n-k} - k\big)^2}{2k(n-k)\big(\log\frac{n}{n-k}\big)^2} \cdot \frac{1}{n} + O(n^{-2}),$$

of order n^{-1}, the numerator being nonzero because $-s\log s < 1 - s$ for $s = \frac{n-k}{n} \in (0,1)$. The exponential value $\frac{k}{2n(n-k-1)}$ is also of order n^{-1}, against $-\frac{1}{2n^2}$ for logistic noise. □

C.6 Randomized Calibrations

Proof of Theorem C.1(b). Both constructions draw a random linear map A of the centered scores with $\mathbb{E}\mathsf{A} = I$ and use the calibration $\eta(\theta) = \frac{1}{c_1}\mathsf{A}\theta + \gamma\,\mathsf{q}(\theta)$ with $\mathsf{q}(\theta)_i = \theta_i^2 - \frac{1}{n}\|\theta\|^2$. For $r > 0$, $\mathsf{A} = I \pm \varepsilon\Omega$ with $\Omega\theta = \theta_j e_i - \theta_i e_j$ for a uniformly random pair $\{i, j\}$ and a uniformly random sign. For $r < 0$, $\mathsf{A} = (1 \pm \varepsilon)I$ with a uniformly random sign. For each draw,

$$Q^{\nu}_{\eta(\theta)} = p_0 + \tfrac{1}{c_1}J\mathsf{A}\theta + \gamma J\mathsf{q}(\theta) + Q_2(\mathsf{A}\theta/c_1) + o(\|\theta\|^2),$$

and averaging over the finitely many draws leaves the first-order term $\frac{1}{c_1}J\theta = J_p\theta$, the Gibbs first-order term.

Averaging the quadratic term. Write $Q_2(\vartheta)(x_B) = \vartheta^\top \mathsf{M}_B\vartheta$ with M_B symmetric. A common shift of the scores leaves the law unchanged, so $\mathsf{M}_B = P\mathsf{M}_B P$ with $P = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$. For the two temperatures, $\mathbb{E}_{\mathsf{A}}Q_2(\mathsf{A}\theta) = \frac{1}{2}[(1+\varepsilon)^2 + (1-\varepsilon)^2]\,Q_2(\theta) = (1+\varepsilon^2)Q_2(\theta)$. For the random pair the average over the sign cancels the cross term, and summing $\theta_j^2 e_i e_i^\top + \theta_i^2 e_j e_j^\top - \theta_i\theta_j(e_i e_j^\top + e_j e_i^\top)$ over the ordered pairs $i \ne j$ gives $2(\|\theta\|^2 I - \theta\theta^\top)$, so that

$$\mathbb{E}\big[\mathsf{A}\theta(\mathsf{A}\theta)^\top\big] = \theta\theta^\top + \frac{\varepsilon^2}{\binom{n}{2}}\big(\|\theta\|^2 I - \theta\theta^\top\big)$$

and, for centered θ,

$$\mathbb{E}_{\mathsf{A}} Q_2(\mathsf{A}\theta)(x_B) = Q_2(\theta)(x_B) + \frac{\varepsilon^2}{\binom{n}{2}} \Big(\|\theta\|^2 \operatorname{tr}(\mathsf{M}_B) - Q_2(\theta)(x_B) \Big).$$

The trace does not depend on B, because a permutation of $[n]$ that maps B to B' conjugates M_B to $\mathsf{M}_{B'}$. The law is normalized, so $\sum_{B'} Q_2(\vartheta)(x_{B'}) = 0$ for every ϑ. Summing the last display over B therefore gives $\operatorname{tr}(\mathsf{M}_B) = 0$, and the average is $\big(1 - \varepsilon^2/\binom{n}{2}\big) Q_2(\theta)$.

The two constants. In both cases $\mathbb{E}_{\mathsf{A}} Q_2(\mathsf{A}\theta/c_1) = \mathsf{t}\, Q_2(\theta/c_1)$, with $\mathsf{t} = 1 + \varepsilon^2$ for the two temperatures and $\mathsf{t} = 1 - \varepsilon^2/\binom{n}{2}$ for the random pair. Since $\binom{n}{k} J\mathsf{q}(\theta)(x_B) = c_1 \theta_B^{(2)}$, the second-order term of the difference between the output law and p_θ is

$$\binom{n}{k}^{-1} \Big[\Big(\frac{\mathsf{t}\alpha}{c_1^2} - \frac{1}{2} \Big) \theta_B^2 + \Big(\frac{\mathsf{t}\beta}{c_1^2} + \gamma c_1 \Big) \theta_B^{(2)} \Big]$$

up to a term that does not depend on B. The first coefficient vanishes if and only if $\mathsf{t}(1 + 2r) = 1$, as $1 + 2r = 2\alpha/c_1^2 > 0$. This is $\varepsilon^2 = \frac{-2r}{1+2r}$ for the two temperatures and $\varepsilon^2 = \frac{2r}{1+2r}\binom{n}{2}$ for the random pair, the stated values, positive exactly when $r < 0$ and when $r > 0$, respectively. Both temperatures $c_1/(1 \pm \varepsilon)$ are positive, because $r \geq -\frac{1}{2n^2}$ gives $\varepsilon^2 \leq \frac{1}{n^2-1}$. With $\mathsf{t} = c_1^2/(2\alpha)$ the second coefficient is $\frac{\beta}{2\alpha} + \gamma c_1$, which vanishes for $\gamma = -\frac{\beta}{2\alpha c_1}$. Both coefficients being zero, the second-order term of the difference is the same for every B, and it sums to zero over the k-subsets, so it is zero. A finite mixture of laws with $o(\|\theta\|^2)$ remainders has an $o(\|\theta\|^2)$ remainder, so the output law is $p_\theta + o(\|\theta\|^2)$. □

C.7 Gumbel Matching to Second Order

Throughout, $\Pi\theta = (I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top)\,\theta\,(I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top)$ is the double centering of a score matrix, so that $\Pi\theta = \theta$ exactly when θ has zero row and column sums.

Corollary C.1 (Gumbel matching to second order)**.** *For square Gumbel matching with $n \geq 2$, every assignment σ and every score matrix θ with $\Pi\theta = \theta$,*

$$n!\, Q_\theta(\sigma) = 1 + \frac{n}{n+1}\langle \theta, P_\sigma \rangle + \frac{(n^2+n-1)\langle \theta, P_\sigma \rangle^2 - (n+1)\|\theta\|_F^2 - \sum_{i,j} \theta_{i\sigma(j)}\theta_{j\sigma(i)} - n\sum_i \theta_{i\sigma(i)}^2}{2(n+1)(n+2)} + O(\|\theta\|_F^3),$$

while $n!\, p_\theta(\sigma) = 1 + \langle \theta, P_\sigma \rangle + \frac{1}{2}\big(\langle \theta, P_\sigma \rangle^2 - \frac{\|\theta\|_F^2}{n-1}\big) + O(\|\theta\|_F^3)$. In particular the first-order temperature is $c_1 = \frac{n}{n+1}$.

Proof of Corollary C.1. The derivatives as clock moments. The expansion comes from the conditional law of the clocks given the optimal assignment, which Appendix B.2 makes explicit at zero scores. Under standard Gumbel noise the perturbed optimizer returns the assignment minimizing $\sum_i \log s_{i\sigma(i)}$ for independent exponential clocks s_{ij} of rate $\exp(\theta_{ij})$. Relabel the columns so that $\sigma = \mathrm{id}$, and let $\mathcal{A}$ be the event that the identity is optimal for unit-rate clocks s. Changing the rates from $\exp(\theta_{ij})$ to one gives the expectation over unit-rate clocks

$$Q_\theta(\mathrm{id}) = \mathbb{E}\Big[\mathbf{1}\{\mathcal{A}\} \prod_{i,j} \exp(\theta_{ij})\, e^{-(\exp(\theta_{ij})-1)s_{ij}} \Big].$$

Against Lebesgue measure the integrand is $\mathbf{1}\{\mathcal{A}\} \prod_{i,j} \exp(\theta_{ij} - e^{\theta_{ij}} s_{ij})$, which with its derivatives in θ is dominated, near any score matrix, by a polynomial in s times the exponential of a negative multiple of $\sum_{i,j} s_{ij}$, so $\theta \mapsto Q_\theta(\mathrm{id})$ is smooth and may be differentiated under the expectation. Ties have probability zero and the law of s is invariant under permutations of the columns, so every assignment is optimal with the same probability and $\mathbb{P}(\mathcal{A}) = 1/n!$. Write $\tilde{s}$ for a clock matrix with the conditional law of s given $\mathcal{A}$. Differentiating at $\theta = 0$ gives $n!\, \partial_{ij} Q_\theta(\mathrm{id})\big|_0 = \mathbb{E}[1 - \tilde{s}_{ij}]$ and

$$\mathsf{Q}''_{(i,j),(i',j')} := n!\, \partial_{ij}\partial_{i'j'} Q_\theta(\mathrm{id})\big|_0 = \mathbb{E}\big[(1 - \tilde{s}_{ij})(1 - \tilde{s}_{i'j'})\big] - \delta\, \mathbb{E}\tilde{s}_{ij},$$

where $\delta = 1$ when the two entries coincide and 0 otherwise.

The conditional law of the clocks. Appendix B.2 parametrizes the optimality event, for each choice of a root column r, by the column potentials y with $y_{\mathsf{r}} = 0$, the tree $\mathcal{T}$ of tight non-assigned entries with its parents p_i ($i \neq \mathsf{r}$), the diagonal clocks and the slacks. Every choice of root parametrizes the same conditional law of $\tilde{s}$, and each moment below is computed with a root outside the rows and the columns of the entries it involves. At zero scores every row of $S_0(y)$ is the same vector w, with $\mathsf{w}_j = e^{y_j} / \sum_{j'} e^{y_{j'}}$, so the diagonal clock of row i has rate $1/\mathsf{w}_i$ in Step 5 of Appendix B.2. Conditionally on $(\mathcal{T}, \mathsf{w})$, with $\mathsf{p}_{\mathsf{r}} = \mathsf{r}$ for the root row,

$$\tilde{s}_{ii} = \mathsf{g}_i \mathsf{w}_i, \qquad \tilde{s}_{i\mathsf{p}_i} = \mathsf{g}_i \mathsf{w}_{\mathsf{p}_i}, \qquad \tilde{s}_{ij} = \mathsf{g}_i \mathsf{w}_j + \mathsf{z}_{ij} \quad (j \notin \{i, \mathsf{p}_i\}),$$

with independent unit-rate exponential slacks z_{ij} and independent g_i of law $\mathrm{Gamma}(2,1)$ for $i \neq \mathsf{r}$ and unit-rate exponential for $i = \mathsf{r}$.

The same step gives the law of $(\mathcal{T}, \mathsf{w})$. The joint density of $(\mathcal{T}, y)$ on $\{y_{\mathsf{r}} = 0\}$ is $\prod_i \mathsf{w}_i \prod_{i \neq \mathsf{r}} \mathsf{w}_{\mathsf{p}_i}$, which conditioning on $\mathcal{A}$ multiplies by $n!$, and the map $y \mapsto \mathsf{w}$ onto the interior of the simplex has Jacobian $\prod_j \mathsf{w}_j$, so $(\mathcal{T}, \mathsf{w})$ has density $n! \prod_{i \neq \mathsf{r}} \mathsf{w}_{\mathsf{p}_i}$ there. Summing the product over the trees rooted at r, by the forest identity of Step 6 applied to the matrix with every row equal to w, gives w_{r}, so w has density $n!\, \mathsf{w}_{\mathsf{r}}$: it is Dirichlet with parameter 2 at r and 1 elsewhere. For distinct $i, j \neq \mathsf{r}$, the trees that contain a prescribed edge $i \to j$ contribute the factor w_j of that edge times the same sum with i and j merged into one column of weight $\mathsf{w}_i + \mathsf{w}_j$, which is again w_{r}. So $\mathbb{P}(\mathsf{p}_i = j \mid \mathsf{w}) = \mathsf{w}_j$. A tree has no cycle, so $\mathsf{p}_i = j$ and $\mathsf{p}_j = i$ cannot both hold.

First order. With a root other than i, $\mathbb{E}\tilde{s}_{ii} = \mathbb{E}\mathsf{g}_i\, \mathbb{E}\mathsf{w}_i = 2 \cdot \frac{1}{n+1}$. Adding a constant to a row of the scores leaves the law unchanged, so $\sum_j \partial_{ij} Q = 0$ and $\sum_j \mathbb{E}\tilde{s}_{ij} = n$, whence, the off-diagonal means of a row being equal by symmetry, $\mathbb{E}\tilde{s}_{ij} = \frac{n+2}{n+1}$ for $i \neq j$. For $\Pi\theta = \theta$ the entries of θ sum to zero, so the first-order term of $n!\, Q_\theta(\mathrm{id})$ is $-\sum_{i,j} \theta_{ij} \mathbb{E}\tilde{s}_{ij} = \frac{n}{n+1} \langle \theta, P_{\mathrm{id}} \rangle$. Compared with the first-order term $\langle \theta, P_{\mathrm{id}} \rangle$ of $n!\, p_\theta$, this gives $c_1 = \frac{n}{n+1}$.

Four second moments. Write $\mathsf{D} = (n+1)(n+2)$, the normalization of the second Dirichlet moments: away from the root $\mathbb{E}\mathsf{w}_j^2 = 2/\mathsf{D}$ and $\mathbb{E}[\mathsf{w}_j \mathsf{w}_{j'}] = 1/\mathsf{D}$ for $j \neq j'$. Take a root outside $\{1, 2\}$, which requires $n \geq 3$. Then $\mathbb{E}\tilde{s}_{11}^2 = \mathbb{E}\mathsf{g}_1^2\, \mathbb{E}\mathsf{w}_1^2 = \frac{12}{\mathsf{D}}$ and $\mathbb{E}[\tilde{s}_{11}\tilde{s}_{22}] = \mathbb{E}\mathsf{g}_1\, \mathbb{E}\mathsf{g}_2\, \mathbb{E}[\mathsf{w}_1 \mathsf{w}_2] = \frac{4}{\mathsf{D}}$, while $\mathbb{P}(\mathsf{p}_1 = 2 \mid \mathsf{w}) = \mathsf{w}_2$ and the absence of two-cycles give

$$\mathbb{E}\tilde{s}_{12}^2 = 6\, \mathbb{E}\mathsf{w}_2^2 + 4\, \mathbb{E}\big[\mathsf{w}_2(1 - \mathsf{w}_2)\big] + 2\, \mathbb{P}(\mathsf{p}_1 \neq 2) = \frac{4}{\mathsf{D}} + \frac{2}{n+1} + 2$$

and

$$\mathbb{E}[\tilde{s}_{12}\tilde{s}_{21}] = 4\, \mathbb{E}[\mathsf{w}_1 \mathsf{w}_2] + 2\, \mathbb{E}\big[\mathsf{w}_1(1 - \mathsf{w}_2)\big] + 2\, \mathbb{E}\big[\mathsf{w}_2(1 - \mathsf{w}_1)\big] + \mathbb{P}(\mathsf{p}_1 \neq 2, \mathsf{p}_2 \neq 1) = 1 + \frac{2}{n+1}.$$

The four values of Q'' that follow are $\frac{(n-1)(n-2)}{\mathsf{D}}$, $\frac{(n+1)(n-2)}{\mathsf{D}}$, $-\frac{n-2}{\mathsf{D}}$ and 0.

The other orbits. The Hessian Q'' is invariant under simultaneous permutations of the rows and the columns, hence constant on each orbit of pairs of entries, of which there are eleven once $n \geq 4$. Score invariance in the rows and in the columns gives $\sum_j \mathsf{Q}''_{(i,j),(i',j')} = 0$ and $\sum_i \mathsf{Q}''_{(i,j),(i',j')} = 0$ for every entry (i', j'). Two entries form a path when the column of one is the row of the other and together they use three indices, and a disjoint pair when they use four, which requires $n \geq 4$. The row and column sums determine the other orbits from these four values:

the two entries	$\mathsf{D}\,\mathsf{Q}''$
(i,i) and (i,i)	$(n-1)(n-2)$
(i,i) and (i',i')	$(n+1)(n-2)$
(i,i) and (i,i') or (i',i)	$-(n-2)$
(i,i) and (i',j')	$-n$
(i,j) and (i,j)	$-(n-2)$
(i,j) and (j,i)	0
(i,j) and (i,j') or (i',j)	2
(i,j) and (j,j') or (i',j')	1

with distinct letters standing for distinct indices. The last line covers both the paths and the disjoint pairs. The first, second, fifth and sixth lines are the four values just computed. The third line follows from the first and the fourth from the second and the third. The seventh follows from the third and the fifth, the paths from the third and the sixth, and the disjoint pairs from the fourth, the seventh and the paths.

Assembly. Write $\mathsf{S} = \langle \theta, P_{\mathrm{id}} \rangle$, $\mathsf{T} = \sum_i \theta_{ii}^2$ and $\mathsf{R} = \sum_{i,j} \theta_{ij}\theta_{ji}$. For $\Pi\theta = \theta$ the row and column sums of θ vanish, and summing $\theta_{ij}\theta_{i'j'}$ over the ordered pairs of entries in each orbit gives, in the order of the table and with its last line split into the paths and the disjoint pairs,

$$\mathsf{T}, \quad \mathsf{S}^2 - \mathsf{T}, \quad -4\mathsf{T}, \quad 2(2\mathsf{T} - \mathsf{S}^2), \quad \|\theta\|_F^2 - \mathsf{T},$$
$$\mathsf{R} - \mathsf{T}, \quad 2(2\mathsf{T} - \|\theta\|_F^2), \quad 2(2\mathsf{T} - \mathsf{R}), \quad \mathsf{S}^2 + \|\theta\|_F^2 + \mathsf{R} - 6\mathsf{T}.$$

Weighting these by the table and adding,

$$\mathsf{D} \sum_{i,j,i',j'} \mathsf{Q}''_{(i,j),(i',j')}\, \theta_{ij}\theta_{i'j'} = (n^2 + n - 1)\mathsf{S}^2 - (n+1)\|\theta\|_F^2 - \mathsf{R} - n\mathsf{T},$$

half of which, divided by D, is the second-order term of $n!\, Q_\theta(\mathrm{id})$. Undoing the relabeling gives the expansion of $n!\, Q_\theta(\sigma)$ in Corollary C.1. For the Gibbs distribution, $n!\, p_\theta(\mathrm{id}) = e^{\mathsf{S}} / \frac{1}{n!} \sum_\sigma e^{\langle \theta, P_\sigma \rangle}$, and under a uniform assignment $\langle \theta, P_\sigma \rangle$ has mean zero and second moment $\frac{\|\theta\|_F^2}{n-1}$, by

$$\mathbb{E}_{p_0}\left[(P_\sigma)_{ij}(P_\sigma)_{i'j'}\right] = \frac{1}{n^2} + \frac{\left(\delta_{ii'} - \frac{1}{n}\right)\left(\delta_{jj'} - \frac{1}{n}\right)}{n-1}.$$

Expanding numerator and denominator gives the stated expansion of $n!\, p_\theta(\sigma)$.

At $n = 2$ no root lies outside $\{1, 2\}$, but $\mathbb{E}\tilde{s}_{11}^2 = \frac{12}{\mathsf{D}} = 1$ still gives $\mathsf{Q}''_{(1,1),(1,1)} = 0$, and the row and column sums then make every second derivative vanish. The numerator of the second-order term of $n!\, Q_\theta$ in Corollary C.1 vanishes as well, every centered 2×2 score matrix being a multiple of $(e_1 - e_2)(e_1 - e_2)^\top$ for the standard basis vectors e_1, e_2. □

C.8 Calibrating Gumbel Matching

Section 4 states the following proposition in prose.

Proposition C.3 (second-order calibration of Gumbel matching)**.** *For square Gumbel matching with $n \geq 3$ and $\Pi\theta = \theta$, the calibration*

$$\eta(\theta) = \frac{n+1}{n}\,\theta + \frac{n+1}{2(n-2)(n+2)}\,\Pi(\theta \circ \theta)$$

is first-order exact and removes the whole second-order discrepancy inside $\operatorname{im} J$*. Outside* $\operatorname{im} J$ *the second-order term of p_θ has two components, proportional to the projections orthogonal to the constants and to* $\operatorname{im} J$ *of the sum and of the difference of $\sigma \mapsto \langle \theta, P_\sigma \rangle^2$ and $\sigma \mapsto \sum_{i,j} \theta_{i\sigma(j)}\theta_{j\sigma(i)}$, the first of which is absent at $n = 3$, and on them the second-order term of the calibrated law is, respectively, $1 + 2r' = 1 - \frac{1}{n^2}$ and $1 + 2r'' = 1 + \frac{1}{n(n+2)}$ times that of p_θ, so no deterministic calibration is second-order exact. The randomized calibration constructed in the proof, which draws a pair of columns and a sign, is second-order exact.*

Proof of Proposition C.3. Three functions of the assignment. The proof is a comparison of coefficients in three functions of σ. Fix θ with $\Pi\theta = \theta$, and carry over $\mathsf{D} = (n+1)(n+2)$ and the three functions of Appendix C.7, now with a score matrix ϑ as their argument:

$$\mathsf{S}(\vartheta) = \langle \vartheta, P_\sigma \rangle, \qquad \mathsf{T}(\vartheta) = \langle \vartheta \circ \vartheta, P_\sigma \rangle, \qquad \mathsf{R}(\vartheta) = \sum_{i,j} \vartheta_{i\sigma(j)}\vartheta_{j\sigma(i)},$$

each written without its argument when $\vartheta = \theta$. By Corollary C.1 the second-order terms of $n!\, Q_\theta$ and of $n!\, p_\theta$ are combinations of S^2, R, T and a constant. The constant-score directions of assignments are the matrices with entries $\mathsf{a}_i + \mathsf{b}_j$, so the first-order term of Corollary C.1 makes $\operatorname{im} J$ the space of functions $\sigma \mapsto \langle \mathsf{C}, P_\sigma \rangle$ with $\Pi\mathsf{C} = \mathsf{C}$, and T lies in $\operatorname{im} J$ up to a constant, $\langle \theta \circ \theta, P_\sigma \rangle = \langle \Pi(\theta \circ \theta), P_\sigma \rangle + \mathrm{const}$. In the notation of Appendix C.5, a calibration with linear part

θ/c_1 and quadratic part η_2 has second-order term $c_1^{-2}Q_2(\theta) + \frac{c_1}{n!}\langle \Pi\eta_2(\theta), P_\sigma\rangle$, so the quadratic part $\eta_2 = \gamma_{\det}\,\Pi(\theta\circ\theta)$ changes only the coefficient of T and the constant.

Independence. For $n \geq 4$ the functions S^2, R and T are independent modulo the constants as θ ranges over the centered score matrices: a combination of them with constant coefficients that is constant in σ for every centered θ has all three coefficients zero. Their coefficients in a quadratic term are therefore well defined. Take the alternating sum of a function of the assignment over id, (12), (34) and $(12)(34)$, with the signs $+,-,-,+$. It annihilates the constants and $\operatorname{im} J$, because $P_{\mathrm{id}} - P_{(12)} - P_{(34)} + P_{(12)(34)} = 0$. On the score matrix supported on the diagonal blocks $\{1,2\}\times\{1,2\}$ and $\{3,4\}\times\{3,4\}$, with entries 1 on the diagonal and -1 off it, the alternating sum of S^2 is 32 and that of R is 0. On the matrix supported on the off-diagonal blocks $\{1,2\}\times\{3,4\}$ and $\{3,4\}\times\{1,2\}$, with ±1 in the pattern of the same 2×2 block, they are 0 and 32. So a combination of S^2, R and T that is constant in σ for every centered θ has no S^2 and no R, and then no T, the second matrix having $\mathsf{T}(\mathrm{id}) = 0$ and $\mathsf{T}\big((13)(24)\big) = 4$.

The two components. The chain rule of Appendix C.5 applies verbatim, Proposition C.1 covering square assignments: a calibration is first-order exact if and only if its differential at 0 is θ/c_1 up to constant-score directions, and its second-order discrepancy is then $c_1^{-2}Q_2(\theta) - p_2(\theta) + J\eta_2(\theta)$, whose class modulo $\operatorname{im} J$ does not depend on η_2. With $c_1 = \frac{n}{n+1}$, Corollary C.1 gives

$$n!\big(c_1^{-2}Q_2(\theta) - p_2(\theta)\big) = -\frac{\mathsf{S}^2 + (n+1)\mathsf{R}}{2n^2(n+2)} - \frac{(n+1)\mathsf{T}}{2n(n+2)} + \text{const.}$$

Modulo $\operatorname{im} J$ and the constants the T term and the constant drop, and rewriting what is left in $\mathsf{S}^2 \pm \mathsf{R}$ gives

$$2r'\cdot\frac{\mathsf{S}^2+\mathsf{R}}{4} + 2r''\cdot\frac{\mathsf{S}^2-\mathsf{R}}{4}, \qquad r' = -\frac{1}{2n^2}, \quad r'' = \frac{1}{2n(n+2)},$$

against $\frac{\mathsf{S}^2+\mathsf{R}}{4} + \frac{\mathsf{S}^2-\mathsf{R}}{4}$ for the second-order term of $n!\,p_\theta$. These are the two components, and for $n \geq 4$ neither is zero: on the first of the two matrices above, the alternating sum gives 32 for both $\mathsf{S}^2+\mathsf{R}$ and $\mathsf{S}^2-\mathsf{R}$.

At $n = 3$ the first component vanishes, and the last of the orbit sums of Appendix C.7 gives the reason: it runs over the pairs of entries with four distinct indices, of which there are none when $n = 3$, and its value is $\mathsf{S}^2 + \|\theta\|_F^2 + \mathsf{R} - 6\mathsf{T}$. Relabeling the columns by σ, which leaves θ centered, gives $\mathsf{S}^2 + \mathsf{R} = 6\mathsf{T} - \|\theta\|_F^2$ for every assignment, so $\mathsf{S}^2 + \mathsf{R}$ lies in $\operatorname{im} J$ up to a constant and the first component is zero. The second is not: the signed sum $\sum_\sigma \operatorname{sgn}(\sigma)\mathsf{f}(\sigma)$ annihilates the constants and $\operatorname{im} J$, because $\sum_\sigma \operatorname{sgn}(\sigma)P_\sigma = 0$, and on the centered θ with rows $(1,0,-1)$, $(0,1,-1)$ and $(-1,-1,2)$ it gives 18 for S^2, hence 36 for $\mathsf{S}^2-\mathsf{R}$. So no deterministic calibration of square Gumbel matching is second-order exact for $n \geq 3$.

The quadratic term. The component inside $\operatorname{im} J$ of a function f of the assignments is read off from the array $\mathbb{E}_{p_0}[\mathsf{f} \mid \sigma(i) = j]$: f is orthogonal to $\operatorname{im} J$ exactly when Π annihilates that array, since $\mathbb{E}_{p_0}\big[\mathsf{f}\,\langle \mathsf{C}, P_\sigma\rangle\big] = \frac{1}{n}\sum_{i,j}\mathsf{C}_{ij}\,\mathbb{E}_{p_0}[\mathsf{f} \mid \sigma(i) = j]$. Condition on $\sigma(i) = j$: the other values form a uniform bijection σ' of the remaining rows onto the remaining columns, and the matrix θ' without row i and column j has row sums $-\theta_{i'j}$, column sums $-\theta_{ij'}$ and total θ_{ij}, because θ is centered. Hence $\mathbb{E}\langle\theta', P_{\sigma'}\rangle = \frac{\theta_{ij}}{n-1}$, and, writing Π' for the double centering of an $(n-1)\times(n-1)$ matrix, the second moment of Appendix C.7 applied to $\Pi'\theta'$ gives

$$\mathbb{E}\langle\theta', P_{\sigma'}\rangle^2 = \frac{\|\Pi'\theta'\|_F^2}{n-2} + \frac{\theta_{ij}^2}{(n-1)^2}, \qquad \|\Pi'\theta'\|_F^2 = \|\theta'\|_F^2 - \frac{\sum_{i'\neq i}\theta_{i'j}^2 + \sum_{j'\neq j}\theta_{ij'}^2}{n-1} + \frac{\theta_{ij}^2}{(n-1)^2},$$

and the second of these is $\frac{n^2}{(n-1)^2}\theta_{ij}^2$ plus terms depending on i alone or on j alone. Since $\mathsf{S} = \theta_{ij} + \langle\theta', P_{\sigma'}\rangle$ on the event, $\mathbb{E}[\mathsf{S}^2 \mid \sigma(i) = j]$ is $\frac{n^2}{(n-1)(n-2)}\theta_{ij}^2$ up to such terms. The same reduction applies to R: on the event it equals $\theta_{ij}^2 + 2\sum_{i'\neq i}\theta_{i\sigma(i')}\theta_{i'j} + \mathsf{R}(\theta')$, the last term being R evaluated at θ' and σ'. The conditional expectation is then $\theta_{ij}^2 + \frac{2\theta_{ij}^2}{n-1} + \mathbb{E}\,\mathsf{R}(\theta')$ with

$$\mathbb{E}\,\mathsf{R}(\theta') = \frac{\|\theta'\|_F^2}{n-1} + \frac{\theta_{ij}^2 - \sum_{i'\neq i}\theta_{i'j}^2 - \sum_{j'\neq j}\theta_{ij'}^2 + \|\theta'\|_F^2}{(n-1)(n-2)},$$

and collecting the coefficient of θ_{ij}^2 gives $\frac{n^2}{(n-1)(n-2)}$ again. For a member of $\operatorname{im} J$ the reduction gives $\mathbb{E}\big[\langle \mathsf{C}, P_\sigma\rangle \mid \sigma(i) = j\big] = \frac{n}{n-1}\mathsf{C}_{ij}$. Comparing the arrays shows that both S^2 and R have component $\frac{n}{n-2}\langle \Pi(\theta\circ\theta), P_\sigma\rangle$ inside $\operatorname{im} J$, while T has $\langle \Pi(\theta\circ\theta), P_\sigma\rangle$. Requiring the discrepancy displayed above, corrected by $c_1\gamma_{\det}\,\langle \Pi(\theta\circ\theta), P_\sigma\rangle$, to have nothing inside $\operatorname{im} J$ gives

$$c_1\gamma_{\det} = \frac{n+1}{2n(n+2)} + \frac{1}{2n(n-2)} = \frac{n}{2(n-2)(n+2)}, \qquad \gamma_{\det} = \frac{n+1}{2(n-2)(n+2)},$$

the quadratic coefficient of Proposition C.3.

The randomized calibration. Let $B_{jj'}$ be the $n\times n$ matrix with entries 1 at (j,j') and (j',j) and 0 elsewhere, so that $\theta B_{jj'}$ carries column j' of θ to position j and column j to position j', and vanishes in the other columns. Draw a pair $\{j,j'\}$ of columns and a sign uniformly at random and take

$$\eta(\theta) = \frac{n+1}{n}\Big[(1\pm\varepsilon\iota)\,\theta \pm \varepsilon\,\theta B_{jj'}\Big] + \gamma_{\mathrm{mix}}\,\Pi(\theta\circ\theta),$$

with $\varepsilon^2 = \frac{n^2}{2(n+1)^2}$, $\iota = \frac{2+\sqrt{2(n+1)}}{n(n-1)}$ and $\gamma_{\mathrm{mix}} = \frac{n+3}{2(n-1)(n+2)}$, determined by the matching of coefficients below. Averaging over the sign leaves the linear part θ/c_1, so the mixture is first-order exact. For the second-order term only the centered part of the linear map counts, so write the linear part as $\mathsf{A}\theta/c_1$ with $\mathsf{A}\theta = \theta \pm \varepsilon\,\theta(\mathsf{K}+\iota P)$, $P = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ and $\mathsf{K} = PB_{jj'}P$: here $\theta(\mathsf{K}+\iota P)$ differs from $\iota\theta + \theta B_{jj'}$ by a matrix with constant rows.

The construction uses three averages over the pair. The first concerns $\mathsf{S}(\theta B_{jj'}) = \theta_{\sigma^{-1}(j)j'} + \theta_{\sigma^{-1}(j')j}$, whose mean is $-\mathsf{S}/\binom{n}{2}$ by the vanishing row sums and whose mean square is $\big[(\|\theta\|_F^2 - \mathsf{T}) + (\mathsf{R}-\mathsf{T})\big]/\binom{n}{2}$. The cross term of R, namely $2\operatorname{tr}\big(\theta P_\sigma^\top\,\theta B_{jj'}P_\sigma^\top\big)$, has mean $-2\mathsf{R}/\binom{n}{2}$, and $\mathsf{R}(\theta B_{jj'})$ has mean $\big[(\|\theta\|_F^2 - \mathsf{T}) + (\mathsf{S}^2-\mathsf{T})\big]/\binom{n}{2}$. Finally $\theta\mathsf{K} = \theta B_{jj'} - \frac{1}{n}\theta B_{jj'}\mathbf{1}\mathbf{1}^\top$ has mean $-\theta/\binom{n}{2}$, and the mean of its entrywise square is $\frac{1}{\binom{n}{2}}\big(\frac{4}{n}-1\big)\theta\circ\theta$ up to a matrix with constant rows. The sign average cancels the terms odd in the perturbation, so with

$$\mathsf{t} = 1 + \varepsilon^2\Big(\iota^2 - \frac{2}{\binom{n}{2}}\,\iota\Big), \qquad \mathsf{t}' = \frac{\varepsilon^2}{\binom{n}{2}},$$

averaging over the sign and the pair gives, up to constants,

$$\mathbb{E}\,\mathsf{S}(\mathsf{A}\theta)^2 = \mathsf{t}\,\mathsf{S}^2 + \mathsf{t}'(\mathsf{R}-2\mathsf{T}), \qquad \mathbb{E}\,\mathsf{R}(\mathsf{A}\theta) = \mathsf{t}\,\mathsf{R} + \mathsf{t}'(\mathsf{S}^2-2\mathsf{T}),$$
$$\mathbb{E}\,\mathsf{T}(\mathsf{A}\theta) = \Big(\mathsf{t} + \mathsf{t}'\Big(\frac{4}{n}-1\Big)\Big)\mathsf{T}.$$

By Corollary C.1 the second-order term of $n!$ times the mixture is, up to a constant,

$$\frac{1}{2\mathsf{D}c_1^2}\Big[(n^2+n-1)\,\mathbb{E}\,\mathsf{S}(\mathsf{A}\theta)^2 - \mathbb{E}\,\mathsf{R}(\mathsf{A}\theta) - n\,\mathbb{E}\,\mathsf{T}(\mathsf{A}\theta)\Big] + c_1\gamma_{\mathrm{mix}}\,\mathsf{T},$$

and it must equal the second-order term $\frac{\mathsf{S}^2}{2}$ of $n!\,p_\theta$. Matching the coefficients of S^2 and of R gives $(n^2+n-1)\mathsf{t} - \mathsf{t}' = \mathsf{D}c_1^2$ and $(n^2+n-1)\mathsf{t}' = \mathsf{t}$, so $\mathsf{t}'\big[(n^2+n-1)^2 - 1\big] = \mathsf{D}c_1^2$, that is, $\mathsf{t}'\,n(n-1)(n+1)(n+2) = \frac{n^2(n+2)}{n+1}$, and

$$\mathsf{t}' = \frac{n}{(n-1)(n+1)^2}, \qquad \mathsf{t} = 1 + \frac{1}{(n-1)(n+1)^2}.$$

Hence $\varepsilon^2 = \binom{n}{2}\mathsf{t}' = \frac{n^2}{2(n+1)^2}$, and ι is the positive root of $\iota^2 - \frac{2}{\binom{n}{2}}\iota = \frac{\mathsf{t}-1}{\varepsilon^2} = \frac{2}{n^2(n-1)}$, namely $\iota = \frac{2+\sqrt{2(n+1)}}{n(n-1)}$. Matching the coefficient of T then gives $c_1\gamma_{\mathrm{mix}} = \frac{\mathsf{t}'\,n^2(n+3)}{2\mathsf{D}c_1^2}$, that is,

$$\gamma_{\mathrm{mix}} = \frac{n+3}{2(n-1)(n+2)}.$$

The three coefficients now agree, so the two second-order terms differ by a constant, which is zero because both sum to zero over the assignments. A finite mixture of laws with $O(\|\theta\|_F^3)$ remainders has an $O(\|\theta\|_F^3)$ remainder, and the mixture is second-order exact. At $n=3$ the identity above makes S^2, R and T dependent modulo constants, so exactness requires fewer than three equalities, which the constants above satisfy. □

D THE REWEIGHTED ESTIMATOR

Appendices D.1 and D.2 state Algorithms 1 and 2 with their settings, Appendix D.3 proves that both are consistent, Appendix D.4 relates the temperature to the effective sample size (Proposition D.1), Appendix D.5 describes the subset dynamic program that computes the true marginals, and Appendix D.6 validates both algorithms against it.

D.1 ALGORITHM 1

The potential samples of Algorithm 1 come from MALA (Roberts & Tweedie, 1996) on $\log \psi_\eta$ in the coordinates $y_2, \dots, y_n$, with $y_1 = 0$, where $\partial \log \psi_\eta(y)/\partial y_j = 1 - \sum_i S_\eta(y)_{ij}$. From y, a step of size h proposes $y' = y + \frac{h}{2}\nabla \log \psi_\eta(y) + \sqrt{h}\,\mathsf{z}$ for a standard Gaussian vector z, and moves to y' with probability

$$\min\Big\{1,\ \frac{\psi_\eta(y')\exp\big(-\|y - y' - \frac{h}{2}\nabla \log \psi_\eta(y')\|^2/(2h)\big)}{\psi_\eta(y)\exp\big(-\|y' - y - \frac{h}{2}\nabla \log \psi_\eta(y)\|^2/(2h)\big)}\Big\},$$

the Metropolis–Hastings probability for this Gaussian proposal. Otherwise the chain stays at y. The 256 chains start at the Sinkhorn column potentials y^* of $\exp(\eta)$ (Proposition B.2) plus independent Gaussian perturbations of standard deviation 0.1 in each coordinate. The step size starts at $h = 0.5$. During a burn-in of 300 steps it is multiplied after every step by $\exp\big(0.3\,(\text{acceptance} - 0.57)\big)$, the acceptance being the fraction of chains that moved at that step, and it is fixed afterwards. Each chain then keeps 128 of its states, 4 steps apart, which gives $T_y = 256 \cdot 128 = 32768$ potential samples. For every potential sample, $P_\sigma^\top S_\eta(y)$ is $S_\eta(y)$ with its rows permuted, so one row-softmax matrix per sample serves all the determinants of line 5.

D.2 ALGORITHM 2

For a nonempty shared-column set $\mathcal{G}$, the chains of Algorithm 2 run in the coordinates $\upsilon_j = \log(-y_j)$, $j \in \mathcal{G}$, which remove the constraints $y_j \le 0$, and every set is padded to n coordinates so that the chains of many sets run in one batch. With padding coordinates $\upsilon' \in \mathbb{R}^{n-|\mathcal{G}|}$, the padded density is

$$\mathsf{f}_\mathcal{G}(\upsilon, \upsilon') = \psi_\eta(y)\, e^{\sum_{j\in\mathcal{G}} \upsilon_j} \prod_{j'=1}^{n-|\mathcal{G}|} \frac{e^{-\upsilon'_{j'}}}{\big(1 + e^{-\upsilon'_{j'}}\big)^2}, \qquad y_j = -e^{\upsilon_j}\ (j \in \mathcal{G}), \quad y_j = 0\ (j \notin \mathcal{G}).$$

The factor $e^{\sum_j \upsilon_j}$ is the Jacobian of $y_j = -e^{\upsilon_j}$, and the padding coordinates carry standard logistic densities, which integrate to one. So $\int \mathsf{f}_\mathcal{G} = \Psi(\mathcal{G})$, and the potentials y of a point drawn from $\mathsf{f}_\mathcal{G}/\Psi(\mathcal{G})$ have the density $\psi_\eta/\Psi(\mathcal{G})$ of Section 5.2. MALA runs on $\log \mathsf{f}_\mathcal{G}$ as in Appendix D.1, with the gradient

$$\frac{\partial \log \mathsf{f}_\mathcal{G}}{\partial \upsilon_j} = 1 - e^{\upsilon_j}\Big(1 - \sum_i S_\eta(y)_{ij}\Big) \quad (j \in \mathcal{G}), \qquad \frac{\partial \log \mathsf{f}_\mathcal{G}}{\partial \upsilon'_{j'}} = -\tanh\big(\upsilon'_{j'}/2\big).$$

Each set has 64 chains. From the origin, 40 gradient steps $(\upsilon, \upsilon') \leftarrow (\upsilon, \upsilon') + 0.3\,\nabla \log \mathsf{f}_\mathcal{G}(\upsilon, \upsilon')$, each taken only if it does not decrease $\log \mathsf{f}_\mathcal{G}$, give the starting point, and the chains start there plus independent Gaussian perturbations of standard deviation 0.05 in each coordinate. Their step size starts at 0.3 and is adapted as in Appendix D.1, from the acceptance of the set's own chains, during a burn-in of 10000 steps. Each chain then keeps 128 of its states, 8 steps apart, which gives $T_y = 64 \cdot 128 = 8192$ potential samples per set. The determinants of line 9 are computed from the matrices $I - P_\sigma^\top S_\eta(y)$ restricted to $\mathcal{G}$ and completed by an identity block to size n, which leaves their values unchanged and lets all assignments run in one batch.

The normalizer $\Psi(\mathcal{G})$ is estimated by annealed importance sampling (Neal, 2001) through the densities $\mathsf{g}^{1-\mathsf{q}}\mathsf{f}_\mathcal{G}^{\mathsf{q}}$ with $\mathsf{q} = 0, (1/300)^{1.5}, (2/300)^{1.5}, \dots, 1$, where g is the Gaussian density with the mean and the covariance of the T_y kept states, the covariance increased by 10^{-6} on the diagonal. In the forward run each of 256 particles starts at a draw from g with weight one. For each nonzero value of q in increasing order, each particle multiplies its weight by the ratio of the density at q to

Algorithm 1 The reweighted estimator on square assignments

Input: scores $\theta \in \mathbb{R}^{n\times n}$; sample sizes T and T_y

Output: the estimate $\hat{M}$ of $M(\theta)$ and ESS/T

1: $\eta \leftarrow \frac{n+1}{n}\,\theta$ ▷ the temperature $c = \frac{n}{n+1}$

2: draw $\xi_1, \dots, \xi_T$ with independent standard Gumbel entries and set $\sigma_t \leftarrow \arg\max_\sigma \langle \eta + \xi_t, P_\sigma \rangle$ by the Hungarian algorithm

3: $T(\sigma) \leftarrow |\{t : \sigma_t = \sigma\}|$ for every sampled σ; σ is frequent if $T(\sigma) \geq 20$

4: run MALA chains on $\log \psi_\eta$ over $y_2, \dots, y_n$ with $y_1 = 0$, started near the Sinkhorn column potentials, and keep T_y samples $y^{(1)}, \dots, y^{(T_y)}$

5: $\mathsf{L}(\sigma) \leftarrow e^{\langle \eta, P_\sigma \rangle}\, \frac{1}{T_y} \sum_{t=1}^{T_y} \det\left(I - P_\sigma^\top S_\eta(y^{(t)}) + \frac{1}{n}\mathbf{1}\mathbf{1}^\top\right)$ for every sampled σ

6: $\log \hat{\Psi}_0 \leftarrow$ the $T(\sigma)$-weighted mean of $\log \frac{T(\sigma)}{T} - \log \mathsf{L}(\sigma)$ over the frequent σ, or over all sampled σ if none is frequent

7: $\hat{Q}_\eta(\sigma) \leftarrow T(\sigma)/T$ if σ is frequent, else $\hat{Q}_\eta^{\mathrm{law}}(\sigma) = \hat{\Psi}_0\, \mathsf{L}(\sigma)$

8: $w_t \leftarrow e^{\langle \theta, P_{\sigma_t} \rangle} / \hat{Q}_\eta(\sigma_t)$ for $t = 1, \dots, T$

9: **return** $\hat{M} = \sum_t w_t P_{\sigma_t} \big/ \sum_t w_t$ and $\mathrm{ESS}/T = (\sum_t w_t)^2 \big/ \left(T \sum_t w_t^2\right)$

Cost: T Hungarian solves, $O(Tn^3)$; $O(n^2)$ per step of a chain; T_y determinants of size n, $O(T_y n^3)$, per distinct sampled assignment.

Algorithm 2 The reweighted estimator on rectangular assignments with private columns

Input: scores $\theta \in (\mathbb{R} \cup \{-\infty\})^{n\times(m+n)}$ with private columns; sample sizes T and T_y; an annealing budget

Output: the estimate $\hat{M}$ of $M(\theta)$ and ESS/T

1: $\eta \leftarrow \frac{n+1}{n}\,\theta$ on the finite entries

2: draw $\xi_1, \dots, \xi_T$ with independent standard Gumbel entries on the finite entries and set $\sigma_t \leftarrow \arg\max_\sigma \langle \eta + \xi_t, P_\sigma \rangle$ by the rectangular Hungarian algorithm

3: $T(\sigma) \leftarrow |\{t : \sigma_t = \sigma\}|$ and $\mathcal{G}(\sigma) \leftarrow$ the shared-column set of σ, for every sampled σ

4: **for** every nonempty shared-column set $\mathcal{G}$ of a sampled assignment **do**

5: run MALA chains for the density proportional to ψ_η on $\{y_j \leq 0,\ j \in \mathcal{G}\}$, $y_j = 0$ for $j \notin \mathcal{G}$, in the coordinates $\log(-y_j)$, and keep T_y samples $y^{(1)}, \dots, y^{(T_y)}$

6: $\log \hat{\Psi}(\mathcal{G}) \leftarrow$ the midpoint of the forward and the reverse annealed importance sampling estimates of $\log \Psi(\mathcal{G})$

7: **end for**

8: $\hat{\Psi}(\varnothing) \leftarrow \psi_\eta(0)$

9: $\mathsf{L}(\sigma) \leftarrow \hat{\Psi}(\mathcal{G})\, e^{\langle \eta, P_\sigma \rangle}\, \frac{1}{T_y} \sum_{t=1}^{T_y} \det\left((I - P_\sigma^\top S_\eta(y^{(t)}))_{\mathcal{G}}\right)$ over the samples of $\mathcal{G} = \mathcal{G}(\sigma)$, the average being 1 if $\mathcal{G} = \varnothing$, for every sampled σ

10: lines 6 to 9 of Algorithm 1 ▷ $\hat{M}$ has n rows and $m + n$ columns

Cost: T rectangular Hungarian solves; per nonempty shared-column set of a sampled assignment, one family of chains and one annealing run in each direction; T_y determinants of size at most n, $O(T_y n^3)$, per distinct sampled assignment.

the density at the previous value, both at its current point, and then takes 6 MALA steps for the density at q with the step size of the chains. The forward estimate is the logarithm of the mean final weight, and the expectation of this logarithm is at most $\log \Psi(\mathcal{G})$ (Grosse et al., 2015). The reverse run starts the 256 particles from the last four states kept by each chain and goes through the values of q in decreasing order in the same way, so that its ratios are the reciprocals of the forward run's. The reverse estimate is minus the logarithm of the mean final weight, and $\log \hat{\Psi}(\mathcal{G})$ is the midpoint of the two estimates. The numbers of particles and of intermediate densities form the annealing budget of Algorithm 2.

D.3 CONSISTENCY

The frequency of a sampled assignment converges to its probability by the strong law of large numbers. The law-based estimate converges as T and the length of the chains grow, by the ergodic theorem for the chains and by the alignment fit, and in Algorithm 2 also under the convergence of the normalizer estimates. The reweighted estimator is a ratio of finite sums, and once every allowed assignment is frequent it is the enumeration of the Gibbs distribution, whatever the chains have produced. The argument is that of self-normalized importance sampling (Owen, 2013, Theorem 9.2), with the proposal probabilities replaced by consistent estimates.

Lemma D.1 (consistency of the reweighted estimator)**.** *Fix the scores θ, and let line 1 of either algorithm use any temperature $c > 0$ in place of $\frac{n}{n+1}$. (a) As T grows, whatever the length of the chains, the output $\hat{M}$ of either algorithm converges to $M(\theta)$ almost surely, and equals $M(\theta)$ once every allowed assignment is frequent. (b) As T and the length of the chains grow, the law-based estimate $\hat{Q}_\eta^{\text{law}}(\sigma)$ of every allowed assignment converges to $Q_\eta(\sigma)$ almost surely in Algorithm 1, and in Algorithm 2 if, in addition, $\log \hat{\Psi}(\mathcal{G}) \to \log \Psi(\mathcal{G})$ almost surely for every shared-column set $\mathcal{G}$.*

Proof. The frequencies. Every allowed assignment is the unique optimum of the perturbed problem on an open set of noise values, and the noise has a positive density on the allowed entries, so $Q_\eta(\sigma) > 0$ for every allowed σ. By the strong law of large numbers, almost surely $T(\sigma)/T \to Q_\eta(\sigma)$ for every allowed σ, and from some T on, every allowed assignment is sampled and frequent.

The chain averages. After its burn-in, a chain of Appendix D.1 is a Metropolis–Hastings chain with a fixed step size. Its invariant probability density and its proposal density are positive everywhere, so it is irreducible and aperiodic. It is then positive Harris recurrent, and its distribution after k steps converges to the invariant distribution in total variation from every starting point (Tierney, 1994, Theorem 1 and Corollary 2). So does the chain of its kept states, which after k steps is distributed as the original chain after $4k$ steps. This chain is therefore positive Harris recurrent and aperiodic as well, and from every starting point the average of a bounded function over the kept states converges almost surely to the expectation of that function under the invariant density (Tierney, 1994, p. 1712 and Theorem 3). The same holds for the chains of Appendix D.2, whose padded density is positive everywhere, and for bounded functions of their potentials y. The determinants of line 5 of Algorithm 1 and line 9 of Algorithm 2 are bounded, since their matrices have entries in $[-1, 2]$, so their averages over the potential samples converge almost surely to their expectations under the densities ψ_η/Ψ and $\psi_\eta/\Psi(\mathcal{G})$.

The law-based estimates. Let $\mathsf{L}(\sigma) = \hat{Q}_\eta^{\text{law}}(\sigma)/\hat{\Psi}_0$ be the law-based estimate without its alignment constant. By Theorem 2, $\mathsf{L}(\sigma) \to \frac{n}{\Psi} Q_\eta(\sigma)$ almost surely in Algorithm 1. By Lemma B.1 and the convergence of the normalizer estimates, with $\hat{\Psi}(\varnothing) = \Psi(\varnothing)$, $\mathsf{L}(\sigma) \to Q_\eta(\sigma)$ almost surely in Algorithm 2. Once every allowed assignment is frequent, $\log \hat{\Psi}_0$ is the average of $\log \frac{T(\sigma)}{T} - \log \mathsf{L}(\sigma)$ over the finitely many allowed assignments with weights $T(\sigma)$, and every term converges to $\log \frac{\Psi}{n}$ in Algorithm 1 and to 0 in Algorithm 2. So $\hat{\Psi}_0 \to \frac{\Psi}{n}$ in Algorithm 1 and $\hat{\Psi}_0 \to 1$ in Algorithm 2, and $\hat{Q}_\eta^{\text{law}}(\sigma) = \hat{\Psi}_0 \, \mathsf{L}(\sigma) \to Q_\eta(\sigma)$ in both.

The estimator. Once every allowed assignment is sampled,

$$\hat{M} = \frac{\sum_\sigma \frac{T(\sigma)}{T \hat{Q}_\eta(\sigma)} e^{\langle \theta, P_\sigma \rangle} P_\sigma}{\sum_\sigma \frac{T(\sigma)}{T \hat{Q}_\eta(\sigma)} e^{\langle \theta, P_\sigma \rangle}},$$

with both sums over the allowed assignments. Once every allowed assignment is frequent, line 7 selects the frequency $T(\sigma)/T$ for each of them, every ratio $T(\sigma)/(T\hat{Q}_\eta(\sigma))$ equals one, and $\hat{M} = \sum_\sigma e^{\langle \theta, P_\sigma \rangle} P_\sigma / Z(\theta) = M(\theta)$ whatever the chains have produced, which proves (a). The convergence of the law-based estimates above proves (b). □

D.4 TEMPERATURE

This subsection relates the temperature to the variance of importance sampling and to the effective sample size, for the weights $p_\theta/Q_{\theta/c}$, which the algorithms estimate up to a common factor. Fix a score matrix θ, square or rectangular, with at least two allowed assignments. At a temperature $c > 0$, let $\tilde{M}$ be the self-normalized estimator from T independent assignments σ_t drawn from $Q_{\theta/c}$ with the weights $p_\theta(\sigma_t)/Q_{\theta/c}(\sigma_t)$. Its summed asymptotic variance is

$$V(c) = \lim_{T\to\infty} T \sum_{i,j} \operatorname{Var} \tilde{M}_{ij}.$$

Let $D(\sigma) = \|P_\sigma - M(\theta)\|_F^2$, the squared distance of an assignment matrix from the marginals, and $\bar{D} = \mathbb{E}_{p_\theta} D$. Let $D_{\min}$ and $D_{\max}$ be the smallest and the largest value of D over the allowed assignments. Here $D_{\min} > 0$, because $M(\theta)$ has an entry strictly between 0 and 1. The χ^2 divergence of p_θ from $Q_{\theta/c}$ is $\chi^2 = \sum_\sigma p_\theta(\sigma)^2/Q_{\theta/c}(\sigma) - 1$. The population effective sample size is $T/(1+\chi^2)$. The covariance under p_θ of the weights and D is

$$R(c) = \mathbb{E}_{p_\theta}\Big[\Big(\frac{p_\theta}{Q_{\theta/c}} - 1\Big)\big(D - \bar{D}\big)\Big].$$

On a fixed compact interval of temperatures, c_E is a temperature that maximizes the population effective sample size and c_V one that minimizes V. Both exist because the law is continuous in the scores.

Proposition D.1 (temperature by effective sample size)**.** *(a) For every temperature $c > 0$,*

$$V(c) = \bar{D}\,(1+\chi^2) + R(c) \qquad \textit{and} \qquad D_{\min}(1+\chi^2) \le V(c) \le D_{\max}(1+\chi^2).$$

Hence $V(c_E) \le \frac{D_{\max}}{D_{\min}} V(c_V)$.

(b) Let the assignments be square with all scores finite and $n \ge 3$, and let $\theta \to 0$ along centered scores, $\Pi\theta = \theta \ne 0$. If the interval contains $\frac{n}{n+1}$ in its interior, then $c_E \to \frac{n}{n+1}$. If it also contains $\frac{(n-1)^2}{(n+1)(n-2)}$ in its interior, then $c_V \to \frac{(n-1)^2}{(n+1)(n-2)}$ and

$$\frac{V(c_E)}{V(c_V)} - 1 = \frac{\|\theta\|_F^2}{(n-1)^5} + O\big(\|\theta\|_F^3\big);$$

the same expansion holds with $\frac{n}{n+1}$ in place of c_E.

Part (a) separates what the temperature controls: the variance is the spread $\bar{D}$ of the assignment matrices, which does not depend on the temperature, times $1+\chi^2$, plus the covariance $R(c)$. The two bounds of part (a) make the effective sample size a measure of the variance of $\tilde{M}$ that needs no true marginals. At a fixed temperature, the fraction ESS/T that the algorithms report tends to $1/(1+\chi^2)$ as T grows and the estimated probabilities converge, by the law of large numbers applied to the weights $p_\theta/Q_{\theta/c}$, whose mean under the proposal is one and whose second moment is $1+\chi^2$, a common factor leaving the fraction unchanged. The temperature c_E with the most stable weights and the temperature c_V with the least variance can differ, but part (a) bounds the variance at c_E by $\frac{D_{\max}}{D_{\min}}$ times the least variance, whatever the scale of the scores. Part (b) locates both as the scores tend to zero: c_E tends to the first-order temperature, and c_V to a temperature higher by $\frac{1}{(n+1)(n-2)}$, whose proposal is slightly more dispersed. At the first-order temperature the variance exceeds the least variance by the factor $1 + \|\theta\|_F^2/(n-1)^5$ to leading order. Appendix E.5 compares the first-order temperature with the maximizer of the effective sample size on instances of Section 6.

Part (a) is an identity for the asymptotic variance of self-normalized importance sampling, and part (b) expands its terms at small scores with Corollary C.1.

Proof of Proposition D.1. (a) Write $\mathsf{w} = p_\theta/Q_{\theta/c}$, positive on the allowed assignments, with $\mathbb{E}_{Q_{\theta/c}}\mathsf{w} = 1$. Fix a function h of the assignment with mean $\bar{\mathsf{h}} = \mathbb{E}_{p_\theta}\mathsf{h}$, and let $\hat{\mathsf{h}} = \sum_t \mathsf{w}(\sigma_t)\mathsf{h}(\sigma_t)/\sum_t \mathsf{w}(\sigma_t)$ be its self-normalized estimate from independent draws σ_t from $Q_{\theta/c}$. With $\bar{\mathsf{w}} = \frac{1}{T}\sum_t \mathsf{w}(\sigma_t)$,

$$\hat{\mathsf{h}} - \bar{\mathsf{h}} = \frac{\bar{\mathsf{a}}}{\bar{\mathsf{w}}}, \qquad \bar{\mathsf{a}} = \frac{1}{T}\sum_{t=1}^{T} \mathsf{w}(\sigma_t)\big(\mathsf{h}(\sigma_t) - \bar{\mathsf{h}}\big).$$

The terms of $\bar{\mathsf{a}}$ are independent and bounded, with mean zero and variance $\mathsf{v} = \mathbb{E}_{Q_{\theta/c}}\big[\mathsf{w}^2(\mathsf{h}-\bar{\mathsf{h}})^2\big]$, so $T\mathbb{E}\bar{\mathsf{a}}^2 = \mathsf{v}$, and $T\bar{\mathsf{a}}^2$ has a bounded second moment and is uniformly integrable. As $\bar{\mathsf{w}} \geq \min \mathsf{w} > 0$ and $\bar{\mathsf{w}} \to 1$ almost surely, $T(\hat{\mathsf{h}}-\bar{\mathsf{h}})^2 - T\bar{\mathsf{a}}^2 = T\bar{\mathsf{a}}^2(\bar{\mathsf{w}}^{-2}-1) \to 0$ in L^1. Since $\mathbb{E}\bar{\mathsf{a}} = 0$,

$$\big|\mathbb{E}(\hat{\mathsf{h}}-\bar{\mathsf{h}})\big| = \big|\mathbb{E}\big[\bar{\mathsf{a}}(\bar{\mathsf{w}}^{-1}-1)\big]\big| \leq \frac{\big(\mathbb{E}\bar{\mathsf{a}}^2\,\mathbb{E}(\bar{\mathsf{w}}-1)^2\big)^{1/2}}{\min \mathsf{w}} = O(1/T).$$

Hence $T \operatorname{Var} \hat{\mathsf{h}} \to \mathsf{v} = \mathbb{E}_{p_\theta}\big[\mathsf{w}(\mathsf{h}-\bar{\mathsf{h}})^2\big]$. For $\mathsf{h}(\sigma) = (P_\sigma)_{ij}$ the estimate $\hat{\mathsf{h}}$ is $\tilde{M}_{ij}$ and $\bar{\mathsf{h}} = M(\theta)_{ij}$, so summing over the entries gives $V(c) = \mathbb{E}_{p_\theta}[\mathsf{w}D]$. With $\mathbb{E}_{p_\theta}\mathsf{w} = \sum_\sigma p_\theta(\sigma)^2/Q_{\theta/c}(\sigma) = 1+\chi^2$ and $\mathbb{E}_{p_\theta}[D-\bar{D}] = 0$,

$$V(c) - \bar{D}\,(1+\chi^2) = \mathbb{E}_{p_\theta}\big[\mathsf{w}\,(D-\bar{D})\big] = \mathbb{E}_{p_\theta}\big[(\mathsf{w}-1)(D-\bar{D})\big] = R(c),$$

and $D_{\min}\mathbb{E}_{p_\theta}\mathsf{w} \leq \mathbb{E}_{p_\theta}[\mathsf{w}D] \leq D_{\max}\mathbb{E}_{p_\theta}\mathsf{w}$ because $\mathsf{w} > 0$. Finally c_E minimizes χ^2 over the interval, which contains c_V, so $V(c_E) \leq D_{\max}\big(1+\chi^2(c_E)\big) \leq D_{\max}\big(1+\chi^2(c_V)\big) \leq \frac{D_{\max}}{D_{\min}}V(c_V)$. Since $D(\sigma) = n + \|M(\theta)\|_F^2 - 2\langle P_\sigma, M(\theta)\rangle$, the constants $D_{\min}$ and $D_{\max}$ follow from the largest and the smallest value of $\langle P_\sigma, M(\theta)\rangle$, the optimal values of two assignment problems with the scores $M(\theta)$.

(b) Let $c_1 = \frac{n}{n+1}$ and $\mathsf{t} = \|\theta\|_F$. The temperature enters through $\mathsf{u} = 1 - c_1/c$, in which the leading terms below are quadratic. Write $\mathsf{V}_\sigma = P_\sigma - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ and $\mathsf{a}_\sigma = \langle \theta, \mathsf{V}_\sigma\rangle$, which equals $\langle \theta, P_\sigma\rangle$ because θ is centered. The law $Q_{\theta/c}$ is a smooth function of (θ, c) for $c > 0$ (Appendix C.7), so the expansions below hold together with their first two derivatives in c, and every $O(\cdot)$ is uniform over the interval and over the directions of θ. Under the uniform distribution p_0, $\mathbb{E}_{p_0}\mathsf{V}_\sigma = 0$, and the second-moment identity of Appendix C.7 gives $\mathbb{E}_{p_0}[\mathsf{a}_\sigma\mathsf{V}_\sigma] = \theta/(n-1)$, hence $\mathbb{E}_{p_0}\mathsf{a}_\sigma^2 = \mathsf{t}^2/(n-1)$. By Corollary C.1, $n!\,Q_{\theta/c}(\sigma) = 1 + \frac{c_1}{c}\mathsf{a}_\sigma + O(\mathsf{t}^2)$ and $n!\,p_\theta(\sigma) = 1 + \mathsf{a}_\sigma + O(\mathsf{t}^2)$, so $\mathsf{w} - 1 = \mathsf{u}\,\mathsf{a}_\sigma + O(\mathsf{t}^2)$ and

$$\chi^2 = \mathbb{E}_{Q_{\theta/c}}\big[(\mathsf{w}-1)^2\big] = \frac{\mathsf{u}^2\mathsf{t}^2}{n-1} + O(\mathsf{t}^3).$$

The marginals satisfy $\mathsf{B} := M(\theta) - \frac{1}{n}\mathbf{1}\mathbf{1}^\top = \mathbb{E}_{p_\theta}\mathsf{V}_\sigma = \theta/(n-1) + O(\mathsf{t}^2)$. Since $D(\sigma) = \|\mathsf{V}_\sigma - \mathsf{B}\|_F^2$ and $\|\mathsf{V}_\sigma\|_F^2 = n-1$, $\bar{D} = n - 1 - \|\mathsf{B}\|_F^2 = n - 1 - \mathsf{t}^2/(n-1)^2 + O(\mathsf{t}^3)$ and $D(\sigma) - \bar{D} = -2\langle \mathsf{V}_\sigma, \mathsf{B}\rangle + 2\|\mathsf{B}\|_F^2 = -2\mathsf{a}_\sigma/(n-1) + O(\mathsf{t}^2)$. Therefore $R(c) = -\frac{2\mathsf{u}}{n-1}\mathbb{E}_{p_\theta}\mathsf{a}_\sigma^2 + O(\mathsf{t}^3) = -2\mathsf{u}\mathsf{t}^2/(n-1)^2 + O(\mathsf{t}^3)$, and part (a) gives

$$V(c) = n - 1 + \mathsf{t}^2\Big[\mathsf{u}^2 - \frac{2\mathsf{u}+1}{(n-1)^2}\Big] + O(\mathsf{t}^3).$$

The variable u increases with c, equals 0 at c_1 and $(n-1)^{-2}$ at $\frac{(n-1)^2}{(n+1)(n-2)}$, and so ranges over an interval containing 0 in its interior, and $(n-1)^{-2}$ as well under the second hypothesis. As functions of u, both χ^2 and V have second derivatives $2\mathsf{t}^2/(n-1) + O(\mathsf{t}^3)$ and $2\mathsf{t}^2 + O(\mathsf{t}^3)$, positive for small t, and first derivatives $2\mathsf{u}\mathsf{t}^2/(n-1) + O(\mathsf{t}^3)$ and $2\mathsf{t}^2(\mathsf{u} - (n-1)^{-2}) + O(\mathsf{t}^3)$, which for small t are negative at the left end of the interval and positive at the right end, for V under the second hypothesis. Each function therefore has a unique minimizer, where its derivative vanishes: $\mathsf{u}_E = O(\mathsf{t})$ and $\mathsf{u}_V = (n-1)^{-2} + O(\mathsf{t})$. So $c_E \to c_1$ and $c_V \to c_1/(1-(n-1)^{-2}) = \frac{(n-1)^2}{(n+1)(n-2)}$. Substituting, $V(c) - V(c_V) = \mathsf{t}^2/(n-1)^4 + O(\mathsf{t}^3)$ both at $c = c_E$ and at $c = c_1$, and $V(c_V) = n - 1 + O(\mathsf{t}^2)$, which gives the expansion of the ratio. □

D.5 Subset Dynamic Program

The true marginals and the samples distributed as p_θ in Section 6 come from a dynamic program over subsets, which computes the permanent on a trellis whose states are the subsets of $[n]$ (Kiah et al., 2021). For square assignments and a set S of i columns, let $Z_i(\mathsf{S})$ be the sum of $\prod_{i' \leq i} e^{\theta_{i'\sigma(i')}}$ over the bijections σ from the first i rows onto S. Then $Z_0(\varnothing) = 1$,

$$Z_i(\mathsf{S}) = \sum_{j \in \mathsf{S}} e^{\theta_{ij}}\, Z_{i-1}(\mathsf{S} \setminus \{j\}),$$

Table D.1: Validation against the true marginals: the MAE over the ambiguous entries (Section 6) of three independent runs of Algorithm 1 or Algorithm 2 on one draw of a scene, their ESS/T, the sampling error of that draw and the number of its ambiguous entries. MAE and sampling error in units of 10^{-3}. The MAE of every run is 0.05 to 0.17 times the sampling error on the formations, 0.49 to 1.31 times on the random and crossing scenes with every target detected, and 1.56 to 1.65 times on the scene with missed detections and clutter.

		MAE ($\times 10^{-3}$)			sampling error	ESS/T			ambiguous
scene	n	run 1	run 2	run 3	($\times 10^{-3}$)	run 1	run 2	run 3	entries
every target detected (Algorithm 1)									
formation	8	0.239	0.239	0.337	1.942	0.98	0.98	0.98	20
formation	12	0.226	0.222	0.119	2.174	0.78	0.79	0.79	24
random	8	1.530	1.985	1.715	2.655	0.94	0.94	0.94	33
random	12	1.656	1.935	1.777	1.633	0.94	0.94	0.94	77
crossing	8	1.797	1.540	1.111	2.263	0.94	0.93	0.94	36
crossing	12	2.204	2.835	2.459	2.170	0.89	0.89	0.89	48
missed detections and clutter (Algorithm 2)									
random	12	2.752	2.913	2.853	1.763	0.57	0.58	0.58	57

and $Z(\theta) = Z_n([n])$. A sample is drawn backward: starting from $\mathsf{S} = [n]$, row $i = n, n-1, \dots, 1$ takes the column $j \in \mathsf{S}$ with probability $e^{\theta_{ij}} Z_{i-1}(\mathsf{S} \setminus \{j\})/Z_i(\mathsf{S})$, and that column is then removed from S. The product of these probabilities telescopes to $e^{\langle \theta, P_\sigma \rangle}/Z(\theta)$.

For rectangular assignments with private columns and a set S of rows, let $Z_j(\mathsf{S})$ be the sum of $\prod_{i \in \mathsf{S}} e^{\theta_{i\sigma(i)} - \theta_{i,m+i}}$ over the injective maps σ from S into the first j shared columns, a forbidden entry contributing zero. Column j is either unused or used by one row of S, so $Z_0(\mathsf{S}) = \mathbf{1}\{\mathsf{S} = \varnothing\}$,

$$Z_j(\mathsf{S}) = Z_{j-1}(\mathsf{S}) + \sum_{i \in \mathsf{S}} e^{\theta_{ij} - \theta_{i,m+i}} Z_{j-1}(\mathsf{S} \setminus \{i\}),$$

and, since the rows outside S take their private columns, $Z(\theta) = e^{\sum_i \theta_{i,m+i}} \sum_{\mathsf{S}} Z_m(\mathsf{S})$. An assignment is sampled by drawing the set S with probability proportional to $Z_m(\mathsf{S})$ and then, for $j = m, m-1, \dots, 1$, leaving column j unused with probability $Z_{j-1}(\mathsf{S})/Z_j(\mathsf{S})$ or giving it to the row $i \in \mathsf{S}$ with probability $e^{\theta_{ij} - \theta_{i,m+i}} Z_{j-1}(\mathsf{S} \setminus \{i\})/Z_j(\mathsf{S})$, in which case that row is then removed from S. The rows that receive no shared column take their private columns.

In both cases $M(\theta)_{ij} = \partial \log Z(\theta)/\partial \theta_{ij}$ for every allowed entry, which one backward pass through the recursion computes. The square recursion has 2^n states and $n2^{n-1}$ transitions, so it takes $O(n2^n)$ operations and stores $O(2^n)$ numbers. The rectangular one takes $O(mn2^n)$ operations and stores $O(m2^n)$ numbers. Both add nonnegative terms only, so they run in the log domain without cancellation.

D.6 Validation

Table D.1 compares three runs of each algorithm with the true marginals, on one draw of each scene. Algorithm 1 runs on the scene families of Section 6 (Appendix E.1) with every target detected at $n \in \{8, 12\}$, the formations at the lag $\tau = \frac{1}{2}$. Algorithm 2 runs on a random scene with missed detections and clutter at $n = 12$. The runs use independent randomness and $T = 2 \cdot 10^4$. The sampling error of a scene here is that of its one draw: the MAE of the empirical marginals of T independent samples from p_θ, drawn by the dynamic program of Appendix D.5. The MAE of every run is 0.05 to 0.17 times the sampling error of its scene on the formations, 0.58 to 1.18 times on the random scenes, 0.49 to 1.31 times on the crossing scenes, and 1.56 to 1.65 times on the scene with missed detections and clutter. In the runs on the random scenes at $n = 12$, no assignment is frequent and the weights come from the law-based estimates alone, while the runs on the formations at $n = 8$ and $n = 12$ have between 34 and 44 frequent assignments.

E EXPERIMENTS

Appendix E.1 gives the scenes of Section 6 and the budgets, Appendix E.2 the baselines, Appendix E.3 the three families at the sizes where the true marginals can be computed (Table E.1) with the collapse of the beliefs of belief propagation and the effective sample sizes, Appendix E.4 the larger scenes and the ranges over the draws of each scene, Appendix E.5 the temperature grid that Appendix D.4 points to, Appendix E.6 the formations too large for the dynamic program, and Appendix E.7 the hardware.

E.1 SCENES

The scenes are two-dimensional. Track i carries a predicted measurement $\hat{z}_i$, and a measurement z_j is a point. A formation scene at prediction lag τ places the n predicted measurements at equal spacing on a circle of circumference $2n$, the first at angle zero and the others counterclockwise, and the targets themselves τ spacings further along it. The measurements are the target positions plus independent Gaussian errors of standard deviation 0.25 in each coordinate. A random scene places the n targets uniformly in $[0, 3]^2$. A crossing scene places $\lfloor n/2 \rfloor$ of them at spacing 1 along the first coordinate axis and $\lceil n/2 \rceil$ at spacing 1 along the second, both lines centered at the origin. In these two families the predicted measurements and the measurements are the target positions plus independent Gaussian errors of standard deviation 0.7 for the predictions and 0.3 for the measurements.

Scenes with every target detected report one measurement per target and no clutter, and their scores are the unit-covariance Gaussian log-likelihoods $\theta_{ij} = -\|z_j - \hat{z}_i\|^2/2$, the constant $-\log 2\pi$ being dropped because it shifts the score of every hypothesis alike. Scenes with missed detections and clutter take ϕ_i to be the Gaussian density with mean $\hat{z}_i$ and covariance $(0.7^2 + 0.3^2)I$, set the detection probability to $P_{\mathrm{D}} = 0.9$, and make the clutter a Poisson process of mean 3 uniform on the bounding box of the predicted measurements padded by 2 on every side, so that $\lambda = 3/\text{area}$. Each target is detected independently with probability P_{D}, and the measurement list is the measurements of the detected targets followed by the clutter points. The order is immaterial to every method. The two detection regimes differ in the covariance of their likelihoods as well as in the miss columns and the clutter. No gating is applied: every track may take every measurement, every shared entry is finite, and with private columns and at least one measurement the reachability condition of Theorem B.1 therefore holds for every hypothesis (Appendix B.4).

Figure 1 follows the formation family at $n \in \{8, 12\}$ and $\tau \in \{0, 0.1, 0.2, \frac{1}{4}, 0.3, 0.4, \frac{1}{2}\}$, and Figure E.1 the same at $n \in \{16, 20\}$ and $\tau \in \{0, \frac{1}{4}, \frac{1}{2}\}$. Table E.1 covers all three families at $n \in \{10, 15, 20\}$, the formations at $\tau = \frac{1}{2}$, and Table E.4 the formations at $\tau = \frac{1}{2}$ and $n \in \{30, 40, 60\}$ with every target detected and $n \in \{30, 40\}$ with missed detections and clutter. Every scene is drawn three times. Where the run-to-run agreement is reported, the reweighted estimator is run a second time with independent randomness, and the agreement is the MAE between the two runs on the ambiguous entries. In every run of Section 6 each sampling method draws $T = 2 \cdot 10^4$ assignments, sequential importance sampling included. Algorithm 1 uses the 256 chains and $T_y = 32768$ potential samples of Appendix D.1. Algorithm 2 uses the annealing budget and the chains of Appendix D.2, which give $T_y = 8192$ potential samples for every shared-column set. The reweighted estimator and the second unweighted estimator run at $c = \frac{n}{n+1}$, except in Appendix E.5. The other methods take the scores as they are. Every run records its seed.

E.2 BASELINES

Belief propagation runs on the association model of Section 6 in its ratio form (Williams & Lau, 2014). Write $\mathsf{a}_{ij} = e^{\theta_{ij} - \theta_{i,m+i}}$ for the likelihood ratio of a shared entry of row i to its miss entry. Each pair (i, j) carries one scalar message in each direction,

$$\mathsf{u}_{ij} = \frac{\mathsf{a}_{ij}}{1 + \sum_{j' \neq j} \mathsf{a}_{ij'} \mathsf{v}_{ij'}}, \qquad \mathsf{v}_{ij} = \frac{1}{1 + \sum_{i' \neq i} \mathsf{u}_{i'j}},$$

and the beliefs are proportional to $\mathsf{a}_{ij}\mathsf{v}_{ij}$ on the shared entries of row i and to 1 on its miss entry, normalized over the row. A sweep computes every u from the current v and then every v from the new u, damping v in the logarithm, $\log \mathsf{v} \leftarrow \frac{1}{2} \log \mathsf{v} + \frac{1}{2} \log \mathsf{v}^{\mathrm{new}}$. Each sum excluding j or i is computed in double precision by subtracting the excluded term from the full sum, and each denominator is

floored at 10^{-300}. The messages start at $\mathsf{v} = 1$, the iteration stops when the largest change of $\log \mathsf{v}$ falls below 10^{-9}, and it is reported as unconverged after 3000 sweeps. On scenes with every target detected no track is missed and no measurement is clutter, which removes the summand 1 from both denominators and leaves $\mathsf{a}_{ij} = e^{\theta_{ij}}$. That iteration, allowed $2 \cdot 10^4$ sweeps, is a damped form of the sum-product algorithm for the permanent (Vontobel, 2013, Lemma 29): without damping or flooring, $\sqrt{\mathsf{a}_{ij}}\,\mathsf{v}_{ij}$ and $\mathsf{u}_{ij}/\sqrt{\mathsf{a}_{ij}}$ satisfy its inverse-likelihood-ratio updates, and $\mathsf{a}_{ij}\mathsf{v}_{ij}$, normalized over the row, is its belief. The stopping rule was met on every scene of the demonstration, including those whose beliefs agree with an assignment matrix to machine precision.

Sinkhorn scaling normalizes the rows and the columns of $\exp\theta$ alternately, 3000 times each, and reads the result as association probabilities. With private columns it runs on the $(m+n) \times (m+n)$ matrix whose upper left block is $\exp\theta$ on the shared entries, whose upper right block is the diagonal matrix of the exponentiated miss entries, whose lower left block is the identity, one clutter row per measurement, and whose lower right block is all ones. The association probabilities are the upper left block of the scaled matrix and the diagonal of its upper right block. An association that uses k shared columns leaves k clutter rows and k private columns free, so it has $k!$ completions in this matrix, each of weight $e^{\langle \theta, P_\sigma \rangle}$: the scaling approximates the marginals of that weighting, whereas sequential importance sampling below divides by its own proposal probabilities and targets p_θ.

Sequential importance sampling builds an assignment one row at a time, in index order. At row i it draws a column not taken by an earlier row with probability proportional to that column's entry in the Sinkhorn scaling, the augmented scaling when there are private columns, renormalized over the columns still free. The weight of the assignment is $e^{\langle \theta, P_\sigma \rangle}$ divided by the product of the probabilities with which its columns were drawn. The weights of the T particles are self-normalized, with no resampling, and their effective sample size is reported as in Section 5. The unweighted estimators average the assignment matrices of T draws of Gumbel matching, at temperature one and at $c = \frac{n}{n+1}$.

E.3 Scene Families

Table E.1 covers the three families at $n \in \{10, 15, 20\}$, the formations at $\tau = \frac{1}{2}$, with every target detected and with missed detections and clutter. On every draw of the formations of Figure 1(a) the converged beliefs of belief propagation agree with the indicator of the most probable hypothesis to within 10^{-14} in every entry, by the mechanism of Vontobel (2013, Theorem 32): when the minimizer of the Bethe free energy over the doubly stochastic matrices is an assignment matrix, the beliefs collapse onto it. Vontobel (2013, Corollary 27) certifies the case: for the most probable hypothesis σ, the assignment matrix is the unique minimizer when the largest eigenvalue of the matrix with entries $e^{\theta_{i\sigma(i')} - \theta_{i\sigma(i)}}$ off the diagonal and zero on it is below one, and is not a minimizer when it is above one. On these draws the eigenvalue lies between 0.30 and 0.95, and over all formation draws with every target detected it exceeds one exactly where the beliefs do not collapse: one draw of Figure E.1(a), one each at $n = 15$ and $n = 20$ in Table E.1, one at $n = 40$ and all three at $n = 60$ in Table E.4. On the formations of Figure 1(b) the reweighted estimator's ESS/T lies between 0.59 and 0.93.

On the rows of Table E.1 the reweighted estimator's ESS/T ranges from 0.11 to 0.97 and sequential importance sampling's from 0.003 to 0.53.

E.4 Larger Scenes and Ranges

Figure E.1 repeats Figure 1 at $n = 16$ and $n = 20$. With every target detected, belief propagation, Sinkhorn scaling and both unweighted estimators are off by 0.06 to 0.30, while the reweighted estimator's error stays between 0.12 and 0.47 times the sampling error. The beliefs of belief propagation agree with the indicator of the most probable hypothesis to within 10^{-14} in every entry on all but one of these eighteen draws, and within 0.08 of it on that one. With missed detections and clutter the reweighted estimator's error is between 0.64 and 1.8 times the sampling error and belief propagation's is between 1.4 and 18 times it. At $n = 20$ and $\tau = \frac{1}{4}$ with every target detected one of the three draws has no ambiguous entry, so that point of the figure is the mean of the other two.

Table E.2 gives the ranges over the three draws of each scene: for the rows of Table E.1 the reweighted estimator's error, belief propagation's, the sampling error and the number of ambigu-

Table E.1: The three scene families at the sizes where the true marginals can be computed, the formations at $\tau = \frac{1}{2}$. Every entry is the mean over three draws of the scene, MAE over the ambiguous entries and sampling error in units of 10^{-3}, and Table E.2 gives the ranges over the draws of the reweighted estimator's error, of belief propagation's error and of the sampling error. The reweighted estimator is more accurate than Sinkhorn scaling, sequential importance sampling and both unweighted estimators in every row, and with every target detected its error is at most 1.1 times the sampling error. The smallest error of each row is in bold.

		MAE ($\times 10^{-3}$)						sampling error	ESS/T	
scene	n	unw.	unw. $n/(n+1)$	Sinkhorn	seq. IS	BP	ours	($\times 10^{-3}$)	ours	seq. IS
every target detected (Algorithm 1)										
formation	10	193	179	254	48.6	202	**0.207**	1.69	0.81	0.02
formation	15	138	132	173	89.1	138	**0.399**	1.71	0.77	0.11
formation	20	205	199	261	176	138	**0.379**	2.02	0.65	0.53
random	10	12.3	7.76	15.9	2.50	3.49	**1.71**	1.90	0.97	0.49
random	15	5.35	3.35	5.94	2.73	**0.856**	1.72	1.71	0.97	0.41
random	20	6.73	5.11	7.71	3.72	2.19	**1.76**	1.69	0.93	0.21
crossing	10	24.7	16.3	33.7	3.41	11.4	**1.55**	1.87	0.90	0.31
crossing	15	23.7	18.8	34.8	4.91	25.9	**2.05**	2.12	0.84	0.13
crossing	20	31.5	26.8	47.9	9.37	25.7	**2.37**	2.23	0.70	0.05
missed detections and clutter (Algorithm 2)										
formation	10	48.4	44.5	94.6	9.82	12.3	**1.90**	2.56	0.59	0.09
formation	15	64.7	55.9	85.0	12.1	14.7	**2.64**	2.57	0.51	0.05
formation	20	66.2	59.2	112	32.0	12.4	**3.37**	2.11	0.49	0.003
random	10	18.9	14.9	30.5	4.17	7.67	**3.71**	1.94	0.23	0.18
random	15	13.8	12.3	25.9	11.3	**5.99**	6.55	1.90	0.15	0.05
random	20	10.7	10.2	16.8	9.14	**3.48**	7.88	1.74	0.11	0.05
crossing	10	25.2	19.7	46.8	5.07	12.0	**3.73**	2.00	0.39	0.16
crossing	15	29.5	24.4	47.9	7.01	18.2	**3.30**	2.24	0.50	0.07
crossing	20	27.3	24.1	51.9	15.8	23.8	**3.64**	1.92	0.32	0.04

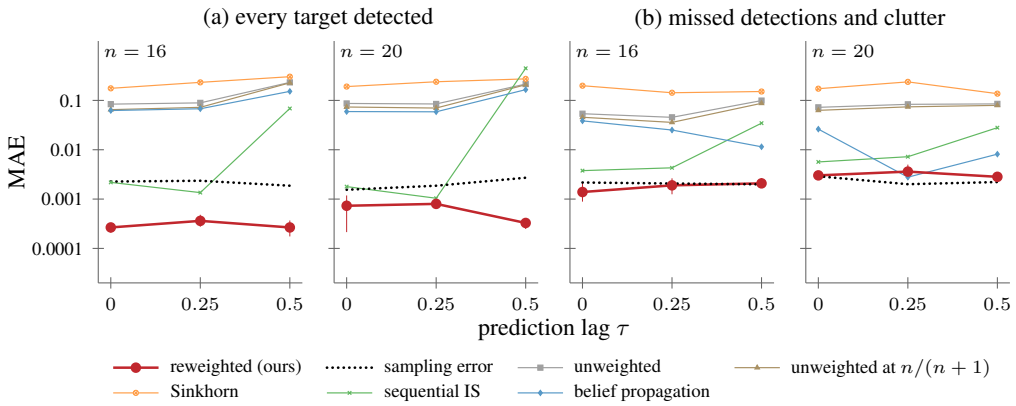


Figure E.1: Formation scenes at $n = 16$ and $n = 20$ against the prediction lag, as in Figure 1: MAE over the ambiguous entries, the reweighted estimator's range over the three draws of each scene, and the sampling error (dotted). The reweighted estimator's mean error stays between 0.12 and 0.47 times the sampling error in (a) and between 0.64 and 1.8 times it in (b).

ous entries, and for the rows of Table E.4 the run-to-run agreement, the effective sample size and the number of ambiguous entries. The number of ambiguous entries varies across the draws of one scene: the three draws of the formation at $n = 40$ with every target detected leave 20, 33 and 80 entries ambiguous, and the two runs of the estimator agree to between 0.0015 and 0.0031 on them.

Table E.2: The mean over the three draws of each scene of Tables E.1 and E.4, with the range over the draws in brackets, and the range of the number of ambiguous entries. MAE, sampling error and run-to-run agreement in units of 10^{-3}. On the formations with every target detected at $n \leq 20$ the reweighted estimator's range lies below the sampling error's.

		MAE ($\times 10^{-3}$)		sampling error	ambiguous
scene	n	BP	ours	($\times 10^{-3}$)	entries
every target detected					
formation	10	202 [173, 226]	0.207 [0.168, 0.274]	1.69 [0.737, 2.63]	[20, 21]
formation	15	138 [15.5, 253]	0.399 [0.344, 0.450]	1.71 [0.948, 2.72]	[30, 32]
formation	20	138 [126, 157]	0.379 [0.275, 0.435]	2.02 [1.24, 2.56]	[40, 43]
random	10	3.49 [2.58, 4.49]	1.71 [1.63, 1.84]	1.90 [1.71, 2.10]	[58, 63]
random	15	0.856 [0.405, 1.46]	1.72 [1.66, 1.77]	1.71 [1.49, 1.95]	[104, 137]
random	20	2.19 [1.42, 2.94]	1.76 [1.51, 2.00]	1.69 [1.52, 1.77]	[137, 159]
crossing	10	11.4 [4.01, 21.4]	1.55 [1.48, 1.58]	1.87 [1.25, 2.44]	[41, 50]
crossing	15	25.9 [14.2, 38.2]	2.05 [1.69, 2.23]	2.12 [1.63, 2.55]	[64, 71]
crossing	20	25.7 [18.8, 34.0]	2.37 [2.27, 2.53]	2.23 [2.00, 2.35]	[77, 80]
missed detections and clutter					
formation	10	12.3 [3.29, 18.4]	1.90 [1.67, 2.21]	2.56 [2.38, 2.72]	[25, 28]
formation	15	14.7 [3.07, 25.1]	2.64 [2.24, 3.31]	2.57 [1.96, 2.96]	[29, 38]
formation	20	12.4 [2.61, 24.3]	3.37 [2.97, 3.95]	2.11 [1.37, 3.02]	[43, 45]
random	10	7.67 [4.79, 11.5]	3.71 [2.54, 4.43]	1.94 [1.84, 2.05]	[39, 60]
random	15	5.99 [3.01, 9.39]	6.55 [3.14, 12.0]	1.90 [1.70, 2.02]	[77, 100]
random	20	3.48 [3.04, 4.21]	7.88 [4.29, 14.8]	1.74 [1.68, 1.86]	[132, 135]
crossing	10	12.0 [7.92, 17.5]	3.73 [2.69, 4.67]	2.00 [1.93, 2.08]	[40, 45]
crossing	15	18.2 [15.7, 21.1]	3.30 [2.88, 3.56]	2.24 [2.20, 2.26]	56
crossing	20	23.8 [16.2, 29.4]	3.64 [3.12, 4.18]	1.92 [1.72, 2.02]	[71, 76]

n	run-to-run agreement ($\times 10^{-3}$)	ESS/T	ambiguous entries
every target detected			
30	0.728 [0.724, 0.730]	0.43 [0.42, 0.45]	[62, 63]
40	2.19 [1.52, 3.15]	0.38 [0.31, 0.51]	[20, 80]
60	17.2 [10.7, 26.5]	0.27 [0.20, 0.32]	[120, 122]
missed detections and clutter			
30	6.85 [4.57, 8.44]	0.29 [0.19, 0.38]	[49, 69]
40	13.4 [7.99, 20.2]	0.28 [0.28, 0.28]	[72, 74]

E.5 Temperature from a Pilot Run

The variance bounds of Appendix D.4 motivate choosing the temperature by maximizing the effective sample size. That choice is tested here against the first-order temperature $c_1 = \frac{n}{n+1}$ on the formations with every target detected at $n \in \{12, 20\}$ and $\tau = \frac{1}{2}$. A pilot run draws $T/5$ assignments at each temperature of the grid $\{0.6, 0.7, 0.8, 0.85, 0.9, 0.95, 1, 1.05\} \cup \{c_1\}$, from the same seed at every grid point and with the production number T_y of potential samples, and $c^\star$ is the temperature whose pilot reports the largest effective sample size. The estimator is then run at $c^\star$ and at c_1 with the production budget and independent randomness.

The maximizer lies below c_1 on every draw (Table E.3), at 0.70 or 0.80 against $c_1 \approx 0.92$ and 0.95. In the production runs, ESS/T at $c^\star$ is 1.15 to 1.31 times its value at c_1, and the error is 1.5 to 6.8 times as large. The reported fraction is computed from the weights of the assignments a run draws, and a lower temperature concentrates the proposal on fewer of them. Proposition D.1 bounds the variance of $\tilde{M}$ at the temperature that maximizes the population effective sample size, whereas a pilot of $T/5$ draws maximizes an estimate of that quantity. Algorithms 1 and 2 therefore fix $c = \frac{n}{n+1}$, and Section 6 reports no tuned run.

Table E.3: The temperature $c^\star$ maximizing the reported effective sample size on a pilot run, against the first-order temperature $c_1 = \frac{n}{n+1}$, on three draws of the formation scene with every target detected at $\tau = \frac{1}{2}$. The pilot draws $T/5$ assignments per grid point. ESS/T and the MAE are those of the two production runs and are both higher at $c^\star$ than at c_1 in every row.

		ESS/T		MAE ($\times 10^{-3}$)	
draw	$c^\star$	at $c^\star$	at c_1	at $c^\star$	at c_1
$n = 12$					
1	0.70	0.81	0.69	0.941	0.138
2	0.70	0.82	0.70	0.805	0.165
3	0.80	0.83	0.72	0.403	0.170
$n = 20$					
1	0.80	0.98	0.78	0.601	0.340
2	0.80	0.77	0.59	0.614	0.317
3	0.80	0.78	0.61	0.528	0.349

Table E.4: Formation scenes at $\tau = \frac{1}{2}$ too large for the dynamic program. The reference is the average of two independent runs of the reweighted estimator, the run-to-run agreement is the MAE between the two runs on the entries the reference places between 0.05 and 0.95, the distances are the MAEs of the other methods from the reference on those entries, and ESS/T is the first run's. Every entry is the mean over three draws of the scene, the agreement and the distances in units of 10^{-3}, and Table E.2 gives the ranges over the draws of the agreement and of ESS/T. The distances of Sinkhorn scaling, sequential importance sampling and the two unweighted estimators from the reference are 8 to 775 times the agreement.

	distance from the two runs' average ($\times 10^{-3}$)					run-to-run	
n	unw.	unw. $n/(n+1)$	Sinkhorn	seq. IS	BP	agreement ($\times 10^{-3}$)	ESS/T
every target detected							
30	290	286	362	564	103	0.728	0.43
40	195	193	264	388	44.0	2.19	0.38
60	154	152	174	391	19.0	17.2	0.27
missed detections and clutter							
30	94.2	89.5	153	104	10.3	6.85	0.29
40	108	107	177	177	12.7	13.4	0.28

E.6 Large Formations

For the formations too large for the dynamic program the reference is the average of two independent runs of the reweighted estimator, with fresh Gumbel, Langevin and annealing randomness. The ambiguous entries are then those it places between 0.05 and 0.95, and the run-to-run agreement, the MAE between the two runs on these entries, measures repeatability rather than accuracy. Table E.4 reports this agreement and the distance of every other method from the two runs' average. With every target detected the two runs agree to 0.0007, 0.0022 and 0.017 on the ambiguous entries, with ESS/T falling from 0.43 to 0.27, and with missed detections and clutter they agree to 0.0069 and 0.013. The distances of Sinkhorn scaling, sequential importance sampling and the two unweighted estimators from the average are 8 to 775 times that agreement. With every target detected, belief propagation's distance is 141, 20 and 1.1 times it at $n = 30$, 40 and 60, and with missed detections and clutter 1.5 and 0.95 times it. At $n = 30$ its beliefs agree with the indicator to machine precision on every draw.

E.7 Hardware

The chains, the annealing, the determinant evaluations and the subset dynamic program run on one NVIDIA RTX 4090. The Hungarian algorithm, belief propagation, Sinkhorn scaling and sequential importance sampling run on an AMD Ryzen 9 7950X with sixteen cores.