# SIX OUTPUTS SUFFICE: UNIQUE-BAYES COUNTEREXAMPLES TO FISHER CONSISTENCY OF THE STRUCTURED-SVM SURROGATE ON MODULAR METRICS

KOREA SUPERINTELLIGENCE LABS

ABSTRACT. Let $\mathcal{Y}$ be a finite output set and let $L$ be a metric on it. The margin-rescaling structured-SVM surrogate is $S_M(v,y) = \max_{y' \in \mathcal{Y}} \{L(y,y') + v_{y'} - v_y\}$, and the question is whether the canonical decoder $v \mapsto \arg\max_y v_y$ is Fisher consistent for $L$. Nowak-Vila, Rudi and Bach showed that it is necessary that every triple of outputs have a common geodesic point — that $L$ be a *modular* metric — and left sufficiency open. Fei and Luo have recently shown that the condition is not sufficient, and that five outputs are necessary and sufficient for a counterexample at a fully supported conditional distribution. Their discussion isolates a strictly harder quantity: the least number $k^*$ of outputs carrying a counterexample that is at once fully supported, has a *single* Bayes output, and whose score maximisers all miss that output. They bracket $5 \le k^* \le 8$ and call the value open.

We prove $k^* \le 6$. On the shortest-path metric of $K_{3,3}$ the distribution $q = (1,1,2,2,3,3)/12$ has the single Bayes output $a_3$, the score vector $v = (-1,-1,-1,0,0,0)$ minimises the conditional surrogate risk exactly — certified by an explicit self-coupling through weak duality — and its maximisers are the whole opposite part. A second witness on the same metric, $q = (1,1,1,2,1,1)/7$, puts the single Bayes output in the heavy part instead, and a seven-output witness on $K_{3,4}$ certifies the cardinality 7 as well, so witnesses now exist at 6, 7 and 8. Two further results are proved on the page rather than machine-checked: a tie-broken complete-bipartite family gives a witness at *every* cardinality $k \ge 6$ and fails exactly at $m = 2$, the case the source itself treats; and an exact decision procedure over all rational distributions and all score vectors, run to completion on every five-point modular metric arising from a graph and on all seventeen with integer distances at most 4, returns no witness, so $k^* = 6$ conditional on that enumeration. The witnesses, the source's three counterexamples and the certificate layer are formally verified in Lean 4 against *Mathlib*.

## 1. INTRODUCTION

1.1. **The surrogate and the consistency question.** Let $\mathcal{Y}$ be a finite set of outputs, $k = |\mathcal{Y}|$, and let $L\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\ge 0}$ be a metric. The *margin-rescaling structured-SVM surrogate* scores a candidate vector $v \in \mathbb{R}^{\mathcal{Y}}$ against a true output $y$ by

$$S_M(v,y) \;=\; \max_{y' \in \mathcal{Y}} \big\{\, L(y,y') + v_{y'} - v_y \,\big\}. \tag{1}$$

For a conditional distribution $q \in \Delta_{\mathcal{Y}}$ write

$$R_q(v) = \sum_{y \in \mathcal{Y}} q_y\, S_M(v,y), \quad V^*(q) = \arg\min_v R_q(v), \quad \ell_q(\hat{y}) = \sum_{y \in \mathcal{Y}} q_y\, L(\hat{y},y), \quad B(q) = \arg\min_{\hat{y}} \ell_q(\hat{y}), \tag{2}$$

and let $H_M(q) = \min_v R_q(v)$. The surrogate is deployed with the *canonical decoder* $v \mapsto \arg\max_y v_y$, and the consistency question is whether that decoder always reads a surrogate-risk minimiser as a Bayes-optimal output:

$$v \in V^*(q) \implies \arg\max_y v_y \subseteq B(q) \qquad \text{for every } q \in \Delta_{\mathcal{Y}}. \tag{3}$$

Nowak-Vila, Rudi and Bach [2] proved a necessary condition for (3) (their Theorem 2.1): writing

$$I(x,y) = \{\, z \in \mathcal{Y} : L(x,y) = L(x,z) + L(z,y) \,\} \tag{4}$$

for the metric interval, one must have

$$I(y_1,y_2) \cap I(y_1,y_3) \cap I(y_2,y_3) \neq \emptyset \qquad \text{for every } y_1,y_2,y_3 \in \mathcal{Y}. \tag{5}$$

As [1] puts it, "In metric geometry, [(5)] is the defining property of a *modular metric space* … Median metrics impose uniqueness of the common point"; we use *modular* for (5) and *median* for the strengthening in which the common point is unique for every triple. Whether (5) is also sufficient was left open in [2], which says verbatim: "It is an open question whether the necessary condition of Thm. 2.1 is also sufficient." Its Appendix Theorem B.9 does assert the positive direction for tree metrics: "If $L$ is a distance defined in a tree, then the Max-Margin loss $S_M$ embeds $L$ with embedding $\psi(y) = -L_y$ and it is consistent under the argmax decoding."

Fei and Luo [1] show that (5) is *not* sufficient, correct the tree claim for branching trees (their Corollary 6.2), and pin the sharp full-support cardinality at five (their Corollary 5.2) in two halves: no metric on three or four outputs satisfying (5) violates (3) at a fully supported $q$ (their Theorem 4.5), while $K_{2,3}$ at the uniform distribution does (their Theorem 5.1 at $m = 2$, $n = 3$). That is the content of their abstract's headline, that "five outputs are necessary and sufficient for a full-support counterexample".

1.2. **The quantity that is open.** The five-point counterexample of [1] has a *two-element* Bayes set. Its eight-point Hamming-cube counterexample (its Theorem 5.3) does not: there the Bayes output is unique and the score maximisers avoid it entirely. The paper's Discussion isolates the gap between the two, verbatim:

> "Second, our five-point full-support witness has multiple Bayes outputs, whereas the Hamming-cube witness has a unique Bayes output and disjoint argmax; the sharp cardinality under these additional requirements remains open."

**Definition 1.1.** A *witness on $k$ outputs* is a triple $(L, q, v)$ with $|\mathcal{Y}| = k$ such that $L$ is a metric on $\mathcal{Y}$ satisfying (5); $q \in \Delta_{\mathcal{Y}}$ has $q_y > 0$ for every $y$; $B(q)$ is a *singleton*; $v \in V^*(q)$; and $\arg\max_y v_y$ is *disjoint* from $B(q)$. We write

$$k^* \;=\; \min\{\, k : \text{a witness on } k \text{ outputs exists} \,\}.$$

(All data in this paper is rational; see the convention opening §2.)

The cardinality results of [1] bracket this quantity. Its Theorem 4.5 — "No four-output full-support obstruction": every metric with $3 \le |\mathcal{Y}| \le 4$ satisfying (5) obeys the argmax property at every full-support distribution — together with its Theorem 3.3 on weighted paths, which covers $|\mathcal{Y}| \le 2$, gives $k^* \ge 5$; its Theorem 5.3 gives $k^* \le 8$. It is $k^*$, and only $k^*$, that the Discussion leaves open.

*Remark* 1.2 (a framing point that is easy to lose). "Five outputs are necessary and sufficient for a full-support counterexample" answers the *plain* full-support question, whose answer is 5 and whose witness is $K_{2,3}$ at the uniform distribution. Definition 1.1 imposes two further requirements — a single Bayes output, and score maximisers disjoint from it — and $k^*$ is a strictly harder quantity. Nothing in this paper contradicts the abstract of [1]; we improve only the upper end of the bracket $5 \le k^* \le 8$ that its Discussion states.

1.3. **Results.** Let $K_{m,n}$ carry its shortest-path metric: distance 2 between two outputs in the same part, 1 across. Write $A = \{a_1, \dots, a_m\}$ and $B = \{b_1, \dots, b_n\}$ for the two parts.

Our main result is that $k^* \le 6$ (Theorem 4.1). On $K_{3,3}$, at the distribution $q = (1, 1, 2, 2, 3, 3)/12$ written in the order $a_1, a_2, a_3, b_1, b_2, b_3$, the score vector $v = (-1, -1, -1, 0, 0, 0)$ minimises the conditional surrogate risk exactly, the Bayes set is the single output $a_3$, and the score maximisers are the whole opposite part $\{b_1, b_2, b_3\}$. This lowers the source's own upper bound by two. Theorem 4.2 gives a second witness on the same metric with the Bayes output in the heavy part, so the phenomenon is not a knife-edge choice of $q$; and Theorem 5.1 gives a seven-output witness on $K_{3,4}$, so witnesses now exist at 6, 7 and 8 outputs. Every claim in these three theorems, in the certificate layer of §2, and in the re-certification of the three counterexamples of [1] in §3, is formally verified (§8).

Two further results are proved on the page and are deliberately outside the formal development; they are gathered in §§5–6 and labelled there. The first (Theorem 5.2) is that breaking the tie inside one part of $K_{m,n}$ gives a witness for every $m \ge 3$, $n \ge m$, hence a witness on $k$ outputs for *every* $k \ge 6$; the construction fails precisely at $m = 2$, which is the case [1] itself treats. The second (§6) is an exact

decision procedure that, for a fixed metric, decides over *all* rational distributions and *all* score vectors at once whether a witness exists; run to completion it returns nothing on any five-point modular metric arising from a connected graph, and nothing on any of the seventeen five-point modular metrics with integer distances at most 4. Together with Theorem 4.1 this gives $k^* = 6$ conditional on that five-point enumeration.

1.4. **What is not claimed.** We do *not* prove $k^* \geq 6$ outright. The lower half of §6 quantifies over two finite classes of five-point modular metrics — those coming from unweighted graphs, and those with integer distances at most 4 — while $k^* \geq 6$ quantifies over the whole cone of five-point modular metrics, a continuum. What is missing is a weight continuum inside finitely many combinatorial types; §7 says why one linear program cannot cover it.

We also do not claim that the surrogate always mis-decodes at a witness distribution. At the distribution of Theorem 4.1 the embedded report of the Bayes output is *also* an exact minimiser, and the canonical decoder reads it correctly (Proposition 4.3 (iii)); the witness is a statement about *some* minimiser. That is the decoder gap [1] announces in its abstract — "an embedding can guarantee the existence of a calibrated link without validating a prescribed argmax link on every surrogate-risk minimizer" — and works out in its Section 6, where the same phenomenon is exhibited on its four-output star: there too the embedded report of the Bayes output is optimal and correctly decoded, while a second optimal report is not.

Finally, the median sub-question is untouched: both witnesses on six outputs here are modular but not median, and the only median witness known is the source's cube on eight outputs.

1.5. **Related work, and where these claims were looked for.** The consistency of convex surrogates for discrete losses has a long literature, and the specific quantity of Definition 1.1 appears in none of it that we could find. We read, in full text, the two versions of [2] (its arXiv version and its conference version) and eight further primary sources in the calibration and structured-prediction line: Osokin, Bach and Lacoste-Julien on calibrated convex surrogates for structured prediction [7]; Ramaswamy and Agarwal on convex calibration dimension [8]; Nowak-Vila, Bach and Rudi on consistent structured prediction [9]; two papers of Finocchiaro, Frongillo and Waggoner on polyhedral surrogate embeddings [10, 11]; Mao, Mohri and Zhong [12]; Tewari and Bartlett [13]; and Liu [14]. Across all ten, and as exact strings, *counterexample*, *geodesic*, *median graph/metric/space*, *modular metric/graph/space*, *tree metric* and *complete bipartite* do not occur; every occurrence of *smallest*, *minimal* or *minimum number* was read in context and refers to something else — a smallest surrogate dimension for a fixed loss matrix, a smallest representative set of reports, a tie-breaking convention, or the first non-binary case $K = 3$ — never to a sharp minimal cardinality of a counterexample. The sharp cardinality studied here is, as far as this search reaches, asked for the first time in [1] and answered from above for the first time here.

## 2. Preliminaries, and the self-coupling certificate

Nothing in this section is new. It is the toolkit of [1] — the easy half of its Proposition 2.1, its complementary-slackness identity (7), its Lemma 2.2, and the argument it runs three times (in its Theorem 3.3 for weighted paths, its Theorem 3.5 for tree metrics in the interior of the simplex, and its Proposition 4.4 for the weighted rectangle) — isolated once for an arbitrary finite metric, so that the witnesses below need only supply data.

All data in this paper is rational, and we work throughout over $\mathbb{Q}$: metrics, distributions, score vectors and couplings take rational values, and $V^*(q)$ means the minimisers of $R_q$ among rational score vectors. This costs nothing. The only optimality tool used below is the weak-duality inequality of Proposition 2.2, whose proof is a rearrangement of finite sums valid over any ordered field; a score vector certified optimal by it therefore minimises $R_q$ among real score vectors as well. (That last sentence is an observation about the proof, not part of the verified statement.)

**Definition 2.1.** A *self-coupling* of $q$ is a matrix $\Pi = (\Pi_{ij})_{i,j \in \mathcal{Y}}$ with $\Pi \geq 0$, $\sum_j \Pi_{ij} = q_i$ and $\sum_i \Pi_{ij} = q_j$ for all $i, j$. Its value is $\langle \Pi, L \rangle = \sum_{i,j} \Pi_{ij} L(i,j)$.

Proposition 2.1 of [1] identifies $H_M(q)$ with the maximum of $\langle \Pi, L\rangle$ over self-couplings of $q$. Only the easy half of that identity is needed here, and only it is proved below.

**Proposition 2.2** (weak duality)**.** *Let $L$ be any function $\mathcal{Y}\times\mathcal{Y}\to\mathbb{Q}$, let $\Pi$ be a self-coupling of $q$ and let $v\in\mathbb{Q}^{\mathcal{Y}}$. Then $\langle \Pi, L\rangle \le R_q(v)$. Consequently, if $R_q(v)=\langle \Pi, L\rangle$ for some self-coupling $\Pi$ of $q$, then $v\in V^*(q)$.*

*Proof.* Since $\sum_j \Pi_{ij} = q_i$ we may write $R_q(v)=\sum_{i,j}\Pi_{ij}S_M(v,i)$. By (1), $S_M(v,i)\ge L(i,j)+v_j-v_i$ for every $j$, and $\Pi_{ij}\ge 0$, so $R_q(v)\ge\sum_{i,j}\Pi_{ij}\big(L(i,j)+v_j-v_i\big)$. The score terms cancel: both $\sum_{i,j}\Pi_{ij}v_j$ and $\sum_{i,j}\Pi_{ij}v_i$ equal $\sum_j q_jv_j$, because $\Pi$ has both marginals equal to $q$. What remains is $\langle \Pi, L\rangle$. □

**Proposition 2.3** (complementary slackness)**.** *If $R_q(v)=\langle \Pi, L\rangle$ for a self-coupling $\Pi$ of $q$, then $\Pi_{ij}>0$ implies $S_M(v,i)=L(i,j)+v_j-v_i$.*

*Proof.* Every term of $\sum_{i,j}\Pi_{ij}\big(S_M(v,i)-L(i,j)-v_j+v_i\big)$ is nonnegative and the sum is zero by the computation in the previous proof, so each term vanishes. □

For $y\in\mathcal{Y}$ let $\varphi(y)\in\mathbb{Q}^{\mathcal{Y}}$ be the *embedded report* $\varphi(y)_z=-L(y,z)$.

**Lemma 2.4** (embedded reports; [1], Lemma 2.2)**.** *If $L$ is a metric then $S_M(\varphi(y),x)=2L(y,x)$ for all $x,y$, and hence $R_q(\varphi(y))=2\,\ell_q(y)$.*

*Proof.* The triangle inequality gives $L(x,z)-L(y,z)+L(y,x)\le 2L(y,x)$ for every $z$, with equality at $z=y$; sum against $q$ for the second claim. □

The next statement is the workhorse of the lower-bound arguments in [1], extracted for an arbitrary metric. It says that mass on the diagonal of an optimal self-coupling forces consistency outright, so no witness can have such a coupling.

**Proposition 2.5** (diagonal criterion)**.** *Let $L$ be a metric, $\Pi$ a self-coupling of $q$ with $R_q(v)=\langle \Pi, L\rangle$, and suppose $\Pi_{yy}>0$ for some $y$. Then* $\arg\max v=\{y\}$ *and $y\in B(q)$.*

*Proof.* Complementary slackness at $(y,y)$ gives $S_M(v,y)=L(y,y)+v_y-v_y=0$. Hence for every $z$, $0\ge L(y,z)+v_z-v_y$, so $v_y\ge v_z+L(y,z)>v_z$ whenever $z\ne y$: the maximiser is unique and equals $y$. The same inequality gives $v_y-v_i\ge L(y,i)$, whence $S_M(v,i)\ge L(i,y)+v_y-v_i\ge 2L(y,i)$ and $R_q(v)\ge 2\ell_q(y)$. But Proposition 2.2 and Lemma 2.4 give $R_q(v)=H_M(q)\le R_q(\varphi(y'))=2\ell_q(y')$ for every $y'$, so $\ell_q(y)\le\ell_q(y')$ for every $y'$, i.e. $y\in B(q)$. □

**Proposition 2.6** (small-surrogate criterion)**.** *If $S_M(v,y)\le L(y,z)$ for every $z\ne y$, then $y\in\arg\max v$.*

*Proof.* $v_z-v_y\le S_M(v,y)-L(y,z)\le 0$ for $z\ne y$. □

Each of the witnesses below is presented as a table of integers: a common denominator $D>0$, an integral metric $L$, integral weights $c$ with $\sum_y c_y=D$ and $q=c/D$, an integral score vector $v$, the integral values $S_M(v,\cdot)$ it produces, and an integral self-coupling $D\,\Pi$. Optimality is then Proposition 2.2 applied to the identity $\sum_{i,j}(D\Pi)_{ij}L(i,j)=\sum_y c_yS_M(v,y)$, an identity between integers.

## 3. The three counterexamples of the source, re-certified

Before anything new, the three counterexamples of [1] were re-derived from its own definitions in exact rational arithmetic and then verified formally. All three reproduce with the paper's printed values. This is what locates the open question, since two of the three fail Definition 1.1 and for different reasons.

**Proposition 3.1** ([1], Corollary 3.2: the four-output unit star)**.** *Let $\mathcal{Y}=\{o,a,b,c\}$ with $L(o,\cdot)=1$ and $L(a,b)=L(b,c)=L(c,a)=2$, let $q=(0,\frac13,\frac13,\frac13)$ and $v=(-1,0,0,0)$. Then $\ell_q=(1,\frac43,\frac43,\frac43)$, $B(q)=\{o\}$, $v\in V^*(q)$ with $R_q(v)=2$, and* $\arg\max v=\{a,b,c\}$ *is disjoint from $B(q)$. The distribution is* not *fully supported: $q_o=0$.*

**Proposition 3.2** ([1], Theorem 5.1 at $m = 2$, $n = 3$: $K_{2,3}$)**.** *Let $L$ be the shortest-path metric of $K_{2,3}$ on $A = \{a_1, a_2\}$, $B = \{b_1, b_2, b_3\}$, let $q$ be uniform and $v = (-1, -1, 0, 0, 0)$. Then $\ell_q = (1, 1, \frac{6}{5}, \frac{6}{5}, \frac{6}{5})$, $B(q) = \{a_1, a_2\}$, $v \in V^*(q)$ with $H_M(q) = 2$, and* $\arg\max v = B$ *is disjoint from $B(q)$. The Bayes set has* two *elements: $B(q) \neq \{y_0\}$ for every $y_0$.*

**Proposition 3.3** ([1], Theorem 5.3: the three-dimensional Hamming cube)**.** *Let $\mathcal{Y} = \{0, 1\}^3$ with the Hamming metric, listed lexicographically, let $q = (2, 1, 1, 1, 1, 2, 2, 1)/11$, and let $v$ be $0$ on the even-parity outputs and $-1$ on the odd-parity ones. Then $v \in V^*(q)$ with $H_M(q) = 30/11$, $B(q) = \{100\}$ is a singleton, and* $\arg\max v$ *is the even-parity class, disjoint from $B(q)$. Hence there is a witness on eight outputs and $k^* \leq 8$. Moreover the Hamming metric on $\{0, 1\}^3$ is* median*: every one of its* 512 *triples has exactly one common geodesic point.*

That the Hamming cube is median is classical; [1] states it and attributes it to the survey [6], and it is re-decided here triple by triple.

The failures are instructive and complementary. The star fails full support; the source's own Theorem 4.5 shows this is unavoidable at four outputs. The five-point $K_{2,3}$ witness is fully supported and has disjoint argmax, but its Bayes set is the whole light part. The cube is a genuine witness, and is the source's upper bound. The three of them are exactly the bracket $5 \leq k^* \leq 8$ made concrete.

## 4. Six outputs suffice

Let $L$ be the shortest-path metric of $K_{3,3}$ on $\mathcal{Y} = A \sqcup B$ with $A = \{a_1, a_2, a_3\}$ and $B = \{b_1, b_2, b_3\}$, so $L = 2$ inside a part, $L = 1$ across, and $L = 0$ on the diagonal. We index $\mathcal{Y}$ in the order $a_1, a_2, a_3, b_1, b_2, b_3$.

**Theorem 4.1** ($k^* \leq 6$)**.** *Let $L$ be the shortest-path metric of $K_{3,3}$, let*

$$q = \tfrac{1}{12}(1, 1, 2, 2, 3, 3), \qquad v = (-1, -1, -1, 0, 0, 0).$$

*Then $L$ is a metric satisfying* (5)*; $q$ is fully supported with $\sum_y q_y = 1$; $\ell_q = \left(\frac{7}{6}, \frac{7}{6}, 1, \frac{4}{3}, \frac{7}{6}, \frac{7}{6}\right)$, so $B(q) = \{a_3\}$ is a singleton; $S_M(v, y) = 2$ for every $y$, so $R_q(v) = 2 = H_M(q)$ and $v \in V^*(q)$; and* $\arg\max v = \{b_1, b_2, b_3\}$ *is disjoint from $B(q)$. Hence there is a witness on six outputs and*

$$k^* \;\leq\; 6.$$

*Moreover $L$ is modular but* not *median.*

*Proof.* The metric axioms and (5) are finitely many identities and inequalities in the $6 \times 6$ integer table; (5) is the statement that each of the 216 triples has a common geodesic point, and it holds: a triple meeting both parts is served by whichever of its outputs is alone in its part, and a triple inside one part by any output of the other. The values $\ell_q$ are the six inner products of $q$ against the rows of $L$; the smallest is $\ell_q(a_3) = 1$, the next $\frac{7}{6}$, so the Bayes set is $\{a_3\}$. For the surrogate values, $S_M(v, y) = 2$ for every $y$: for $y \in A$ the loss-augmented maximum is $\max(0 - 1,\, 2 - 1,\, 1 + 0) + 1 = 2$ and for $y \in B$ it is $\max(1 - 1,\, 2 + 0,\, 0) - 0 = 2$. Hence $R_q(v) = 2$.

Optimality is Proposition 2.2 applied to the self-coupling $\Pi$ with

$$12\,\Pi_{a_3a_1} = 12\,\Pi_{a_1a_3} = 12\,\Pi_{a_3a_2} = 12\,\Pi_{a_2a_3} = 1,$$
$$12\,\Pi_{b_1b_2} = 12\,\Pi_{b_2b_1} = 12\,\Pi_{b_1b_3} = 12\,\Pi_{b_3b_1} = 1, \qquad 12\,\Pi_{b_2b_3} = 12\,\Pi_{b_3b_2} = 2,$$

and all other entries zero. Its row and column sums are $12q$, and every supported pair lies inside a part, so $L = 2$ on the support and $\langle \Pi, L \rangle = 2 \sum_{i,j} \Pi_{ij} = 2 = R_q(v)$. Then $v \in V^*(q)$ and $H_M(q) = 2$. Finally $\arg\max v = B$ because $v$ is 0 on $B$ and $-1$ on $A$, and $B \cap \{a_3\} = \emptyset$.

For the last sentence: $I(a_1, a_2) = \{a_1, a_2, b_1, b_2, b_3\}$ and likewise for the other two pairs inside $A$, so the triple $(a_1, a_2, a_3)$ has all three of $b_1, b_2, b_3$ as common geodesic points — the intersection has three elements, not one. □

Every assertion in Theorem 4.1 except the optimality of $v$ is a finite check on the integer table, and is decided by kernel evaluation (§8). Optimality is *not* a finite check — it quantifies over all score vectors — and is discharged by Proposition 2.2 from the single certificate $\Pi$.

**Theorem 4.2** (a second witness, with the Bayes output in the heavy part)**.** *On the same metric let* $q' = \frac{1}{7}(1,1,1,2,1,1)$ *and* $v' = (1,1,1,0,0,0)$*. Then* $q'$ *is fully supported,* $B(q') = \{b_1\}$ *is a singleton,* $v' \in V^*(q')$ *with* $R_{q'}(v') = 2$*, and* $\arg\max v' = \{a_1, a_2, a_3\}$ *is disjoint from* $B(q')$*.*

*Proof.* As before: $\ell_{q'}(b_1) = 1$ while $\ell_{q'}(a_i) = \frac{8}{7}$ and $\ell_{q'}(b_2) = \ell_{q'}(b_3) = \frac{9}{7}$; $S_M(v', y) = 2$ for every $y$; and the self-coupling putting $14\,\Pi_{ij} = 1$ on each ordered pair of distinct elements of $A$, $14\,\Pi_{b_1 b_2} = 14\,\Pi_{b_1 b_3} = 14\,\Pi_{b_2 b_1} = 14\,\Pi_{b_3 b_1} = 2$ and zero elsewhere has both marginals $14q'$, support inside the parts, and value $2 = R_{q'}(v')$. □

The two witnesses differ in more than their arithmetic: the single Bayes output sits in the light part in Theorem 4.1 and in the heavy part in Theorem 4.2, so the failure of (3) on $K_{3,3}$ is not a knife-edge feature of one distribution. The second witness was found by the decision procedure of §6, and only then verified.

4.1. **Why the source's theorems do not reach** $K_{3,3}$**.** Theorem 5.1 of [1] is stated for $K_{m,n}$ with $2 \le m < n$ at the *uniform* distribution, and both hypotheses fail at $m = n = 3$ at once. The following control makes the second failure precise, and part (iv) is the reason the tie must be broken by hand.

**Proposition 4.3** (controls)**.**
(i) *There is no witness on* 0 *and none on* 1 *output.*
(ii) *The three-point equilateral metric is a metric and fails* (5)*, so* (5) *is not vacuous.*
(iii) *At the metric and distribution of Theorem 4.1, the embedded report* $\varphi(a_3)$ *is* also *an exact minimiser of* $R_q$*, and* $\arg\max \varphi(a_3) = B(q)$*: the canonical decoder reads that minimiser correctly.*
(iv) *At the metric of Theorem 4.1 and the* uniform *distribution on six outputs,* $B(q) = \mathcal{Y}$*: every output is Bayes.*
(v) *The embedded report of the non-Bayes output* $b_1$ *has* $R_q(\varphi(b_1)) = \frac{8}{3} > 2$*, so it is not a minimiser and the optimality hypothesis of Definition 1.1 is not vacuous.*

Part (iv) says exactly why the uniform distribution — the one Theorem 5.1 of [1] uses — can carry no witness on $K_{3,3}$: with the two parts of equal size the Bayes set is everything, so the disjointness requirement is unsatisfiable there for every $v$. A witness on $K_{3,3}$ must break the tie, which is the step the source, holding the $K_{m,n}$ family, does not take.

*Remark* 4.4 (the denominator 7 of Theorem 4.2 is smallest possible). A fully supported rational distribution on six outputs with common denominator $D$ has six positive integer numerators summing to $D$, so $D \ge 6$, and $D = 6$ forces the uniform distribution, where by Proposition 4.3 (iv) every output is Bayes and no score vector can have disjoint argmax. So no witness on $K_{3,3}$ has a distribution of denominator below 7, and Theorem 4.2 attains 7. The counting step here is an elementary argument on the page; only the input — part (iv) — is machine-checked.

## 5. Witnesses at every cardinality at least six

The source's own infinite family (its Theorem 5.1) is $K_{m,n}$ with $2 \le m < n$ at the uniform distribution, whose Bayes set is the whole light part and never a singleton. Breaking the tie *inside* one part gives an infinite family of single-Bayes witnesses instead. We first record the instance that is verified formally.

**Theorem 5.1** (a seven-output witness)**.** *Let* $L$ *be the shortest-path metric of* $K_{3,4}$ *on* $A = \{a_1, a_2, a_3\}$*,* $B = \{b_1, b_2, b_3, b_4\}$*, let* $q = \frac{1}{8}(2,1,1,1,1,1,1)$ *in the order* $a_1, a_2, a_3, b_1, \ldots, b_4$ *and let* $v$ *be* 0 *on* $A$ *and* 1 *on* $B$*. Then* $q$ *is fully supported,* $\ell_q(a_1) = 1$ *while* $\ell_q(y) = \frac{5}{4}$ *for every other* $y$*, so* $B(q) = \{a_1\}$*;* $S_M(v, y) = 2$ *for every* $y$ *and* $v \in V^*(q)$*; and* $\arg\max v = B$ *is disjoint from* $B(q)$*. Hence there is a witness on seven outputs.*

*Proof.* As in Theorem 4.1, with the self-coupling that splits $a_1$'s mass equally against $a_2$ and $a_3$ and pairs $b_1$ with $b_3$ and $b_2$ with $b_4$ in both directions; every supported pair is inside a part, so the value is 2, which equals $R_q(v)$. □

Together with Theorem 4.1 and Proposition 3.3, this certifies witnesses at cardinalities 6, 7 and 8. The general statement is the following; it is proved here on the page and is *not* part of the formal development, for the reason given after the proof.

**Theorem 5.2** (the tie-broken complete-bipartite family; proved on the page, not machine-checked)**.** *Let $m \geq 3$ and $n \geq m$, let $\mathcal{Y} = A \sqcup B$ with $|A| = m$, $|B| = n$, and let $L$ be the shortest-path metric of $K_{m,n}$. Put $N = m + n + 1$ and*

$$q(a_1) = \frac{2}{N}, \qquad q(a_i) = \frac{1}{N} \ (i \geq 2), \qquad q(b_j) = \frac{1}{N}, \qquad v = 0 \text{ on } A, \quad v = 1 \text{ on } B.$$

*Then $q$ is fully supported with $\sum_y q_y = 1$, $B(q) = \{a_1\}$, $v \in V^*(q)$ and $\arg\max v = B$. Hence there is a witness on $m + n$ outputs, and since $\{m + n : m \geq 3,\ n \geq m\} = \{k : k \geq 6\}$, there is a witness on $k$ outputs for every $k \geq 6$.*

*Proof.* That $L$ is a metric satisfying (5) is the first paragraph of the proof of Theorem 5.1 of [1]. Writing $P_y$ for the part containing $y$, one has $\ell_q(y) = 2\big(q(P_y) - q_y\big) + \big(1 - q(P_y)\big) = 1 + q(P_y) - 2q_y$; with $q(A) = (m + 1)/N$ and $q(B) = n/N$ this gives

$$\ell_q(a_1) = 1 + \frac{m - 3}{N}, \qquad \ell_q(a_i) = 1 + \frac{m - 1}{N}, \qquad \ell_q(b_j) = 1 + \frac{n - 2}{N},$$

so $B(q) = \{a_1\}$, using $m - 3 < m - 1$ and (since $m \leq n$) $m - 3 < n - 2$. Next $S_M(v, y) = 2$ for every $y$: for $y \in A$ the loss-augmented maximum is $\max(0 + 0,\ 2 + 0,\ 1 + 1) = 2$ and $v_y = 0$; for $y \in B$ it is $\max(0 + 1,\ 2 + 1,\ 1 + 0) = 3$ and $v_y = 1$. Hence $R_q(v) = 2$.

For the dual certificate, note that a complete graph whose vertex capacities have maximum at most half their total carries a perfect fractional matching: lay the capacities out as arcs on a circle whose circumference is the total and match every point to its antipode; no arc meets its own image, because its length is at most half the circumference. Inside $A$ this needs $2/N \leq (m + 1)/(2N)$, i.e. $m \geq 3$; inside $B$ it needs $1/N \leq n/(2N)$, i.e. $n \geq 2$. Let $\Pi$ be the union of the two matchings: it is a symmetric self-coupling of $q$ with zero diagonal, supported inside the parts, so $L = 2$ on its support and $\langle \Pi, L \rangle = 2 \sum_{i,j} \Pi_{ij} = 2 = R_q(v)$. Proposition 2.2 gives $v \in V^*(q)$. Finally $\arg\max v = B$, disjoint from $\{a_1\}$. □

The hypothesis $m \geq 3$ is sharp in the construction, and its failure at $m = 2$ is exactly the source's own case: with $|A| = 2$ the capacities $(2, 1)/N$ inside $A$ have maximum $2/N$, more than half of $3/N$, and no perfect fractional matching inside $A$ exists. Theorem 5.2 was checked in exact arithmetic at the nine pairs $m = 3, 4, 5$, $n = m, \ldots, 6$ — metric, modularity, full support, singleton Bayes set, exact optimality through the self-coupling and disjoint argmax all confirmed. At $m = 2$ it correctly reports that this $q$ and $v$ are not a witness, because the score vector is no longer optimal: a self-coupling has value 2 only if all its mass sits on pairs at distance 2, that is inside the parts, and inside $A$ the row and column sums at $a_1$ and $a_2$ then force $\Pi_{a_2 a_1}$ to be both $2/N$ and $1/N$. So $H_M(q) < 2 = R_q(v)$ there. Only the instances $m = n = 3$ (Theorem 4.2) and $m = 3$, $n = 4$ (Theorem 5.1) are verified formally: a certificate is a finite table, whereas the general case needs the circle construction for an arbitrary capacity vector, which is a different kind of argument.

## 6. Five points: an exact decision over all distributions and all scores

This section is a computation with a proof attached; like Theorem 5.2 it is outside the formal development, and we say at the end exactly what it does and does not establish. Fix a metric $L$ on $n$ points. We decide, for that $L$, whether *any* fully supported rational $q$ and *any* score vector $v$ form a witness.

The starting point is Proposition 2.5: in a witness no optimal self-coupling has a positive diagonal entry. Two further steps make the search finite.

**Lemma 6.1** (canonical support)**.** *Suppose $(L, q, v)$ is a witness. The set of optimal self-couplings of $q$ is a face of the transportation polytope $\Gamma(q, q)$, closed under transposition (because $L$ is symmetric) and under averaging, so it contains a symmetric member $\Pi$, necessarily with zero diagonal. Then $m_{ij} = \Pi_{ij}$*

*is a fractional perfect matching with capacities $q$, and an extreme point $m^*$ of the intersection of the optimal face with that polytope is an extreme point of the fractional-perfect-matching polytope itself, so its support is a spanning subgraph each of whose components is a tree or is unicyclic with an odd cycle. Complementary slackness (Proposition 2.3) applies to $m^*$, since it is optimal.*

**Lemma 6.2** (decoupling)**.** *Fix $L$ and a candidate Bayes output $y^*$. A witness with Bayes set $\{y^*\}$ exists if and only if for some support $S$ as in Lemma 6.1 both of the following hold.*

- (V) *The system $L(i,j)+v_j \geq L(i,z)+v_z$ — for every $i$, every $j$ with $\{i,j\} \in S$ and every $z$ — has a solution $v$ with $v_{y^*} < v_z$ for some $z$.*
- (Q) *There are $q > 0$ with $\sum_y q_y = 1$ and a symmetric $m \geq 0$ supported on $S$ with $\sum_j m_{ij} = q_i$, such that $\ell_q(y^*) < \ell_q(y)$ for every $y \neq y^*$.*

*Proof sketch.* ($\Leftarrow$) The constraints in (V) say that for $\{i,j\} \in S$ the loss-augmented maximum in row $i$ is attained at $j$, so $S_M(v,i) = L(i,j)+v_j-v_i$ there. Summing against $m$ and cancelling the score terms as in Proposition 2.2 gives $\sum_i q_i S_M(v,i) = \langle m, L\rangle$, so $v$ is optimal for $q$; (Q) makes $q$ fully supported with $B(q) = \{y^*\}$, and (V) puts $y^*$ outside the argmax. ($\Rightarrow$) is Proposition 2.5 and Lemma 6.1. □

The point of the decoupling is that (V) does not involve $q$ and (Q) does not involve $v$, and each is small. System (V) is a system of difference constraints $v_j - v_z \geq L(i,z) - L(i,j)$: it is feasible if and only if its constraint digraph has no negative cycle, and $\max(v_z - v_{y^*})$ over its solutions is a shortest-path distance, so one exact-rational Floyd–Warshall decides it for all pairs $(y^*, z)$ at once. System (Q) is a small exact rational linear program with strict inequalities; infeasibility is certified by a Motzkin transposition certificate and feasibility by an explicit rational point, and in this implementation both are re-verified in exact arithmetic independently of the simplex code that produced them, so the linear-programming code affects completeness of the search but never soundness of its verdicts. Supports are enumerated exhaustively — 327 of them at $n = 5$ and 4411 at $n = 6$ — and reduced modulo the isometry group of $L$ acting on the pair $(S, y^*)$.

6.1. **What the procedure returns at five points.** Two finite classes of five-point modular metrics were enumerated and decided.

- Of the 21 connected graphs on five vertices, exactly 5 have a modular shortest-path metric: $K_{1,4}$, the spider with arms $2,1,1$, the path $P_5$, the 4-cycle with a pendant vertex, and $K_{2,3}$. Three are trees and four are median. Every one of the five returns *no witness*, in 266 seconds in total.
- Of the 228,502 labelled metrics on five points with integer distances at most 4, exactly 780 are modular, falling into 17 isometry classes (15 of them median), so twelve genuinely weighted metrics beyond the five graph metrics. Every one of the seventeen returns *no witness*, in 1,060 seconds of certificates.

Here "no witness" means: there is no fully supported rational $q$ with a single Bayes output and no score vector $v \in V^*(q)$ whose argmax misses it — over *all* rational $q$ and all score vectors — with a Motzkin or Farkas certificate for each of the (support, $y^*$) orbits.

The procedure was validated three ways, and the third is the one that matters. (i) On four four-point modular metrics — the unit star, the unit rectangle, the weighted rectangle with sides 1 and 2, and the four-point path — it returns no witness, agreeing on those instances with Theorem 4.5 of [1]; that paper proves the general four-point statement by a classification argument the procedure knows nothing about. (ii) On $K_{3,3}$ it *finds* a witness from scratch in about four seconds, namely the one of Theorem 4.2; a procedure that only ever said "none" would prove nothing. (iii) Dropping the unique-Bayes requirement alone — replacing "$B(q)$ is a singleton" by "$B(q)$ equals a prescribed set" — and re-running on $K_{2,3}$, the very metric that returns nothing above, recovers the source's own Theorem 5.1 witness with its printed numbers: the uniform $q$, $v = (-1,-1,0,0,0)$, $H_M = 2$, $\ell_q = (1,1,\frac{6}{5},\frac{6}{5},\frac{6}{5})$, $B(q) = \{a_1,a_2\}$ and $\arg\max v = \{b_1,b_2,b_3\}$. So the absence of a five-point witness is caused by the unique-Bayes requirement and by nothing else.

**Corollary 6.3** (conditional)**.** $k^* = 6$*, conditional on the enumeration above: no five-point modular metric arising from a connected graph, and no five-point modular metric with integer distances at most* 4*, carries a witness, and Theorem 4.1 exhibits one on six outputs.*

The condition is not decorative. A lower bound $k^* \geq 6$ quantifies over the whole cone of five-point modular metrics, a continuum; the two classes above are the unweighted skeleton and a distance-$\leq 4$ grid on that cone. Every finite metric is the shortest-path metric of a weighted graph, so what is missing is a weight continuum inside finitely many combinatorial types.

## 7. What remains

7.1. **The five-point continuum.** The obstruction to closing Corollary 6.3 is not computational cost. The witness system is *biaffine*: it is jointly linear in $(L, v, \xi)$ where $\xi_i = S_M(v, i)$, and jointly linear in $(q, m)$, and the only bilinear coupling is through the Bayes constraints $\sum_j q_j\big(L(y^*, j) - L(y, j)\big) < 0$, which mix $q$ with $L$. One linear program therefore cannot decide a cone of metrics. Two routes are open: classify five-point modular metric spaces in the style of the four-point classification of [1] (its Lemma 4.3: star, path, weighted rectangle) and settle each type by hand; or decide bilinear feasibility per type, in small dimension but with machinery we do not have here.

7.2. **The median sub-question.** Among graph shortest-path metrics the median ones are, by [3], exactly the modular ones with no induced $K_{2,3}$: verbatim, "median graphs are modular graphs not containing induced $K_{2,3}$". Both small witnesses known — the source's $K_{2,3}$ and our $K_{3,3}$ — therefore sit exactly at the minimal obstruction to being median, while the source's cube is median (Proposition 3.3). So the same sharp cardinality restricted to median metrics is at most 8, and at least 6 under the same condition as Corollary 6.3; whether it is 6, 7 or 8 is open. Deciding it at six points is a finite job that we did not finish: of the 112 connected graphs on six vertices, 16 have a modular metric (11 median, six of them trees), and 8 of the 16 were decided — seven returning no witness, and $K_{3,3}$ returning Theorem 4.2 — leaving 8 undecided. We report this because it bounds what is known, not because it settles anything.

*Remark* 7.1 (counts, reported as a measurement). The number of connected graphs on $n$ vertices whose shortest-path metric is modular, and the number whose metric is median, for $n = 1, \ldots, 8$:

$$\text{modular}: \ 1, 1, 1, 3, 5, 16, 41, 156; \qquad \text{median}: \ 1, 1, 1, 3, 4, 11, 23, 69.$$

The median row is OEIS `A292623` [4] ("Number of median graphs on $n$ nodes"), matched on all eight terms and on its definition. The modular row is not in OEIS; its six-term prefix collides with two bipartite-graph sequences that separate from it at $n = 8$ and differ in definition. Every term of both rows has been produced by at least two enumerations that share no code: one built on `nauty`'s isomorphism-class generator, and one that enumerates *labelled* graphs exhaustively — a modular graph is triangle-free, so that search can be grown one vertex at a time — and recovers the class counts as orbit counts of the symmetric group acting on the labelled set. They agree on all sixteen terms, and the labelled run's intermediate counts of triangle-free graphs, $1, 2, 7, 41, 388, 5789, 133501, 4682270$, are OEIS `A213434` [5]. These counts support no theorem in this paper; they say how large the six- and seven-point censuses are.

7.3. **Formalising the exhaustion.** The decision procedure of §6 is not formalised, and would need two things absent from the library we build on: extreme points of transportation polytopes, and Motzkin transposition certificates. The certificate side of this paper needs neither, which is why every certificate in §§2–5 is verified and §6 is not.

## 8. Verification

Theorems 4.1, 4.2 and 5.1, Propositions 2.2, 2.3, 2.5, 2.6, 3.1, 3.2, 3.3 and 4.3, and Lemma 2.4 have been formally verified in Lean 4 against *Mathlib* [15, 16]. The development is six modules and about 1,070 lines: the definitions of §1 and the certificate layer of §2; a structure carrying one instance's integer certificate data together with the bridge from integer to rational statements; the two

six-output witnesses; the seven-output witness; the source's three counterexamples; and the controls. There is no `sorry` and no axiom is declared by hand: a print of the axiom set returns `propext`, `Classical.choice` and `Quot.sound`, and nothing else, for all forty declarations. In particular there is no compiled-evaluation axiom — no `native_decide` — so every finite check below is performed by the kernel itself.

What is decided by exhaustive evaluation, and what is not, is worth stating precisely. Decided by evaluation over the finite index set: the metric axioms; the common-geodesic condition (5), as $6^3 = 216$ triples for $K_{3,3}$, $7^3 = 343$ for $K_{3,4}$ and $8^3 = 512$ for the cube, and the same 512 triples for the medianness of the cube — while the non-medianness of $K_{3,3}$ is decided on the single triple $(a_1, a_2, a_3)$, after evaluating the three metric intervals it involves; the surrogate values $S_M(v, y)$, through an upper bound valid for every competitor and one competitor attaining it; the marginals and the value of each self-coupling; and the Bayes and argmax sets, read off the integer tables. *Not* decided by evaluation, and proved once for an arbitrary finite metric: everything in §2, in particular the weak-duality inequality that turns a coupling into a proof of optimality against all score vectors. So the quantifier over score vectors is never searched — it is discharged by Proposition 2.2.

The numerical values quoted alongside these statements for orientation — the vectors $\ell_q$ and the risks $R_q(v)$ and $H_M(q)$ — were recomputed in exact rational arithmetic, and for the three counterexamples of [1] they agree with that paper's printed values; what the formal development certifies about each instance is the list in the previous paragraph together with the exact optimality of its score vector.

One implementation note, because it shapes every witness above. Certificate data is carried as integers scaled by a common denominator, not as rationals: in this library rational addition normalises through a greatest-common-divisor computation that the kernel does not reduce, so kernel evaluation is unusable on rational tables — even $2+2=4$ over the rationals fails to evaluate — while the identical table over the integers evaluates without difficulty. All rational reasoning therefore happens once, in the bridge lemmas, and every instance is a table of integers plus a kernel evaluation.

Deliberately outside the formal development, and labelled where they appear: Theorem 5.2 and the whole of §6, together with the exact-arithmetic checks of Theorem 5.2 at nine parameter pairs; the counting argument of Remark 4.4, whose input is verified but whose one-line conclusion is not; the counts of Remark 7.1; the six-point census of §7; and every statement attributed to [1] or [2] that is quoted rather than re-proved — in particular $k^* \geq 5$, which is the source's Theorem 4.5 and is used here only as the lower end of the bracket.

The source of the development is currently held in a private repository: repository reference to be added at submission.

*On the reference list.* Every theorem, corollary, lemma, proposition, equation and section number attributed to [1] above is the printed numbering of its **v1** e-print, read off a compilation of the v1 e-print itself and cross-checked against the PDF arXiv serves for v1. That numbering is version-sensitive, and v1 is the only version: the arXiv record was fetched again on 3 September 2026 and shows one submission, 27 August 2026 14:47:51 UTC, no withdrawal, no erratum, no journal reference and no publisher DOI (only the arXiv-issued DataCite DOI). Should a v2 appear, every number attributed to [1] must be re-read against it. Two further notes. The proceedings volumes of NeurIPS 2017 and NeurIPS 2019 print no page ranges, so [7] and [10] carry none. The published versions of [2] and [9] print their first author as A. Nowak; the arXiv versions, and this paper, use A. Nowak-Vila.

## References

[1] J. Fei and J. Luo, *Common Geodesics Do Not Guarantee Fisher Consistency of the Structured SVM: Minimal Counterexamples and a Tree-Metric Classification*, arXiv:2608.27203v1 [cs.LG], submitted 27 August 2026. The source of the problem and of everything attributed above: the surrogate (1), the risks (2), the consistency notion (3), the common-geodesic condition (5) and the modular/median terminology, the self-coupling dual and its complementary slackness, its Lemma 2.2 on embedded reports, its Corollary 3.2, Theorems 3.3, 3.5, 4.5, 5.1 and 5.3, its Lemma 4.3 and Proposition 4.4, the decoder discussion of its Section 6 and its Corollary 6.2, and the open question quoted in §1.2. arXiv record checked live on 3 September 2026: one version, primary category cs.LG with no cross-lists, journal-reference and publisher-DOI fields empty, no withdrawal.

[2] A. Nowak-Vila, A. Rudi and F. Bach, *On the consistency of max-margin losses*, arXiv:2105.15069; Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS 2022), PMLR **151**, 4612–4633. Both the arXiv version and the PMLR version were read on 3 September 2026. Cited for its Theorem 2.1 (the common-geodesic condition is necessary), for the open problem quoted in §1, and for its Appendix Theorem B.9 on tree metrics.

[3] J. Chalopin, M. Changat, V. Chepoi and J. Jacob, *First-order logic axiomatization of metric graph theory*, arXiv:2203.01070. Read on 3 September 2026; cited for the single sentence quoted in §7, that median graphs are modular graphs with no induced $K_{2,3}$.

[4] The On-Line Encyclopedia of Integer Sequences, sequence `A292623`, *Number of median graphs on n nodes*. Consulted on 3 September 2026; its terms $1, 1, 1, 3, 4, 11, 23, 69, 178, 567$ agree with the median row of Remark 7.1 on all eight terms computed there.

[5] The On-Line Encyclopedia of Integer Sequences, sequence `A213434`, *$a(n)$ is the number of labeled triangle-free simple graphs on n vertices*. Consulted on 3 September 2026; its terms $1, 2, 7, 41, 388, 5789, 133501, 4682270$ agree with the intermediate counts of the labelled enumeration behind Remark 7.1.

[6] H.-J. Bandelt and V. Chepoi, *Metric graph theory and geometry: a survey*, in: Surveys on Discrete and Computational Geometry: Twenty Years Later, Contemporary Mathematics **453**, American Mathematical Society, 2008, 49–86, doi:10.1090/conm/453/08795. Cited by [1] for the classical medianness of the Hamming cube; the entry here was checked against the Crossref record for that DOI and against the zbMATH record, which agree on title, authors, series volume, pages and year.

[7] A. Osokin, F. Bach and S. Lacoste-Julien, *On structured prediction theory with calibrated convex surrogate losses*, Advances in Neural Information Processing Systems **30** (NeurIPS 2017), Curran Associates; arXiv:1703.02403.

[8] H. G. Ramaswamy and S. Agarwal, *Convex calibration dimension for multiclass loss matrices*, Journal of Machine Learning Research **17** (2016), no. 14, 1–45.

[9] A. Nowak-Vila, F. Bach and A. Rudi, *Consistent structured prediction with max-min margin Markov networks*, Proceedings of the 37th International Conference on Machine Learning (ICML 2020), PMLR **119**, 7381–7391.

[10] J. Finocchiaro, R. Frongillo and B. Waggoner, *An embedding framework for consistent polyhedral surrogates*, Advances in Neural Information Processing Systems **32** (NeurIPS 2019), Curran Associates; arXiv:1907.07330.

[11] J. Finocchiaro, R. M. Frongillo and B. Waggoner, *An embedding framework for the design and analysis of consistent polyhedral surrogates*, Journal of Machine Learning Research **25** (2024), no. 63, 1–60.

[12] A. Mao, M. Mohri and Y. Zhong, *Structured prediction with stronger consistency guarantees*, Advances in Neural Information Processing Systems **36** (NeurIPS 2023), Curran Associates, 46903–46937, doi:10.52202/075280-2032.

[13] A. Tewari and P. L. Bartlett, *On the consistency of multiclass classification methods*, Journal of Machine Learning Research **8** (2007), no. 36, 1007–1025.

[14] Y. Liu, *Fisher consistency of multicategory support vector machines*, Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS 2007), PMLR **2**, 291–298.

[15] L. de Moura and S. Ullrich, *The Lean 4 theorem prover and programming language*, in: Automated Deduction — CADE 28, Lecture Notes in Computer Science, Springer, 2021, 625–635, doi:10.1007/978-3-030-79876-5_37.

[16] The mathlib Community, *The Lean mathematical library*, in: Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs (CPP 2020), ACM, 2020, 367–381, doi:10.1145/3372885.3373824.