GraphMaxEnt

Arthur Ramer

School of Computer Science and Engineering
University of New South Wales
Sydney 2052, Australia

Abstract. Assume a undirected graph G on a finite domain X and a probability distribution P on X. Graph entropy, defined in terms of the vertex packing polytope, is recast as

$$H(G,P) = \min_R -\sum p_i \log \mathrm{Pl}^{(R)}(x_i),$$

with suitably defined plausibility $Pl^{(R)}$ wrt probability distribution R on the independent subsets of G.

The plausibility which results from the minimising R is called the *plausibility wrt* P on X serves to define the *graph information distance*

$$D(G,Q||P) = \sum q_i \log \frac{\mathrm{Pl}_Q(x_i)}{\mathrm{Pl}_P(x_i)}$$

for two distributions Q and P on X, given the graph structure G.

One verifies the usual properties of additivity and subadditivity wrt weak products of the supporting graphs. The method of GraphMaxEnt can be formulated accordingly. It is postulated that it admits an axiomatisation akin to that for MaxEnt.

Applied to probability kinematics, it permits interpreting a wide range of probability reassignments as the result of minimisation of graph information distance. This leads to defining Jeffrey rules for all these reassignments, along with inverse Jeffrey conditioning.

MaxGraphEnt is often successful in including new conditional information in situations when the standard MaxEnt may produce unsatisfactory and counterintuitive results. Such resolution, using entropies, of 'Private Benjamin problem' is presented.

Key Words: Graph entropy, probability kinematics, belief revision, inverse conditioning.

MOTIVATION

Entropy has been a favourite tool for probability kinematics problem. Although entropy originated in communication theory, it soon found applications in several other areas. Principle of maximum entropy (and minimum information distance) was very successfully employed in decision making, hereupon in probability kinematics and, by extension, in general belief kinematics in AGM model [1, 3] of belief revision. Application of these principles to probability revisions is discussed in detail in our earlier paper [14].

MaxEnt principle easily leads to *defining* conditional distributions as well to Jeffrey conditioning. It can also be adapted to resolve a problem of inverse conditioning - undoing the conditioning operation in the most 'reasonable' way. However, there are important models of reassigning probability and belief weights that do not conform to the proportional re-weighting. They are known under the name of *imaging*, and represent a selective transfer of beliefs. Given P on the universe of worlds $\{w_1, \ldots, w_n\}$ and a new logical restriction A such that P_A^+ must be zero on $w_{k+1}, \ldots, w_n$, we stipulate that each of these worlds transfers its probability to a specific, most-like-itself world; there is an implicit graph structure that determines, for every w_i, where we are permitted to transfer its weight of belief.

Under the AGM principles such process would be best justified as some minimal change. Standard entropy cannot accomplish it, but its extension to *graph entropy* succeeds.

A famous example of such situation is 'Private Benjamin Problem' posed by Bas van Fraasen [16]. He uses it to question the suitability of entropy (standard) to revision of probabilistic beliefs, not recognising that this is a problem for imaging and not just a simple conditioning.

The presence of graph structure on the elements of the domain accords a certain semantic structure to its domain. (In contradistinction, entropy is 'context-free'.) It restricts the permissible flows of probability assignments. Problem of van Fraasen has an implicit, built-in semantic restriction; when modeled as maximisation of graph entropy, the most natural solution ensues.

We propose to formalise this approach and formulate GraphMaxEnt - a family of two decision rules

- Given constraints on probabilist values and a graph of permitted probability transfers, assume the distribution that maximises the corresponding graph entropy.
- In the above setting, if a prior distribution is known, the posterior distribution is selected by minimising graph information distance.

While graph entropy [7] was already defined in 1973, information distance has not been defined. We do it in this paper. Similarly, the decision rules based on graph entropy are only defined here.

CONSTRUCTIVE DEFINITIONS

Shannon definition of entropy serves to measure various limits on communication capacities in channels where outputs are error-free: symbols received are unambiguous in that no two outputs may ever be confused. Once the information about $x_i \in X$ is transmitted there is no doubt about the identity of that x_i. (Such a transmission or choice from X is obviously subject to a probabilistic chance.) Allowing for such confusion should *lower* the entropy—indistinguishable elements could be, in a sense, transmitted together. A formal model [7, 8] recognises graph G on the vertices $\{x_i\}$, where an edge (x_i, x_j) is formed whenever these two vertices *cannot* be confused. Thus the standard entropy corresponds to the complete graph $K_n, n = |X|$. Contrariwise, a fully confusable arrangement consists of n isolated vertices, with the presumed entropy 0.

Given the distribution P on X, the definition of $H(G, P)$ requires considering probability distributions on the collection $\mathcal{I}$ of the *maximal independent* sets of vertices.[1]

[1] *Independent* means that no two vertices form an edge.

Denoting $\mathcal{I} = \{Y \subseteq X, Y\text{- max ind}\}$, we first need a *joint* probability distribution S on $\mathcal{I} \times X$, such that

- $S(Y,x) = 0$ if $x \notin Y$
- S projected onto X is precisely P

Let R be its projection onto $\mathcal{I}$. We put

$$H(G,P) = \min_S (H(P) + H(R) - H(S)).$$

Although the expression may seem convoluted, it is actually quite easy to work with; in particular, there is a simple algorithm finding the minimising S and computing the entropy. There is an equivalent definition due to Simonyi [15], based on the notion of *vertex packing polytope*. This is less suitable for computations, but better for generalisations. However, it can be recast into a very useful formula using the notion of plausibility. With notation as above, we first consider arbitrary probability distribution R defined on $\mathcal{I}$ and put

$$\mathrm{Pl}^{(R)}(x) = \sum_{Y : x \in Y} R(Y).$$

We have a fairly easy result

$$H(G,P) = \min_R - \sum p_i \log \mathrm{Pl}^{(R)}(x_i).$$

It holds that R that minimises the expression above is the same distribution as in Korner and Simonyi [15] definitions. We use it to define the *plausibility wrt* P on X

$$\mathrm{Pl}_P(x) = \mathrm{Pl}^{(R)}(x), \quad R = \arg\min H(G,P).$$

It serves to define the *graph information divergence*

$$D(G,Q||P) = \sum q_i \log \frac{\mathrm{Pl}_Q(x_i)}{\mathrm{Pl}_P(x_i)}$$

for two distributions Q and P on X, given a (fixed) graph structure G. It is straightforward to offer a similar definition wrt the change of G, but it produces useful results only in restricted cases.

GRAPH ENTROPY COMPUTATIONS

We discuss here the question of obtaining *closed form* expressions for fairly simple graphs. The problem of numerical solutions, whether for graph entropy or graph distance is easily tractable by any better package for convex optimisation. For our purposes, esp. to compute graph information distance, one would like to have just one probability distribution on independent subsets that would minimise the expression required to define graph entropy. It is always the case - we reiterate, after Simonyi [15] that entropy computation has always a unique minimising argument. It means that when probability distribution on vertices is Q and on independent sets is R then

$$\arg\min_R H(G,Q) = \arg\min_R -\sum q_i \log \mathrm{Pl}_Q^{(R)}(x_i)$$

is uniquely defined.

A similar argument applies to the case when probability distribution on vertices remains Q, while plausibility is computed wrt distribution P giving a unique minimising argument

$$\arg\min_R \; -\sum q_i \log \mathrm{Pl}_P^{(R)}(x_i).$$

Lastly, the distance is the difference of these solutions

$$D(G,Q||P) = -\sum q_i \log \mathrm{Pl}_P(x_i) - H(G,Q).$$

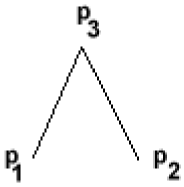
We start by demonstrating that even the original definition based on forming explicitly a joint distribution on $V(G)$ and $\mathcal{J}$ - set of all independent maximal sets, is quite workable. We recall that we need to form S - joint probability on $(V,\mathcal{J})$ supported on incident pairs

$$S(v,I) = 0 \text{ if } v \notin I, \quad S_{\downarrow 1} = P$$

Then we can define probability on $\mathcal{J}$ as projection $Q := S_{\downarrow 2}$ and find

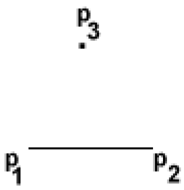
$$H(G,P) = \min_S (H(P) + H(Q) - H(S)) = \min_S I(P,Q).$$

We are now ready to look at few examples.



$\mathcal{J} \setminus V$	p_1	p_2	p_3
$\{p_1,p_2\}$	p_1	p_2	0
$\{p_3\}$	0	0	p_3

$$\begin{aligned} H(G) &= H(P) + H(Q) - H(S) \\ &= H(p_1,p_2,p_3) + H(p_1+p_2,p_3) - H(p_1,p_2,p_3) \\ &= -(p_1+p_2)\log(p_1+p_2) - p_3 \log p_3 \end{aligned}$$

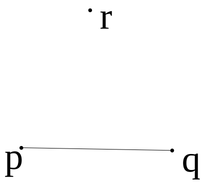


$\mathcal{J} \setminus V$	p_1	p_2	p_3
$\{p_1,p_3\}$	p_1	0	x
$\{p_2,p_3\}$	0	p_2	y

Under $x+y=p_3$ the solution to $\arg\min_{<x,y>}(H(p_1+x,p_2+y)-H(p_1,p_2,x,y))$ becomes

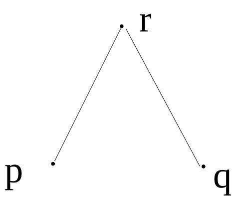
$$x=\frac{p_1p_3}{p_1+p_2},\quad y=\frac{p_2p_3}{p_1+p_2}$$

For comparison, computations based on plausibility are only one line long



$$\begin{aligned}\mathcal{J} &= \{pr,qr\},\ I=\mathcal{P}(pr),\ II=\mathcal{P}(qr)\\ H &= \min(-p\log I-q\log II-r\log(I+II))\\ &= -p\log\frac{p}{p+q}-q\log\frac{q}{p+q}\end{aligned}$$

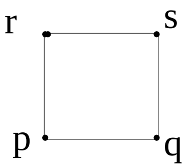
and the companion case



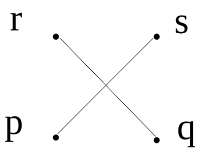
$$\begin{aligned}\mathcal{J} &= \{pq,r\},\ I=\mathcal{P}(pq),II=\mathcal{P}(r)\\ H &= \min(-(p+q)\log I-r\log II\\ &= -p\log(p+q)-q\log(p+q)-r\log r\end{aligned}$$

We can confirm that K_3 - complete graph on three vertices indeed has the entropy-splitting property (it is perfect), as the sum of these two entropies is the ordinary Shannon entropy on (p,q,r).

For four vertices most cases are easily handled; for example

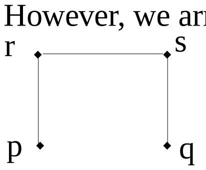


$$\begin{aligned}\mathcal{J} &= \{ps,qr\}\\ H &= \min(-(p+s)\log I-(q+r)\log II)\\ &= -(p+s)\log(p+s)-(q+r)\log(q+r)\end{aligned}$$



$$\begin{aligned}H &= H(p,q,r,s)-H(\bar{G})\\ &= -p\log\frac{p}{p+s}-q\log\frac{q}{q+r}-r\log\frac{r}{q+r}-s\log\frac{s}{p+s}\end{aligned}$$

However, we arrive for the first time at a more difficult case when computing



$$\begin{aligned}\mathcal{J} &= \{pq,ps,qr\}\\ H(G) &= \min(-p\log(I+II)-q\log(I+III)-r\log III-s\log II)\end{aligned}$$

If $pq\leq rs$, the minimum lies on the boundary $I=0$, giving

$$H=-(p+s)\log(p+s)-(q+r)\log(q+r)$$

For the case $pq \geq rs$ we first compute $H(\bar{G})$ and take advantage of the splitting property of K_4, namely $H(G) = H(p,q,r,s) - H(\bar{G})$. We find

$$H(G) = -p\log\frac{p}{p+r} - q\log\frac{q}{q+s} - r\log\frac{r}{p+r} - s\log\frac{s}{q+s}$$

APPLYING GRAPH-MAX-ENT

Ordinary conditioning can be justified in a number of ways [13, 17], all leading to the same numerical result

$$p_i' = \frac{p_i}{p_1 + \ldots + p_k}, \quad 1 \leq i \leq k < n$$

which can be obtained as

$$\arg\min D(P' \| P)$$

subject only to $\sum p_i' = 1$; the solution is unique and $p_i' \geq p_i$.

When a graph structure is present, thus probability transfer restricted, even defining the conditional assignment becomes nontrivial. We propose it be treated along the lines of minimum change principle and demonstrate how graph entropies lead to attractive results. We present these in symbolic form, so as to be able to extract some insight into the conditioning process.

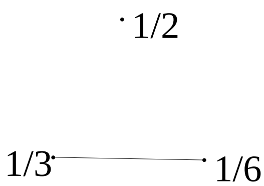
For the general case one should require

$$p_i' \geq p_i;$$

still the solution may be nonunique, leading to only partly unspecified probabilities. This is due to a 'free' transfer of mass between the nondistinguishable nodes.

The results can usually be interpreted as various forms of imaging [9, 10, 11]. An extreme case is the completely disconnected graph where nodes are indistinguishable; then all entropies and distances are 0.

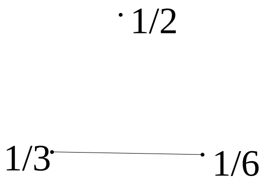
For the first case we take a three element distribution $p = \frac{1}{3}, q = \frac{1}{6}, r = \frac{1}{2}$ and aim to reduce $r \to 0$.



$$D \;=\; p\log\frac{\frac{p}{p+q}}{\frac{1/3}{1/3+1/6}} + q\log\frac{\frac{q}{p+q}}{\frac{1/6}{1/3+1/6}}$$

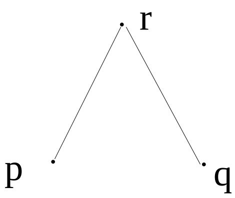
Lagrange multipliers give $\frac{p}{2} = q$, thus the new values become $p = \frac{2}{3}, q = \frac{1}{3}$, which represents a proportional allocation of r to the other nodes.

To reduce $p \rightarrow 0$ requires imposing a boundary condition: q must remain $\geq \frac{1}{6}$. Now



$$D = q \log 3$$

and p is transfered to r: $r = \frac{5}{6}$. Companion cases are handled similarly. Consider the request to reduce $r \rightarrow 0$, hence $p' + q' = 1$ for the graph



$$D \quad = \quad q' \log \frac{p' + q'}{p + q} + q' \log \frac{p' + q'}{p + q} = -\log(p + q)$$

Minimising distance D tells us only that $p' \geq p,\ q' \geq q$. However, both reducing $p \rightarrow 0$ and minimising $D = q' \log{(q'/\frac{q}{p+q})} + r' \log \frac{r'}{r}$ gives a unique answer $q' = p + q,\ r' = r$.

CONCLUDING EXAMPLE

A somewhat controversial issue is the suitability of entropy for conditionalisation based itself on conditional premises [4, 16]. A prototypical situation is usually framed as the JB problem (after the film 'Private (Judy) Benjamin'). In one scene there JB commands a unit that becomes totally lost during the military games. As van Fraasen puts it, she is totally disoriented and assigns (implicitly) probability $\frac{1}{4}$ to being in any of the four sectors

- *BH*: 'Blue' headquarters - friendly area
- *B2*: 'Blue' support - also friendly
- *RH*: 'Red' headquarters - enemy
- *R2*: 'Red' support - also hostile

Her aim is to secure the *RH* sector. She receives a garbled radio message "…if in Red area, it is $3 \div 1$ that you are in the Headquarters area …" Receiving a message she needs to reassess the probabilities so that $P(RH|RH \vee R2) = 0.75$. The intuitive answer $(\frac{1}{2}, \frac{1}{2}, \frac{3}{8}, \frac{1}{8})$ preserving *independence* between the 'Red' and 'Blue' states, does not follow from the use of any unstructured entropy. A direct conditional reasoning [4] restores the independence, but cannot be reduced to the basic MaxEnt.

We show that the 'obvious' answer is obtained if the entropy on an *incomplete* graph is used. It appears that similar 'successes' can be generated for majority of like cases. However, it is the ease of creating such solutions that cautions against the automatic use of entropy as the *normative* decision rule. The JB problem is tackled by omitting a specific edge from the complete graph. This can be given a logical basis, but it feels more like an explanation *ex post*, and suggests that the MaxEnt and MinInf are best kept as *descriptive rules*, occasional successes to the contrary notwithstanding [12]. Their prescriptive use would require a supporting logical framework that could decide *ex ante* on choice of the graph of 'information' transfers.

We recover the answer by *removing* the edge $(RH, R2)$ from the interconnection graph. It can be justified as the explication of the fact that states RH and R2 can be *confused*. We feel it is best viewed as simply explaining the success of GraphMaxEnt in this instance.

ACKNOWLEDGEMENTS

This research was supported by the Australia Research Council grant DP0210999 "Asynchronous Continuous Time Conditioning".

Technical preparation of this article was assisted by Christopher Petrov, Computing Support Officer of the School of Computer Science and Engineering.

REFERENCES

1. CE Alchourron, P Gardenfors, D Makinson. *On the logic of theory change: partial meet contractions and revision functions.* J. Symbolic Logic 50(1985), 510–530.
2. I Csiszar, J Korner. Information Theory: Coding Theorems for Discrete Memoryless Systems. Academic Press, New York 1981.
3. P Garderfors. Knowledge in Flux. The MIT Press, Cambridge MA 1988.
4. AJ Grove, JY Halpern. *Probability update: conditioning vs cross-entropy.* Proc. 13th Annual Conf. Uncertainty in Artificial Intelligence, Providence, RI, August 1997.
5. RC Jeffrey. The Logic of Decision. McGraw-Hill, New York 1965.
6. JN Kapur, HK Kesavan. Entropy Optimization Principles with Applications. Academic Press, New York 1992.
7. J Korner. *Coding of an information source having ambiguous alphabet and the entropy of graphs.* Trans. Sixth Prague Conf. Information Theory. Academia, Prague 1973, pp. 411–425.
8. J Korner, A Orlitsky. *Zero-error information theory.* IEEE Trans. Information Theory 44(1998), 6:2207–2229.
9. I Levi. *personal communication to P Gardenfors, in [3].*
10. I Levi. The Enterprise of Knowledge. The MIT Press, Cambridge, MA 1980.
11. DK Lewis. *Probabilities of conditionals and conditional probabilities.* Phil. Review 85(1976, 297–315.
12. A Ramer. *Conditional possibility measures.* Int. J. Cybernetics and Systems 20(1989), 233–247.
13. A Ramer. *Note on defining conditional probability,* Amer. Math. Monthly, 97(1990), 336–337.
14. A Ramer. *Belief revision as combinatorial optimisation,* IPMU-2002 - 9th Int. Conf. Information Processing and Management of Uncertainty, Paris, France, July 2002.
15. G Simonyi. *Graph entropy: a survey.* DIMACS Series in Discr. Math. and Theor. Comp. Science 20(1995), pp. 399–441.
16. BC van Fraasen. *Symmetries of personal probability kinematics.* In N Rescher (ed). Scientific Enquiry in Philosophical Perspective. University Press of America, Lanham, MD 1987, pp. 183–223.
17. PM Williams. *Bayesian conditionalisation and the principle of minimum information.* British J. Phil. Science 31(1980), 131–144.