Robust Estimators under the Imprecise Dirichlet Model

MARCUS HUTTER
IDSIA, Switzerland

Abstract

Walley's Imprecise Dirichlet Model (IDM) for categorical data overcomes several fundamental problems which other approaches to uncertainty suffer from. Yet, to be useful in practice, one needs efficient ways for computing the imprecise=robust sets or intervals. The main objective of this work is to derive exact, conservative, and approximate, robust and credible interval estimates under the IDM for a large class of statistical estimators, including the entropy and mutual information.

1 Introduction

This work derives interval estimates under the Imprecise Dirichlet Model (IDM) [Wal96] for a large class of statistical estimators. In the IDM one considers an i.i.d. process with unknown chances[1] π_i for outcome i. The prior uncertainty about $\boldsymbol{\pi}$ [2] is modeled by a set of Dirichlet priors[3] $\{p(\boldsymbol{\pi}) \propto \prod_i \pi_i^{st_i-1} : \boldsymbol{t} \in \Delta\}$, where[4] $\Delta := \{\boldsymbol{t} : t_i \geq 0, \sum_i t_i = 1\}$, and s is a hyper-parameter, typically chosen between 1 and 2. Sets of probability distributions are often called Imprecise probabilities, hence the name IDM for this model. We avoid the term *imprecise* and use *robust* instead, or capitalize *Imprecise*. IDM overcomes several fundamental problems which other approaches to uncertainty suffer from [Wal96]. For instance, IDM satisfies the representation invariance principle and the symmetry principle, which are mutually exclusive in a pure Bayesian treatment with proper prior [Wal96]. The counts n_i for i form a minimal sufficient statistic of the data of size $n = \sum_i n_i$. Statistical estimators $F(\boldsymbol{n})$ usually also depend on the chosen

[1] Also called *objective* or *aleatory* probabilities.

[2] We denote vectors by $\boldsymbol{x} := (x_1, ..., x_d)$ for $\boldsymbol{x} \in \{\boldsymbol{n}, \boldsymbol{t}, \boldsymbol{u}, \boldsymbol{\pi}, ...\}$.

[3] Also called *second order* or *subjective* or *belief* or *epistemic* probabilities.

[4] Strictly speaking, Δ should be the open simplex [Wal96], since $p(\boldsymbol{\pi})$ is improper for $\boldsymbol{t}$ on the boundary of Δ. For simplicity we assume that, if necessary, considered functions of $\boldsymbol{t}$ can and are continuously extended to the boundary of Δ, so that, for instance, minima and maxima exist. All considerations can straightforwardly, but cumbersomely, be rewritten in terms of an open simplex. Note that open/closed Δ result in open/closed robust intervals, the difference being numerically/practically irrelevant.

prior: so a set of priors leads to a set of estimators $\{F_{\boldsymbol{t}}(\boldsymbol{n}) : \boldsymbol{t} \in \Delta\}$. For instance, the expected chances $E_{\boldsymbol{t}}[\pi_i] = \frac{n_i+st_i}{n+s} =: u_i(\boldsymbol{t})$ lead to a robust interval estimate $[\frac{n_i}{n+s}, \frac{n_i+s}{n+s}] \ni E_{\boldsymbol{t}}[\pi_i]$. Robust intervals for the variance $\text{Var}[\pi_i]$ [Wal96] and for the mean and variance of linear-combinations $\sum_i \alpha_i \pi_i$ have also been derived [Ber01]. Bayesian estimators (like expectations) depend on $\boldsymbol{t}$ and $\boldsymbol{n}$ only through $\boldsymbol{u}$ (and $n+s$ which we suppress), i.e. $F_{\boldsymbol{t}}(\boldsymbol{n}) = F(\boldsymbol{u})$. The main objective of this work is to derive approximate, conservative, and exact intervals $[\min_{\boldsymbol{t}\in\Delta} F(\boldsymbol{u}), \max_{\boldsymbol{t}\in\Delta} F(\boldsymbol{u})]$ for general $F(\boldsymbol{u})$, and for the expected (also called predictive) entropy and the expected mutual information in particular. These results are key building blocks for applying IDM. Walley suggests, for instance, to use $\min_{\boldsymbol{t}} P_{\boldsymbol{t}}[\mathcal{F} \geq c] \geq \alpha$ for inference problems and $\min_{\boldsymbol{t}} E_{\boldsymbol{t}}[\mathcal{F}] \geq c$ for decision problems [Wal96], where $\mathcal{F}$ is some function of $\boldsymbol{\pi}$. One application is the inference of robust tree-dependency structures [Zaf01, ZH03], in which edges are partially ordered based on Imprecise mutual information.

Section 2 gives a brief introduction to IDM and describes our problem setup. In Section 3 we derive exact robust intervals for concave functions F, such as the entropy. Section 4 derives approximate robust intervals for arbitrary F. In Section 5 we show how bounds of elementary functions can be used to get bounds for composite function, especially for sums and products of functions. The results are used in Section 6 for deriving robust intervals for the mutual information. The issue of how to set up IDM models on product spaces is discussed in Section 7. Section 8 addresses the problem of how to combine Bayesian credible intervals with the robust intervals of the IDM. Conclusions are given in Section 9.

2 The Imprecise Dirichlet Model

Random i.i.d. processes. We consider discrete random variables $\imath \in \{1,...,d\}$ and an i.i.d. random process with outcome $i \in \{1,...,d\}$ having probability π_i. The chances $\boldsymbol{\pi}$ form a probability distribution, i.e. $\boldsymbol{\pi} \in \Delta := \{\boldsymbol{x} \in I\!R^d : x_i \geq 0 \forall i,\, x_+ = 1\}$, where we have used the abbreviation $\boldsymbol{x} = (x_1,...,x_d)$ and $x_+ := \sum_{i=1}^d x_i$. The likelihood of a specific data set $\boldsymbol{D}$ with n_i observations i and total sample size $n = n_+ = \sum_i n_i$ is $p(\boldsymbol{D}|\boldsymbol{\pi}) = \prod_i \pi_i^{n_i}$. The chances π_i are usually unknown and have to be estimated from the sample frequencies n_i. The frequency estimate $\frac{n_i}{n}$ for π_i is one possible point estimate.

Second order p(oste)rior. In the Bayesian approach one models the initial uncertainty in $\boldsymbol{\pi}$ by a (second order) prior "belief" distribution $p(\boldsymbol{\pi})$ with domain $\boldsymbol{\pi} \in \Delta$. The Dirichlet priors $p(\boldsymbol{\pi}) \propto \prod_i \pi_i^{n'_i-1}$, where n'_i comprises prior information, represent a large class of priors. n'_i may be interpreted as (possibly fractional) virtual number of "observation". High prior belief in i can be modeled by large n'_i. It is convenient to write $n'_i = s \cdot t_i$ with $s := n'_+$, hence $\boldsymbol{t} \in \Delta$. Having no initial bias one should choose a prior in which all t_i are equal, i.e. $t_i = \frac{1}{d} \forall i$.

Examples for s are 0 for Haldane's prior [Hal48], 1 for Perks' prior [Per47], $\frac{d}{2}$ for Jeffreys' prior [Jef46], and d for Bayes-Laplace's uniform prior [GCSR95]. From the prior and the data likelihood one can determine the posterior $p(\boldsymbol{\pi}|\boldsymbol{D}) = p(\boldsymbol{\pi}|\boldsymbol{n}) \propto \prod_i \pi_i^{n_i+st_i-1}$.

The posterior $p(\boldsymbol{\pi}|\boldsymbol{D})$ summarizes all statistical information available in the data. In general, the posterior is a very complex object, so we are interested in summaries of this plethora of information. A possible summary is the expected value or mean $E_{\boldsymbol{t}}[\pi_i] = \frac{n_i+st_i}{n+s}$ which is often used for estimating π_i. The accuracy may be obtained from the covariance of $\boldsymbol{\pi}$.

Usually one is not only interested in an estimation of the whole vector $\boldsymbol{\pi}$, but also in an estimation of scalar functions $\mathcal{F} : \Delta \to I\!R$ of $\boldsymbol{\pi}$, such as the entropy $\mathcal{H}(\boldsymbol{\pi}) = -\sum_i \pi_i \log \pi_i$, where log denotes the natural logarithm. Since $\mathcal{F}$ is itself a random variable we could determine the posterior distribution $p(\mathcal{F}_0|\boldsymbol{n}) = \int_\Delta \delta(\mathcal{F}(\boldsymbol{\pi}) - \mathcal{F}_0) p(\boldsymbol{\pi}|\boldsymbol{n}) d\boldsymbol{\pi}$ of $\mathcal{F}$, which may further be summarized by the posterior mean $E_{\boldsymbol{t}}[\mathcal{F}] = \int_\Delta \mathcal{F}(\boldsymbol{\pi}) p(\boldsymbol{\pi}|\boldsymbol{n}) d\boldsymbol{\pi}$ and possibly the posterior variance $\text{Var}_{\boldsymbol{t}}[\mathcal{F}]$. A simple, but crude approximation for the mean can be obtained by exchanging E with $\mathcal{F}$ (exact only for linear functions): $E_{\boldsymbol{t}}[\mathcal{F}(\boldsymbol{\pi})] \approx \mathcal{F}(E_{\boldsymbol{t}}[\boldsymbol{\pi}])$. The approximation error is typically of the order $\frac{1}{n}$.

The Imprecise Dirichlet Model. The classical approach, which consists of selecting a single prior, suffers from a number of problems. Firstly, choosing for example a uniform prior $t_i = \frac{1}{d}$, the prior depends on the particular choice of the sampling space. Secondly, it assumes exact prior knowledge of $p(\boldsymbol{\pi})$. The solution to the second problem is to model our ignorance by considering sets of priors $p(\boldsymbol{\pi})$, often called Imprecise probabilities. The specific *Imprecise Dirichlet Model* (IDM) [Wal96] considers the set of *all* $\boldsymbol{t} \in \Delta$, i.e. $\{p(\boldsymbol{\pi}|\boldsymbol{n}) : \boldsymbol{t} \in \Delta\}$ which solves also the first problem. Walley suggests to fix the hyperparameter s somewhere in the interval $[1,2]$. A set of priors results in a set of posteriors, set of expected values, etc. For real-valued quantities like the expected entropy $E_{\boldsymbol{t}}[\mathcal{H}]$ the sets are typically intervals, which we call robust intervals

$$E_{\boldsymbol{t}}[\mathcal{F}] \in [\min_{\boldsymbol{t}\in\Delta} E_{\boldsymbol{t}}[\mathcal{F}], \max_{\boldsymbol{t}\in\Delta} E_{\boldsymbol{t}}[\mathcal{F}]].$$

Problem setup and notation. Consider any statistical estimator F. F is a function of the data $\boldsymbol{D}$ and the hyperparameters $\boldsymbol{t}$. We define the general correspondence

$$u_i^{\cdots} = \frac{n_i + st_i^{\cdots}}{n+s}, \quad \text{where } ^{\cdots} \text{ can be various superscripts.} \tag{1}$$

F can, hence, be rewritten as a function of $\boldsymbol{u}$ and $\boldsymbol{D}$. Since we regard $\boldsymbol{D}$ as fixed, we suppress this dependence and simply write $F = F(\boldsymbol{u})$. This is further motivated by the fact that all Bayesian estimators of functions $\mathcal{F}$ of $\boldsymbol{\pi}$ only depend on $\boldsymbol{u}$ and the sample size $n+s$. It is easy to see that this holds for the mean, i.e. $E_{\boldsymbol{t}}[\mathcal{F}] = F(\boldsymbol{u}; n+s)$, and similarly for the variance and all higher (central) moments. The main focus of this work is to derive exact and approximate expressions for upper and lower F values

$$\overline{F} := \max_{\boldsymbol{t}\in\Delta} F(\boldsymbol{u}) \quad \text{and} \quad \underline{F} := \min_{\boldsymbol{t}\in\Delta} F(\boldsymbol{u}), \qquad \underline{\overline{F}} := [\underline{F}, \overline{F}]$$

$\boldsymbol{t} \in \Delta \Leftrightarrow \boldsymbol{u} \in \Delta'$, where $\Delta' := \{\boldsymbol{u} : u_i \geq \frac{n_i}{n+s},\, u_+ = 1\}$. We define $\boldsymbol{u}^{\overline{F}}$ as the $\boldsymbol{u} \in \Delta'$ which maximizes F, i.e. $\overline{F} = F(\boldsymbol{u}^{\overline{F}})$, and similarly $\boldsymbol{t}^{\overline{F}}$ through relation (1). If the maximum of F is assumed in a corner of Δ' we denote the index of the corner by $i^{\overline{F}}$, i.e. $t_i^{\overline{F}} = \delta_{ii^{\overline{F}}}$, where δ_{ij} is Kronecker's delta function. Similarly $\boldsymbol{u}^{\underline{F}}$, $\boldsymbol{t}^{\underline{F}}$, $i^{\underline{F}}$.

3 Exact Robust Intervals for Concave Estimators

In this section we derive exact expressions for $\underline{\overline{F}}$ if $F : \Delta \to \mathbb{R}$ is of the form

$$F(\boldsymbol{u}) = \sum_{i=1}^{d} f(u_i) \quad \text{and concave} \quad f : [0,1] \to \mathbb{R}. \tag{2}$$

The expected entropy is such an example (discussed later). Convex f are treated similarly (or simply take $-f$).

The nature of the solution. The approach to a solution of this problem is motivated as follows: Due to symmetry and concavity of F, the global maximum is attained at the center $u_i = \frac{1}{d}$ of the probability simplex Δ, i.e. the more uniform $\boldsymbol{u}$ is, the larger $F(\boldsymbol{u})$. The nearer $\boldsymbol{u}$ is to a vertex of Δ, i.e. the more unbalanced $\boldsymbol{u}$ is, the smaller is $F(\boldsymbol{u})$. The constraints $t_i \geq 0$ restrict $\boldsymbol{u}$ to the smaller simplex

$$\Delta' = \{\boldsymbol{u} : u_i \geq u_i^0,\, u_+ = 1\} \quad \text{with} \quad u_i^0 := \frac{n_i}{n+s},$$

which prevents setting $u_i^{\overline{F}} = \frac{1}{d}$ and $u_i^{\underline{F}} = \delta_{i1}$. Nevertheless, the basic idea of choosing $\boldsymbol{u}$ as uniform / as unbalanced as possible still works, as we will see.

Greedy $F(\boldsymbol{u})$ minimization. Consider the following procedure for obtaining $\boldsymbol{u}^{\underline{F}}$. We start with $\boldsymbol{t} \equiv \boldsymbol{0}$ (outside the usual domain Δ of F, which can be extended to $[0,1]^d$ via (2)) and then gradually increase $\boldsymbol{t}$ in an axis-parallel way until $t_+ = 1$. With axis-parallel we mean that only one component of $\boldsymbol{t}$ is increased, which one possibly changes during the process. The total zigzag curve from $\boldsymbol{t}^{start} = \boldsymbol{0}$ to $\boldsymbol{t}^{end}$ has length $t_+^{end} = 1$. Since all possible curves have the same (Manhattan) length 1, $F(\boldsymbol{u}^{end})$ is minimized for the curve which has (on average) smallest F-gradient along its path. A greedy strategy is to follow the direction i of currently smallest F-gradient $\frac{\partial F}{\partial t_i} = f'(u_i)\frac{s}{n+s}$. Since f' is monotone decreasing ($f'' < 0$), $\frac{\partial F}{\partial t_i}$ is smallest for largest u_i. At $\boldsymbol{t}^{start} = \boldsymbol{0}$, $u_i = \frac{n_i}{n+s}$ is largest for $i = i^{min} := \arg\max_i n_i$. Once we start in direction i^{min}, $u_{i^{min}}$ increases even further whereas all other u_i ($i \neq i^{min}$) remain constant. So the moving direction is never changed and finally we reach a local minimum at $t_i^{end}=\delta_{ii^{min}}$. In [Hut03] we show that this is a global minimum, i.e.

$$t_i^{\underline{F}}=\delta_{ii^{\underline{F}}} \quad \text{with} \quad i^{\underline{F}}:=\arg\max_i n_i. \tag{3}$$

Greedy $F(\boldsymbol{u})$ maximization. Similarly we maximize $F(\boldsymbol{u})$. Now we increase $\boldsymbol{t}$ in direction $i=i_1$ of maximal $\frac{\partial F}{\partial t_i}$, which is the direction of smallest $u_i\propto n_i+st_i$. Again, (only) u_{i_1} increases, but possibly reaches a value where it is no longer the smallest one. We stop if it becomes equal to the second smallest u_i, say $i=i_2$. We now have to increase u_{i_1} and u_{i_2} with same speed (or in an ε-zigzag fashion) until they become equal to u_{i_3}, etc or until $u_+=1=t_+$ is reached. Assume the process stops with direction i_m and minimal u being $\tilde{u}$, i.e. finally $u_{i_k}=\tilde{u}$ for $k\leq m$ and $t_{i_k}=0$ for $k>m$. From the constraint $1=u_+=\sum_{k\leq m}u_{i_k}+\sum_{k>m}u_{i_k}=m\tilde{u}+\sum_{k>m}\frac{n_{i_k}}{n+s}$ we obtain $\tilde{u}(m)=\frac{1}{m}[1-\sum_{k>m}\frac{n_{i_k}}{n+s}]=[s+\sum_{k\leq m}n_{i_k}]/[m(n+s)]$. One can show that $\tilde{u}(m)$ has one global minimum (no local ones) and that the final m is the one which minimizes $\tilde{u}$, i.e.

$$\tilde{u}=\min_{m\in\{1\ldots d\}}\frac{s+\sum_{k\leq m}n_{i_k}}{m(n+s)}, \quad \text{where } n_{i_1}\leq n_{i_2}\leq\ldots\leq n_{i_d}, \quad u_i^{\overline{F}}=\max\{u_i^0,\tilde{u}\}. \tag{4}$$

If there is a unique minimal n_{i_1} with gap $\geq s$ to the second smallest n_{i_2} (which is quite likely for not too small n), then $m=1$ and the maximum is attained at a corner of Δ (Δ').

Theorem 1 (Exact extrema for concave functions on simplices) *Assume $F:\Delta'\to I\!R$ is a concave function of the form $F(\boldsymbol{u})=\sum_{i=1}^d f(u_i)$. Then F attains the global maximum $\overline{F}$ at $\boldsymbol{u}^{\overline{F}}$ defined in (4) and the global minimum $\underline{F}$ at $\boldsymbol{u}^{\underline{F}}$ defined in (3).*

Proof. What remains to be shown is that the solutions obtained in the last paragraphs by greedy minimization/maximization of $F(\boldsymbol{u})$ are actually global minima/maxima. For this assume that $\boldsymbol{t}$ is a local minimum of $F(\boldsymbol{u})$. Let $j:=\arg\max_i u_i$ (ties broken arbitrarily). Assume that there is a $k\neq j$ with non-zero t_k. Define $\boldsymbol{t}'$ as $t_i'=t_i$ for all $i\neq j,k$, and $t_j'=t_j+\varepsilon$, $t_k'=t_k-\varepsilon$, for some $0<\varepsilon\leq t_k$. From $u_k\leq u_j$ and the concavity of f we get[5]

$$\begin{aligned} F(\boldsymbol{t}')-F(\boldsymbol{t}) &= [f(u_j')+f(u_k')]-[f(u_j)+f(u_k)] \\ &= [f(u_j+\sigma\varepsilon)-f(u_j)]-[f(u_k)-f(u_k-\sigma\varepsilon)] \;<\; 0 \end{aligned}$$

where $\sigma:=\frac{s}{n+s}$. This contradicts the minimality assumption of $\boldsymbol{t}$. Hence, $t_i=0$ for all i except one (namely j, where it must be 1). (Local) minima are attained in the vertices of Δ. Obviously the global minimum is for $t_i^{\underline{F}}=\delta_{ii^{\underline{F}}}$ with $i^{\underline{F}}:=\arg\max_i n_i$. This solution coincides with the greedy solution. Note that the global minimum

[5] Slope $\frac{f(u+\varepsilon)-f(u)}{\varepsilon}$ is a decreasing function in u for any $\varepsilon>0$, since f is concave.

may not be unique, but since we are only interest in the value of $F(\boldsymbol{u}^{\underline{F}})$ and not its argument this degeneracy is of no further significance.

Similarly for the maximum, assume that $\boldsymbol{t}$ is a (local) maximum of $F(\boldsymbol{u})$. Let $j := \arg\min_i u_i$ (ties broken arbitrarily). Assume that there is a $k \neq j$ with non-zero t_k *and* $u_k > u_j$. Define $\boldsymbol{t}'$ as above with $0 < \varepsilon < \min\{t_k, t_k - t_j\}$. Concavity of f implies

$$F(\boldsymbol{t}') - F(\boldsymbol{t}) = [f(u_j + \sigma\varepsilon) - f(u_j)] - [f(u_k) - f(u_k - \sigma\varepsilon)] > 0,$$

which contradicts the maximality assumption of $\boldsymbol{t}$. Hence $t_i = 0$ if u_i is not minimal ($\tilde{u}$). The previous paragraph constructed the unique solution $\boldsymbol{u}^{\overline{F}}$ satisfying this condition. Since this is the only local maximum it must be the unique global maximum (contrast this to the minimum case). □

Theorem 2 (Exact extrema of expected entropy) *Let $\mathcal{H}(\boldsymbol{\pi}) = -\sum_i \pi_i \log \pi_i$ be the entropy of $\boldsymbol{\pi}$ and the uncertainty of $\boldsymbol{\pi}$ be modeled by the Imprecise Dirichlet Model. The expected entropy $H(\boldsymbol{u}) := E_{\boldsymbol{t}}[\mathcal{H}]$ for given hyperparameter $\boldsymbol{t}$ and sample $\boldsymbol{n}$ is given by*

$$H(\boldsymbol{u}) = \sum_i h(u_i) \quad \textit{with} \quad h(u) = u\cdot[\psi(n+s+1) - \psi((n+s)u+1)] = u\sum_{k=(n+s)u+1}^{n+s} k^{-1} \qquad (5)$$

where $\psi(x) = d\log\Gamma(x)/dx$ is the logarithmic derivative of the Gamma function and the last expression is valid for integral s and $(n+s)u$. The lower $\underline{H}$ and upper $\overline{H}$ expected entropies are assumed at $\boldsymbol{u}^{\underline{H}}$ and $\boldsymbol{u}^{\overline{H}}$ given in (3) and (4) (with $F \rightsquigarrow H$, see also (1)).

A derivation of the exact expression (5) for the expected entropy can be found in [WW95, Hut02]. The only thing to be shown is that h is concave. This may be done by exploiting special properties of the digamma function ψ (see [AS74]).

There are fast implementations of ψ and its derivatives and exact expressions for integer and half-integer arguments

Example. For $d = 2$, $n_1 = 3$, $n_2 = 6$, $s = 1$ we have $n = 9$, $u_1 = \frac{3+t_1}{10}$, $u_2 = \frac{6+t_2}{10}$, $\boldsymbol{t}^0 = 0$, $\boldsymbol{u}^0 = \binom{.3}{.6}$, see (1). From (3), $i^{\underline{H}} = 2$, $\boldsymbol{t}^{\underline{H}} = \binom{0}{1}$, $\boldsymbol{u}^{\underline{H}} = \binom{.3}{.7}$. From (4), $i_1 = 1$, $i_2 = 2$, $\tilde{u} = \min\{\frac{1+3}{9+1}, \frac{1+3+6}{2\cdot(9+1)}\} = \frac{4}{10}$, $\boldsymbol{u}^{\overline{H}} = \max\{\boldsymbol{u}^0, \tilde{u}\} = \binom{.4}{.6} \Rightarrow \boldsymbol{t}^{\overline{H}} = \binom{1}{0}$ is in corner. From (5), $h(\frac{3}{10}) = \frac{2761}{8400}$, $h(\frac{4}{10}) = \frac{2131}{6300}$, $h(\frac{6}{10}) = \frac{1207}{4200}$, $h(\frac{7}{10}) = \frac{847}{3600}$, hence $\underline{\overline{H}} = [H(\boldsymbol{u}^{\underline{H}}), H(\boldsymbol{u}^{\overline{H}})] = [h(\frac{3}{10}) + h(\frac{7}{10}), h(\frac{4}{10}) + h(\frac{6}{10})] = [0.5639..., 0.6256...]$, so $\overline{H} - \underline{H} = O(\frac{1}{10})$.

4 Approximate Robust Intervals

In this section we derive approximations for $\underline{\overline{F}}$ suitable for arbitrary, twice differentiable functions $F(\boldsymbol{u})$. The derived approximations for $\underline{\overline{F}}$ will be robust in the sense of covering set $\overline{\underline{F}}$ (for any n), and the approximations will be "good" if n is not too small. In the following, we treat $\sigma := \frac{s}{n+s}$ as a (small) expansion parameter. For $\boldsymbol{u},\boldsymbol{u}^* \in \Delta'$ we have

$$u_i - u_i^* \;=\; \sigma\cdot(t_i - t_i^*) \quad\text{and}\quad |u_i - u_i^*| \;=\; \sigma|t_i - t_i^*| \;\leq\; \sigma \quad\text{with}\quad \sigma := \tfrac{s}{n+s}. \tag{6}$$

Hence we may Taylor-expand $F(\boldsymbol{u})$ around $\boldsymbol{u}^*$, which leads to a Taylor series in σ. This shows that F is approximately linear in $\boldsymbol{u}$ and hence in $\boldsymbol{t}$. A linear function on a simplex assumes its extreme values at the vertices of the simplex. This has already been encountered in Section 3. The consideration above is a simple explanation for this fact. This also shows that the robust interval $\overline{\underline{F}}$ is of size $\overline{F} - \underline{F} = O(\sigma)$.[6] Any approximation to $\overline{\underline{F}}$ should hence be at least $O(\sigma^2)$. The expansion of F to $O(\sigma)$ is

$$F(\boldsymbol{u}) \;=\; \overbrace{F(\boldsymbol{u}^*)}^{F_0 = O(1)} + \overbrace{\sum_i [\partial_i F(\check{\boldsymbol{u}})](u_i - u_i^*)}^{F_R = O(\sigma)} \tag{7}$$

where $\partial_i F(\check{\boldsymbol{u}})$ is the partial derivative $\frac{\partial_i F(\check{\boldsymbol{u}})}{\partial \check{u}_i}$ of $F(\check{\boldsymbol{u}})$ w.r.t. $\check{u}_i$. For suitable $\check{\boldsymbol{u}} = \check{\boldsymbol{u}}(\boldsymbol{u},\boldsymbol{u}^*) \in \Delta'$ this expansion is exact (F_R is the exact remainder). Natural points for expansion are $t_i^* = \frac{1}{d}$ in the center of Δ, or possibly also $t_i^* = \frac{n_i}{n} = u_i^*$. See [Hut03] for such a general expansion. Here, we expand around the improper point $t_i^* := t_i^0 \equiv 0$, which is outside(!) Δ, since this makes expressions particularly simple.[7] (6) is still valid in this case, and F_R is exact for some $\check{\boldsymbol{u}}$ in

$$\Delta'_e := \{\boldsymbol{u} : u_i \geq u_i^0 \forall i,\, u_+ \leq 1\}, \quad\text{where}\quad u_i^0 = \frac{n_i}{n+s}.$$

Note that we keep the exact condition $\boldsymbol{u} \in \Delta'$. F is usually already defined on Δ'_e or extends from Δ' to Δ'_e without effort in a natural way (analytical continuation). We introduce the notation

$$F \sqsubseteq G \quad :\Leftrightarrow \quad F \leq G \quad\text{and}\quad F = G + O(\sigma^2) \tag{8}$$

stating that G is a "good" upper bound on F. The following bounds hold for arbitrary differentiable functions. In order for the bounds to be "good," F has to be Lipschitz differentiable in the sense that there exists a constant c such that

$$|\partial_i F(\boldsymbol{u})| \leq c \quad\text{and}\quad |\partial_i F(\boldsymbol{u}) - \partial_i F(\boldsymbol{u}')| \leq c|\boldsymbol{u} - \boldsymbol{u}'|$$
$$\forall \boldsymbol{u},\boldsymbol{u}' \in \Delta'_e \quad\text{and}\quad \forall\, 1 \leq i \leq d. \tag{9}$$

[6] $f(\boldsymbol{n},\boldsymbol{t},s) = O(\sigma^k) \;:\Leftrightarrow\; \exists c \forall \boldsymbol{n} \in I\!N_0^d, \boldsymbol{t} \in \Delta, s > 0 : |f(\boldsymbol{n},\boldsymbol{t},s)| \leq c\sigma^k$, where $\sigma = \frac{s}{n+s}$.

[7] The order of accuracy $O(\sigma^2)$ we will encounter is for all choices of $\boldsymbol{u}^*$ the same. The concrete numerical errors differ of course. The choice $\boldsymbol{t}^* = \boldsymbol{0}$ can lead to $O(d)$ smaller F_R than the natural center point $\boldsymbol{t}^* = \frac{1}{d}$, but is more likely a factor $O(1)$ larger. The exact numerical values depend on the structure of F.

If F depends also on $\boldsymbol{n}$, e.g. via σ or $\boldsymbol{u}^0$, then c shall be independent of them.

The Lipschitz condition is satisfied, for instance, if the curvature $\partial^2 F$ is uniformly bounded. This is satisfied for the expected entropy H (see (5)), but violated for the approximation $E_{\boldsymbol{t}}[\mathcal{H}] \approx \mathcal{H}(\boldsymbol{u})$ if $n_i = 0$ for some i.

Theorem 3 (Approximate robust intervals) *Assume $F : \Delta'_e \to I\!R$ is a Lipschitz differentiable function (9). Let $[\underline{F},\overline{F}]$ be the global [minimum,maximum] of F restricted to Δ'. Then*

$$F(\boldsymbol{u}^1) \sqsubseteq \overline{F} \sqsubseteq F_0 + F_R^{ub} \;\; where \;\; F_R^{ub} = \max_i F_{iR}^{ub} \;\; and \;\; F_{iR}^{ub} \; = \; \sigma \max_{\boldsymbol{u}\in\Delta'_e}[\partial_i F(\boldsymbol{u})]$$

$$F_0 + F_R^{lb} \sqsubseteq \underline{F} \sqsubseteq F(\boldsymbol{u}^2) \;\; where \;\; F_R^{lb} = \min_i F_{iR}^{lb} \;\; and \;\; F_{iR}^{lb} \; = \; \sigma \min_{\boldsymbol{u}\in\Delta'_e}[\partial_i F(\boldsymbol{u})]$$

$F_0 = F(\boldsymbol{u}^0)$, and $u_i^1 = \delta_{ii^1}$ with $i^1 = \arg\max_i F_{iR}^{ub}$, and $u_i^2 = \delta_{ii^2}$ with $i^2 = \arg\min_i F_{iR}^{lb}$, and $\sqsubseteq$ defined in (8) means $\leq$ and $=$ *$+O(\sigma^2)$, where $\sigma = 1 - u_+^0$.*

For conservative estimates, the lower bound on $\underline{F}$ and the upper bound on $\overline{F}$ are the interesting ones.

Proof. We start by giving an $O(\sigma^2)$ bound on $\overline{F}_R = \max_{\boldsymbol{u}\in\Delta'} F_R(\boldsymbol{u})$. We first insert (6) with $\boldsymbol{t}^* = \boldsymbol{t}^0 \equiv \boldsymbol{0}$ into (7) and treat $\check{\boldsymbol{u}}$ and $\boldsymbol{t}$ as separate variables:

$$F_R(\check{\boldsymbol{u}},\boldsymbol{t}) \; = \; \sigma \sum_i [\partial_i F(\check{\boldsymbol{u}})] \cdot t_i \sqsubseteq \max_{\check{\boldsymbol{u}}\in\Delta'_e} \left\{ \sigma \sum_i [\partial_i F(\check{\boldsymbol{u}})] \cdot t_i \right\} \sqsubseteq \sum_i F_{iR}^{ub} \cdot t_i$$

$$\text{with} \quad F_{iR}^{ub} \; := \; \sigma \max_{\check{\boldsymbol{u}}\in\Delta'_e} [\partial_i F(\check{\boldsymbol{u}})] \tag{10}$$

The first inequality is obvious, the second follows from the convexity of max. From assumption (9) we get $\partial_i F(\boldsymbol{u}) - \partial_i F(\boldsymbol{u}') = O(\sigma)$ for all $\boldsymbol{u},\boldsymbol{u}' \in \Delta'_e$, since Δ'_e has diameter $O(\sigma)$. Due to one additional σ in (10) the expressions in (10) change only by $O(\sigma^2)$ when introducing or dropping $\max_{\check{\boldsymbol{u}}}$ anywhere. This shows that the inequalities are tight within $O(\sigma^2)$ and justifies $\sqsubseteq$. We now upper bound $F_R(\boldsymbol{u})$:

$$\overline{F}_R = \max_{\boldsymbol{u}\in\Delta'} F_R(\boldsymbol{u}) \sqsubseteq \max_{\boldsymbol{t}\in\Delta} \max_{\check{\boldsymbol{u}}\in\Delta'_e} F_R(\check{\boldsymbol{u}},\boldsymbol{t}) \sqsubseteq \max_{\boldsymbol{t}\in\Delta} \sum_i F_{iR}^{ub} \cdot t_i = \max_i F_{iR}^{ub} =: F_R^{ub} \tag{11}$$

A linear function on Δ is maximized by setting the t_i component with largest coefficient to 1. This shows the last equality. The maximization over $\check{\boldsymbol{u}}$ in (10) can often be performed analytically, leaving an easy $O(d)$ time task for maximizing over i.

We have derived an upper bound F_R^{ub} on $\overline{F}_R$. Let us define the corner $t_i = \delta_{ii^1}$ of Δ with $i^1 := \arg\max_i F_{iR}^{ub}$. Since $\overline{F}_R \geq F_R(\boldsymbol{u})$ for all $\boldsymbol{u}$, $F_R(\boldsymbol{u}^1)$ in particular is a lower bound on $\overline{F}_R$. A similar line of reasoning as above shows that that $F_R(\boldsymbol{u}^1) = \overline{F}_R + O(\sigma^2)$. Using $\overline{F + const.} = \overline{F} + const.$ we get $O(\sigma^2)$ lower and upper bounds on $\overline{F}$, i.e. $F(\boldsymbol{u}^1) \sqsubseteq \overline{F} \sqsubseteq F_0 + F_R^{ub}$. $\underline{F}$ is bound similarly with all max's replaced by min's and inequalities reversed. Together this proves the Theorem 3. □

5 Error Propagation

Approximation of $\underline{F}$ (special cases). For the special case $F(\boldsymbol{u}) = \sum_i f(u_i)$ we have $\partial_i F(\boldsymbol{u}) = f'(u_i)$. For concave f like in case of the entropy we get particularly simple bounds

$$F_{iR}^{ub} = \sigma \max_{\boldsymbol{u}\in\Delta'_e} f'(u_i) = \sigma f'(u_i^0), \qquad F_R^{ub} = \sigma \max_i f'(u_i^0) = \sigma f'(\tfrac{\min_i n_i}{n+s}),$$

$$F_{iR}^{lb} = \sigma \min_{\boldsymbol{u}\in\Delta'_e} f'(u_i) = \sigma f'(u_i^0+\sigma), \quad F_R^{lb} = \sigma \min_i f'(u_i^0+\sigma) = \sigma f'(\tfrac{\max_i n_i+s}{n+s}),$$

where we have used $\max_{\boldsymbol{u}\in\Delta'_e} f'(u_i) = \max_{u_i\in[u_i^0,u_i^0+\sigma]} f'(u_i) = f'(u_i^0)$, and similarly for min. Analogous results hold for convex functions. In case the maximum cannot be found exactly one is allowed to further increase Δ'_e as long as its diameter remains $O(\sigma)$. Often an increase to $\square' := \{\boldsymbol{u} : u_i^0 \leq u_i \leq u_i^0+\sigma\} \supset \Delta'_e \supset \Delta'$ makes the problem easy. Note that if we were to perform these kind of crude enlargements on $\max_{\boldsymbol{u}} F(\boldsymbol{u})$ directly we would loose the bounds by $O(\sigma)$.

Example (continued). $\sigma = \frac{1}{10}$, $h'(\frac{3}{10}) = \frac{13051}{2520} - \frac{1}{2}\Pi^2$, $h'(\frac{7}{10}) = \frac{91717}{8400} - \frac{7}{6}\Pi^2$, $H_0 = H(\boldsymbol{u}^0) = h(\frac{3}{10}) + h(\frac{6}{10})$, $H_R^{ub} = \frac{1}{10}h'(\frac{3}{10})$, $H_R^{lb} = \frac{1}{10}h'(\frac{7}{10}) \Rightarrow [H_0 + H_R^{lb}, H_0 + H_R^{ub}] = [0.5564...,0.6404...]$, hence $H_0 + H_R^{ub} - \overline{H} = 0.0148 = O(\frac{1}{10^2})$, $\underline{H} - H_0 - H_R^{lb} = 0.0074... = O(\frac{1}{10^2})$.

Error propagation. Assume we found bounds for estimators $G(\boldsymbol{u})$ and $H(\boldsymbol{u})$ and we want now to bound the sum $F(\boldsymbol{u}) := G(\boldsymbol{u}) + H(\boldsymbol{u})$. In the direct approach $\overline{F} \leq \overline{G} + \overline{H}$ we may lose $O(\sigma)$. A simple example is $G(\boldsymbol{u}) = u_i$ and $H(\boldsymbol{u}) = -u_i$ for which $F(\boldsymbol{u}) = 0$, hence $0 = \overline{F} \leq \overline{G} + \overline{H} = u_i^0 + \sigma - u_i^0 = \sigma$, i.e. $\overline{F} \not\sqsubseteq \overline{G} + \overline{H}$. We can exploit the techniques of the previous section to obtain $O(\sigma^2)$ approximations.

$$F_{iR}^{ub} \;=\; \sigma \max_{\boldsymbol{u}\in\Delta'_e} \partial_i F(\boldsymbol{u}) \;\sqsubseteq\; \sigma \max_{\boldsymbol{u}\in\Delta'_e} \partial_i G(\boldsymbol{u}) + \sigma \max_{\boldsymbol{u}\in\Delta'_e} \partial_i H(\boldsymbol{u}) \;=\; G_{iR}^{ub} + H_{iR}^{ub}$$

Theorem 4 (Error propagation: Sum) *Let $G(\boldsymbol{u})$ and $H(\boldsymbol{u})$ be Lipschitz differentiable and $F(\boldsymbol{u}) = \alpha G(\boldsymbol{u}) + \beta H(\boldsymbol{u})$, $\alpha, \beta \geq 0$, then $\overline{F} \sqsubseteq F_0 + F_R^{ub}$ and $\underline{F} \sqsupseteq F_0 + F_R^{lb}$, where $F_0 = \alpha G_0 + \beta H_0$, and $F_{iR}^{ub} \sqsubseteq \alpha G_{iR}^{ub} + \beta H_{iR}^{ub}$, and $F_{iR}^{lb} \sqsupseteq \alpha G_{iR}^{lb} + \beta H_{iR}^{lb}$.*

It is important to notice that $F_R^{ub} \not\sqsubseteq G_R^{ub} + H_R^{ub}$ (use previous example), i.e. $\max_i[G_{iR}^{ub} + H_{iR}^{ub}] \not\sqsubseteq \max_i G_{iR}^{ub} + \max_i H_{iR}^{ub}$. $\max_i$ can not be pulled in and it is important to propagate F_{iR}^{ub}, rather than F_R^{ub}.

Every function F with bounded curvature can be written as a sum of a concave function G and a convex function H. For convex and concave functions, determining bounds is particularly easy, as we have seen. Often F decomposes naturally into convex and concave parts as is the case for the mutual information, addressed later. Bounds can also be derived for products.

Theorem 5 (Error propagation: Product) *Let $G,H:\Delta'_e\to[0,\infty)$ be non-negative Lipschitz differentiable functions (9) with non-negative derivatives $\partial_i G,\partial_i H\geq 0\;\forall i$ and $F(\boldsymbol{u})=G(\boldsymbol{u})\cdot H(\boldsymbol{u})$, then $\overline{F}\sqsubseteq F_0+F_R^{ub}$, where $F_0=G_0\cdot H_0$, and $F_{iR}^{ub}\sqsubseteq G_{iR}^{ub}(H_0+H_R^{ub})+(G_0+G_R^{ub})H_{iR}^{ub}$, and similarly for $\underline{F}$.*

Proof. We have

$$F_{iR}^{ub}\;=\;\sigma\max\partial_i F\;=\;\sigma\max\partial_i(G{\cdot}H)\;=\;\sigma\max[(\partial_i G)H+G(\partial_i H)]\;\sqsubseteq$$

$$\sigma(\max\partial_i G)(\max H)+\sigma(\max G)(\max\partial_i H)\sqsubseteq G_{iR}^{ub}(H_0{+}H_R^{ub})+(G_0{+}G_R^{ub})H_{iR}^{ub}$$

where all functions depend on $\boldsymbol{u}$ and all max are over $\boldsymbol{u}\in\Delta'_e$. There is one subtlety in the last inequality: $\max G\neq\overline{G}\sqsubseteq G_0+G_R^{ub}$. The reason for the $\neq$ being that the maximization is taken over Δ'_e, not over Δ' as in the definition of $\overline{G}$. The correct line of reasoning is as follows:

$$\max_{\boldsymbol{u}\in\Delta'_e}G_R(\boldsymbol{u})\sqsubseteq\max_{\boldsymbol{t}\in\Delta_e}\sum_i G_{iR}^{ub}\cdot t_i=\max\{0,\max_i G_{iR}^{ub}\}=G_R^{ub}\;\Rightarrow\;\max G\sqsubseteq G_0+G_R^{ub}$$

The first inequality can be proven in the same way as (11). In the first equality we set the $t_i=1$ with maximal G_{iR}^{ub} *if* it is positive. If all G_{iR}^{ub} are negative we set $\boldsymbol{t}\equiv\boldsymbol{0}$. We assumed $G\geq 0$ and $\partial_i G\geq 0$, which implies $G_R\geq 0$. So, since $G_R\geq 0$ anyway, this subtlety is ineffective. Similarly for $\max H_R$. □

It is possible to remove the rather strong non-negativity assumptions. Propagation of errors for other combinations like ratios $F=G/H$ may also be obtained.

6 Robust Intervals for Mutual Information

Mutual Information. We illustrate the application of the previous results on the Mutual Information between two random variables $\imath\in\{1,...,d_1\}$ and $\jmath\in\{1,...,d_2\}$. Consider an i.i.d. random process with outcome $(i,j)\in\{1,...,d_1\}\times\{1,...,d_2\}$ having joint probability π_{ij}, where $\boldsymbol{\pi}\in\Delta:=\{\boldsymbol{x}\in I\!R^{d_1\times d_2}:x_{ij}\geq 0\forall ij,\;x_{++}=1\}$. An important measure of the stochastic dependence of $\imath$ and $\jmath$ is the mutual information

$$I(\boldsymbol{\pi})=\sum_{i=1}^{d_1}\sum_{j=1}^{d_2}\pi_{ij}\log\frac{\pi_{ij}}{\pi_{i+}\pi_{+j}}=\sum_{ij}\pi_{ij}\log\pi_{ij}-\sum_i\pi_{i+}\log\pi_{i+}-\sum_j\pi_{+j}\log\pi_{+j}\quad(12)$$

$$=\;\mathcal{H}(\boldsymbol{\pi}_{\imath+})+\mathcal{H}(\boldsymbol{\pi}_{+\jmath})-\mathcal{H}(\boldsymbol{\pi}_{\imath\jmath})$$

$\pi_{i+}=\sum_j\pi_{ij}$ and $\pi_{+j}=\sum_i\pi_{ij}$ are row and column marginal chances. Again, we assume a Dirichlet prior over $\boldsymbol{\pi}_{\imath\jmath}$, which leads to a Dirichlet posterior $p(\boldsymbol{\pi}_{\imath\jmath}|\boldsymbol{n})\propto\prod_{ij}\pi_{ij}^{n_{ij}+st_{ij}-1}$ with $\boldsymbol{t}\in\Delta$. The expected value of π_{ij} is

$$E_{\boldsymbol{t}}[\pi_{ij}]=\frac{n_{ij}+st_{ij}}{n+s}=:u_{ij}$$

The marginals $\boldsymbol{\pi}_{i+}$ and $\boldsymbol{\pi}_{+j}$ are also Dirichlet with expectation u_{i+} and u_{+j}. The expected mutual information $I(\boldsymbol{u}) := E_{\boldsymbol{t}}[I]$ can, hence, be expressed in terms of the expectations of three entropies $H(\boldsymbol{u}) := E_{\boldsymbol{t}}[\mathcal{H}]$ (see (5))

$$\begin{aligned} I(\boldsymbol{u}) = H(\boldsymbol{u}_{\imath+}) + H(\boldsymbol{u}_{+\jmath}) - H(\boldsymbol{u}_{\imath\jmath}) \;&=\; H_{row} + H_{col} - H_{joint} \\ &=\; \sum_i h(u_{i+}) + \sum_j h(u_{+j}) - \sum_{ij} h(u_{ij}) \end{aligned}$$

where here and in the following we index quantities with *joint*, *row*, and *col* to denote to which distribution the quantity refers.

Crude bounds for $I(\boldsymbol{u})$. Estimates for the robust IDM interval $[\min_{\boldsymbol{t}\in\Delta} E_{\boldsymbol{t}}[I],$ $\max_{\boldsymbol{t}\in\Delta} E_{\boldsymbol{t}}[I]]$ can be obtained by [minimizing,maximizing] $I(\boldsymbol{u})$. A crude upper bound can be obtained as

$$\overline{I} \;:=\; \max_{\boldsymbol{t}\in\Delta} I(\boldsymbol{u}) \;=\; \max[H_{row} + H_{col} - H_{joint}] \;\leq$$

$$\max H_{row} + \max H_{col} - \min H_{joint} \;=\; \overline{H}_{row} + \overline{H}_{col} - \underline{H}_{joint},$$

where exact solutions to $\overline{H}_{row}$, $\underline{H}_{row}$ and $\underline{H}_{joint}$ are available from Section 3. Similarly $\underline{I} \geq \underline{H}_{row} + \underline{H}_{col} - \overline{H}_{joint}$. The problem with these bounds is that, although good in some cases, they can become arbitrarily crude. The following $O(\sigma^2)$ bound can be derived by exploiting the error sum propagation Theorem 4.

Theorem 6 (Bound on lower and upper Mutual Information) *The following bounds on the expected mutual information $I(\boldsymbol{u}) = E_{\boldsymbol{t}}[I]$ are valid:*

$$\begin{aligned} &I(\boldsymbol{u}^1) \sqsubseteq \overline{I} \sqsubseteq I_0 + I_R^{ub} \quad \textit{and} \quad I_0 + I_R^{lb} \sqsubseteq \underline{I} \sqsubseteq I(\boldsymbol{u}^2), \quad \textit{where} \\ &I_0 = I(\boldsymbol{u}^0) = H_{0row} + H_{0col} - H_{0joint} = h(u_{i+}^0) + h(u_{+j}^0) - h(u_{ij}^0), \\ &I_{ijR}^{ub} \sqsubseteq H_{iRrow}^{ub} + H_{jRcol}^{ub} - H_{ijRjoint}^{lb} = h'(u_{i+}^0) + h'(u_{+j}^0) - h'(u_{ij}^0{+}\sigma), \\ &I_{ijR}^{lb} \sqsupseteq H_{iRrow}^{lb} + H_{jRcol}^{lb} - H_{ijRjoint}^{ub} = h'(u_{i+}^0{+}\sigma) + h'(u_{+j}^0{+}\sigma) - h'(u_{ij}^0), \end{aligned}$$

with h defined in (5), and $t_{ij}^0 = 0$, and $t_{ij}^1 = \delta_{(ij)(ij)^1}$ with $(ij)^1 = \arg\max_{ij} I_{ijR}^{ub}$, and $t_{ij}^2 = \delta_{(ij)(ij)^2}$ with $(ij)^2 = \arg\min_{ij} I_{ijR}^{lb}$.

7 IDM for Product Spaces

Product spaces $\Omega = \Omega_1 \times ... \times \Omega_m$ with $\Omega_k = \{1,...d_k\}$ occur frequently in practical problems, e.g. in the mutual information ($m = 2$), in robust trees ($m = 3$), or in Bayesian nets in general (m large). Without loss of generality we only discuss the $m = 2$ case in the following. Ignoring the underlying structure in Ω, a Dirichlet prior in case of unknown chances $\pi_{\imath\jmath}$ and an IDM as used in Section 6 with

$$\boldsymbol{t} \in \Delta := \{\boldsymbol{t} \in I\!R^{d_1 \times d_2} \equiv I\!R^{d_1} \otimes I\!R^{d_2} \,:\, t_{ij} \geq 0 \forall ij,\; t_{++} = 1\} \tag{13}$$

seems natural. On the other hand, if we take into account the structure of Ω and go back to the original motivation of IDM this choice is far less obvious. Recall that one of the major motivations of IDM was its reparametrization invariance in the sense that inferences are not affected when grouping or splitting events in Ω. For unstructured spaces like Ω_k this is a reasonable principle. For illustration, let us consider objects of various *shape* and *color*, i.e. $\Omega = \Omega_1 \times \Omega_2$, $\Omega_1 = \{ball, pen, die, ...\}$, $\Omega_2 = \{yellow, red, green, ...\}$ in generalization to Walleys bag of marbles example. Assume we want to detect a potential dependency between *shape* and *color* by means of their mutual information I. If we have no prior idea on the possible kind of colors, a model which is independent of the choice of Ω_2 is welcome. Grouping red and green, for instance, corresponds to $(x_{i1}, x_{i2}, x_{i3}, x_{i4}, ...) \rightsquigarrow (x_{i1}, x_{i2}+x_{i3}, x_{i4}, ...)$ *for all shapes i*, where $\boldsymbol{x} \in \{\boldsymbol{n}, \boldsymbol{\pi}, \boldsymbol{t}, \boldsymbol{u}\}$. Similarly for the different shapes, for instance we could group all round or all angular objects. The "smallest IDM" which respects this invariance is the one which considers all

$$\boldsymbol{t} \in \boldsymbol{\Delta} := \Delta_{d_1} \otimes \Delta_{d_2} \; \subsetneq \; \Delta. \tag{14}$$

The tensor or outer product $\otimes$ is defined as $(\boldsymbol{v} \otimes \boldsymbol{w})_{ij} := v_i w_j$ and $V \otimes W := \{\boldsymbol{v} \otimes \boldsymbol{w} : \boldsymbol{v} \in V, \boldsymbol{w} \in W\}$. It is a bilinear (not linear!) mapping. This "small tensor" IDM is invariant under arbitrary grouping of columns and rows of the chance matrix $(\boldsymbol{\pi}_{ij})_{1 \leq i \leq d_1, 1 \leq j \leq d_2}$. In contrast to the larger Δ IDM model it is not invariant under arbitrary grouping of matrix cells, but there is anyway little motivation for the necessity of such a general invariance. General non-column/row cross groupings would destroy the product structure of Ω and with that the mere concepts of shape and color, and their correlation. For $m > 2$ as in Bayes-nets cross groupings look even less natural. Whether the $\boldsymbol{\Delta}$ or the larger simplex Δ is the more appropriate IDM model depends on whether one regards the structure $\Omega_1 \times \Omega_2$ of Ω as a natural prior knowledge or as an arbitrary a posteriori choice. The smaller IDM has the potential advantage of leading to more precise predictions (smaller robust sets).

Let us consider an estimator $F : \Delta \to I\!\!R$ and its restriction $F_\otimes : \boldsymbol{\Delta} \to I\!\!R$. Robust intervals $[\underline{F}, \overline{F}]$ for Δ are generally wider than robust intervals $[\underline{F}_\otimes, \overline{F}_\otimes]$ for $\boldsymbol{\Delta}$. Fortunately not much. Although $\boldsymbol{\Delta}$ is a *lower-dimensional* subspace of Δ, it contains all vertices of Δ. This is possible since $\boldsymbol{\Delta}$ is a *nonlinear* subspace. The set of "vertices" in both cases is $\{\boldsymbol{t} : t_{ij} = \delta_{ii_0} \delta_{jj_0},\ i_0 \in \Omega_1,\ j_0 \in \Omega_2\}$. Hence, *if* the robust interval boundaries $\overline{\underline{F}}$ are assumed in the vertices of Δ *then* the interval for the $\boldsymbol{\Delta}$ IDM model is the same ($\overline{\underline{F}} = \overline{\underline{F}}_\otimes$). Since the condition is "approximately" true, the conclusion is "approximately" true. More precisely:

Theorem 7 (IDM bounds for product spaces) *The $O(\sigma^2)$ bounds of Theorem 3 on the robust interval $\overline{\underline{F}}$ in the full IDM model Δ (13), remain valid for $\overline{\underline{F}}_\otimes$ in the product IDM model $\boldsymbol{\Delta}$ (14).*

Proof.

$$F(\boldsymbol{u}^1) \leq \overline{F}_\otimes \leq \overline{F} \leq F_0 + F_R^{ub} = F(\boldsymbol{u}^1) + O(\sigma^2),$$

where $\overline{F}_\otimes := \max_{\boldsymbol{t}\in\mathbb{\Delta}} F(\boldsymbol{u})$ and $\boldsymbol{u}^1$ was the “F_R maximizing” vertex as defined in Theorem 6 ($F(\boldsymbol{u}^1) \sqsubseteq \overline{F}$). The first inequality follows from the fact that all Δ vertices also belong to $\mathbb{\Delta}$, i.e. $\boldsymbol{t}^1 \in \mathbb{\Delta}$. The second inequality follows from $\mathbb{\Delta} \subset \Delta$. The remaining (in)equalities follow from Theorem 3. This shows that $|\overline{F}_\otimes - \overline{F}| = O(\sigma^2)$, hence $F_0 + F_R^{ub}$ is also an $O(\sigma^2)$ upper bound to $\overline{F}_\otimes$. This implies that to the approximation accuracy we can achieve, the choice between Δ and $\mathbb{\Delta}$ is irrelevant. □

8 Robust Credible Intervals

Bayesian credible sets/intervals. For a probability distribution $p : I\!\!R^d \to [0,1]$, an α-credible region is a measurable set A for which $p(A) := \int p(x)\chi_A(x)d^dx \geq \alpha$, where $\chi_A(x) = 1$ if $x \in A$ and 0 otherwise, i.e. $x \in A$ with probability at least α. For given α, there are many choices for A. Often one is interested in “small” sets, where the size of A may be measured by its volume $\text{Vol}(A) := \int \chi_A(x)d^dx$. Let us define a/the smallest α-credible set

$$A^{min} := \arg\min_{A:p(A)\geq\alpha} \text{Vol}(A)$$

with ties broken arbitrarily. For unimodal p, A^{min} can be chosen as a connected set. For $d = 1$ this means that $A^{min} = [a,b]$ with $\int_a^b p(x)dx = \alpha$ is a minimal length α-credible interval. If, additionally p is symmetric around $E[x]$, then $A^{min} = [E[x] - a, E[x] + a]$ is also symmetric around $E[x]$.

Robust credible sets. If we have a set of probability distributions $\{p_t(x),\, t \in T\}$, we can choose for each t an α-credible set A_t with $p_t(A_t) \geq \alpha$, a minimal one being $A_t^{min} := \arg\min_{A:p_t(A)\geq\alpha}\text{Vol}(A)$. A robust α-credible set is a set A which contains x with p_t-probability at least α for *all* t. A minimal size robust α-credible set is

$$A^{min} := \arg\min_{A=\bigcup_t A_t : p_t(A_t)\geq\alpha \forall t\in T} \text{Vol}(A) \tag{15}$$

It is not easy to deal with this expression, since A^{min} is *not* a function of $\{A_t^{min} : t \in T\}$, and especially does not coincide with $\bigcup_t A_t^{min}$ as one might expect.

Robust credible intervals. This can most easily be seen for univariate symmetric unimodal distributions, where t is a translation, e.g. $p_t(x) = \text{Normal}(E_t[x] = t, \sigma = 1)$ with 95% credible intervals $A_t^{min} = [t-2, t+2]$. For, e.g. $T = [-1,1]$ we get $\bigcup_t A_t^{min} = [-3,3]$. The credible intervals *move* with t. One can get a smaller union if we take the intervals $A_t^{sym} = [-a_t, a_t]$ symmetric around 0. Since A_t^{sym} is a non-central interval w.r.t. p_t for $t \neq 0$, we have $a_t > 2$, i.e. A_t^{sym} is larger than A_t^{min}, but one can show that the increase of a_t is smaller than the shift of A_t^{min} by t, hence we save something in the union. The optimal choice is neither A_t^{sym} nor A_t^{min}, but something in-between. In the extended version [Hut03] this is illustrated for the triangular distribution $p_t(x)=\max\{0\,,1-|x-t|\}$ with $t\in T:=[-\gamma,\gamma]$, where closed form solutions can be given.

An interesting open question is under which general conditions we can expect $A^{min}\subseteq\bigcup_t A_t^{min}$. In any case, $\bigcup_t A_t$ can be used as a conservative estimate for a robust credible set, since $p_t(\bigcup_{t'}A_{t'})\geq p_t(A_t)\geq\alpha$ for all t.

A special (but important) case which falls outside the above framework are one-sided credible intervals, where only A_t of the form $[a,\infty)$ are considered. In this case $A^{min}=\bigcup_t A_t^{min}$, i.e. $A^{min}=[a_{min},\infty)$ with $a_{min}=\max\{a:p_t([a,\infty])\geq\alpha\forall t\}$.

Approximations. For complex distributions like for the mutual information we have to approximate (15) somehow. We use the following notation for shortest α-credible *intervals* w.r.t. a univariate distribution $p_t(x)$:

$$\underset{\sim}{\widetilde{x}}_t \;\equiv\; [\underset{\sim}{x}_t,\widetilde{x}_t] \;\equiv\; [E_t[x]-\Delta\underset{\sim}{x}_t\,,\,E_t[x]+\Delta\widetilde{x}_t] \;:=\; \mathop{\arg\min}_{[a,b]:p_t([a,b])\geq\alpha}\;(b-a),$$

where $\Delta\widetilde{x}_t:=\widetilde{x}_t-E_t[x]$ ($\Delta\underset{\sim}{x}_t:=E_t[x]-\underset{\sim}{x}_t$) is the distance from the right boundary $\widetilde{x}_t$ (left boundary $\underset{\sim}{x}_t$) of the shortest α-credible interval $\underset{\sim}{\widetilde{x}}_t$ to the mean $E_t[x]$ of distribution p_t. We can use $\underset{\simeq}{\overline{\widetilde{x}}}\equiv[\underset{\simeq}{x},\overline{\widetilde{x}}]:=\bigcup_t\underset{\sim}{\widetilde{x}}_t$ as a (conservative, but not shortest) robust credible interval, since $p_t(\underset{\simeq}{\overline{\widetilde{x}}})\geq p_t(\underset{\sim}{\widetilde{x}}_t)\geq\alpha$ for all t. We can upper bound $\overline{\widetilde{x}}$ (and similarly lower bound $\underset{\simeq}{x}$) by

$$\overline{\widetilde{x}} \;=\; \max_t(E_t[x]+\Delta\widetilde{x}_t) \;\leq\; \max_t E_t[x]+\max_t\Delta\widetilde{x}_t \;=\; \overline{E[x]}+\overline{\Delta\widetilde{x}}. \qquad (16)$$

We have already intensively discussed how to compute upper and lower quantities, particularly for the upper mean $\overline{E[x]}$ for $x\in\{\mathcal{F},\mathcal{H},I,...\}$, but the linearization technique introduced in Section 4 is general enough to deal with all in t differentiable quantities, including $\Delta\widetilde{x}_t$. For example for Gaussian p_t with variances σ_t we have $\Delta\widetilde{x}_t=\kappa\sigma_t$ with κ given by $\alpha=\mathrm{erf}(\kappa/\sqrt{2})$, where erf is the error function (e.g. $\kappa=2$ for $\alpha\approx 95\%$). We only need to estimate $\max_t\sigma_t$.

For non-Gaussian distributions, exact expression for $\Delta\widetilde{x}_t$ are often hard or impossible to obtain and to deal with. Non-Gaussian distributions depending on some sample size n are usually close to Gaussian for large n due to the central limit theorem. One may simply use $\kappa\sigma_t$ in place of $\Delta\widetilde{x}_t$ also in this case, keeping in mind that this could be a non-conservative approximation. More systematically, simple (and for large n good) upper bounds on $\Delta\widetilde{x}_t$ can often be obtained and should preferably be used.

Further, we have seen that the variation of sample depending differentiable functions (like $E_t[x]=E_t[x|\boldsymbol{n}]$) w.r.t. $t\in\Delta$ are of order $\frac{s}{n+s}$. Since in such cases the standard deviation $\sigma_t\sim n^{-1/2}\sim\Delta\widetilde{x}_t$ is itself suppressed, the variation of $\Delta\widetilde{x}_t$

with t is of order $n^{-3/2}$. If we regard this as negligibly small, we may simply fix some $t^* \in \Delta$:

$$\max_t \Delta\widetilde{x}_t = \kappa\sigma_{t^*} + O(n^{-3/2})$$

Since $\Delta\widetilde{x}_t$ is "nearly" constant, this also shows that we lose at most $O(n^{-3/2})$ precision in the bound (16) (equality holds for $\Delta\widetilde{x}_t$ independent of t). Expressions for the variance of I, for instance, have been derived in [WW95, Hut02].

9 Conclusions

This is the first work, providing a systematic approach for deriving closed form expressions for interval estimates in the Imprecise Dirichlet Model (IDM). We concentrated on exact and conservative *robust* interval ([lower,upper]) estimates for concave functions $F = \sum_i f_i$ on simplices, like the entropy. The conservative estimates widened the intervals by $O(n^{-2})$, where n is the sample size. Here is a dilemma, of course: For large n the approximations are good, whereas for small n the bounds are more interesting, so the approximations will be most useful for intermediate n. More precise expressions for small n would be highly interesting. We have also indicated how to propagate robust estimates from simple functions to composite functions, like the mutual information. We argued that a reduced IDM on product spaces, like Bayesian nets, is more natural and should be preferred in order to improve predictions. Although improvement is formally only $O(n^{-2})$, the difference may be significant in Bayes nets or for very small n. Finally, the basics of how to combine robust with credible intervals have been laid out. Under certain conditions $O(n^{-3/2})$ approximations can be derived, but the presented approximations are not conservative. All in all this work has shown that IDM has not only interesting theoretical properties, but that explicit (exact/conservative/approximate) expressions for robust (credible) intervals for various quantities can be derived. The computational complexity of the derived bounds on $F = \sum_i f_i$ is very small, typically one or two evaluations of F or related functions, like its derivative. First applications of these (or more precisely, very similar) results, especially the mutual information, to robust inference of trees look promising [ZH03].

Acknowledgements. I want to thank Peter Walley for introducing IDM to me and Marco Zaffalon for encouraging me to investigate this topic. This work was supported by SNF grant 2000-61847.00 to Jürgen Schmidhuber.

References

[AS74] M. Abramowitz and I. A. Stegun, editors. *Handbook of mathematical functions*. Dover publications, inc., 1974.

[Ber01] J.-M. Bernard. Non-parametric inference about an unknown mean using the Imprecise Dirichlet Model. In G. de Cooman, T. Fine, and T. Seidenfeld, editors, *Proceedings of the 2nd International Symposium on Imprecise Probabilities and Their Application (ISIPTA-2001)*, pages 40–50, The Netherlands, 2001. Shaker Publishing.

[GCSR95] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis.* Chapman, 1995.

[Hal48] J. B. S. Haldane. The precision of observed values of small frequencies. *Biometrika*, 35:297–300, 1948.

[Hut02] M. Hutter. Distribution of mutual information. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 399–406, Cambridge, MA, 2002. MIT Press.

[Hut03] M. Hutter. Robust estimators under the Imprecise Dirichlet Model (extended version). Technical report, IDSIA, Manno(Lugano), Switzerland, 2003. http://www.idsia.ch/∼marcus/ai/idmx.ps.

[Jef46] H. Jeffreys. An invariant form for the prior probability in estimation problems. In *Proc. Royal Soc. London (A)*, volume 186, pages 453–461, 1946.

[Per47] W. Perks. Some observations on inverse probability. *J. Inst. Actuar.*, 73:285–312, 1947.

[Wal96] P. Walley. Inferences from multinomial data: Learning about a bag of marbles. *Journal of the Royal Statistical Society B*, 58(1):3–57, 1996.

[WW95] D. H. Wolpert and D. R. Wolf. Estimating functions of distributions from a finite set of samples. *Physical Review E*, 52(6):6841–6854, 1995.

[Zaf01] M. Zaffalon. Robust discovery of tree-dependency structures. In G. de Cooman, T. Fine, and T. Seidenfeld, editors, *Proceedings of the 2nd International Symposium on Imprecise Probabilities and Their Application (ISIPTA-2001)*, pages 394–403, The Netherlands, 2001. Shaker Publishing.

[ZH03] M. Zaffalon and M. Hutter. Robust inference of trees. Technical Report IDSIA-11-03, IDSIA, Manno (Lugano), CH, 2003.

Marcus Hutter is with the AI research institute IDSIA, Galleria 2, CH-6928 Manno-Lugano, Switzerland. E-mail: marcus@idsia.ch, HP: http://www.idsia.ch/∼marcus/idsia