Game Theory, Maximum Generalized Entropy, Minimum Discrepancy, Robust Bayes and Pythagoras

P. D. Grünwald[1]
CWI Amsterdam,
The Netherlands.
e-mail: pdg@cwi.nl

A. P. Dawid
University College London,
United Kingdom.
e-mail: dawid@stats.ucl.ac.uk

Suppose that, for purposes of inductive inference or choosing an optimal decision, we wish to select a single distribution P^* to act as representative of a class Γ of such distributions. The Maximum Entropy Principle (Jaynes 1989; Csiszár 1991) is widely applied for this purpose, but its rationale has often been controversial (Shimony 1985; Seidenfeld 1986). Here we emphasize and generalize a reinterpretation of the Maximum Entropy Principle (Topsøe 1979): that the distribution P^* that maximizes the entropy over Γ also minimizes the worst-case expected logarithmic score (log loss). In the terminology of Decision Theory (Berger 1985), P^* is a *robust Bayes*, or Γ-*minimax*, act, when loss is measured by the log loss. This gives a decision-theoretic justification for Maximum Entropy.

This Work We extend this result to apply to a generalized concept of entropy, tailored to whatever loss function L is regarded as appropriate, not just log loss. We show that, under regularity conditions, maximizing this generalized entropy constitutes the major step towards finding the minimax decision against Γ with respect to L. This leads to a notion of *generalized exponential families* of distributions, which, for the case of the logarithmic loss, reduce to the usual exponential families. We extend generalized entropy to *generalized relative entropy*, and show how this is essentially the same as a general decision-theoretic definition of *discrepancy*. We show that those divergences between probability measures that belong to the family of *Bregman divergences* are special cases of such discrepancies. A discrepancy can also be used as a loss function in its own right: we show that a minimax result for relative entropy (Haussler 1997) can be extended to this more general case. Finally, in what is perhaps our most intriguing result, we show that a 'Pythagorean property' (Csiszár 1991) known to hold for relative entropy and for Bregman divergences in fact applies much more generally, and is intimately connected to game theory: roughly speaking, the Pythagorean property holds for a discrepancy D if and only if the minimax theorem holds for the loss function on which D is based. This abstract is a brief overview of our results in (Grünwald and Dawid 2002). For ease of exposition, we make several strong yet simplifying assumptions here, such as a finite sample space. In (Grünwald and Dawid 2002) our assumptions are considerably relaxed.

[1]Peter Grünwald was supported in part by the EU Fourth Framework BRA NeuroCOLT II Working Group EP 27150, and the European Science Foundation Programme on Highly Structured Stochastic Systems. Philip Dawid received support from Eurandom and the Gatsby Charitable Foundation.

I Maximum Entropy and game theory

Let $\mathcal{X}$ be a finite sample space, and Γ a family of distributions over $\mathcal{X}$. Consider a Decision Maker DM, who has to make a decision whose consequences will depend on the outcome of a random variable X defined on $\mathcal{X}$. She is willing to assume that X is distributed according to some $P \in \Gamma$, a known family of distributions over $\mathcal{X}$, but she does not know which such distribution applies. DM would like to pick a single $P^* \in \Gamma$ to base decisions on. One way of selecting such a P^* is to apply the *Maximum Entropy Principle* (Jaynes 1989), which advises DM to pick that distribution $P^* \in \Gamma$ maximizing $H(P)$ over all $P \in \Gamma$. Here $H(P)$ denotes the *Shannon entropy* of P, $H(P) := -\sum_{x \in \mathcal{X}} p(x) \log p(x) = \mathrm{E}_P\{-\log p(X)\}$, where p is the probability mass function of P.

Let $\mathcal{A}$ be the set of all probability mass functions defined over $\mathcal{X}$. By the information inequality (Cover and Thomas 1991), we have that, for any distribution P, $\inf_{q \in \mathcal{A}} \mathrm{E}_P\{-\log q(X)\}$ is achieved, uniquely, at $q = p$, where it takes the value $H(P)$. That is,

$$H(P) = \inf_{q \in \mathcal{A}} \mathrm{E}_P\{-\log q(X)\}, \qquad (1)$$

and so the maximum entropy can be written as

$$\sup_{P \in \Gamma} H(P) = \sup_{P \in \Gamma} \inf_{q \in \mathcal{A}} \mathrm{E}_P\{-\log q(X)\}.$$

Now consider the 'log loss game', in which DM has to specify some $q \in \mathcal{A}$, and her ensuing loss, if Nature then reveals $X = x$, is measured by $-\log q(x)$. As is well known, this loss function has a coding-theoretic interpretation: $L(x) = -\log q(x)$ may be interpreted as the number of bits needed to describe x when x is encoded based on the code corresponding to probability mass function q (Cover and Thomas 1991). By analogy with minimax results of game theory, one might conjecture that

$$\sup_{P \in \Gamma} \inf_{q \in \mathcal{A}} \mathrm{E}_P\{-\log q(X)\} = \inf_{q \in \mathcal{A}} \sup_{P \in \Gamma} \mathrm{E}_P\{-\log q(X)\}. \qquad (2)$$

As we have seen, P achieving the supremum on the left-hand side of (2) is a maximum entropy distribution in Γ.

But, just as important, q achieving the infimum on the right-hand side of (2) is a *robust Bayes* act against Γ, or a Γ-*minimax* act (Berger 1985), for the log loss decision problem.

Indeed, it turns out that, when Γ is closed and convex,

(i). (2) holds under very general conditions.

(ii). The infimum on the right-hand side is achieved, uniquely, for $q = p^*$, the probability mass function of the maximum entropy distribution P^*.

Thus, in this game between DM and Nature, the maximum entropy distribution P^* may be viewed, simultaneously, as defining both Nature's maximin and – in our view more interestingly – DM's minimax strategy. In other words, *Maximum Entropy is Robust Bayes.* Note that we did not *not* restrict the acts q available to DM to those corresponding to a distribution in the restricted set Γ: that the optimal act p^* does indeed turn out to have this property is a consequence of, not a restriction on, the analysis.

Summarizing, if DM is interested in coding outcomes distributed according to one of the distributions in Γ, then he can minimize the worst-case expected code length by using the code based on the maximum entropy distribution P^*. For finite sample spaces, proving (i) and (ii) is easy (Grünwald and Dawid 2002). Using much more advanced techniques, Topsøe (1979) showed that (i) and (ii) continue to hold in much more general sample spaces (with appropriately adjusted definitions of entropy).

II Our work: Generalized entropy

The above Robust Bayes view of Maximum Entropy might be regarded as justifying its use in those decision problems, such as *discrete coding* and *Kelly gambling* (Cover and Thomas 1991), where the log loss is an appropriate loss function to use. But what if we are interested in other loss functions? This is the principal question we address here. Our basic tool will be a natural generalization of the concept of entropy (or 'uncertainty inherent in a distribution'), tailored to the specific loss function we are interested in.

A Basic Result

Our Game First we need to generalize our setting. We assume a decision maker DM has to take some action a selected from a given 'action space' $\mathcal{A}$, after which Nature will reveal the value $x \in \mathcal{X}$ of a quantity X. DM will then suffer a real-valued loss $L(x, a)$. We suppose that Nature takes no account of the action chosen by DM. Then this can be considered as a zero-sum game between Nature and DM. For now, we assume $\mathcal{X}$ to be finite and L to be uniformly bounded, *i.e.* there exists K such that for all x, a, $|L(x, a)| < K$.

Both Nature and DM are allowed to make randomized moves, such a move being described by a probability distribution P over $\mathcal{X}$ (for Nature) or ζ over $\mathcal{A}$ (for DM). We denote by $\mathcal{P}_0$ the set of all distributions over $\mathcal{X}$ and by $\mathcal{Z}$ the set of all distributions over $\mathcal{A}$. We define the *expected* loss of a against P as $L(P, a) := \mathrm{E}_{X \sim P}\{L(X, a)\}$ and analogously, the loss of ζ against P as $L(P, \zeta) := \mathrm{E}_{A \sim \zeta}\{L(P, A)\}$. Let $\mathcal{X} = \{1, \ldots, k\}$. We think of each $P \in \mathcal{P}$ as a vector $(p(x_1), \ldots, p(x_k))$ in $\mathbf{R}^k$. Each randomized act $\zeta \in \mathcal{Z}$ is fully determined by the vector $(L(x_1, \zeta), \ldots, L(x_k, \zeta))$. We can therefore think of each ζ as a vector in $\mathbf{R}^k$ and of $\mathcal{Z}$ as a subset of $\mathbf{R}^k$. Using these identifications we endow $\mathcal{P}$ and ζ with the standard Euclidean topology on $\mathbf{R}^k$.

In our setting, Nature is typically restricted to pick a distribution $P \in \Gamma$ where Γ is some convex subset of $\mathcal{P}_0$. The resulting scenario can be viewed as a game $(\Gamma, \mathcal{A}, L)$ between DM and Nature, where Nature selects a distribution $P \in \Gamma$, DM selects an act a from $\mathcal{A}$ and the ensuing loss to DM is taken to be $L(P, a)$. DM is still allowed to randomize, i.e. pick randomized acts $\zeta \in \mathcal{Z}$. In principle, Nature is also allowed to randomize. However, as is not hard to show, by convexity of Γ every randomized act for Nature can be replaced by a non-randomized act having the identical loss function. Therefore we shall not consider randomized acts for nature.

Generalized Entropy The *generalized entropy* of a distribution P (with respect to loss function L) is defined as the minimum loss achieveable against P:

$$H(P) := \inf_{a \in \mathcal{A}} L(P, a),$$

Thus, $H(P)$ is equal to the *Bayes loss* of P. It is well-known that $H(\cdot)$ is always concave (in the appropriate sense, (Grünwald and Dawid 2002)). The interpretation of the Bayes loss H relative to loss function L as a generalized notion of entropy goes back to DeGroot (1962).

Example II.1 We briefly consider entropies corresponding to three loss functions that often arise in practice:

(i). The *log loss*: Let $\mathcal{A}$ be the set of all probability mass functions over $\mathcal{X}$ and let L be the log loss, $L(x, q) = -\log q(x)$ (this loss function is not uniformly bounded but this poses no problems, see below). Then by (1) generalized entropy reduces to Shannon entropy; the entropy $H(P)$ as a function of P for $\mathcal{X} = \{0, 1\}$ is shown in Figure 1 on the left.

(ii). The *Brier score*: let $\mathcal{A}$ be as above and let $L(x, q) := ||\delta^x - q(x)||^2$, where δ^x is the point mass distribution on $X = x$: $\delta^x(y) = 1$ if $x = y$, $\delta^x(y) = 0$ otherwise. One easily shows that $L(P, q) = \sum_x q(x)^2 - 2\sum_x p(x)q(x) + 1$, which is uniquely minimized for $q = p$. Hence, the Bayes act for P is the probability mass function for p itself, and it follows that the generalized entropy for the Brier score (loss) is given by $H(P) = L(P, P) = 1 - \sum_x p(x)^2$. The Brier entropies for $\mathcal{X} = \{0, 1\}$ are shown in Figure 1.

(iii). The *0/1 loss*: Let $\mathcal{A} = \mathcal{X}$. Define $L(x, a) = 0$ if $x = a$ and $L(x, a) = 1$ otherwise. It can be shown

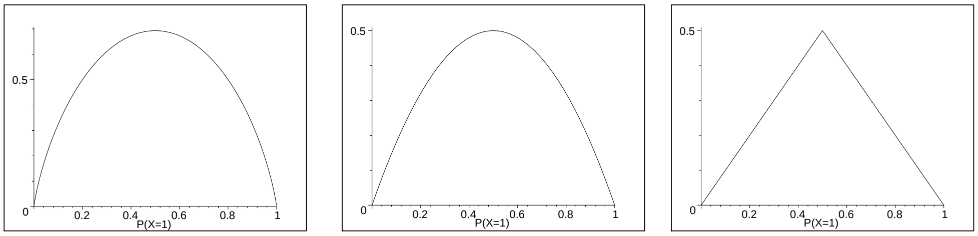

Figure 1: log, Brier and zero-one entropies for the case $\mathcal{X} = \{0, 1\}$

that in this case, the generalized entropy $H(P) = \inf_{a \in \mathcal{X}} L(P, a)$ is given by $H(P) = 1 - p_{\max}$, where $p_{\max} = \max_{x \in \mathcal{X}} P(X = x)$. An example is shown in Figure 1. □

Basic Theorem If Γ is convex, then

(i). The game $(\Gamma, \mathcal{A}, L)$ has a value, *i.e.* $\sup_{P \in \Gamma} H(P) = \sup_{P \in \Gamma} \inf_{\zeta \in \mathcal{Z}} L(P, \zeta) = \inf_{\zeta \in \mathcal{Z}} \sup_{P \in \Gamma} L(P, \zeta)$.

(ii). If, furthermore, Γ and $\mathcal{Z}$ are closed, then there exists a distribution P^* maximizing, over Γ, the generalized entropy $H(\cdot)$ associated with the loss function L. Moreover, P^* has a Bayes act $\zeta^* \in \mathcal{Z}$ (achieving $\inf_{\zeta \in \mathcal{Z}} L(P^*, \zeta)$) that is a robust Bayes (Γ-minimax) decision relative to L. In other words, there is an ζ^* achieving $\inf_{\zeta \in \mathcal{Z}} L(P^*, \zeta)$ such that (P^*, ζ^*) is a *saddle-point* in the game $(\Gamma, \mathcal{A}, L)$, *i.e.*

$$\begin{array}{llll} \text{(a)} & L(P^*, \zeta^*) & \leq & L(P^*, \zeta), \quad \text{for all } \zeta \in \mathcal{Z}. \\ \text{(b)} & L(P^*, \zeta^*) & \geq & L(P, \zeta^*), \quad \text{for all } P \in \Gamma. \end{array}$$

In the simple form stated here, our result is not directly applicable to the (unbounded) log loss function. To make it so, we may slightly restrict both Γ and $\mathcal{A}$: fix an (arbitrarily small) $\epsilon > 0$, let $\mathcal{P}_\epsilon$ be the set of all distributions P over $\mathcal{X}$ with for all x, $p(x) > \epsilon$. Let $\Gamma_\epsilon = \Gamma \cap \mathcal{P}_\epsilon$ and let $\mathcal{A}_\epsilon$ be the set of probability mass functions corresponding to $\mathcal{P}_\epsilon$. Then Γ_ϵ is convex and our basic theorem is applicable to $(\Gamma_\epsilon, \mathcal{A}_\epsilon, L)$. Thus, our basic theorem implies a version of the result for the log loss described in Section I above, restricted to distributions in $\mathcal{P}_\epsilon$. For part (i) this is clear; for part (ii), note that, when the log loss is used, then the Bayes act $q \in \mathcal{A}_\epsilon$ for a distribution $P \in \Gamma_\epsilon$ is unique and equal to the probability mass function of P. Hence, specialized to the log loss, (ii) says that the probability mass function p^* of the maximum Shannon entropy distribution P^* must be Γ-minimax.

Generalizations In the simple form stated here, our basic theorem is an immediate consequence of a well-known generalization of Von Neumann's minimax theorem (Ferguson 1967, Theorem 2, page 85). In (Grünwald and Dawid 2002) we provide several generalizations and variations of our result, which do require several new proof techniques. Broadly, we show that, under appropriate regularity conditions on the loss functions, (variations of) parts (i) and (ii) continue to hold for arbitrary sample spaces, unbounded loss functions and situations where Γ and/or $\mathcal{Z}$ are not closed. As a special case, we get that for finite sample spaces, (i) and (ii) do always hold for the log loss– there is in fact no need to restrict to $\mathcal{P}_\epsilon$.

Remark In the simple form stated here, there is in a sense nothing 'really new' about our basic theorem, since it is an immediate consequence of Ferguson's. However, it does provide a new *interpretation* of the minimax strategy P^* as a maximum generalized entropy distribution. It is this interpretation which leads to several truely novel results which we summarize below.

B Consequences of Basic Result

Generalized relative entropy and discrepancy Given an (arbitrary) loss function L and an (arbitrary) *reference act* $e \in \mathcal{A}$, we define the *relative* (to e) loss as

$$L_e(x, a) := L(x, a) - L(x, e).$$

The generalized entropy relative to L_e is denoted $H_e(P)$:

$$H_e(P) := \inf_{a \in \mathcal{A}} L_e(P, a) = \inf_{a \in \mathcal{A}} L(P, a) - L(P, e).$$

Maximizing this generalized *relative* entropy has an interesting interpretation in terms of minimizing a decision-theoretic 'discrepancy', as we now indicate: for an arbitrary loss function L, we define the associated *discrepancy* $D : \mathcal{P} \times \mathcal{A} \to [0, \infty]$ as

$$D(P, e) := L(P, e) - \inf_{a \in \mathcal{A}} L(P, a) = L(P, e) - H(P). \quad (3)$$

Maximum Generalized Relative Entropy may now be reinterpreted as a 'Minimum Discrepancy method': since

$$H_e(P) = H(P) - L(P, e) = -D(P, e),$$

the distribution achieving $\sup_{P \in \Gamma} H_e(P)$ coincides with the distribution achieving $\inf_{P \in \Gamma} D(P, e)$. The definition of discrepancy is extended to randomized acts in the obvious manner.

Example II.2 For the log loss, the associated discrepancy function is just the familiar Kullback-Leibler (KL) divergence, and the method then coincides with the 'classical' Minimum Relative Entropy method (Jaynes 1989) (note that, for Jaynes, 'relative entropy' is the same as KL divergence; for us it is the negative of this.) For the Brier score, straightforward calculation shows that the associated discrepancy between a distribution P and unrandomized act q is just the Euclidean distance $D(P, q) = \sum_x (p(x) - q(x))^2$. For the 0/1-loss, we get $D(P, x) = p_{\max} - p(x)$, with $p_{\max} = \max_{x \in \mathcal{X}} p(x)$. □

The Pythagorean property The KL divergence has a celebrated property reminiscent of squared Euclidean distance: it satisfies an analogue of the Pythagorean theorem (Csiszár 1975). This property takes the form of the following inequality: let Q_0 be any distribution over $\mathcal{X}$. Let Γ be a convex set of distributions over $\mathcal{X}$. Let $P^* := \arg\min_{P \in \Gamma} D(P, Q_0)$ be the distribution in Γ minimizing KL divergence to Q_0. Then, under weak regularity conditions, for all $P \in \Gamma$,

$$D(P||P^*) + D(P^*||Q_0) \leq D(P||Q_0), \tag{4}$$

D denoting KL divergence. It has been noted (Csiszár 1991; Lafferty 1999) that a version of (4) is shared by the broader class of Bregman divergences. In (Grünwald and Dawid 2002) we show that Bregman divergences that are defined over the unit simplex form a special case of our decision-based discrepancies defined by (3). Therefore one may conjecture that some form of the 'Pythagorean inequality' (4) may hold for more general decision-based discrepancies. We show that this is indeed so, and we connect it to our basic game-theoretic result. Specifically, the following holds for any discrepancy $D(\cdot, \cdot)$ defined in terms of an arbitrary loss function L as in (3): let $\zeta_0 \in \mathcal{Z}$ be an arbitrary (possibly randomized) act, and let $L_0(x, a) := L(x, a) - L(x, \zeta_0)$ be the loss relative to ζ_0. Let $\Gamma \subseteq \mathcal{P}_0$ (note that Γ is *not* necessarily convex!). Let $P^* \in \Gamma$, $\zeta^* \in \mathcal{Z}$. Then we have the following game-theoretic interpretation of the Pythagorean inequality:

THEOREM (P^*, ζ^*) *is a saddle-point in* $(\Gamma, \mathcal{A}, L_0)$ *if and only if* for all $P \in \Gamma$:

$$D(P, \zeta^*) + D(P^*, \zeta_0) \leq D(P, \zeta_0). \tag{5}$$

Example II.3 For the Brier score, the associated discrepancy $D(P, q)$ is the Euclidean distance between the mass function of P and q. As long as we take ζ_0 unrandomized, ζ^* will be unrandomized and equal to p^*, the mass function corresponding to P^* (Ex. II.2). Then, if additionally Γ is convex, (5) is just Pythagoras' theorem applied to projections of points on convex sets.

For the log-loss, the associated discrepancy is the KL divergence, and again $\zeta^* = p^*$ (Ex. II.2). By plugging in a probability mass function q_0 for ζ_0, we see that in this special case, our result (5) says that (4) holds as long as the game with loss function $L_0(x, p) = -\log p(x) - \log q_0(x)$ admits a saddle-point – even if Γ is not convex. If we take q_0 uniform, then $-\log q_0(x)$ is constant and (5) holds if and only if the game with the original, non-relative loss function $L(x, p) = -\log p(x)$ has a saddle-point. $\square$

GENERALIZATIONS In the full paper we show that the Pythagorean theorem above holds not just for finite $\mathcal{X}$ and bounded L. In fact, the theorem holds as long as $D(P, \zeta_0) < \infty$ for all $P \in \Gamma$, *under no further conditions on* $\mathcal{X}, \mathcal{A}$ *and* L *whatsoever*. We also give conditions under which (5) holds with equality.

Further Results In the full paper we consider several further results based on generalized entropy:
1. generalized exponential families We consider in detail constraints of the form $\Gamma = \{P : \mathrm{E}_P(T) = \tau\}$. For fixed loss function L and statistic T, as τ varies we obtain a family of maximum generalized entropy distributions, one for each value of τ. For Shannon entropy, this turns out to coincide with the *exponential family* having natural sufficient statistic T (Csiszár 1975). In close analogy, we define the collection of maximum generalized entropy distributions, as we vary τ, to be the *generalized exponential family* determined by L and T, and we give several examples of such generalized exponential families. In particular, the additive models introduced by Lafferty (1999) are special cases of our generalized exponential families.
2. a generalized redundancy-capacity theorem It is often more natural to seek minimax decisions with respect to the discrepancy associated with a loss, rather than with respect to the loss directly. With any game we thus associate a new 'derived game', in which the discrepancy constructed from the original loss function now serves as a new loss function. It turns out that our basic (minimax) theorem applies to games of this form too: broadly, whenever the conditions for such a theorem hold for the original game, they also hold for the derived game. As a special case, we reprove the 'redundancy-capacity theorem', a well-known minimax theorem for the KL divergence (Haussler 1997; Merhav and Feder 1995).

References

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, (2nd edn). Springer, New York.

Cover, T. and Thomas, J. A. (1991). *Elements of Information Theory.* Wiley Interscience, New York.

Csiszár, I. (1975). *I*-divergence geometry of probability distributions and minimization problems. *Ann. Probab.*, **3**, 146–58.

Csiszár, I. (1991). Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. *Ann. Statist.*, **19**, 2032–66.

DeGroot, M. H. (1962). Uncertainty, information and sequential experiments. *Ann. Math. Stat.*, **33**, 404–19.

Ferguson, T. S. (1967). *Mathematical Statistics. A Decision-Theoretic Approach.* Academic Press, New York.

Grünwald, P. and Dawid, A. (2002). Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. Technical Report 223, University College London, Dept. of Statistical Science.

Haussler, D. (1997). A general minimax result for relative entropy. *IEEE Trans. Inform. Theory*, **43**, 1276–80.

Jaynes, E. T. (1989). *Papers on Probability, Statistics and Statistical Physics.* Kluwer Academic Publishers.

Lafferty, J. (1999). Additive models, boosting and inference for generalized divergences. In *Proc. COLT '99*, pp. 125–33. University of California at Santa Cruz.

Merhav, N. and Feder, M. (1995). A strong version of the redundancy-capacity theorem of universal coding. *IEEE Trans. Inform. Theory*, **41**, 714–22.

Seidenfeld, T. (1986). Entropy and uncertainty. *Philosophy of Science*, **53**, 467–91.

Shimony, A. (1985). The status of the principle of maximum entropy. *Synthèse*, **63**, 35–53.

Topsøe, F. (1979). Information-theoretical optimization techniques. *Kybernetika*, **15**, 8–27.