Generalized Entropy, Game Theory and Pythagoras

Peter Grünwald
EURANDOM
www.cwi.nl/~pdg

Joint work with A.P. Dawid, University College, London

Overview

1. Maximum Entropy (MaxEnt)
2. A Game-Theoretic Characterization of Maximum Entropy
3. Generalized Entropy and Game Theory
4. Pythagoras

Overview

1. Maximum Entropy (MaxEnt)
2. A Game-Theoretic Characterization of Maximum Entropy
3. Generalized Entropy and Game Theory
4. Pythagoras

Setting

$\mathcal{X}$	Finite (for now) Sample Space
$\mathcal{P}$	Set of all distributions over $\mathcal{X}$
$\mathcal{C} \subseteq \mathcal{P}$	'Convex' Closed Subset of $\mathcal{P}$
$\mathbf{H}$	Entropy:

$$\mathbf{H}(P) := E_P[-\ln P(X)] = -\sum_{x \in \mathcal{X}} P(x) \ln P(x)$$

Maximum Entropy Principle

Jaynes 1957

- Suppose we only know that $X \sim P, P \in \mathcal{C}$
- We are asked to make probabilistic predictions/ decisions about X
- According to 'MaxEnt', we should predict using the $\tilde{P} \in \mathcal{C}$ that maximizes entropy under the constraint $\mathcal{C}$:

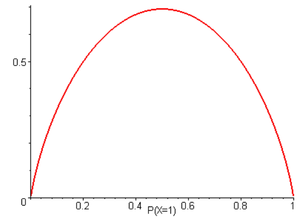
$$\tilde{P} := \arg\max_{P \in \mathcal{C}} \mathbf{H}(P).$$

$$\tilde{P} := \arg\max_{P \in \mathcal{C}} \mathbf{H}(P).$$

Since entropy is concave and $\mathcal{X}$ is finite
$\mathcal{C}$ is closed and convex :

Unique MaxEnt $\tilde{P}$ always exists!

Example 1: if $\mathcal{C} = \mathcal{P}$ then $\tilde{P}$ is uniform



MaxEnt generalizes Laplace's (1812) Principle of Indifference

Example 2: independence

if

$$\begin{aligned} \mathcal{X} &= \mathcal{X}_1 \times \mathcal{X}_2 \\ \mathcal{X}_1 &= \mathcal{X}_2 = \{0,1\} \\ \mathcal{C} &= \{P \; : \; P(X_1 = 1) = p; P(X_2 = 1) = q\} \end{aligned}$$

then

$$\tilde{P}(X_1 = x_1 \mid X_2 = x_2) = \tilde{P}(X_1 = x_1)$$

Rule of thumb: if consistent with constraint, MaxEnt renders variables independent

Example 3: Brandeis Dice
(Jaynes 1962)

$$\mathcal{X} = \{1,2,3,4,5,6\}$$
$$\mathcal{C} = \{P \; : \; E_P[X] = 4.5\}$$
$$\tilde{P}(X = x) = \frac{1}{Z(\beta)} e^{-\beta x}$$
$$Z(\beta) = \sum_{x \in \mathcal{X}} e^{-\beta x}$$
$$\beta = -0.345\ldots$$

Example 3: Brandeis Dice,continued
(Jaynes 1962)

In *practice*, given $X_1, X_2, \ldots, X_n$

Observe *empirical averages of some function(s) of* X :

$$\frac{1}{n}\sum_{i=1}^{n} \phi(X_i) = t$$

in dice case:

$$\frac{1}{n}\sum_{i=1}^{n} X_i = 4.5$$

Motivation

Rule of Thumb: as symmetric, uniform and independent as possible

Prime Motivation: the MaxEnt distribution for a constraint is the least committal, most *inherently uncertain* distribution, making the smallest number of additional assumptions beyond what is known etc.

Does it make any sense?

Philosophers, Probabilists, Statisticians, Physicists and Logicians have been arguing about that for 200 years now! (and still don't agree)

Laplace, Venn, Boltzmann, Keynes, Ehrenfest, Pearson,...

Pros and Contras

PRO

- Axiomatic characterizations
  (Csiszar '89, `only rational inference procedure')
- Concentration Phenomenon
  (Jaynes '78, Sanov property)
- Often quite good results!
  (e.g. Stutzer, econometrics)
- Game-Theoretic Robustness properties
  (Topsøe '79/Dawid & Grünwald now)

Pros and Contras

CONTRA

- *Ex Nihilo Nihil* : Suppose $X \sim P^*$. In general, of course, $P^* \neq \tilde{P}$
  (Ellis, 1842)
- In continuous case, MaxEnt can give arbitrary results
  depends on choice of coordinate system
  Bertrand's Paradox (1900)
- Sometimes very counterintuitive results
  Judy Benjamin problem (Van Fraassen, 1981)

Overview

1. Maximum Entropy (MaxEnt)
2. A Game-Theoretic Characterization of Maximum Entropy
3. Generalized Entropy and Game Theory
4. Pythagoras

Overview

1. Maximum Entropy (MaxEnt)
2. A Game-Theoretic Characterization of Maximum Entropy
   - Some Game/Decision Theory
   - Basic Result
3. Generalized Entropy and Game Theory
4. Pythagoras

Decision Theory

$\mathcal{A}$ Set of Actions/Decisions

$L : \mathcal{X} \times \mathcal{A} \rightarrow \mathbf{R}^+ \cup \{\infty\}$ Loss Function

$L(x, a)$ Loss incurred by Statistician who has decided a when actual outcome is x .

$L(P, a) := E_P[L(X, a)]$ Abbreviation

Logarithmic Loss

$\mathcal{A} = \mathcal{P}$

Here actions are formally same as probability distributions

$L_{\text{lg}}(x, P) := -\ln P(X = x)$ $[= -\ln p(x)]$

Measures how well P fits x

Logarithmic loss is a proper scoring rule, i.e. for all P :

$$P = \arg\min_{Q \in \mathcal{A}} E_P[-\ln Q(X)] = \arg\min_{Q \in \mathcal{A}} L_{\text{lg}}(P, Q)$$

(follows by information inequality)

Basic Result

Information Inequality: $\mathbf{H}(P) = \inf_{Q \in \mathcal{P}} E_P[-\ln Q(X)]$

$$\mathbf{H}(\tilde{P}) = \sup_{P \in \mathcal{C}} \mathbf{H}(P) = \sup_{P \in \mathcal{C}} \inf_{Q \in \mathcal{P}} E_P[-\ln Q(X)]$$
$$= \inf_{Q \in \mathcal{P}} \sup_{P \in \mathcal{C}} E_P[-\ln Q(X)]$$

??? Von Neumann 1928 ???

Basic Result

$$\mathbf{H}(\tilde{P}) = \sup_{P \in \mathcal{C}} \mathbf{H}(P) = \sup_{P \in \mathcal{C}} \inf_{Q \in \mathcal{P}} E_P[-\ln Q(X)]$$
$$= \inf_{Q \in \mathcal{P}} \sup_{P \in \mathcal{C}} E_P[-\ln Q(X)]$$

Grünwald 1998 / Topsøe 1979 !!!

Basic Result, cont.

MaxEnt as a game between Nature and Statistician

MaxEnt $\tilde{P}$ worst-case optimal strategy for Nature:

$$\sup_{P \in \mathcal{C}} \mathbf{H}(P) = \sup_{P \in \mathcal{C}} \inf_{Q \in \mathcal{P}} E_P[-\ln Q(X)]$$

achieved for $P = \tilde{P}$

Basic Result, cont.

MaxEnt as a game between Nature and Statistician

MaxEnt $\tilde{P}$ worst-case optimal strategy for Nature:

$$\sup_{P \in \mathcal{C}} \mathbf{H}(P) = \sup_{P \in \mathcal{C}} \inf_{Q \in \mathcal{P}} E_P[-\ln Q(X)]$$

achieved for $P = \tilde{P}$

MaxEnt $\tilde{P}$ worst-case optimal strategy for Statistician:

surprising!

$$\inf_{Q \in \mathcal{P}} \sup_{P \in \mathcal{C}} E_P[-\ln Q(X)]$$

achieved for $Q = \tilde{P}$

Basic Result, cont.

MaxEnt $\tilde{P}$ worst-case optimal strategy for Statistician:

$$\inf_{Q \in \mathcal{P}} \sup_{P \in \mathcal{C}} E_P[-\ln Q(X)]$$

achieved for $Q = \tilde{P}$

Basic Result, cont.

MaxEnt $\tilde{P}$ worst-case optimal strategy for Statistician:

$$\inf_{Q \in \mathcal{P}} \sup_{P \in \mathcal{C}} E_P[-\ln Q(X)]$$

achieved for $Q = \tilde{P}$

Nature has to satisfy constraint

Statistician can choose anything she likes

Example: Brandeis Dice

Jaynes 1962

$$\mathcal{X} = \{1,2,3,4,5,6\}$$

$$\mathcal{C} = \{P \ : \ E_P[X] = 4.5\}$$

$$\tilde{P}(X = x) = \frac{1}{Z(\beta)} e^{-\beta x}$$

$$Z(\beta) = \sum_{x \in \mathcal{X}} e^{-\beta x}$$

$$\beta = -0.345\ldots$$

Brandeis Dice, cont.

Jaynes 1962

$$\mathcal{C} = \{P \ : \ E_P[X] = 4.5\}$$

$$\tilde{P}(X = x) = \frac{1}{Z(\beta)} e^{-\beta x}$$

$$\begin{aligned} E_P[-\ln \tilde{P}(X)] &= E_P[\beta X + \ln Z(\beta)] = \beta 4.5 + \ln Z(\beta) = \\ &= E_{\tilde{P}}[\beta X + \ln Z(\beta)] = \mathbf{H}(\tilde{P}) = \text{const.} \end{aligned}$$

Hence no matter what P is, as long as it is in $\mathcal{C}$ our average log loss will be **just as large as we expect it to be** (i.e. as if $\tilde{P}$ were `true`) (e.g. $P(X=4) = P(X=5) = \frac{1}{2}$)

Brandeis Dice, cont.

Jaynes 1962

$$\mathcal{C} = \{P \ : \ E_P[X] = 4.5\}$$

$$\tilde{P}(X = x) = \frac{1}{Z(\beta)} e^{-\beta x}$$

$$\begin{aligned} E_P[-\ln \tilde{P}(X)] &= E_P[\beta X + \ln Z(\beta)] = \beta 4.5 + \ln Z(\beta) = \\ &= E_{\tilde{P}}[\beta X + \ln Z(\beta)] = \mathbf{H}(\tilde{P}) = \text{const.} \end{aligned}$$

Hence no matter what $\tilde{P}$ is, as long as it is in $\mathcal{C}$ our average log loss will be **just as large as we expect it to be** (i.e. as if $\tilde{P}$ were `true`) (e.g. $P(X=4) = P(X=5) = \frac{1}{2}$)

$\tilde{P}$ is an **equalizer strategy**

Brandeis Dice, cont.

Jaynes 1962

$$\mathcal{C} = \{P \ : \ E_P[X] = 4.5\}$$

$$\tilde{P}(X = x) = \frac{1}{Z(\beta)} e^{-\beta x}$$

$$\begin{aligned} E_P[-\ln \tilde{P}(X)] &= E_P[\beta X + \ln Z(\beta)] = \beta 4.5 + \ln Z(\beta) = \\ &= E_{\tilde{P}}[\beta X + \ln Z(\beta)] = \mathbf{H}(\tilde{P}) = \text{const.} \end{aligned}$$

On the other hand,

$E_{\tilde{P}}[-\ln Q(X)] > E_{\tilde{P}}[-\ln \tilde{P}(X)] = \mathbf{H}(\tilde{P})$ if $Q \neq \tilde{P}$

Hence if we use any $Q \neq \tilde{P}$ for prediction, Nature can make us suffer by choosing $P = \tilde{P}$

$\tilde{P}$ is **uniquely minimax**

Large Samples: MaxEnt as `**maximum probability** principle'

$$\sup_{P \in \mathcal{C}} E_P[-\ln \tilde{P}(X)] = \mathbf{H}(\tilde{P})$$

$$\sup_{P \in \mathcal{C}} E_P[-\ln Q(X)] = \mathbf{H}(\tilde{P}) + \epsilon$$

Hence for all $P \in \mathcal{C}$

$\tilde{P}(X_1, \ldots, X_n) \approx e^{-n\mathbf{H}(\tilde{P})}$ with P -prob. 1

but for all Q there exists a $P \in \mathcal{C}$ and such that

$Q(X_1, \ldots, X_n) \approx e^{-n(\mathbf{H}(\tilde{P})+\epsilon)}$ with P - prob. 1

and hence $\frac{\tilde{P}(X_1,\ldots,X_n)}{Q(X_1,\ldots,X_n)} \approx e^{n\epsilon}$

Application: Kelly Gambling

- Statistician can buy (arbitrary nr) of tickets for each outcome, at price $1 / ticket
- If actual outcome is x , ticket on x pays $K. Otherwise it pays nothing
- Statistician puts fraction P(x) of her capital on outcome (ticket) x
- Statistician plays game n times; at each round, she reinvests all her capital

Application: Kelly Gambling

- Statistician can buy (arbitrary nr) of tickets for each outcome, at price \$1 / ticket
- If actual outcome is x , ticket on x pays \$K. Otherwise it pays nothing
- Statistician puts fraction P(x) of her capital on outcome (ticket) x
- Statistician plays game n times; at each round, she reinvests all her capital
- Gain after n rounds:

$$G_P^{(n)} = K^n P(x_1) P(x_2) \cdot P(x_n)$$

Application: Kelly Gambling

Sequentially gambling as if data were distributed according to MaxEnt $\tilde{P}$ leads to worst-case optimal expected growth-rate (and hence, for large n, maximal end-capital, with P -probability 1)

Applications: Coding and Gambling

CODING

use (Shannon-Fano) code based on $\tilde{P}$ to encode outcomes. By LLN, with P -probability 1, for large enough sample you minimize the maximum nr of bits needed to encode the sample.

KELLY GAMBLING

when sequentially gambling on outcomes, by hedging your bets according to $\tilde{P}$, you maximize worst-case expected optimal growth rate of your capital (and, by LLN, for large samples, with high P - probability, end capital)

Three Directions ↗

MaxEnt as a `maximum probability principle'
connection to `concentration phenomenon'
Grünwald 2001, Strong Entropy Concentration, Game Theory, Coding and Randomness

Three Directions ↗ →

MaxEnt as a `maximum probability principle'
connection to `concentration phenomenon'
Grunwald 2001, Strong Entropy Concentration, Game Theory, Coding and Randomness

MaxEnt (and exponential families in general) are robust for certain prediction tasks – they may be suitable for some, but unsafe for other tasks (safe statistics)
Grünwald 2000, `Maximum Entropy and the Glasses You are Looking Through'

Three Directions ↗ → ↘

MaxEnt as a `maximum probability principle'
connection to `concentration phenomenon'
Grünwald 2001, Strong Entropy Concentration, Game Theory, Coding and Randomness

MaxEnt (and exponential families in general) are robust for certain prediction tasks – they may be suitable for some, but unsafe for other tasks (safe statistics)
Grünwald 2000, `Maximum Entropy and the Glasses You are Looking Through'

What if we are interested in another loss function???
Dawid & Grünwald

The Clue

Same Story Can Still Be Told!

Overview

1. Maximum Entropy (MaxEnt)
2. A Game-Theoretic Characterization of Maximum Entropy
3. Generalized Entropy and Game Theory
4. Pythagoras

Game/Decision Theory

$\mathcal{A}, \mathcal{X}, \mathcal{C}$ Action Space, Sample Space, Constraint Set

$\mathcal{A}^r$ **Randomized** actions (set of distributions over $\mathcal{A}$)

$L : \mathcal{X} \times \mathcal{A} \to \mathbf{R}^+ \cup \{\infty\}$ Loss Function

$L(P, \mathbf{a}) := E_P E_{\mathbf{a}}[L(X, A)]$

$(\mathcal{C}, \mathcal{A}^r, L)$ Our Game!

Game/Decision Theory

$\mathcal{A}, \mathcal{X}, \mathcal{C}$ Action Space, Sample Space, Constraint Set

$\mathcal{A}^r$ **Randomized** actions (set of distributions over $\mathcal{A}$)

$L : \mathcal{X} \times \mathcal{A} \to \mathbf{R}^+ \cup \{\infty\}$ Loss Function

$L(P, \mathbf{a}) := E_P E_{\mathbf{a}}[L(X, A)]$

$(\mathcal{C}, \mathcal{A}^r, L)$ Our Game!

Statistician's Choice

Nature's Choice

Generalized Entropy

CENTRAL DEFINITION

For (arbitrary) loss function L , the 'L -entropy of P' is defined by

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

De Groot 1962

Generalized Entropy

CENTRAL DEFINITION

For (arbitrary) loss function L , the 'L -entropy of P' is defined by

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

Shannon Entropy is special case:

$$H_{\lg}(P) = \inf_{Q \in \mathcal{A}} L_{\lg}(P, Q) = \inf_{Q \in \mathcal{A}} E_P[-\ln Q(X)]$$

Generalized Entropy

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

always concave

often differentiable

Example: Brier (squared) Loss

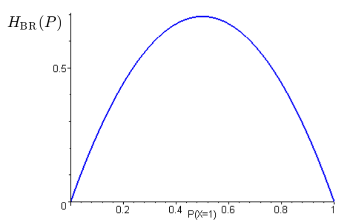
$\mathcal{X} = \{1, \ldots, k\}$
$\mathcal{A} = \mathcal{P}$

$$L_{\mathrm{BR}}(i,P) := ||\vec{e}_i - \vec{p}||^2 =$$
$$(P(1))^2 + \ldots + (P(i-1))^2 + (1-P(i))^2 + (P(i+1))^2 + \ldots + (P(k))^2$$
$$H_{\mathrm{BR}}(P) = \inf_{Q \in \mathcal{A}} L_{\mathrm{BR}}(P,Q) = L_{\mathrm{BR}}(P,P)$$

Brier loss is proper scoring rule

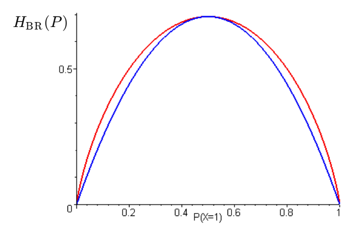
Example: Brier (squared) Loss

$\mathcal{X} = \{0,1\}$ $\quad H_{\mathrm{BR}}(P) = 2P(1)(1-P(1))$



Example: Brier (squared) Loss

$$H_{\mathrm{BR}}(P) = 2P(1)(1-P(1))$$



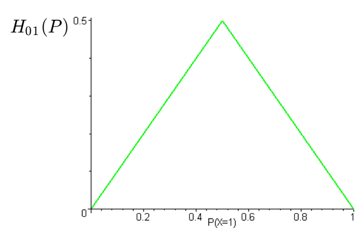
Example: 0/1 - Loss

$\mathcal{X} = \{1, \ldots, k\}$
$\mathcal{A} = \mathcal{X}$
$L_{01}(i,a) = 1$ if $i \neq a$, and 0 if $i = a$

$$H_{01}(P) = \inf_{a \in \mathcal{X}} L_{01}(P,a) =$$
$$= \inf_{a \in \mathcal{X}} P(X \neq a) = 1 - \sup_{x \in \mathcal{X}} P(x)$$

Example: 0/1 - Loss

$\mathcal{X} = \{0,1\}$ $\quad H_{01}(P) = 1 - \sup\{P(0), P(1)\}$



Main Theorem

Assume

- $\mathcal{C}$ convex, tight and closed in weak topology;

AND

- L is bounded from above OR
- $a_P := \arg\inf_{a \in \mathcal{A}} L(P,a)$ is unique for all P
  AND
  $L(Q, a_P)$ is lower semi-continuous as a function of P for all fixed Q

Main Theorem

...then:

$$\underline{V} \;:=\; \sup_{P \in \mathcal{C}} \mathbf{H}_L(P) = \sup_{P \in \mathcal{C}} \inf_{a \in \mathcal{A}} L(P,a)$$

is reached for some $\tilde{P}_L$

$$\overline{V} \;:=\; \inf_{\mathbf{a} \in \mathcal{A}^r} \sup_{P \in \mathcal{C}} L(P, \mathbf{a})$$

is reached for som $\tilde{a}_{\tilde{P}_L}$ **achieving** $\inf_{\mathbf{a} \in \mathcal{A}^r} L(\tilde{P}_L, \mathbf{a})$

$$\underline{V} \;=\; \overline{V}$$ **Game has a value!**

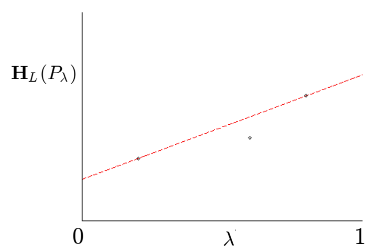
Proof Sketch

$\mathbf{H}_L(P)$ is always concave, i.e. for all $P_0, P_1 \in \mathcal{C}$ we have:

$$\mathbf{H}_L(\lambda P_1 + (1-\lambda)P_0) \geq \lambda \mathbf{H}_L(P_1) + (1-\lambda)\mathbf{H}_L(P_0)$$

Proof of Concavity

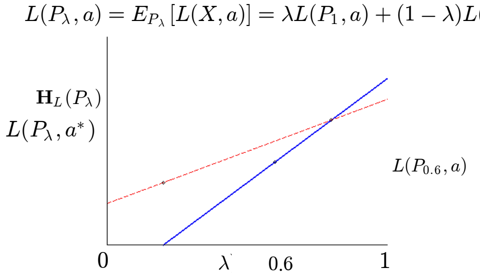
$P_\lambda := \lambda P_1 + (1-\lambda) P_0$



Proof of Concavity

$P_\lambda := \lambda P_1 + (1-\lambda) P_0$

$L(P_\lambda, a) = E_{P_\lambda}[L(X,a)] = \lambda L(P_1, a) + (1-\lambda) L(P_0, a)]$



Under differentiability assumption:

For all $P_0, P_1 \in \mathcal{C}$, $\frac{d}{d\lambda} H_L(P_\lambda)$ exists for all $0 \leq \lambda \leq 1$

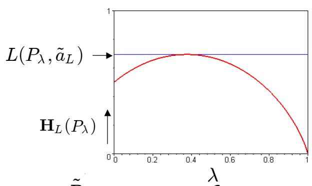
Trivially,

$$\inf_{\mathbf{a} \in \mathcal{A}^r} \sup_{P \in \mathcal{C}} L(P, \mathbf{a}) \geq L(\tilde{P}_L, \tilde{a}_L)$$

We will show that the inequality is an equality.

Under differentiability assumption:

$$L(P_\lambda, a) = E_{P_\lambda}[L(X,a)] = \lambda L(P_1, a) + (1-\lambda) L(P_0, a)$$



If $\tilde{P}_L$ in interior of $\mathcal{C}$ then for all $P \in \mathcal{C}$

$$L(P, \tilde{a}_L) = L(\tilde{P}_L, \tilde{a}_L) = \mathbf{H}_L(\tilde{P}_L)$$

And hence $\inf_{\mathbf{a} \in \mathcal{A}^r} \sup_{P \in \mathcal{C}} L(P, \mathbf{a}) \leq L(\tilde{P}_L, \tilde{a}_L)$

Overview

1. Maximum Entropy (MaxEnt)
2. A Game-Theoretic Characterization of Maximum Entropy
3. Generalized Entropy and Game Theory
4. Pythagoras

Discrepancy (= generalized relative entropy)

For given loss function L, we can define the discrepancy $D_L(P, a)$ by

$$D_L(P, a) = L(P, a) - \inf_{a \in \mathcal{A}} L(P, a)$$

Relative Entropy is special case:

$$\begin{aligned} D(P||Q) &= \sum_x P(x) \ln \frac{P(x)}{Q(x)} \\ &= E_P[-\ln Q(X) - [-\ln P(X)]] \\ &= E_P[-\ln Q(X)] - \inf_{Q' \in \mathcal{P}} E_P[-\ln Q'(X)]. \end{aligned}$$

Example Discrepancy: Brier score

$$L_{\mathrm{BR}}(x, Q) := ||\vec{e}_x - \vec{q}||^2$$

$$L_{\mathrm{BR}}(P, Q) = E_{X \sim P} L_{\mathrm{BR}}(X, Q)$$

$$D_{\mathrm{BR}}(P, Q) = L_{\mathrm{BR}}(P, Q) - \inf_{Q' \in \mathcal{P}} L_{\mathrm{BR}}(P, Q') =$$

$$||\vec{p} - \vec{q}||^2 = \sum_x (P(x) - Q(x))^2$$

- This is just the squared Euclidean distance!

Minimum Relative Entropy Principle

For a given 'prior' distribution Q and constraint $\mathcal{C}$ pick distribution $\tilde{P}$ achieving

$$\inf_{P \in \mathcal{C}} D(P||Q) = \inf_{P \in \mathcal{C}} \sum P(X) \ln \frac{P(X)}{Q(X)}$$

- Interpretation: Q is the member of $\mathcal{C}$ that is closest to $\tilde{P}$, i.e. it is the projection of Q on $\mathcal{C}$

Pythagorean Property

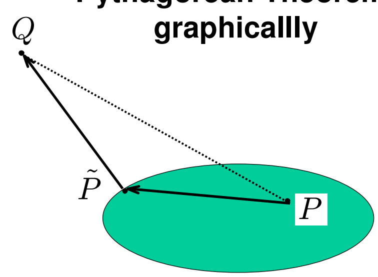
As noted by Csiszár, relative entropy behaves in some ways like squared Euclidean distance: for all priors Q and all $P \in \mathcal{C}$ we have

$$D(P||\tilde{P}) + D(\tilde{P}||Q) \leq D(P||Q)$$

Under some extra conditions we have equality.

Csiszár 1975, 1991, many others

Pythagorean Theorem, graphicallly



$$D(P||\tilde{P}) + D(\tilde{P}||Q) \leq D(P||Q)$$

Relative Games

For every loss function L and reference act e, we can define the relative loss $L_e(X, a)$ by

$$L_e(X, a) := L(X, a) - L(X, e)$$

Main Theorem

Grünwald and Dawid, 2002

For all e , $\mathcal{C}$ such that $D_L(P, e)$ is finite for all $P \in \mathcal{C}$ the game $(\mathcal{C}, \mathcal{A}^r, L_e)$ has a value, i.e.

$$\sup_{P \in \mathcal{C}} \inf_{a \in \mathcal{A}} L_e(P, a) = \inf_{\mathbf{a} \in \mathcal{A}^r} \sup_{P \in \mathcal{C}} L_e(P, \mathbf{a})$$

reached for saddlepoint $(\tilde{P}_L, \tilde{a}_L)$

if and only if, for all $P \in \mathcal{C}$:

$$D_L(P, \tilde{a}_L) + D_L(\tilde{P}_L, e) \leq D_L(P, e)$$

If $\tilde{P}_L$ has full support, then equality holds

Pythagoras = Von Neumann

Who could have guessed?

In words:

The Pythagorean Property holds iff the minimax theorem applies to the loss function under consideration

For example:

minimax theorem holds for squared loss ;
Pythagorean property reduces to high-school Pythagorean theorem

Conclusions/What is this good for?

- Applications in
  - `Robust Bayesian' inference Berger 1985
  - Iterative Scaling (uses Pythagorean property)
- Theoretical Developments:
  - Generalized Exponential Families
  - Generalized Sufficient Statistics (!!!)
  - Generalized Concentration Phenomenon!?

Thank you for your attention!