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

Peter Grünwald
CWI Amsterdam
www.cwi.nl/~pdg

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

CWI is the National Research Institute for Mathematics and Computer Science in the Netherlands.

Overview

1. Maximum Entropy and Game Theory
2. Maximum Generalized Entropy
3. M.G.E. and Robust Bayes (Result I)
4. Minimum Discrepancy
5. Pythagoras (Result II)
6. Conclusion

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 $\mathcal{P}$
$\mathbf{H}$	Shannon (for now) 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 P, $P \in \mathcal{C}$

We are asked to make probabilistic predictions/ decisions about X

According to 'MaxEnt', we should predict using the unique $\tilde{P} \in \mathcal{C}$ that maximizes entropy under the constraint $\mathcal{C}$:

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

Does it make any sense?

- MaxEnt applied in speech recognition, computer vision, stock market prediction…
- …but not always clear why it would be a good idea to use it!
- Various rationales and criticisms have been given over time
- Topsøe (1979) offers a game-theoretic interpretation/rationale

Basic Result

Information Inequality:

If $P \neq Q$ then
$E_P[-\ln Q(X)] > E_P[-\ln P(X)]$
therefore we can write
$\mathbf{H}(P) = \inf_{Q \in \mathcal{P}} E_P[-\ln Q(X)]$

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

Basic Result

Information Inequality:

If $P \neq Q$ then
$E_P[-\ln Q(X)] > E_P[-\ln P(X)]$
so that
$\mathbf{H}(P) = \inf_{Q \in \mathcal{P}} E_P[-\ln Q(X)]$

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

??? Von Neumann 1928 ???

Basic Result

Information Inequality:

If $P \neq Q$ then
$E_P[-\ln Q(X)] > E_P[-\ln P(X)]$
so that
$\mathbf{H}(P) = \inf_{Q \in \mathcal{P}} E_P[-\ln Q(X)]$

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

Topsøe 1979! (in great generality)

Basic Result, cont.

MaxEnt as a game between Nature and Statistician
with loss measured by 'log loss' $L(x, Q) := -\ln Q(x)$

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

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

achieved for $P = \tilde{P}$

Basic Result, part II

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

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

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, part II

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, part II

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

Basic Result, part II

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 $Q = \tilde{P}$

- This justifies the Maximum Entropy Principle when the 'log loss' is the proper loss function to use:
  - **Coding, Kelly Gambling**

The Clue

…but what if we are interested in another loss function?

Similar Story Can Still Be Told!

Overview

1. Maximum Entropy and Game Theory
2. Maximum Generalized Entropy
3. M.G.E. and Robust Bayes (Result I)
4. Minimum Discrepancy
5. Pythagoras (Result II)
6. Conclusion

Game/Decision Theory

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

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

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

Example: Logarithmic Loss

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

Here actions are formally same as probability distributions

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

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

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

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
Rao 1982

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:

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

Generalized Entropy

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

always concave

(infimum of linear functions)

often differentiable

Example: Brier (squared) Loss

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

$L_{\mathrm{BR}}(i, P) := ||\vec{e}_i - \vec{p}||^2 =$

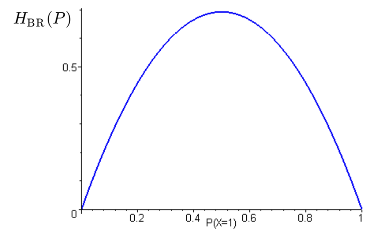
$(P(1))^2 + \ldots + (P(i-1))^2 + (1 - P(i))^2 + (P(i+1))^2 + \ldots + (P(k))^2$

$H_{\mathrm{BR}}(P) = \inf_{Q \in \mathcal{A}} L_{\mathrm{BR}}(P, Q) = L_{\mathrm{BR}}(P, P)$

Brier loss is proper scoring rule

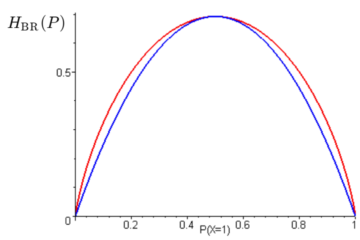
Example: Brier (squared) Loss

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



Example: Brier (squared) Loss

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



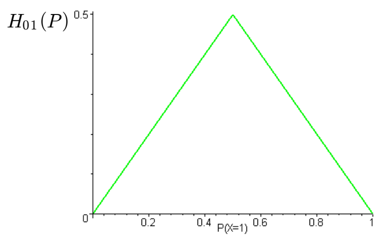
Example: 0/1 - Loss

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

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



Overview

1. Maximum Entropy and Game Theory
2. Maximum Generalized Entropy
3. M.G.E. and Robust Bayes (Result I)
4. Minimum Discrepancy
5. Pythagoras (Result II)
6. Conclusion

Main Theorem (baby version)

Assume

- $\mathcal{X}$ finite
- $\mathcal{C}$ convex and closed
- $\mathcal{A}$ closed
- L bounded from above

Main Theorem (baby version)

...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 some $\tilde{a}_{\tilde{P}_L}$ **achieving** $\inf_{\mathbf{a} \in \mathcal{A}^r} L(\tilde{P}_L, \mathbf{a})$

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

But what is new here?

- Mathematically:
  - Nothing new in baby version
  - In paper we present an adult version
    - General sample spaces, unbounded loss functions, non-compact sets of constraints…
    - New proof technique
- Conceptually:

  'maximum generalized entropy is robust Bayes'
    - New view leads to new math results later

Overview

1. Maximum Entropy and Game Theory
2. Maximum Generalized Entropy
3. M.G.E. and Robust Bayes (Result I)
4. Minimum Discrepancy
5. Pythagoras (Result II)
6. Conclusion

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

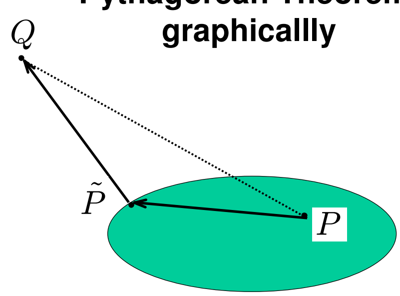
As noted by Csiszár, relative entropy behaves in some ways like squared Euclidean distance: for all priors Q and all $P \in \mathcal{C}$ we have

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

Under some extra conditions we have equality.

Csiszár 1975, 1991, many others

Pythagorean Theorem, graphicallly



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

Overview

1. Maximum Entropy and Game Theory
2. Maximum Generalized Entropy
3. M.G.E. and Robust Bayes (Result I)
4. Minimum Discrepancy
5. Pythagoras (Result II)
6. Further Developments

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 $P \in \mathcal{C}$:

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

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

Conclusion

- We have shown:
  - Maximum (Gen.) Entropy = Robust Bayes
  - Pythagoras = Von Neumann
- Three further results in full paper:
  - Relation to Bregman divergences
  - Generalized Exponential Families
  - Generalized Redundancy-Capacity (Gallagher-Ryabko-Haussler) Theorem

Thank you for your attention!

How general is Pythagorean property?

- Both squared Euclidean distance and relative entropy are examples of **Bregman divergences**
- Pythagoras known to hold for such divergences