Elicitation and Machine Learning

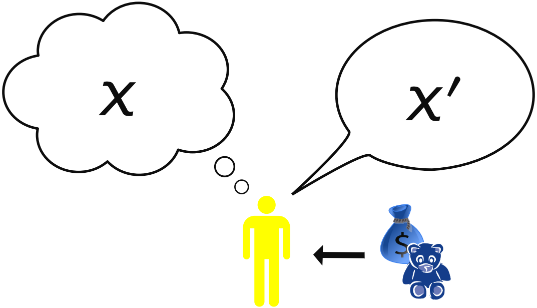
a tutorial at EC 2016

Part I

Rafael Frongillo and Bo Waggoner

25 July 2016

Information Elicitation



Contracts to exchange information for money/goods.

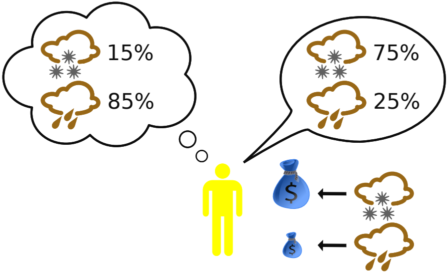
Scoring rules, peer prediction, prediction markets, mechanism design, ...

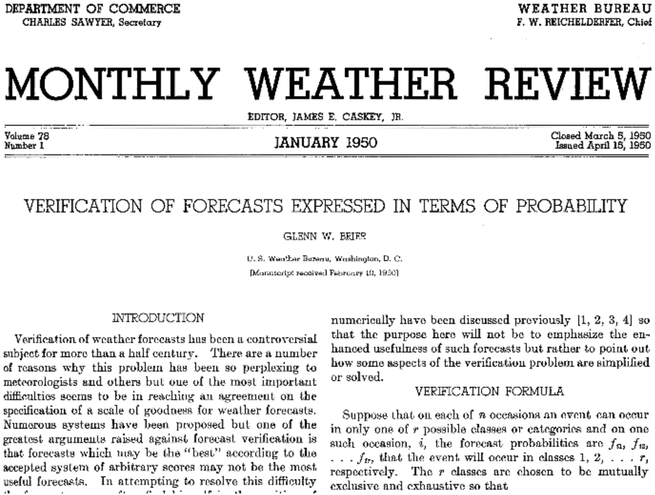
A Challenge

Want to know chance of rain tomorrow.

Approach meteorologist Bob, need a contract:
 If he says p: paid $S(p,1)$ if rain & $S(p,0)$ if snow.

How should you choose S?





DEPARTMENT OF COMMERCE
CHARLES SAWYER, Secretary

WEATHER BUREAU
F. W. REICHELDERFER, Chief

MONTHLY WEATHER REVIEW

EDITOR, JAMES E. CASKEY, JR.

Volume 78
Number 1

JANUARY 1950

Closed March 5, 1950
Issued April 15, 1950

VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY

GLENN W. BRIER

U. S. Weather Bureau, Washington, D. C.

[Manuscript received February 10, 1950]

INTRODUCTION

Verification of weather forecasts has been a controversial subject for more than a half century. There are a number of reasons why this problem has been so perplexing to meteorologists and others but one of the most important difficulties seems to be in reaching an agreement on the specification of a scale of goodness for weather forecasts. Numerous systems have been proposed but one of the greatest arguments raised against forecast verification is that forecasts which may be the "best" according to the accepted system of arbitrary scores may not be the most useful forecasts. In attempting to resolve this difficulty

numerically have been discussed previously [1, 2, 3, 4] so that the purpose here will not be to emphasize the enhanced usefulness of such forecasts but rather to point out how some aspects of the verification problem are simplified or solved.

VERIFICATION FORMULA

Suppose that on each of n occasions an event can occur in only one of r possible classes or categories and on one such occasion, i, the forecast probabilities are f_{i1}, f_{i2}, . . . f_{ir}, that the event will occur in classes 1, 2, . . . r, respectively. The r classes are chosen to be mutually exclusive and exhaustive so that

$$S(p, y) = 2py - p^2$$

Questions Raised

- Generalization to > 2 outcomes? *Yep.*
- Are there other truthful functions?

  Here's one: $S(p, y) = \log p(y)$
- Statistics or *properties*?

  Avg inches of rain? Variance?
- Multiple-choice questions?
- A hypothesis about future data points?

Also: how does all this relate to **mechanism design**?

Goals for you

Part I: proper scoring rules + mechanism design

- **Role of convexity**
- MD applications
- Set up Part II

Part II: property elicitation + machine learning

- **Role of convexity**
- Known results / frontiers
- ML applications

Outline

Part I

1. Convex analysis primer
2. Scoring rule characterization
3. Common AGT applications

   break

4. Mechanism design and general truthfulness
5. Truthful multiple-choice quizes

Part II

1. Eliciting properties: advanced
2. Connections to machine learning

Historical Aside

- 1931: de Finetti

  “Probability does not exist”

  Subjective probability as defined by fair price of a gamble, like a prediction market.

  I.e., define probability via “scoring rule”

- 1660: Pascal and Fermat

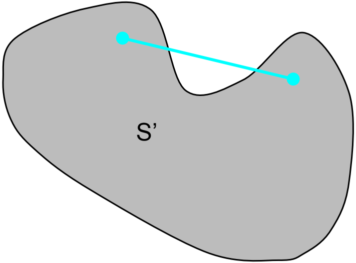
  Foundation of probability via fair splitting of tournament prizes.

Takeaway: link between scoring rules and subjective probability is as old as probability itself.

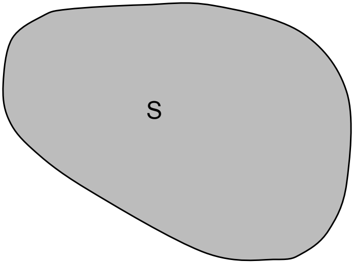
I.1. Convex Analysis Primer

Convex Set

Def. $S \subseteq \mathbb{R}^n$ is *convex* if, for any $a, b \in S$, the line segment between a and b lies in S.

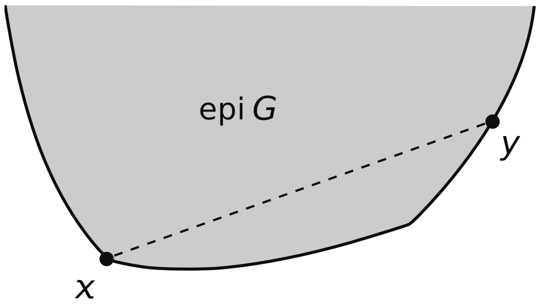


Not convex



Convex

Convex Function

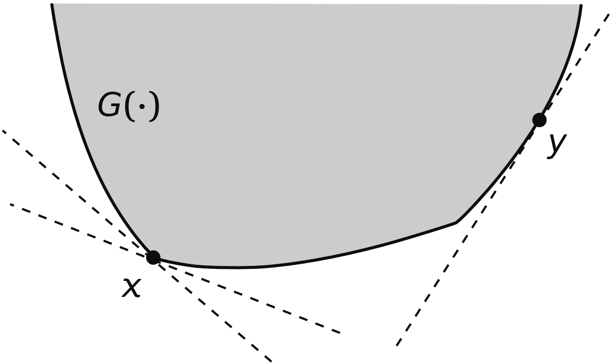


Def. $G : \mathbb{R}^n \to \mathbb{R}$ is *convex* if its epigraph is convex

equivalently: if for all $x, y \in \mathbb{R}^n$ and all $\alpha \in [0, 1]$

$$\alpha G(x) + (1 - \alpha) G(y) \geq G(\alpha x + (1 - \alpha) y)$$

Subgradient



Def. A vector $dG_x \in \mathbb{R}^n$ is a *subgradient* to G at x if

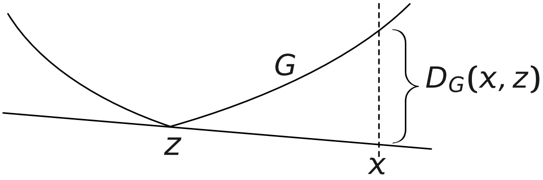
$$\forall z \in \mathbb{R}^n \quad G(z) \geq G(x) + dG_x \cdot (z - x)$$

Bregman Divergence

Def. The *Bregman divergence* of a convex G is

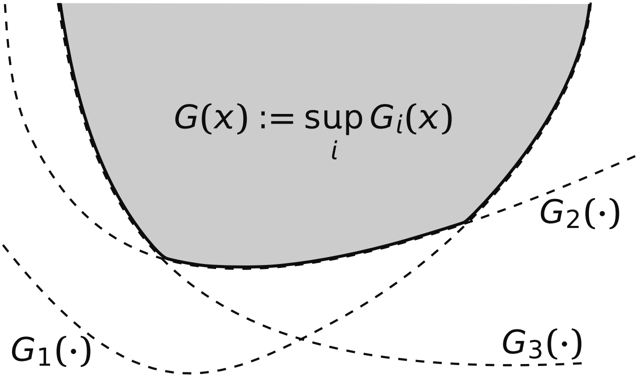
$$D_G(x,z) = G(x) - \left[G(z) + dG_z \cdot (x - z) \right]$$

"difference between $G(x)$ and its linear approx from z".



$D_G(x,z) \geq 0$ by definition of subgradient (or by picture).

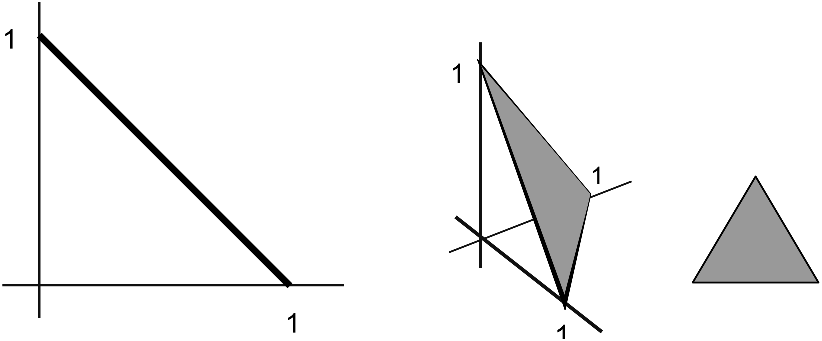
Pointwise Supremum



Prop. A pointwise supremum of convex functions is convex

The simplex

$\Delta_{\mathcal{Y}} =$ set of probability distributions on $\mathcal{Y}$.



Note: will often draw the line or triangle; keep in mind that really they live in 2d and 3d.

I.2. Scoring Rule Characterization

Scoring rule foundations

Goal 1: Understand geometry of proper scoring rules.

Goal 2: Characterize the full set of proper scoring rules.

Goal 3: Develop tools for constructing them.

Scoring Rules

- *Outcome space* $\mathcal{Y}$ *e.g. weather*
- *Private belief* $p \in \mathcal{P}$ *set of distributions*
- *Scoring rule* $S : \mathcal{P} \times \mathcal{Y} \to \mathbb{R}$
  $S(p, y) =$ "score for report p when outcome is y"

Let $$S(p; q) := \mathop{\mathbb{E}}_{q} [S(p, Y)]$$

"expected score for report p with belief q"

Properness

S is *proper* if for all p, q,

$$S(q;q) \geq S(p;q).$$

S is *strictly proper* if for all $p \neq q$,

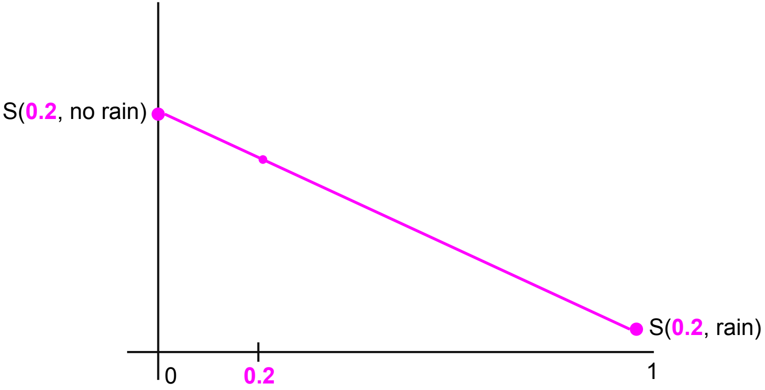
$$S(q;q) > S(p;q).$$

Why consider strict properness?

- Otherwise, use rule $S(p,y) = 1$.
- Can allow for costs of reporting: $S(\cdot,\cdot)$ is (strictly) proper $\iff$ $aS(\cdot,\cdot) + b$ is.

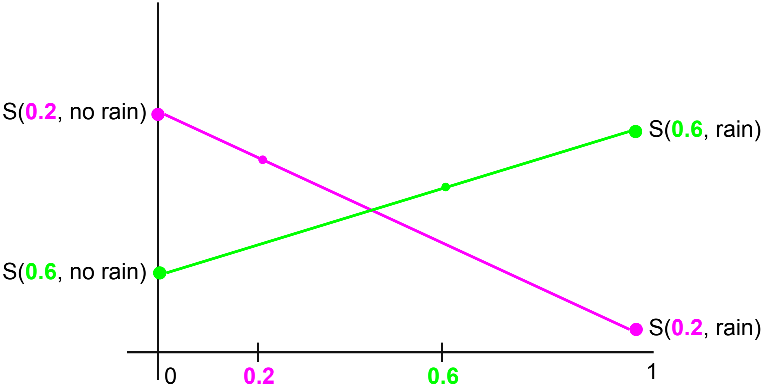
Geometry of scoring rules

Question: What does $S(p;q)$ look like *as a function of* q?
Linear! $S(p;q) = S(p,\cdot) \cdot q$.



Geometry of scoring rules

Question: What does $S(p;q)$ look like *as a function of q*?
Linear! $S(p;q) = S(p,\cdot) \cdot q$.



Geometry continued

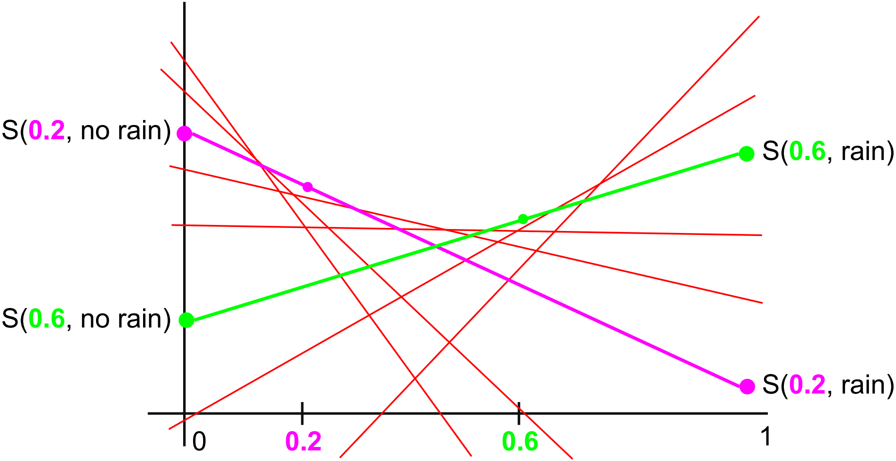
What can we say about the *expected truthful score function*

$$G(q) = S(q;q)?$$

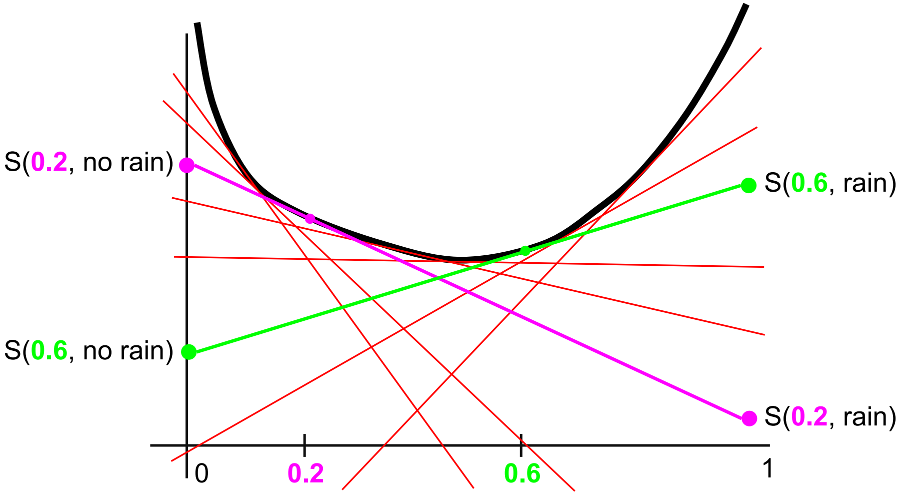
Properness: $G(q) = \sup_p S(p;q)$.

G is a pointwise maximum of linear functions
$\implies$ G is convex!

Picturing *G*



Picturing *G*



How does S relate to G?

$S(p;\cdot)$ is:

- a linear function,
- equal to G at p,
- everywhere below G.

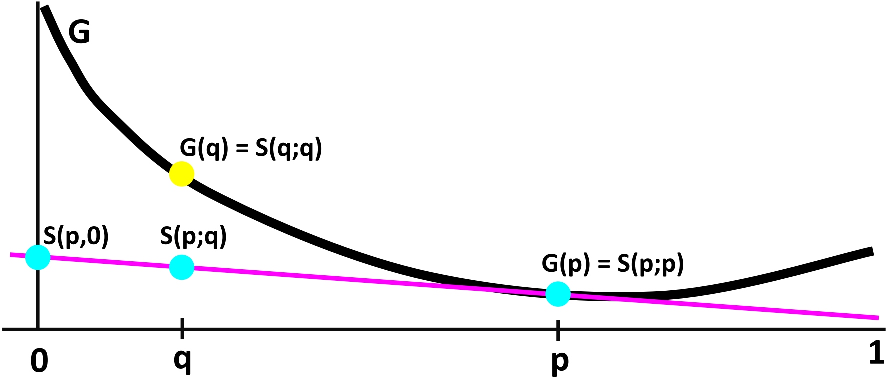
In other words, a linear approximation to G at p.

Hence, $S(p;\cdot)$ can be written

$$S(p;q) = G(p) + dG_p \cdot (q - p).$$

where dG_p is a subgradient of G at p.

S is a linear approximation to *G*



Theorem (Scoring Rule Characterization)

A scoring rule S is (strictly) proper ***if and only if*** *there exists a (strictly) convex G with*

$$S(p,y) = G(p) + dG_p \cdot (\mathbb{1}_y - p).$$

$\mathbb{1}_y$ is the distribution putting probability 1 on y.
Note $S(p;\mathbb{1}_y) = S(p,y)$.

[McCarthy 1956, Savage 1971, Gneiting and Raftery 2007]

Proof

(→) done! (check: strictness)

(←) Given $G(p)$, let $S(p,y) = G(p) + dG_p \cdot \left(\mathbb{1}_y - p\right).$

Note that $S(p;q) = G(p) + dG_p \cdot (q - p)$.

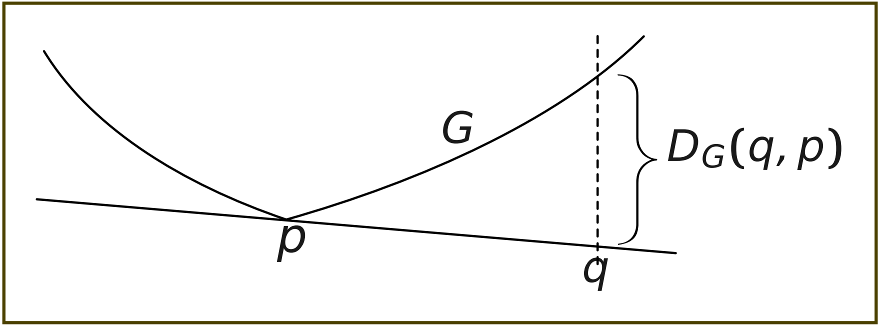
Proper:

$$\begin{aligned} S(q;q) - S(p;q) &= G(q) - \Big[G(p) + dG_p \cdot (p - q) \Big] \\ &= D_G(q,p) \\ &\geq 0 \end{aligned}$$

by the nonnegativity property of Bregman divergence.

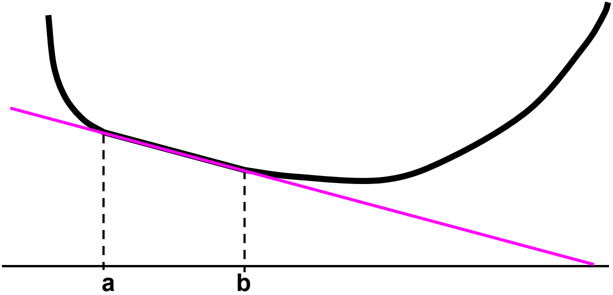
Proof by picture

$D_G(q,p) =$ “with belief q, difference in expected utility between truthfulness and misreporting p”



Picturing Strictness

What's happening to this scoring rule between a and b?



All reports in this region map to the same scores!
→ Agent is indifferent between any reports in $[a, b]$
→ $D_G(p, q) = 0$ for all $p, q \in [a, b]$.

Discussion

Previous goals for scoring rules:

1) Geometry: $S(p;\cdot)$ is a linear approximation to $G(p)$.

2) Characterize full set: exact correspondence with set of convex functions.

3) Tools for constructing: build a convex function with the desired shape.

- *e.g.* incentive not to deviate is Bregman divergence.
- linear segments represent **indifference regions** (useful later).

I.3. Scoring Rule Applications in AGT

AGT applications

1. Peer prediction [Miller, Resnick, Zeckhauser 2005].

Suppose two agents receive private, correlated signals s_i, s_j from a known prior distribution.
How can we incentivize both agents to truthfully report without knowing any ground truth?

Ask i for signal, compute posterior probability distribution q_i over signals of j. Pay $S(p_i, s_j)$.
Meanwhile for j, pay $S(p_j, s_i)$.

AGT applications continued

2. Complexity theory [Azar, Micali 2012].

What is the complexity of solving problems given access to an infinitely powerful, but rational, self-interested oracle?

e.g., #P problems can be solved in one round of interaction (how?).

Hint: a #P-complete problem is “what fraction of inputs to this circuit produce output 1?”

AGT applications continued

3. Prediction markets [Hanson 2003, . . .].

How to generalize a proper scoring rule to both **elicit** and **aggregate** beliefs of multiple agents?

1. Market designer chooses initial prediction $p^{(0)}$.
2. Agent 1 arrives, updates market prediction to $p^{(1)}$.

$$\vdots$$

Agent i arrives, updates market prediction to $p^{(i)}$.

$$\vdots$$

3. Outcome y is observed.
4. Each agent i is paid $S(p^{(i)}, y) - S(p^{(i-1)}, y)$.

AGT applications continued

4. Mechanism design.

Many opportunities, esp. involving both **valuations** and **beliefs**.

Example: Bidders in an ad auction know their own click-through rates.
Truthfully elicit CTR (**prediction** of click) and **valuation**, run an auction to assign ad slots.

Outline

Part I

1. Convex analysis primer
2. Scoring rule characterization
3. Common AGT applications

   break ⇐
4. Mechanism design and general truthfulness ⇐
5. Truthful multiple-choice quizes

Part II

1. Eliciting properties: advanced
2. Connections to machine learning

I.4. Mechanism Design and General Truthfulness

Mechanism Design

- Outcome space $\mathcal{Y}$ *possible allocations*
- Private type $t : \mathcal{Y} \rightarrow \mathbb{R} \;\; \in \mathcal{T}$ *valuation ftn*
- Allocation rule $f : \mathcal{T} \rightarrow \Delta_{\mathcal{Y}}$
- Payment rule $\pi : \mathcal{T} \rightarrow \mathbb{R}$

Let: $U(t', t) = \mathop{\mathbb{E}}_{y \sim f(t')} \left[\, t(y) \, \right] - \pi(t')$

Truthfulness condition

$$\forall t, t' \in \mathcal{T} \quad U(t', t) \leq U(t, t)$$

Scoring Rule ≈ Mechanism?

$p \in \Delta_{\mathcal{Y}}$	Private type	$t \in \mathcal{T} \subseteq \mathbb{R}^{\mathcal{Y}}$
$S(p', p) \leq S(p, p)$	Truthfulness	$U(t', t) \leq U(t, t)$
$\mathbb{E}_{y \sim p}\left[S(p', y)\right]$	Utility?	$\mathbb{E}_{y \sim f(t')}\left[t(y)\right] - \pi$
$=$		$=$
$\left\langle S(p', \cdot), \boxed{\text{p}} \right\rangle$	Affine!	$\left\langle f(t'), \boxed{\text{t}} \right\rangle - \pi$

Generalized Model: Affine Score

1. Type space $\mathcal{T}$ *$\subseteq$ vector space*
2. Utility/score $S(t', t)$ is *affine* in t
3. Truthfulness: $\forall t, t' \in \mathcal{T} \;\; S(t', t) \leq S(t, t)$

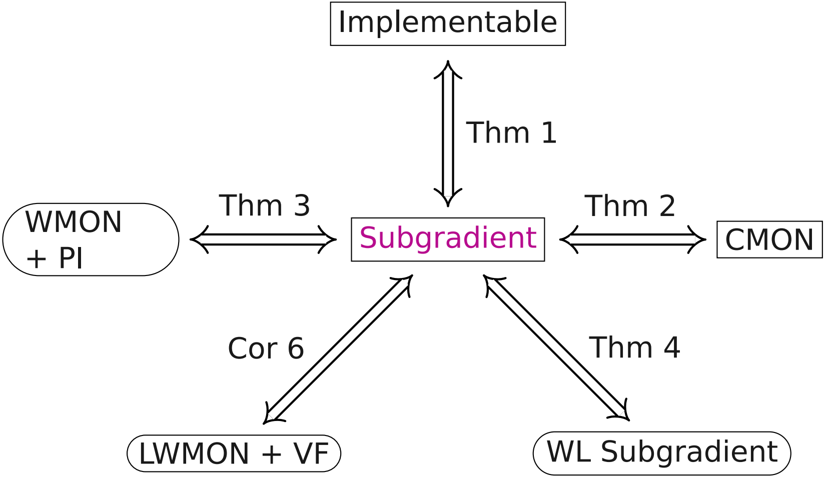
Characterization Theorem [F & Kash 2014]

Char. affine scores in terms of convexity / subgradients.

- Scoring rules: recovers characterization
- Mechanism design: implementability, revenue equiv

  subgradient = allocation
- Combinations of the two!

Implementability Conditions



Other Application Domains

- Decision scoring rules

  Othman & Sandholm 2010, Chen & Kash 2011

- Proper losses for partial labels

  Cid-Suero 2012

- Responsive lotteries

  Feige & Tennenholtz 2010

- Mechanisms with partial allocation

  Cai, Mahdian, Mehta, Waggoner 2013

- Crowdsourcing mechanisms for data labeling

  Shah and Zhou 2015

Ex: Decision scoring rules

- Principal must choose some action $\alpha \in \{1, \ldots, m\}$
- After choosing the action, she will face some outcome $y \in \mathcal{Y}$
- Wants to know the conditional distributions $\Pr(y|\alpha)$
- Decides to ask an agent...

- Agent reports values $P_{\alpha,y} = \Pr(y|\alpha)$ for all α, y
- Principal then chooses action distribution $d_P \in \Delta_m$
- Agent paid $S(P, \alpha, y)$ upon action α and outcome y

Ex: Decision scoring rules

Expected score of report P' is

$$\mathop{\mathbb{E}}_{a\sim d_{P'}} \mathop{\mathbb{E}}_{y\sim P_{a,*}} S(r,a,y) = \overset{\textit{type}}{\begin{bmatrix} P_{1,*} \\ P_{2,*} \\ \vdots \\ P_{m,*} \end{bmatrix}} \cdot \begin{bmatrix} d_{P'}(1)\, S(P',1,\cdot) \\ d_{P'}(2)\, S(P',2,\cdot) \\ \vdots \\ d_{P'}(m)\, S(P',m,\cdot) \end{bmatrix}$$

Affine (here linear) in type!

Recover previous characterization.

Same type for crowdsourcing example.

Duality

Back to mechanisms and scoring rules...

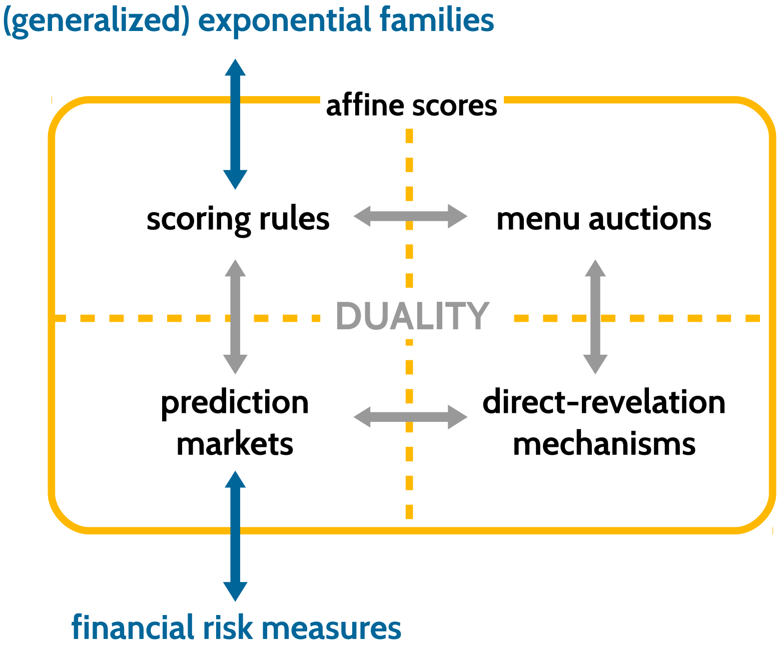
$$\begin{array}{ll} \text{Mechanism:} & f : \mathbb{R}^{\mathcal{Y}} \to \Delta_{\mathcal{Y}} \\ \text{Scoring rule:} & S : \Delta_{\mathcal{Y}} \to \mathbb{R}^{\mathcal{Y}} \end{array}$$

Swap type and allocation!

Via convex conjugate duality, connection goes deeper:

Theorem

A mechanism's consumer surplus function is conjugate to its price function, which is the G of a scoring rule.



(generalized) exponential families

affine scores

scoring rules

menu auctions

DUALITY

prediction markets

direct-revelation mechanisms

financial risk measures

I.5. Eliciting Answers to Multiple-Choice Questions

AKA “finite properties”

Part II (after lunch): more general properties

Asking a Multiple Choice Question

Tomorrow, will it:

- Rain?
- Not rain, but be sunny?
- Not rain, but be cloudy?

Question: What is “truthfulness”?

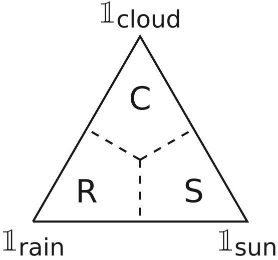
Answer: Explicitly specify the truthful answer for each belief.

Truthfulness example

Let $\Gamma : \Delta_{\{\text{rain,sun,cloud}\}} \to \{\text{rain,sun,cloud}\}$.

$\Gamma(p) =$ *the "truthful" report for belief p.*

Probability simplex on 3 outcomes: C = set of distributions that should report "clouds", etc


$\mathbb{1}_{\text{cloud}}$
C
R
S
$\mathbb{1}_{\text{rain}}$
$\mathbb{1}_{\text{sun}}$

"If you believe __ is most likely, report __."

Some Definitions

- *outcome space* $\mathcal{Y}$ *same as before*
- *report space* $\mathcal{R}$ *a finite set*
- *property* $\Gamma : \Delta_{\mathcal{Y}} \to \mathcal{R}$
- *(generalized) scoring rule* $S : \mathcal{R} \times \mathcal{Y} \to \mathbb{R}$

  $S(r, y) =$ *"score for report r when outcome is y"*

S *elicits* Γ if

$$\Gamma(p) = \arg\max_{r \in \mathcal{R}} \mathbb{E}_{p} S(r, Y).$$

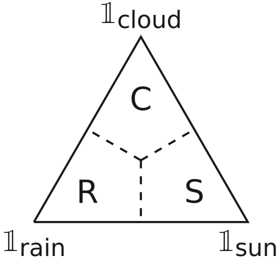
If there exists such an S, then Γ is *elicitable.*

Here Y is distributed according to p.

Key Concept: Level Sets

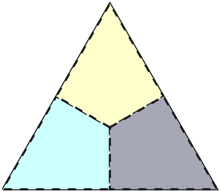
A *level set* is a set of distributions that have the same "right answer", *i.e.* the level set of r is $\{p : \Gamma(p) = r\}$.

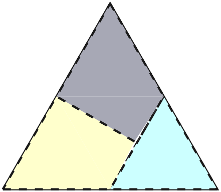
Observe: For elicitation, suffices to consider level sets. (The label of each level set is irrelevant.)

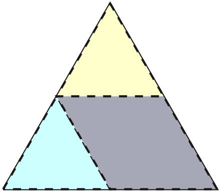

$\mathbb{1}_{\text{cloud}}$
C
R
S
$\mathbb{1}_{\text{rain}}$
$\mathbb{1}_{\text{sun}}$

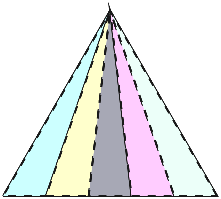
Example properties

Which do you think are elicitable?





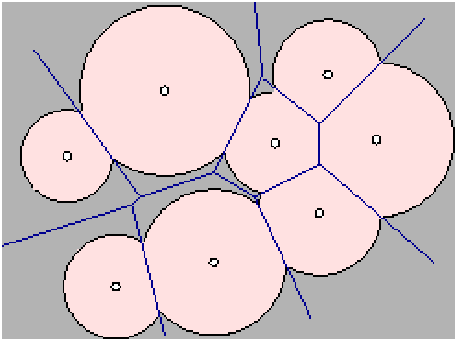




Theorem (Finite prop. characterization)

Γ *is elicitable* $\iff$ Γ *is a* ***power diagram****.*

Power diagram = weighted Voronoi diagram.



[[Lambert et al. 2008], [F & Kash 2014]]

Key Intuition

For incentives, always think about **expected utility as a function of type** (here, belief p).

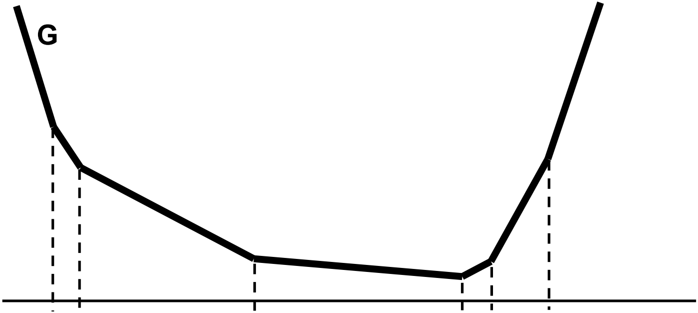
The expected score for truthfully reporting given belief p is some convex function:

$$G(p) = \mathop{\mathbb{E}}_{p} S\left(\Gamma(p), Y\right).$$

Note: If we constructed a proper scoring rule S from G, agents would be indifferent between beliefs in the same level set.
$\implies$ G is **flat** (matches a hyperplane) on the level set.

Picture for binary outcome

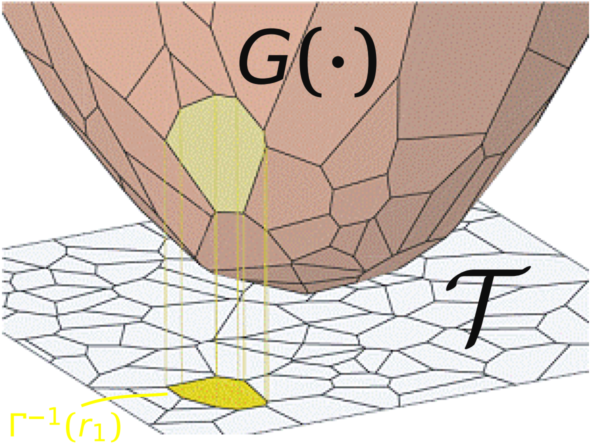
When $\mathcal{Y} = \{0, 1\}$:



Can elicit any finite collection of intervals!

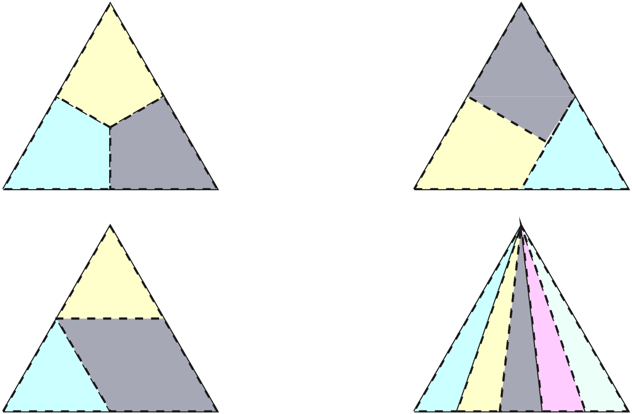
Proof by picture(s)

Claim: Γ is elicitable $\iff$ it is a *power diagram*:



[Image credit: Pooran Memari]

Revisiting property examples



Can elicit all but top right. Try to see this by picturing the G constructed from hyperplanes for each picture.

Example Applications

Mechanism design: Saks, Yu (2005) showed that an allocation rule over a **finite** set of allocations is truthful if it satisfies “weak monotonicity”.

Simplified proof using finite property characterization.

Peer prediction: Characterizations and constructions of minimal peer prediction mechanisms by viewing them as finite properties. [[F & Witkowski]]

End of Part I.

Don’t miss the Nobel Lecture!
Come back for Part II @ 13:00!