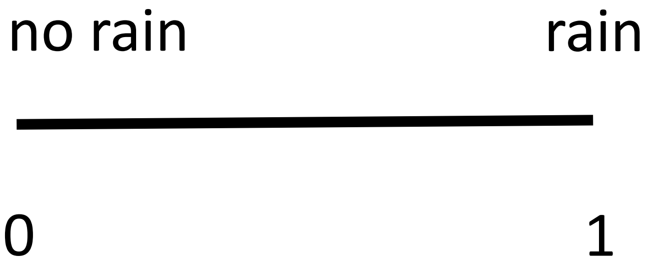
Proper Scoring Rules

conitzer@cs.duke.edu

Probability forecasts

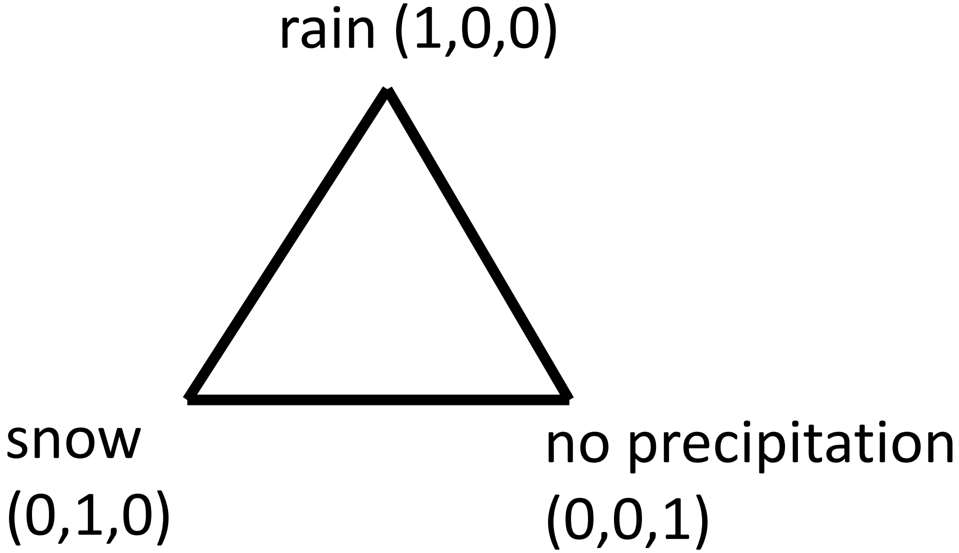


no rain

rain

0

1



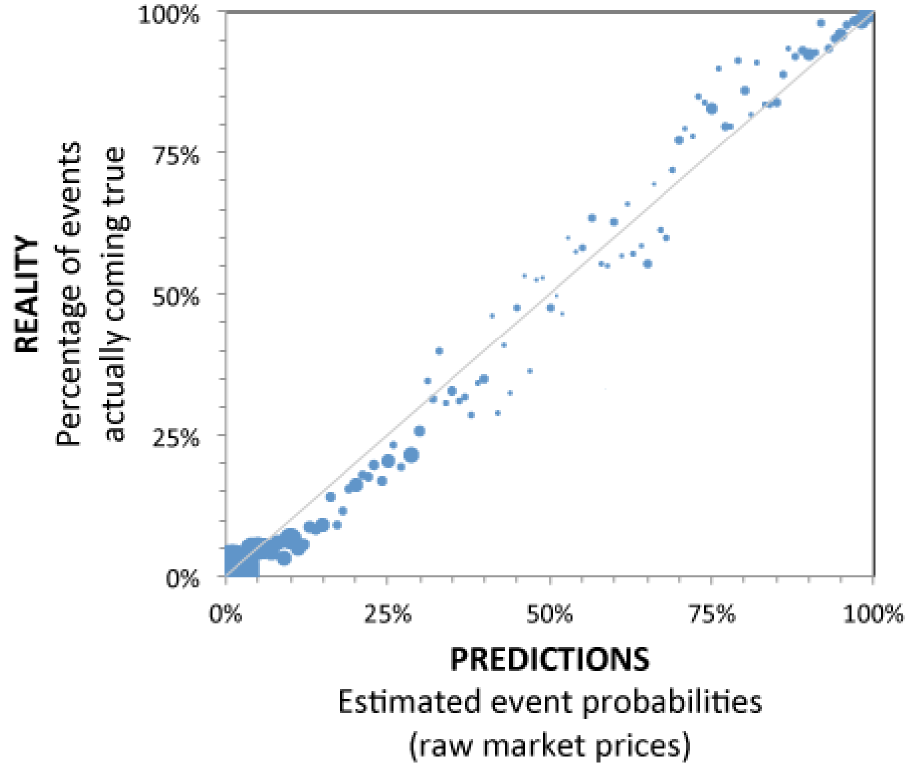
rain (1,0,0)

snow (0,1,0)

no precipitation (0,0,1)

What makes a probability forecaster good?

- **Calibration:** in the long run, of all the times you forecasted *x*%, roughly *x*% should turn out “yes” (for all *x*)
- What’s an easy way to be calibrated?
- **Sharpness:** more extreme forecasts are preferred
- How to trade these off?



“Hypermind forecast calibration over 2 years on 181 question[s] and 472 possible event outcomes. Every day at noon, the estimated probability of each outcome was recorded. Once all the questions are settled, we can compare, at each level of probability, the percentage of events predicted to occur and the percentage that actually occurred. The size of data points indicates the number of forecasts recorded at each level of probability.”
(https://blog.hypermind.com/2016/06/25/lessons-from-brexit/)

Definitions

- Let $S(\mathbf{p}, \omega)$ denote the reward for outcome ω after reporting $\mathbf{p}$
  - If the outcomes are $\{0, 1\}$, just report $p = p_1$
- S is proper if for all $\mathbf{p}$, $\mathbf{p}$ is in $\arg\max_{\mathbf{p}'} E_{\omega\sim\mathbf{p}}\, S(\mathbf{p}', \omega)$
- S is strictly proper if for all $\mathbf{p}$, $\{\mathbf{p}\} = \arg\max_{\mathbf{p}'} E_{\omega\sim\mathbf{p}}\, S(\mathbf{p}', \omega)$

A scoring rule

- Let $S(\mathbf{p}, \omega) = p_{\omega}$.
- Is this proper?

Some example scoring rules (proofs that they are strictly proper on the board)

- Quadratic: $S(p, 1) = 2p - p^2 - (1-p)^2$, $s(p, 0) = 2(1-p) - p^2 - (1-p)^2$
  - Generally: $S(\mathbf{p}, \omega) = 2p_\omega - \Sigma_{\omega'} (p_{\omega'})^2$
- Logarithmic: $S(p, 1) = \ln(p)$, $s(p, 0) = \ln(1-p)$
  - Generally: $S(\mathbf{p}, \omega) = \ln(p_\omega)$
- Vince's crazy proper scoring rule:

  $S(p, 1) = (2-p)e^p$

  $S(p, 0) = (1-p)e^p$

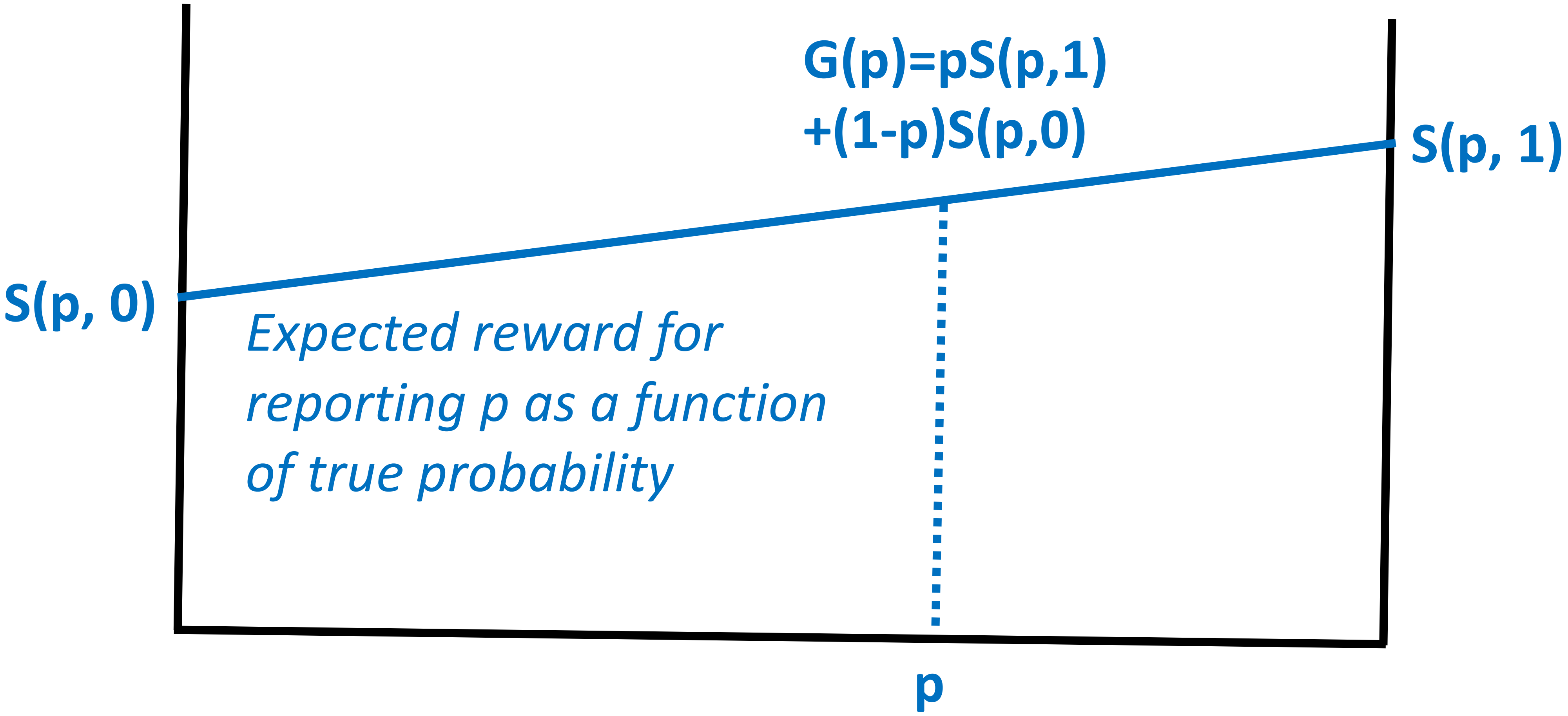
  (Do you want to make your own?)
- What's nice / not nice about some of these rules?

Can we go beyond individual examples?

- Can we come up with a way of generating more proper scoring rules?
- How would we ever know we have found the optimal proper scoring rule (according to some metric)?
- If we have constraints, how do we know that a proper scoring rule exists that satisfies them?

Expected value of reporting truthfully

- Let G(p) denote your expected reward if you believe **and** report p

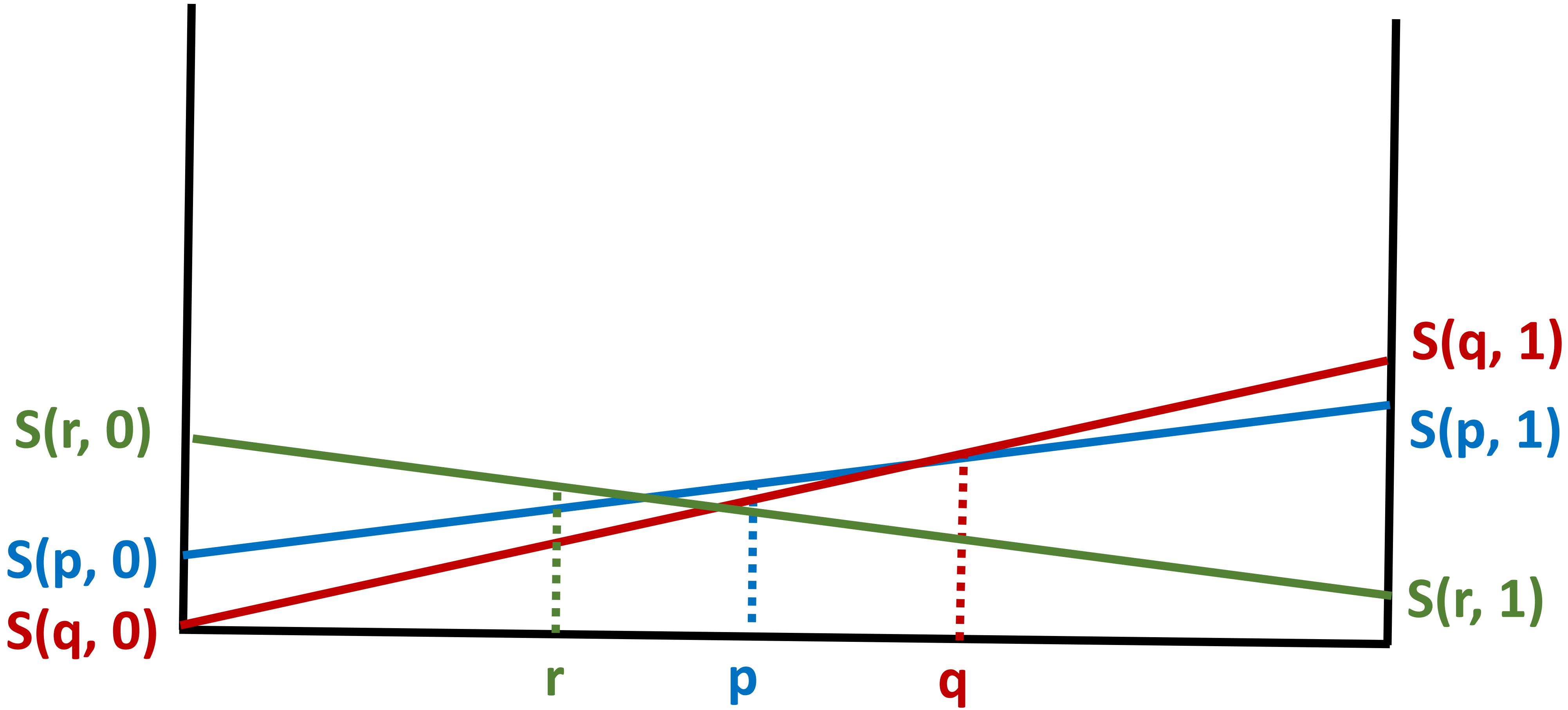

G(p)=pS(p,1)
+(1-p)S(p,0)
S(p, 1)
S(p, 0)
Expected reward for reporting p as a function of true probability
p

Proper scoring rules must have convex G

- Let $q = p+\varepsilon$ and $r = p-\varepsilon$. Consider the following manipulation.
- Sometimes, when you believe q, report p.
- Similarly (equally often, say also at rate α), when you believe r, report p
- For all these misreports, the actual probability is $(\alpha(p+\varepsilon)+\alpha(p-\varepsilon))/2\alpha = p$
- So these misreports on average give you $G(p)$
- Reporting truthfully, you would have received on average $(G(q)+G(r))/2$
- So for the rule to be proper, need $G(p) \leq (G(q)+G(r))/2$ (strictly proper: $<$)
  - Interpretation: **Destroying information should not help you!**
  - Sharpness is valued
- Can generalize this argument to conclude that $G(p)$ must be (strictly) convex for a (strictly) proper scoring rule (on the board)

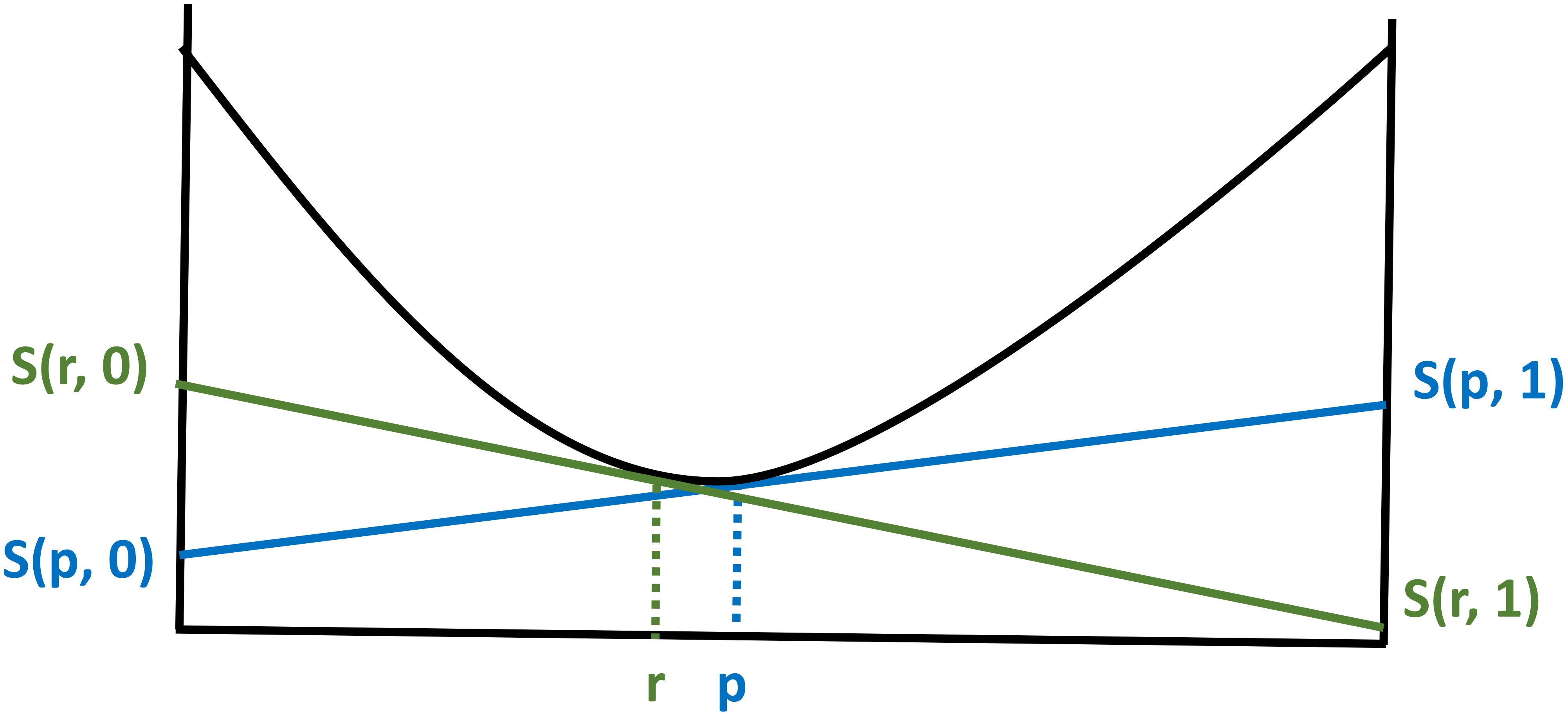
Convexity of G…

- Let G(p) denote your expected reward if you believe and report p



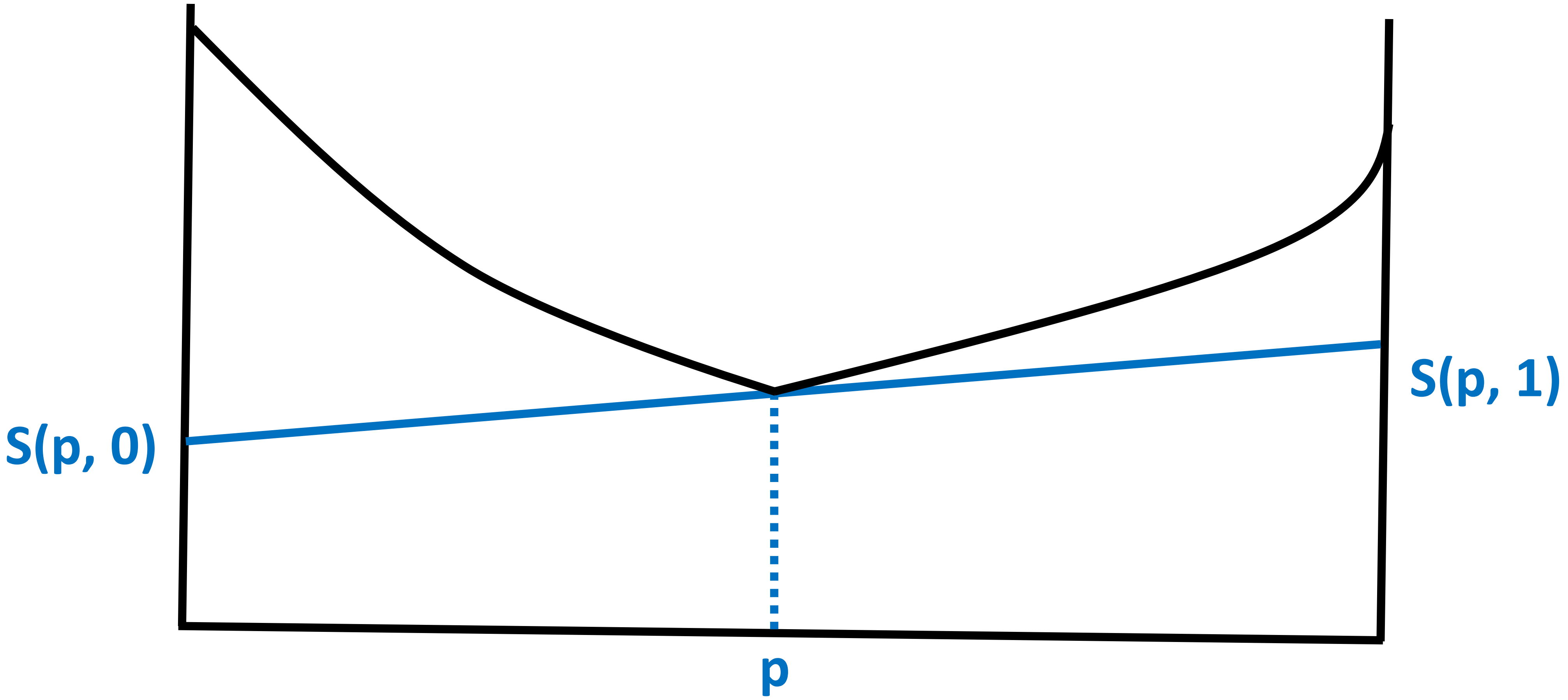
Conversely: for any (strictly) convex G there is a (strictly) proper scoring rule

- Just use tangent lines!



Slight caveat

- Some convex functions may not have a well defined derivative everywhere
  - … but any subtangent line will do



General characterization

[Savage 1971, Gneiting and Raftery 2007]

- **Theorem.** A scoring rule is (strictly) proper if and only if there exists a (strictly) convex function G such that

  $$S(\mathbf{p}, \omega) = G(\mathbf{p}) + G^*(\mathbf{p}) \cdot (\mathbf{e}_\omega - \mathbf{p})$$

  where the vector $G^*(\mathbf{p})$ is a subgradient of G at **p**, that is,

  $$\text{for all } \mathbf{r},\ G^*(\mathbf{p}) \cdot (\mathbf{r} - \mathbf{p}) \leq G(\mathbf{r}) - G(\mathbf{p})$$

Principal-aligned proper scoring rules [Shi, Conitzer, Guo 2009]

Spending millions of dollars on some kind of fantasy league terror game is absurd and, frankly, ought to make every American angry. ***–Senators Wyden, Dorgan 2003***

- Suppose we are worried about the forecaster taking undesirable actions to affect the outcome
  - Asking a developer to predict when the product will be ready
  - Asking someone capable of committing terrorist acts whether there will be a terrorist act
- Let u_ω denote the principal's utility for outcome ω. We would like the proper scoring rule to be aligned with the principal's utility, i.e., not create incentives to reduce it
- **Theorem.** A proper scoring rule is aligned with principal utility function u if and only if it corresponds to $G(\mathbf{p}) = g(\mathbf{p} \cdot \mathbf{u})$ where g is convex and non-decreasing.
- What examples of such a function have we seen?
- Proof – can you figure it out?