Homework 1
CS 236r, Spring 2016

Due: Tuesday, March 1, 11:59pm

1. **Practice and Geometric Interpretation of Proper Scoring Rules.**
   Consider the following convex "expected score function" for a prediction $p \in [0, 1]$ over a binary event: $G(p) = p^2$.

   (a) Give a proper scoring rule S whose expected score for reporting truthfully is G. Is S *strictly* proper?

   (b) Draw (roughly) the graph of G on $[0, 1]$. Consider the prediction $p = 0.5$ and indicate on the drawing the payoffs for reporting 0.5 when each outcome occurs.

   (c) Add to the drawing the expected utility for reporting $p = 0.25$ when one's true belief is equal to $q = 0.25$. Indicate on the drawing the difference in payoff between this case and misreporting $p = 0.5$ when one's true belief is $q = 0.25$. Calculate the difference.

   (d) Write the Bregman divergence function of G, and calculate the divergence $D_G(0.25, 0.5)$.

   (e) What value of p maximizes $G(p)$? Since G is the "expected score function", why doesn't an agent maximize expected score by always reporting this value of p? Explain briefly.

2. **The Savage Characterization of Proper Scoring Rules.**
In this problem we will walk through the proof of a fundamental correspondence between proper scoring rules and convex functions. Some useful definitions:

- Given a proper scoring rule S, let $S(p;q) = \mathbb{E}_{\omega \sim q} S(p,\omega)$, the expected score for reporting p with belief q.
- $f : \mathbb{R}^d \to \mathbb{R}$ is *convex* if, for all $\alpha \in [0,1]$, we have $\alpha f(x) + (1-\alpha) f(y) \geq f(\alpha x + (1-\alpha) y)$.
- The notation δ_ω refers to the probability distribution putting mass one on outcome ω and 0 on all other outcomes. (Hence δ_ω is a vector consisting of all zeros except for one one.) The notation $\langle a, b \rangle$ is the dot-product between vectors a and b.
- The vector r is a *subgradient* of a function f at point p if, for all q, $f(q) \geq f(p) + \langle r, q-p \rangle$. Not all functions have subgradients at all points. We use $f'(p)$ to denote a subgradient of f at p if it exists (even if f is not differentiable).

A key fact we will use is that f is convex if and only if it has a subgradient $f'(x)$ at every point x in the interior of its domain. (Prove this for bonus points.)

We'll prove:

Theorem 1. *For every convex function G, there exists a proper scoring rule S such that $G(p) = S(p;p)$ and $S(p,\omega) = G(p) + \langle G'(p), \delta_\omega - p \rangle$. For every proper scoring rule S, there exists a convex function G such that the above holds.*

(a) **Direction $G \to S$.** Given a convex function G, define S by letting $S(p,\omega) = G(p) + \langle G'(p), \delta_\omega - p \rangle$. Show that $S(p;p) = G(p)$. Then show that S is proper, *i.e.* $S(q;q) \geq S(p;q)$. (Hint: use the definition of the subgradient.)

(b) **Direction $S \to G$.** Given a proper scoring rule S, define G by letting $G(p) = S(p;p)$.

   i. Let S_p be the vector $(S(p,\omega_1), \ldots, S(p,\omega_n))$. Argue that $S(p;q) = \langle S_p, q \rangle$.

   ii. Show that S_p is a subgradient of G at point p. (Hint: use properness of S.)

   iii. Argue that G is convex and that $S(p,\omega) = G(p) + \langle G'(p), \delta_\omega - p \rangle$.

3. **Properties and Machine Learning.**

(a) Consider the loss function $\ell(h, x) = |h - x|$, where $h, x \in \mathbb{R}$, h is a hypothesis, and x is a data point. Given that x is drawn from distribution P, what hypothesis h minimizes expected loss? (In other words, what property does the scoring rule $s(h, x) = -\ell(h, x)$ elicit?)

(b) We saw in class the theorem that if a property is directly elicitable, then its level sets are convex. (*Directly elicitable*: there exists a scoring rule $s(h, x)$ such that the property maximizes expected score. *Level set*: the set of probability distributions having the same value of the property.) Prove this in the special case where the property is the mean of the distribution, *i.e.* prove that the level sets of the mean are convex.

(c) Using the previously-mentioned theorem, give a counterexample showing that the variance is not directly elicitable.

(d) Now consider a scoring rule $s(h, x_1, x_2)$ that takes a hypothesis h and two i.i.d. observations, x_1 and x_2, and outputs a score. Show that variance *is* directly elicitable by such a scoring rule.

(e) Using the previous result, describe a new type of loss function that provides a "consistent" estimator for the variance: As the amount of data collected goes to infinity, the minimizer of this loss function converges to the variance of the distribution.

4. **Elicitation and/or Peer Prediction.**
A lepidopterist in Shanghai observes whether or not a butterfly there flaps its wings. A meteorologist in Houston would like to elicit this observation truthfully, but the meteorologist cannot observe it directly. Instead, the meteorologist can only observe whether or not there is a hurricane in the following week. It is common knowledge that, if the butterfly does not flap its wings, there is a 0.1 chance of the hurricane occurring, whereas if it does, there is a 0.11 chance.

   (a) Give a payment rule whereby the meteorologist strictly incentivizes the lepidopterist to report truthfully. The payment should be a function of the lepidopterist's report (flap or no flap) and the meteorologist's observation (hurricane or no hurricane).

   (b) The meteorologist is concerned that your payment rule may not give good enough incentives. Now give a payment rule which always gives a nonnegative payment, and where the expected payment for truthfulness is always at least 1 unit higher than expected payment for untruthfulness.

   (c) A third party in Cairo observes neither the hurricane nor butterfly, but wishes to elicit both observations from the respective experts. Suppose that the *a priori* probability of the butterfly flapping its wings is 0.5. Give payment rules for the two experts, based on both of their reports, so that it is a strict equilibrium for both to truthfully report their observations.