Scoring Rules for Distribution Statistics

Sébastien Lahaie
Microsoft Research, NYC

1 Preliminaries

We consider a state space $\mathcal{X}$ which may be discrete or continuous, and a measure ν over sets of states. We will typically take the state space to be $\mathbf{R}$ or $\mathbf{R}^d$ together with the Lebesgue measure (perhaps restricted to some subset of states), or some finite set together with the counting measure.[1] A *probability density* is a ν-measurable function $p : \mathcal{X} \to \mathbf{R}_+$ that integrates to 1 over $\mathcal{X}$.

Let $\mathcal{P}$ denote the set of all probability densities over the given measure space. We are interested in eliciting information about some unknown density $p \in \mathcal{P}$ which corresponds to the beliefs of an expert. The expert may be familiar with p in its entirety, or only with certain properties of p. Given a set of densities $\mathcal{D} \subseteq \mathcal{P}$, a *property* is a real (possibly vector-valued) function $\Gamma : \mathcal{D} \to \mathbf{R}^k$. With a slight abuse of terminology we call elements of its image $\Theta = \Gamma(\mathcal{D})$ *properties* and we say that p has property θ when $\Gamma(p) = \theta$. If Γ is one-to-one, we instead call Γ a *parametrization* and refer to the elements of its image Θ as *parameters*.

To be concrete, our outcome space could for instance be $\mathbf{R}$ with ν the Lebesgue measure restricted to $[0, +\infty)$. The set $\mathcal{P}$ then consists of all probability densities with support contained in the non-negative reals, and $\mathcal{D} \subset \mathcal{P}$ might consist of the densities of exponential distributions. An example of a property over $\mathcal{P}$ is the mapping from a density to its mean.[2] When restricted to the exponential densities $\mathcal{D}$, this mapping is a parametrization. In much of this work, as in this example, the purpose of the measure ν is to implicitly encode bounds on the support of possible densities. The reason

[1] We will not need to make reference to the underlying σ-algebra, but to complete the specification of the measure space we take it to be the collection of Borel sets in the case of $\mathbf{R}$ and $\mathbf{R}^d$, and the power set in the case of a discrete set.

[2] The property may not be defined for all $p \in \mathcal{P}$. In this case we implicitly restrict it to the subset of $\mathcal{P}$ for which it is defined whenever its domain $\mathcal{D}$ is not made explicit.

we use this approach, rather than explicitly stating bounds, is that it allows for clearer connections with maximum entropy estimation later on.

Given the set of relevant densities $\mathcal{D} \subseteq \mathcal{P}$, which contains the expert's belief p, and a property Γ over this set, the expert is asked to report $\theta = \Gamma(p)$. The expert is rewarded according to a *scoring rule* $S : \Theta \times \mathcal{X} \to \mathbf{R} \cup \{-\infty\}$ which pays it $S(\hat{\theta}, x)$ according to how well its report $\hat{\theta} \in \Theta$ agrees with the eventual outcome $x \in \mathcal{X}$. The idea is to design the scoring rule so that the expert is incentivized to report its true belief, setting $\hat{\theta} = \theta$.

Definition 1 *Given a property $\Gamma : \mathcal{D} \to \mathbf{R}^k$, a scoring rule S is* proper *over domain $\mathcal{D}$ if for each $\theta \in \Theta$ and $p \in \mathcal{D}$ with property θ, we have*

$$\mathrm{E}_p[S(\theta, x)] > \mathrm{E}_p[S(\hat{\theta}, x)] \tag{1}$$

for all $\hat{\theta} \neq \theta$. If the domain is $\mathcal{D} = \mathcal{P}$, the set of all possible densities, then we say the scoring rule is strongly proper*.*

- Note that the inequality is strict, in the literature these rules are called *strictly* proper, but we only consider strictly proper rules here so don't make this qualification.

- Lambert et al.'s rules are actually strongly proper, however we consider scoring rules over vector valued properties. Lambert et al only develop these by adding together different scoring rules for each property. We go further in our constructions and develop rules that are not necessarily separable.

- It is useful to consider rules that are not strongly proper because we might have constructions for these for properties where no strongly proper rule is known.

- It is important to understand the informational requirements placed on the expert. Under a proper scoring rule, the expert does not need to be aware of the entire density p to be incentivized to report its property θ, since θ is an optimal report no matter what the density might be. The agent only needs to agree that the density indeed lies in $\mathcal{D}$.

- For instance, if we are considering densities with support on $[0, +\infty)$, and $\mathcal{D}$ is the set of lognormal densities, then a proper scoring rule over $\mathcal{D}$ for the mean does not require the agent to know anything beyond the mean (e.g., it does not need to know the variance).

- A strongly proper scoring rule goes further; with such a rule we do not require the expert to believe the density is drawn from some subset $\mathcal{D}$, only that it agrees with the support implied by the base measure ν. Therefore, a strongly proper scoring rule for the mean here would elicit the mean no matter what kind of density over the positive reals might apply (e.g., exponential, Pareto, lognormal).

- In practical scenarios this seems like an important extension, because while it might be clear which states are possible, it might be difficult to ensure the expert agrees that the probability density for states falls within some restricted family $\mathcal{D}$.

2 Maximum Likelihood

Theorem 1 *The logarithmic scoring rule defined by*

$$S(\theta, x) = \log p(x; \theta)$$

is weakly proper for $\mathcal{F}$ over Θ if and only if the maximum likelihood estimate for θ is consistent. Furthermore, if $\mathcal{M}$ is an alternative parameter space such that there exists a bijection $t : \mathcal{M} \to \Theta$, then $\bar{S}(\mu, x) = S(t(\mu), x)$ is weakly proper for $\mathcal{F}$ over $\mathcal{M}$.

Example 1. Suppose X is supported on $[0, +\infty)$ and follows an exponential distribution with density $f(x; \lambda) = \lambda e^{-\lambda x}$ parametrized by a rate $\lambda > 0$. The mean μ is related to the rate via the one-to-one mapping $\mu = 1/\lambda$, so the density can alternatively be parametrized by the mean. According to Theorem 1 the following is a weakly proper scoring rule for the mean of an exponential distribution:

$$S(\mu, x) = -\frac{x}{\mu} - \log \mu. \tag{2}$$

To verify this, suppose the agent believes X follows an exponential distribution with rate λ_0, equivalently mean $\mu_0 = 1/\lambda_0$. By the one-to-one relation between rate and mean, choosing μ to maximize the expected score (2) is equivalent to choosing λ to maximize $\mathrm{E}[-\lambda x + \log \lambda] = -\lambda\mu_0 + \log \lambda$. The latter is strictly concave in λ, and using basic calculus its unique global maximum is $\lambda^* = 1/\mu_0$, which therefore leads the agent to report $\mu = \mu_0 = 1/\lambda^*$ under scoring rule (2).

Example 2. Suppose that X is supported on $[1, +\infty)$ and follows a Pareto distribution with density $f(x; \alpha) = \alpha / x^{\alpha+1}$ parametrized by an index $\alpha > 0$. The mean μ is related to the index via the one-to-one mapping $\mu = \frac{\alpha}{\alpha - 1}$, so the density can alternatively be parametrized by the mean. Theorem 1 gives the following weakly proper scoring rule for the mean of a Pareto distribution with support on $[1, +\infty)$:

$$S(\mu, x) = \log \frac{\mu}{\mu - 1} - \left(\frac{\mu}{\mu - 1} + 1 \right) \log x. \tag{3}$$

To verify this, suppose the agent believes X follows a Pareto distribution with index α_0, equivalently mean $\mu_0 = \frac{\alpha_0}{\alpha_0 - 1}$. By the one-to-one relation between the index and mean, choosing μ to maximize the expected score (3) is equivalent to choosing α to maximize $\mathrm{E}[\log \alpha - (\alpha + 1) \log x] = \log \alpha - (\alpha + 1)\mathrm{E}[\log x]$. The latter is strictly concave in α, and using basic calculus its unique global maximum is $\alpha^* = 1/\mathrm{E}[\log X] = \alpha_0$, where the last equality follows from the fact that $\log X$ has an exponential distribution with rate α when X is Pareto with rate α. Therefore (3) leads the agent to report $\mu = \mu_0 = \frac{\alpha^*}{\alpha^* - 1}$.

- Explain that the scoring rule from Example 2 does not elicit the mean when the agent's distribution is exponential (with change of variable so that it lies in $[1, +\infty)$.
- Observe that scoring rule from Example does elicit the mean when the agent's distribution is exponential, Pareto, or in fact any other distribution with support on $[0, +\infty)$ such as the lognormal.
- This occurs because the scoring rule is linear in x, so its expectation is the same for any distribution with mean μ.
- Another feature to note is that under an appropriate parametrization (rate instead of mean), the scoring rule (1) is convex in the parameter being elicited.
- These two observations taken together lay the groundwork for the characterization of strongly proper scoring rules for distribution statistics.
- Before we proceed should note that the arguments behind Theorem 1 apply not just to maximum likelihood but to any *m-estimator*, which are notions from robust statistics. Explain what an m-estimator looks like and how there are general consistency results for those as well.

- The point of these alternative estimators is to be more robust to outliers in the sample, and so these kinds of estimators could translate into scoring rule more robust to inaccuracies in the agent's assessment of the underlying probability distribution or its parameters. (Just a hunch.)