Ponderings on Judgemental Estimation

Ozzie Gooen

June 20, 2026

Abstract

abstract-text

Contents

1 Introduction

1.1 Judgemental Estimation

Much of the existing literature on Bayesian Statistics relies on idealized models of agent understanding and experimental design. For instance, many models assume logical omniscience, which has been criticized as unrepresentative for human reasoning [1].

Judgemental estimation refers to the kinds of estimates made by forecasters in research on judgemental forecasting; where it is distinguished from statistical estimation. Judgemental estimations do not assume logical omniscience. Reasoners have initial intuitions on probability statements that may be inconsistent with each other.

This idea can be mixed with decision theory such that we can find utility-optimal methods for reasoners to make updates. This is useful for making choices about what kinds of consistency to strive for where there are costs for probabilistic improvements.

Here we are interested in developing a descriptive model of judgemental probabilistic reasoning that can be used for normative purposes. This work may be best viewed in terms of information economics rather than probability theory or epistemology. The concept of "Value of Information" is typically used within information economics. Here we argue that updating should be done in ways that maximize the value of information.

Judgemental estimates can be looked at as equivalent to credences. It's not

clear what terminology is best to use here. Credences go by a few definitions in the literature and are typically used for epistemic questions that are typically relatively unrelated for the purposes of this work.

Other related fields are those of judgemental forecasting, from which the term "judgemental" is derived from, and that of cognitive science, which does creates partially realistic models of human learning but is generally more about descriptive details than normative theorizing.

[1] `https://plato.stanford.edu/entries/logic-epistemic/#LogiOmni`

1.2 Theory Status & Request for Feedback

This work is quite early, somewhat poorly researched, and relatively unstructured. It's relatively dense. There aren't many full examples, and where there are, there typically isn't corresponding theoretical rigor. The main goal of this document is to further the development of these ideas, rather than to optimize the explanations or rigor.

I personally don't have much experience writing academic math papers or understanding much of the relevant literature. Arguably, there's a whole lot of relevant literature.

At this point my main concern is to do the following:

1. Make sure these ideas on this are online to some capacity.
2. Get feedback regarding what aspects may be interesting to others.
3. Get advice and assistance what existing literature, notations, and terminology is best for much of this.
4. Use this work as something of a theoretical foundation in the development of judgemental prediction services.

2 Definitions & Key Assumptions

2.1 Information Value and Decision Value

A rational agent with preferences that can be modeled as a Von Neumann–Morgenstern utility function should generally strive to obtain information for the sake of optimizing that utility function.

Here, the obtaining of information is viewed as an instrumental goal. If knowledge of information were to be considered a terminal goal, that could be stated as part of said utility function, and optimization could still be understood as sequences of Expected-Value maximization decisions.

Generally, we expect that for most humans, the vast majority of the benefit of information is instrumental.

Even if the knowledge of information were considered a terminal value, it seems likely that information would have to be prioritized. There are some kinds of information that are generally regarded as more interesting than others.

It's difficult to get precise estimates of value. One estimation method used in Economics is the *Willingness to Pay* (WTP). A naive version of this may make predictable mistakes, as people may naturally have a poor idea of how much they would value information on much further reflection. *Information value* refers to the value agents actually get, rather than what they initially believe they get.

One modification could be to incoorporate *Enlightened Preferences*, as have been discussed by Bryan Caplan. Then we could imagine the *Enlightened Willingness to Pay*, or EWTP. One could attempt to imagine the EWTP for different individuals and different levels of enlightenment, to best estimate information value or decision value.

2.2 Information Value Theoretic Interpretations of Human Phenomena

Information-theoretic interpretations have been theorized as mathematical foundations for many important technical and mathematical phenomena.

To get up-to-date examples, just search "Information-theoretic" in Google Scholar.

Information Theory was derived to assist in technical problems around the transmission of specified information. In "A Mathematical Theory of Communication", Claude E Shannon discussed the necessary theory to understand how to send complete signals given possible sources of noise. Much of future work in Information Theory held a few assumptions:

1. The information source is fixed and relatively limited.
2. The necessary entropy probabilities can be known in advance and are generally objective.

In many human examples these assumptions do not hold. Humans have the

choice of obtaining and sharing vast amounts of information and must select a very small fraction of it for learning. They also have great uncertainty regarding entropy amounts. In conversations, for example, the act of selecting which information to transmit is arguably a higher-variance decision than deciding how to best convey that information to best minimize noise loss.

— Notes

https://ezproxy-prd.bodleian.ox.ac.uk:2461/chapter/10.1007/978-1-84882-491-1_8 "Information-Theoretic Interpretations" "An Information-Theoretic Interpretation of Thresholds in Probabilistic Rough Sets"

2.3 Academic Work as Instrumental Information Value

Academic progress could be viewed through the lens of instrumental value. Doing so would offer interpretations on the efficacies of various strategies, methods, and developments. Arguably, many existing cost-benefit analyses of scientific work have primarily used similar assumptions.

There are other claims that academic work is terminally valuable. This should be expressible in utility functions, but this conversion is highly complicated. Some questions would emerge:

1. How can we distinguish highly terminally valuable academic work from non terminally valuable academic work? If we cannot distinguish, should we optimize academic work to get as much information as possible? It's possible this could lead to highly unintuitive results.
2. What is the time-weighting of the value of humanity's terminally valuable knowledge? If it were more efficient to spend 500 years growing the economy before engaging in terminal academic work, would that be an optimal trade-off?

Even if instrumental value makes up less value than terminal value, it seems significantly more estimable, or tractable to analyze.

If much of Academic value were to come from instrumental information value, then many high-level questions of science and academic could be partially explained in terms of instrumental information value.

2.4 Belief-Relativity

Claim: Two judgemental agents with different priors, and without significant time to communicate with each other (as in Aumann's agreement theorem), should at least occasionally disagree on probabilistic statements in realistic settings; though often to very minor extents.

There are a few assumptions here:

1. Both agents have intuitions that are functions of a lot of data about the world.
2. It would be infeasible to explicitly describe all of the data responsible for such intuitions, due to a combination of the fact that it may be impossible or very expensive.
3. The data from both agents has a lot of divergence; agent A has witnessed a lot of data that agent B hasn't, and vice versa.
4. The beliefs of both agents when conditioned on specific new information would not be dominated by that information, in ways influenced by their respective differences in information.

This can be simply stated that agents can be expected to have different priors and likelihood functions, and that these differences can be expected to lead to differences in posteriors.

Here we use the term *estimation dominance* to refer to an estimate that should be used in place of, instead of in addition to, other estimates of the same variable.

$$P(A|p_{dominant}, p_{i...n}) = p_{dominant}$$

If one takes this claim to be true, we can say that probabilistic statements are *belief-relative*, meaning that they vary agent to agent based on each agents' existing priors.

Where belief-relativity holds, then there is no such thing as an *objective* probabilistic statement; in the sense that no probabilistic statement could be expected to be dominant for all recipients.

Because entropy and information quantities rely on probabilities, those would be considered belief-relative.

Some Bayesian writing clarifies this by specifying "general" priors in all Bayesian equations.

$$P(A|\omega)$$

Where ω is a generic term representing the worldly prior of the given agent.

3 Mathematical Theory

3.1 Decision-Relevant Information

In one sense a noisy picture (just random pixels of black & white) has a high information content, because it is difficult to predict. In a different sense, it has a very low information content, because the only decision-relevant piece of information could be that it's "noise".

Related, say we are interested in how many coins of a set of coins are heads. We can compress all of the physical information we know about the coins into a very small string, like, "8 coins, each with a negligible bias". When we do this we don't lose any information that we could predict would be important for the sake of the calculation.

Arguably one important aspect of prediction question operationalization / parameterization is to wind up with a "maximally compressed" representation that contains all of the information that matters, with as little as possible of what doesn't.

For instance, if one had two choices, A, and B, and wanted to decide between them, they could estimate Utility(A) and Utility(B), but this likely involves unnecessary information. What one really cares about is something like:

$$P(\mathbb{E}(A) > \mathbb{E}(B))$$

Where we can expect that the entropy, H,

$$H(P(\mathbb{E}(A) > \mathbb{E}(B))) < H(\mathbb{E}(A), \mathbb{E}(B))$$

A very simple example of this would be to say that if someone offered you "either \$1 or \$2", you wouldn't need to estimate the total impact of each on your utility function; rather, you just need adequate confidence on two subquestions:

1. Is \$1 < \$2?
2. Is the expected value of money likely to be positive?

Related to this, we can think in terms of "value of information" to show where information reductions are costly. If we believe our own total expected value conditional on having information I_2 is the same as that on us having information $I_1 \supset I_2$ then this information loss didn't cost any information value.

$$E(Utility|I_1) = E(Utility|I_2 \subset I_1) \implies Value(I_2) = Value(I_1)$$

If information can come with some cost in some sense, then one wants to seek the most "compressed" representations.

Note 1: Instead of writing

$$P(\mathbb{E}(A) > \mathbb{E}(B))$$

we may want to use the equation

$$P(\mathbb{E}(A - B) > 0)$$

Note 2: On the equation:

$$P(\mathbb{E}(A) > \mathbb{E}(B))$$

This is a bit gnarly because probability is handled both in the P parameter and the $\mathbb{E}$ parameters, so calculation would require careful handling of expectations of possible knowledge gains.

3.2 Selecting of Predictions to Minimize Expected Loss

Say we have a set of calibrated predictions $p_{1..n}$ on some claim ϕ, and we can only choose one to personally use. We would like to minimize the expected loss function that's based on a logarithmic scoring rule.

A naive to do would be to select the prediction with the lowest self-expected loss.

$i_{optimal} = argmin_{i \in 1..n} \mathbb{E}(S(P_i)|P_i)$

However, there are some situations where this would fail, because the presence of all of the predictions $p_{1..n}$ contains information that would lead one to believe that some predictions are effectively overconfident. Notice the last term of this corrected equation.

$i_{optimal} = argmin_{i \in 1..n} \mathbb{E}(S(P_i)|P_{i..n})$

The important thing here is that:

$\mathbb{E}(S(P_i)|P_i) \neq \mathbb{E}(S(P_i)|P_{i..n})$

To give a concrete example, say that there are two predictions of the "mean number of a quantity with a population of 50,000" from two predictors that are perfectly modeled as beta distributions from the finding of binary evidence. Both predictors begin with priors of

$$P_{prior} = beta(1, 1)$$

The first predictor saw 10 points of data and provides the distribution

$$P_1 \sim beta(1, 12)$$

The second predictor saw all points that the first predictor saw, plus 2 more points:

$$P_1 \sim beta(3, 12)$$

In this case, if they independently estimated their own expected losses, the first predictor would expect a lower expected loss. The differential entropy (the same as the expected value of a log score) of beta(1,12) is less than that of beta(3,12).

$$H(beta(1, 12)) < H(beta(3, 12))$$

That said, note that this situation should be expected to be unusual.

One can generally expect that an increase in information leads to a decrease in expected loss, but this is not always the case.

One selection strategy that wouldn't be vulnerable to this specific failure would be to select the prediction that came from the information with the lowest entropy, rather than the prediction that itself had the lowest entropy. Of course, for this to work, we'd only care about the aspects of the entropy of the information that are relevant for the prediction, so this needs to be specified.

Questions:

1. Is there are formulaic procedure to specify and then estimate the information content that generated probability estimates, in a way that could be used for prediction selection?

2. In some situations, $H(p_1) < H(p_2) \implies p_{optimal} = p_1$. In others, it may be predictable with some predictably probability q. Calculating the likelihood

function would help tell us the specifics. It would be interesting to understand this relationship in many kinds of common formulations of data discovery.

3. It seems like when we are setting up prediction systems, we may not want to minimize calculated expected loss, but rather minimize uncertainty of the data generating processes that lead to the relevant forecasts. How can we best model this?

— Notes - Add information about how to estimate the entropy of the source. Have a good description of the necessary notation.

3.3 Selection of Models Informed by Judgmental Intuitions

Say you have 3 ways to estimate the same thing, and these methods produce different answers. What should you do?

Many models of agent behavior assume that the agent is logically omniscient, but real people don't have this property. Real people have judgemental intuitions that are inconsistent with each other. Further, complete consistency of judgemental intuitions can probably be shown to be computationally intractable given any reasonable constraints.

We can start this problem by discussing credences. My impression is that most discussion around credences happens in epistemology.

Say an agent believes:

$$cred(Y) = uniform(10, 20)$$

$$cred(X) = uniform(5, 7)$$

$$cred(Y = 2X) = 1$$

$$P(Y|cred(Y)) = uniform(10, 20)$$

$$P(Y|cred(Y = 2X), cred(X)) = uniform(10, 14)$$

The agent here has two different available methods to calculate Y; one using their direct credence, and one using a simple calculation. What procedure should this agent use for deciding?

Arguably, this would follow the same logic as in the previous example. We can first assume that these estimates are calibrated (and if not, we can apply a transformation for this to be the case).

Let's call these various estimation methods, and possible ways of combining them while staying calibrated, $P_{1..n}$.

Then,

$i_{optimal} = argmin_{i \in 1..n} \mathbb{E}(S(P_i)|P_{i..n})$

When in doubt of how to combine them, and when in doubt on how $\mathbb{E}(S(P_i)|P_{i..n})$ differs from $\mathbb{E}(S(P_i)|P_i)$, then we can make the simpler selection:

$i_{optimal} = argmin_{i \in 1..n} \mathbb{E}(S(P_i)|P_i)$

If we aren't sure how to combine different models, then our options for i will be that much smaller.

In the case above,

$$\mathbb{E}(S(uniform(10, 14))) < \mathbb{E}(S(uniform(10, 20)))$$

for a log scoring rule S.

Therefore, this agent should generally prefer this low-expected-cost option over the alternative.

I believe some work around Bayesian Epistemology may be relevant here. Some work on Conditional Credences is discussed following this link: `http://fitelson.org/bayes/titelbaum_ch3.pdf`

Questions:

1. I'm sure that there's much cleaner notation and terminology to use for this idea, but I'm not sure what it is or where to search for it.

4 Prediction and Language

4.1 Uncertainty in Question Definitions

In real life, question predictors and evaluators don't completely agree on question definitions. For example, the possible prediction question "How positive will intervention X be for the United States" would generally be considered mediocre, arguably because it would be difficult for predictors and clients to estimate exactly how an evaluator would interpret that question.

This problem could be expanded to the much more general problem that in people don't share exact matches of most definitions or sentences. Not only

do people disagree on terminology, but they also cannot perfectly estimate the terminology that others believe.

We can refer to the ability of a person to understand a statement as their *comprehension*, and to loss that could come out of a failure to do this as *comprehension loss*.

The big question here is how to best model this explicitly in ways that would be useful for selecting definitions that would minimize comprehension loss.

Interestingly, the ability of an individual to comprehend a statement is very similar of that to estimate a variable. Arguably, comprehension could be defined as a type of estimation over a complex parameterization, which could thus trivially be made into a type of prediction. Therefore we can use much of the same terminology as is used in predictions. For instance, an individual is overconfident in a comprehension if they mistakenly believe a false interpretation with more confidence than is supported by the evidence. Their comprehension could hypothetically be scored using a proper scoring rule against a true statement definition.

Ironically enough, the main area I know of where definition disagreement is well parameterized is that of "Words of estimate probability." For our purposes we can assume that people will always use exact numbers for probabilities, so we don't need to use this work, but a generalization of it could help with the answers of understanding of common definitions and sentences.

`https://en.wikipedia.org/wiki/Words_of_estimative_probability`

Hypothetically we can come up with some distance functions between interpretations of a given statement. This is easy when the statement could be neatly parameterized into a single parameter.

Consider the statement:

John Fillmore Smith will drink a lot of water tomorrow

One may make the simplification that all of the uncertainty in this phrase lies in the meaning of "a lot", which could be estimated as a probability distribution over a unit of volume. If there is a "correct" definition, this could be compared with any other distribution using the KL divergence. A large KL divergence between interpretations would indicate a substantial disagreement or error.

Similarly, one could also imagine that there may be uncertainty in who John Fillmore Smith is referring to, how "water" may be defined, and how "tomorrow" may be defined.

One way to compare these specific uncertainties would be to imagine how they

would influence a forecasters' total estimate on how likely this claim is to be true.

4.2 Effective Parameterization and Parameterization Loss

The challenge of defining specific forecasting questions has been discussed in literature around the Good Judgement Project and basically anyone who has tried to operationalize forecasting questions for others to use.

The Good Judgement Project has discussed the trade-off of *rigor vs. relevance*; the idea being that questions can often be either be explicit and easy to verify, or ambiguous though more relevant. For instance, the question, *If I eat potatoes, will that help my health?* may be useful, but is highly ambiguous. The question, *If I eat one potato today, will I report a stomach ache tomorrow?* is more specific and verifiable, but also less important than the first question, especially if done to attempt to estimate the first question.

We can define *parameterization* here as the process of converting an uncertainty into a set of parameterized statements that can be directly estimated using probabilities or probability distributions.

For example, the question: *What will happen to me next year?* is not well parameterized, but the subquestion, *How many hours will I be in REM sleep next year?* is.

Often in forecasting setups we have a high-level question we would like forecasters to help provide the answer to, but there is no straightforward way to fully parameterize it in a reasonable matter. In these cases we may choose to parameterize some specific subparts of it. The above example *What will happen to me next year?* is one example of this.

Say the true question we have is ϕ_1 and the subquestion can be considered as $\phi_2 \subset \phi_1$. If this was a highly *lossy* parameterization, then,

$\mathbb{E}(U|\phi_1) \gg \mathbb{E}(U|\phi_2)$

The difference of information about ϕ_1 between having people forecast ϕ_1 and phi_2 can be considered the *parameterization loss*.

4.3 Parameterization Loss and Comprehension Loss

Some parameterizations come with statement definitions that we can predict will be poorly comprehended. We can look at information losses due to miscomprehension as *comprehension loss*.

This leads to a possible conflict of *parameterization loss* vs. *comprehension loss* for various possible statements. Vague parameterizations may have low parameterization loss, but would have comprehension loss. Narrow parameterizations would likely have high parameterization loss, but hopefully low comprehension loss.

It's often possible to use a 'brute force' strategy of using very large number of very specific parameterizations. This could lead to low parameterization losses and comprehension losses. However, it may be very expensive. Parameterization and comprehension losses can be looked at as limitations of benefits, but benefits typically have to be considered with costs.

5 Future Work & Questions

5.1 Expected Loss

In this presentation, I discuss the concept of "Expected Loss" and how it can be used to make decisions about forecasting setups. Does this make sense? Is there any related literature I should be familiar with?

`https://www.youtube.com/watch?v=zTBp0Lmw4ZE&t=1721s`

One general assumption here is that we don't have to worry about all kinds of calibration errors; the main one that we need to worry about is systematic and predictable overconfidence, so that's really the main element to estimate and adjust for.

5.2 Entropy of Distributions

Differential entropy is the naive version of entropy to use for distributions, but it's "not nice". Differential entropy is unit-dependent and can be negative. Arguably one better construct is to use the limiting density of discrete points, but this seems to require the use of a uniform distribution with unspecified bounds. In many judgemental cases, it's not clear what these bounds should be.

Overall I haven't been able to find much discussion or many examples of how to best use the limiting density of discrete points, or other approaches to entropy besides that and differential entropy.

`https://en.wikipedia.org/wiki/Limiting_density_of_discrete_points`

5.2.1 Resulting Issues

If a forecaster reduces the uncertainty of $X \sim Uniform(100, 200)$ by 30%, they would have a much higher score than one that reduces the uncertainty of $X \sim Uniform(1, 2)$ by 30%.

If I have a variable $X \sim Uniform(1, 2)$ I would like to understand how transformations to X result in information loss about X. The transformation

$$Y = 2X$$

should hypothetically lead to 0 information loss if specified correctly; but the differential entropy would obviously increase. The specification here may include the fact that one knows that this transformation occurred.

Say that that the information that "Y came from X" is described by the variable i. Then,

$$P(Y|i) \neq P(Y)$$

Related, if H is the information entropy on a variable, then,

$$H(X) = H(Y|i)$$