Bayes nets and the dynamics of probabilistic language

Representing confidence & ignorance. You are asked to predict the outcome of a competition—say, a chess game between **A** and **B**—based on your background knowledge. Consider the following cases:

i) You know nothing at all about **A** and **B**.

ii) After watching many matches between **A** and **B**, you are confident that they are evenly matched.

In both cases, ***A** and **B** are equally likely to win* is an appropriate judgment—but for intuitively very different reasons. It is widely assumed that the classical Bayesian theory, where an agent's uncertainty is represented by a unique measure, cannot account for this difference. Halpern ['03, §2.3] takes a similar case to motivate the use of sets of measures or related enrichments: "probability is not good at representing ignorance". Such considerations have also led to widespread use of these devices in formal epistemology [Joyce'10, Elga'10].

However, de Finetti ['77]; Pearl ['88, §7.3] point out that the availability of **hierarchical** models, e.g. Bayes nets, undercuts the intuitive motivation for this more complex representation. These models are used in many modern applications in statistics, AI, and psychology. Probabilities derive from graphs representing causal relations sets of variables, together with the conditional distribution on each variable given its parents. Bayes nets readily represent the two kinds of uncertainty: see Fig.1a. Player i's performance is a Gaussian with parameters μ_i (**skill**) and σ_j (**consistency**). Each parameter's parents represent uncertainty about causal factors. $P(\mathbf{A}\ \mathbf{beats}\ \mathbf{B})$ is the probability that **A**'s noisy performance exceeds **B**'s, here equal to $P(\mu_{\mathbf{A}} > \mu_{\mathbf{B}})$. There is always a precise best guess about $\mu_{\mathbf{A}} - \mu_{\mathbf{B}}$, but confidence depends on the amount of evidence: low when only general domain knowledge is available, and high when evidence indicates equal skill (Fig. 1b).

Communicating uncertainties. Recent work suggests that *likely*, *probable*, and perhaps other epistemic operators have a probabilistic semantics as in (1) [Yalcin'05,'07, Swanson'06, Lassiter'10]. If so, we need a model of the way that people take up the information in probability statements, as in dialogue (2). Carl has learned something here, but what? Conditioning won't do: '$P(\mathbf{rain}) > .5$' does not pick out a set of worlds, and $P(\cdot \mid P(\mathbf{rain}) > .5)$ is not even defined. Yalcin's ['12] approach involves a dynamic semantics/pragmatics using sets of measures. Update by ϕ *is likely* eliminates measures μ where $\mu(\phi) \not> .5$, cf. (3).

There are several reasons to revisit this update procedure. The ability of Bayes nets to represent degrees of ignorance calls into question whether the added complexity of sets of measures is **ever** needed; decision-theoretic considerations also create difficulties [Elga'10]. Second, Yalcin's update procedure is obliged to use special devices for probabilistic language. Third, the procedure deals with the communication of known uncertainties—e.g., the color of a ball drawn with even probability from one of two urns of known composition—only if we use sets of measures to represent these cases as well, violating the spirit of the argument from ignorance. I propose to modify the probabilistic semantics, relativizing probability statements not to a measure P but to a Bayes net B, as in (4). Suppose ϕ is a value of variable $\mathcal{V}$. Each world w is associated with a 'local probability' function P_w taking ϕ to its conditional probability given the values, at w, of the parents of $\mathcal{V}$'s closest non-deterministic ancestors. Associating probabilistic statements with sets of worlds makes it possible to condition on them, simplifying the dynamics dramatically: all update is conditionalization. For example, conditioning on "**B** is likely to win" (2b) assigns zero mass to worlds where the parents of the closest non-deterministic ancestors of **winner**—$\mu_{\mathbf{A}}, \sigma_{\mathbf{A}}, \mu_{\mathbf{B}}$, and $\sigma_{\mathbf{B}}$—do not interact so as to favor **B**'s victory: see (5) and Fig.2. The result is that we condition on **B**'s strength exceeding **A**'s—the intuitively correct result—and the choice of this conditioning expression is generated by a fully compositional procedure. Ongoing work suggests that this semantics also yields reasonable predictions for epistemics in conditionals and under other epistemics, such as *definitely possible* and *probably unlikely* (cf. [Moss'15]).

While many formal and empirical questions remain, we suggest that formal models of communication & reasoning could benefit from increased engagement with the representations used in applied Bayesian work.

(1) $[\![likely]\!]^P = \lambda\phi_{\langle s,t\rangle}.P(\phi) > .5$

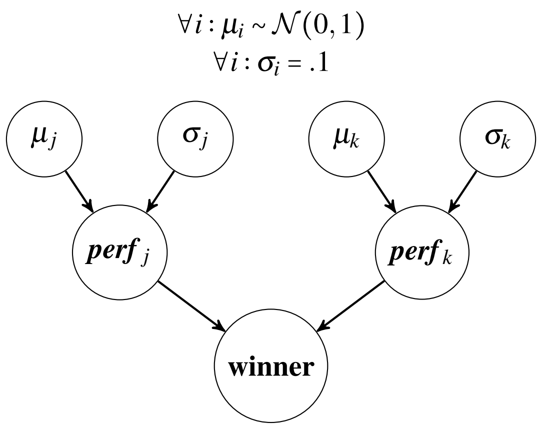
(2) a. **Carl**: Who is going to win the chess game between **A** and **B**?

b. **David**: **B** is likely to win.

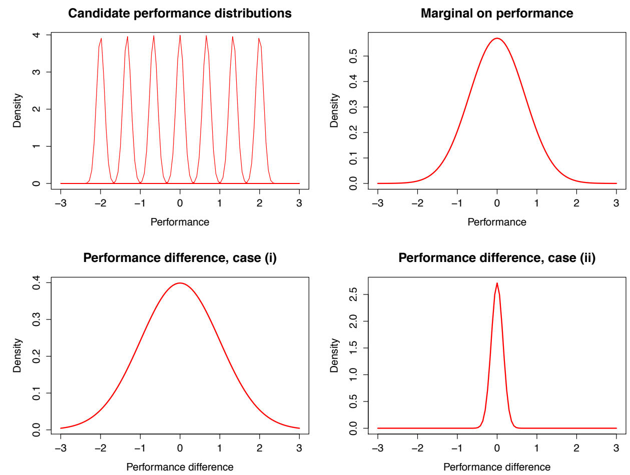
(3) $I_{\textbf{Carl}} \underset{\text{learn (2b)}}{\Longrightarrow} I_{\textbf{Carl}} \cap \{\mu \mid \mu(\textbf{B wins}) > .5\}$

(4) $[\![likely]\!]^B = \lambda\phi_{\langle s,t\rangle}.\{w \mid P_w(\phi) > .5\}$

(5) $P_{\textbf{Carl}} \underset{\text{learn (2b)}}{\Longrightarrow} P_{\textbf{Carl}}(\cdot \mid \{w \mid P_w(\textbf{B wins}) > .5\})$



(a) Hierarchical model of player performance in a single match, inspired by the Microsoft Trueskill system used to rank Xbox Live players [Bishop'13]. Each $perf_i$ has a $\mathcal{N}(\mu_i, \sigma_i)$ distribution. **B wins** is a deterministic node, true iff $perf_{\textbf{B}} > perf_{\textbf{A}}$.



(b) Top left: some of the ∞ candidate performance distributions varying with μ_i. Right: marginal on performance. Bottom: distributions on $perf_j - perf_k$. Both are centered on 0, so that $P(\textbf{A wins}) = P(\textbf{B wins}) = .5$. Case (i) [left] is the prior; case (ii) [right] is conditional on each player winning 15 of 30 matches. Increased confidence is reflected in lower variance of the expected difference.

Figure 1

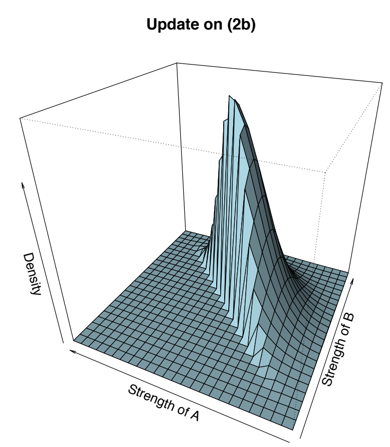


Figure 2: See right for caption.

Figure 2 caption: Conditioning on $\{w \mid P_w(\textbf{B wins}) > .5\}$ eliminates combinations of immediately causally relevant variables that do not make **B wins** likely, given the conditional probability information encoded by the Bayes net.

References Bishop'13, Model-based machine learning ▽ Elga'10, Subjective probabilities should be sharp ▽ de Finetti'77, Probabilities of probabilities ▽ Halpern'03, *Reasoning about Uncertainty* ▽ Joyce'10, Defense of imprecise credences ▽ Lassiter'10, Gradable epistemic modals ▽ Moss'15b, Semantics & pragmatics of epistemic vocabulary ▽ Pearl'88, *Probabilistic Reasoning in Intelligent Systems* ▽ Swanson '06, *Interactions With Context* ▽ Yalcin'05, A puzzle about epistemic modals ▽ Yalcin'07, Epistemic modals ▽ Yalcin'12, Context probabilism