Implicit Probabilities in Update Semantics

Henk Zeevat

Update semantics is both the first ([Karttunen, 1976]) and the final proposal in discourse semantics/dynamic semantics. Update semantics consists in a definition of information states and an update function or relation specified recursively for some update language. All proposals in dynamic semantics and discourse semantics have an interpretation for some instantiation of these two parameters. A proposal in dynamic semantics that cannot be interpreted in an update semantics should presumably be given up.

Update semantics moreover has a naturalistic interpretation as a formal treatment of what goes on when a human interpreter interprets and accepts a natural language utterance or some other perceptual input. It is a good basis for a family of consequence relations, the simplest one being that an update of the consequence on all information state updated with the premises is a fix point. Update semantics is preluded by [Karttunen, 1976] and [Heim, 1983], but [Veltman, 1996] first turned it into proper logical proposal.

Unfortunately, the naturalistic interpretation of update semantics is flawed by the absence of probabilistic reasoning in most versions of update semantics[1]. This means in particular that the process of interpretation for natural language utterances and perhaps the most important source of information for humans: visual signals cannot be integrated in the model. While non-stochastic visual interpretation does not exist, the underdetermination of meaning by form also known as the massive ambiguity of NL utterances, the main motivation behind modern computational linguistics, has not received sufficient theoretical attention in natural language semantics and pragmatics. It would seem to follow from the underdetermination of meaning by form that the problem of meaning selection is at least as important as the problem of meaning combination which is standardly taken to be the central problem of the field. It should —as in vision— be one of the two central problems.

This note is an attempt to show that these probabilities are already implicit in eliminative update semantics as such.[2] One merely needs to make some assumptions about sampling and correlations and one has the basic probabilistic notions. This paper develops the probabilities in the most basic way to establish that eliminative update semantics comes with a natural assignment of subjective probabilities to the sentences that are not yet true or false.

[Haas-Spohn, 1995] defines an epistemic alternative as one that the subject cannot rule out as being the actual world on the basis of the information that the subject has at her disposition about the actual world, given unlimited opportunities of investigating the alternative. This

[1] An exception is [Kooi, 2003] or [Van Benthem et al., 2009] that however follow a quite different route.

[2] The construction below is my reaction to the doubts of some colleagues that the scheme for Bayesian pragmatics briefly discussed towards the end of this note can be formalised. In fact —as I discovered very recently— [Bacchus et al., 1996] already defends a similar position about "subjective" probabilities underpinned by a similar formalisation and aiming for a similar goal: the reduction of default logic to probabilities.

criterion can be generalised to probabilities. The subject believes on the basis of her experiences that a correlation in the actual world can be estimated as having a certain strength. Sampling in the given alternative provides an imperfect check on whether the correlation in the alternative is as strong as in the actual world, as far as the subject knows. The check can then be used to adapt the probability that the alternative is the actual world. The alternative can still be the actual world but with a certain probability that goes down if the the strengths of its correlations do not match the observations of the actual world within the experience of the subject. If the updates conform better with the correlations in the world, its probability will go up again.

This means that all sentences that are not already true or false in the information state receive a probability by summing up the probability of all the worlds in which they are true. In the finite case —the case that will be worked out below— that gives $p_\sigma(\varphi) = \Sigma_{w \in [[\varphi]] \cap 1[\sigma]} p(w)$. .

Can this be formalised? The answer is that one needs assumptions about sampling and correlations and needs to deal with the relation of probability and the set of elements of the update semantics. The simplest version results by assuming (1) that all worlds in information states have a domain that has a cardinality smaller than some given n, (2) that sampling for a correlation is a question of counting all relevant observations for the correlation in the world and (3) that the subject assumes a finite set of correlations.

Assumption (3) can be seen as one kind of beliefs of the subject. She would believe that certain event types are causally connected, i.e. that a correlation exists, while other event types are not correlated. [Bacchus, 1990] formalises correlations as $[\varphi|\psi]_{x_1,\ldots,x_n}$ and under his semantics it is guaranteed that such correlations have a value q as soon as $\exists x_1, \ldots, x_n \psi$. This is unnatural for arbitrary formulas, but the semantics is a good approximation if the correlations are beliefs of the subject. The belief of the subject that a correlation exists is the belief that q can be approximated with arbitrary precision by taking large enough samples of instances of ψ. The belief in a correlation therefore goes together with assuming that there are enough instances for the correlation to be measured appropriately so that belief alternatives in an update semantics will have enough instances for a correlation the subject believes in. (3) can also be seen as a restatement of the important point of [Pearl, 2000] that people have reliable judgments about the dependence and independence of stochastic variables.

(2) gives Bacchus' approach to probabilistic first order logic using finite worlds in which correlations $[A|X]_{(1)_1,\ldots,(1)_n}$ are assigned the value $\frac{|\{<x_1,\ldots,x_n>:w \models X \wedge A[x_1,\ldots,x_n]\}|}{|\{<x_1,\ldots,x_m>:w \models X[x_1,\ldots,x_n]\}|}$ in worlds w. (1) would seem to be the most problematic assumption. It is however defensible since one is modelling finite creatures making generalisations about a finite universe based on finite set of observations and not mathematical reasoning or mathematical physics. One can still add mathematics by constructing the worlds as the combination of an invariant infinite mathematical structure and a finite model with certain axioms about their connections, as long as the correlations one assumes are based on the finite model alone. If this is not sufficiently convincing yet, (1) can be given up, switching to probability density functions and infinite information states made up from infinite worlds. In that case, one can no longer rely on Bacchus' simple definition of sampling and needs to come up with a more contentful notion of sampling.

Think of an information state σ as a sequence of updates $\sigma =< \varphi_1, \ldots, \varphi_k >$ with the φ_i taken from a first order language L. Let the empty information state $1 = W$, the set of all models for some finite L with domains of cardinality $k \leq n$, with n chosen sufficiently high. W is finite modulo isomorphism. $1[\sigma]$ is the set of alternative worlds that survive the sequence of updates. Let p be an assignment of probabilities to W, assigning values $p(w) \in [0,1]$ and such that $\Sigma_{w \in W} = 1$. p will automatically assign a probability to sentences of the language L by setting $p(\varphi) = \Sigma_{w \models \varphi} p(w)$ Assume further that K is a finite set of correlations $[A|X]_{x_1,\ldots,x_n}$, the correlations the subject believes in.

The probability assignments needed are assignments p_σ which have been updated with σ by means of Bayes' rule (1).

(1) $p_\sigma(w) = p(w|\sigma) = \frac{p(w)p(\sigma|w)}{p(\sigma)}$

$p(\sigma)$ can be defined as the sum of the $p(w)$ for $w \models \sigma$, i.e. $p(\sigma) = \Sigma_{w \in 1[\sigma]} p(w)$.

The central step is determining $p(\sigma|w)$.

Crucially, some consequences of σ will be observations that are relevant to correlations in K. Let φ be a consequence of σ. If there is a correlation $[A|X]_{x_1,\ldots,x_n} \in K$ such that $\sigma \models \varphi \leftrightarrow \theta A$ or $\sigma \models \neg\varphi_i \leftrightarrow \theta A$ where θ is a unifier for $x_1, \ldots, x_n$ and $\sigma \models \theta X$, φ is a relevant observation.

The strongest condition X is needed that would predict φ and is a consequence of σ. For this, we look at all correlations $[A|X]_{x_1,\ldots,x_n} \in K$ such that $\sigma \models \varphi \leftrightarrow \theta(\neg)A$ for which $\sigma \models \theta X$. We now form the new correlation $[A|X_1 \ldots X_n]_{x_1 \ldots x_m}$, normalising to the positive case. One can now determine the value v of the new correlation in w[3] by counting $\{< x_1, \ldots, x_m >: w \models X_1 \wedge \ldots \wedge X_n \wedge A[x_1, \ldots, x_m]\}$ and $\{< x_1, \ldots, x_m >: w \models X_1 \wedge \ldots \wedge X_n[x_1, \ldots, x_m]\}$ and taking the quotient.

$p(\sigma|w) = \Pi_{\varphi \text{ is relevant}} p(\varphi|w)$. We can now compute the result of Bayes' rule for the probability of w on the basis of σ and the prior.

Let P_0 be the set of assignments to W. Let $Q = \{p(w|\sigma) : p \in P_o\}$. Let $q_\sigma(w) = \frac{q(w)}{\Sigma_{w \in 1[\sigma]} q(w)}$ and let $P_\sigma = \{q_\sigma : q(\bigwedge \sigma) = 1 \wedge q \in Q\}$. P_σ contains the rational assignments for σ that assign 1 to the elements of σ and that have learnt from the observations contained in σ.

The construction can be refined in many ways. It is possible to shift to probability density functions to accommodate an infinite number of worlds, basing the needed σ-algebra on the values of the correlations.[4] It is possible to allow infinite models by introducing a notion of finite sampling as the basis for values to regularities[5]. It is sensible to also allow correlations whose value changes with time in a continuous way. It is possible —and necessary for many

[3] There is at least one instance of the conjunction on the right hand side. A correct worry is that there may be too little instances for the integrated correlation in w to be interesting in helping to determine the probability that w is the actual world. A way out is to stipulate that there should be a minimum number of instances of the right hand side for a correlation to have a value with respect to w. If this is enforced, certain relevant observations do not change the prior probability and they should not.

[4] This produces the rather puzzling notion of being epistemically possible with zero probability.

[5] This is preferable even for the finite case, but there many options for filling in the notion of a finite sample

applications— to develop accounts of how to come to believe in new regularities and how to lose faith in old ones: it is not rational to keep variables independent if one observes a correlation, it is not rational to maintain dependencies if they do not show up in the observations.

There are also many ways to come to a combination of σ with a set of probabilities. Information states σ determine a combination $(1[\sigma], P_\sigma)$ as defined above and can be described as remembering all the incoming information that is used to refine the probability assignments. But one could also first refine the assignments, forgetting the information on which the refinement is based. In that case, one obtains information states $(1[\sigma], P_{\tau\circ\sigma})$. One can even assume that there are innate constraints on the assignments before any learning starts.

The construction allows the comparison of sentences under rational probability assignments. This can be written as $\varphi < \psi$. There are two obvious interpretations for this comparison. For all rational p, $p(\varphi) < p(\psi)$ and the even stricter: for all rational p and q, $p(\varphi) < q(\psi)$.

The notion can be added to the first order update language L that we considered so far. If φ and ψ are formulas of L, $\varphi < \psi$ is a formula of $L(<)$. For L-formulas, one already has $\sigma \models \varphi$ iff $1[\sigma][\varphi] = 1[\sigma]$. Proper $L(<)$-formulas do not define updates, but are interpreted in an extension of $\sigma \models \varphi$ to $L(<)$ defined by P_σ, as above. This seems right. One can come to believe that the probabilities are different from what one thinks by finding out more about what should have been the case if one's experience were more like a proper sample of a correlation, e.g. by more observations or proper statistical evidence, but not by learning facts about inequalities of stochastic parameters, except where these can be interpreted as statements about evidence.

My preference is for the second stricter notion of $\varphi < \psi$, since the first still allows a uniform small advantage to be decisive. Suppose I see somebody and cannot quite decide whether it is Mary or Sue, but have a small preference for it being Mary. Then it does not seem rational to decide it is Mary, since there is still a serious possibility that it is Sue. Under the second interpretation, it will be rational to decide for Mary: the worst probability for Mary is better than the best one for Sue.

The comparison relation is a properly dynamic relation, because of ongoing learning. But it is also non-monotonic for classical reasons. Suppose $\sigma \models \varphi < \psi$. It is still quite possible that one discovers that ψ is not the case and that φ is true. Then $\sigma[\neg\psi \wedge \varphi] \not\models \varphi < \psi$. The contribution to the probability of φ and ψ from the worlds where φ was false and ψ true has been eliminated.

The construction can be applied in semantics, in the formalisation of Bayesian interpretation, in intention recognition, in belief revision and in defaults.

In semantics, it gives new operators $\varphi > \perp$ and $\neg(\varphi < \top)$ and $\varphi > \neg\varphi$ that can perhaps be used in the analysis of epistemic modalities like *may, might* and *must* and for operators like *likely* and *unlikely.*[6] We will not explore this connection further at this point.

Bayesian interpretation as such can be formalised directly in probability theory if the aim is

[6] The observation that there are connections between probability and modality and that gradability of modal judgments is a good argument for exploring these connections is due to Dan Lassiter (talk, Paris, January 2013).

limited to stochastic interpretation. [Zeevat, 2014] makes the case that linguistics provides a largely symbolically definable and largely deterministic causal model of language production that can be used in conjunction with prior probability in Bayesian natural language interpretation. What is described above is a logical reconstruction of such prior probabilities. An axiomatisation of the causal model as a relation $produce(''\varphi'', s)$[7] that can be added to σ interprets a stronger version of Bayesian interpretation as in (2).

(2) $\sigma \models interpret(s, ''\varphi'')$ iff
$\sigma \models \varphi \wedge produce(''\varphi'', s) > \psi \wedge produce(''\psi'', s)$ for all ψ such that
$\sigma \not\models \varphi \leftrightarrow \psi$

The formulation using the stricter definition of $\varphi < \psi$ is a formal version of Gricean intention recognition. In a recognition of e.g. Bill approaching in the corridor there is not just the activation of Bill by the signal and the probability that Bill caused the signal, but the realisation that competing activated objects that by the fact of their activation have a certain probability cannot have enough probability to win from the winner. And that is precisely what the inequality says. Using the weaker interpretation of the inequality, one obtains classical Bayesian interpretation with uncertainty. The strong version can be seen as giving a probabilistic logic of interpretation (comparable to e.g. abductive reasoning as in [Hobbs et al., 1990]) in which one can conduct pragmatics and semantics using linguistic knowledge, knowledge about the speaker, the context and the world and stochastic knowledge, including stochastic relations emerging from linguistics. Pursuing this further is however not the subject of this note.

A third application is belief revision and counterfactuals. The problem is of the same kind as Bayesian interpretation: an optimisation problem. If $\sigma \models \neg\varphi$ and the subject learns φ what should the subject do? She should find the substate τ of σ that is consistent with φ, maximal (there are no larger substates consistent with φ) and has the highest prior probability. With strong inequalities, this puts a presupposition on the use of a counterfactual: that the best revision can be recognised. There will be cases in which belief revision is not determined, because none of the possibilities gain the upperhand in prior probability. The weak interpretation will nearly always give a result, but would not make coordination on the same revision probable in interpreting counterfactuals.

Finally, the strong version of $\varphi > \psi$ can be used as a semantics for defaults.

Under the current view, a default statement like “quakers are pacifists” would come out as “it is more likely beyond uncertainty that a is a quaker and a pacifist than that a is a quaker and not a pacifist”. A set of such statements gives a partial order over propositions.

A reconstruction of default logic should be possible[8]. What is not possible however is a reconstruction of an update logic of defaults, since that would require updates with $\varphi > \psi$ and those are not determinate. We could at most constrain P_0 to meet certain inequalities before refining by σ.

[7] Between codes for the formula expressing the intention “φ” and strings of words s. The classical formalisation is easier on infinite models. One could however interpret production rules as causal correlations with a high probability instead of using implications.

[8] See also the treatment of default reasoning in [Bacchus et al., 1996]

It seems posisble to reconstruct a default logic within the current set-up as in (3). Further work is however necessary.

(3) Hard updates are added to 1.
Inequalities are treated as tests.
presumably φ is defined as $\varphi > \neg\varphi$.
$\varphi_1 \dots \varphi_n, \mathit{ineq}_1 \dots \mathit{ineq}_m \models$ **presumably** ψ
iff
$\forall\sigma(\sigma[\varphi_1]\dots[\varphi_n] \models \mathit{ineq}_1 \dots \mathit{ineq}_m \;\Rightarrow \sigma[\varphi_1]\dots[\varphi_n] \models \psi > \neg\psi)$

The main thesis of this note is that there is a sense in which subjective probabilities are already present in information states for eliminative update semantics. Computing them in order to formalise probabilistic reasoning has a serious potential in semantics, the theory of interpretation and in default reasoning. Such information states have uncertain prior probabilities and can do Bayesian reasoning uniformly to deal with a number of at first sight quite different applications. The second thesis is that the strong criterion for choosing between alternatives is important in all of these applications. It is the difference between gambling on the most likely candidate and rationally discarding candidates that could not win in the light of the currently available evidence.

References

[Bacchus, 1990] Bacchus, F. (1990). Lp, a logic for representing and reasoning with statistical knowledge. *Computational Intelligence*, 6(4):209–231.

[Bacchus et al., 1996] Bacchus, F., Grove, A. J., Halpern, J. Y., and Koller, D. (1996). From statistical knowledge bases to degrees of belief. *Artificial intelligence*, 87(1):75–143.

[Haas-Spohn, 1995] Haas-Spohn, U. (1995). *Indexikalität und subjektive Bedeutung*. Akademie-Verlag, Berlin.

[Heim, 1983] Heim, I. (1983). On the projection problem for presuppositions. In Barlow, M., Flickinger, D., and Westcoat, M., editors, *Second Annual West Coast Conference on Formal Linguistics*, pages 114–126. Stanford University.

[Hobbs et al., 1990] Hobbs, J., Stickel, M., Appelt, D., and Martin, P. (1990). Interpretation as abduction. Technical Report 499, SRI International, Menlo Park, California.

[Karttunen, 1976] Karttunen, L. (1976). Discourse referents. In *Syntax and Semantics Vol. 7*, pages 363–386. Academic Press.

[Kooi, 2003] Kooi, B. P. (2003). Probabilistic dynamic epistemic logic. *Journal of Logic, Language and Information*, 12(4):381–408.

[Pearl, 2000] Pearl, J. (2000). *Causality: Models, Reasoning, and Inference.* Cambridge University Press.

[Van Benthem et al., 2009] Van Benthem, J., Gerbrandy, J., and Kooi, B. (2009). Dynamic update with probabilities. *Studia Logica*, 93(1):67–96.

[Veltman, 1996] Veltman, F. (1996). Defaults in update semantics. *Journal of Philosophical Logic*, 25:221–261.

[Zeevat, 2014] Zeevat, H. (2014). *Production and Interpretation of Natural Language. Linguistics meets Cognition.* Jacob Brill.