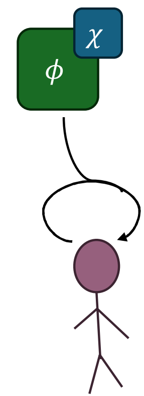
Learning with Confidence

Oliver Richardson

Uncertainty in Artificial Intelligence (UAI) 2025

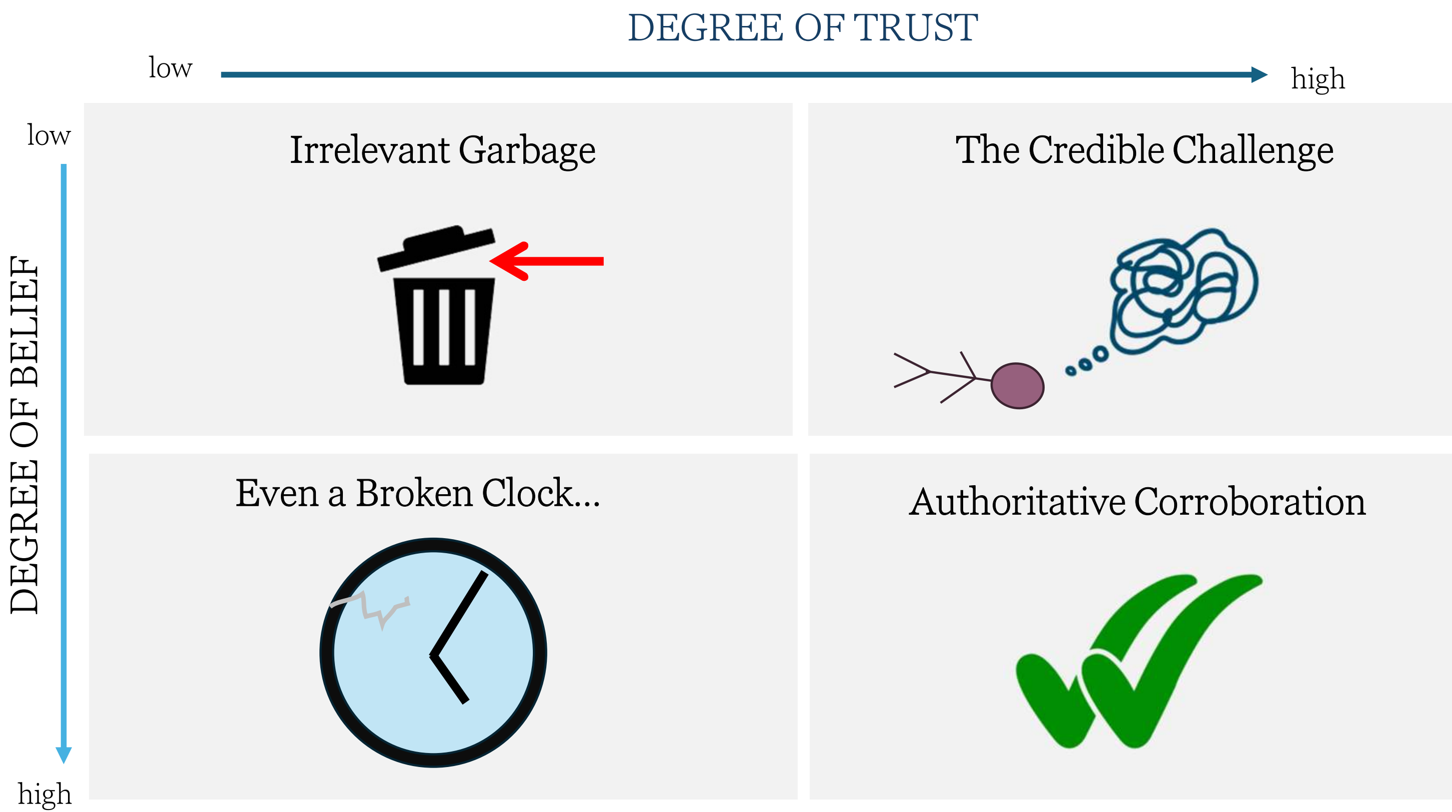
Mila / Université de Montréal



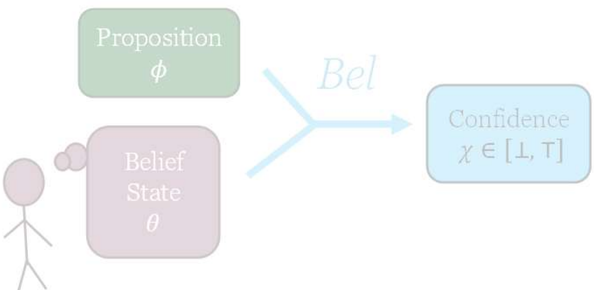
What does it mean (not) to have *confidence* in a statement ϕ?

Two interpretations:

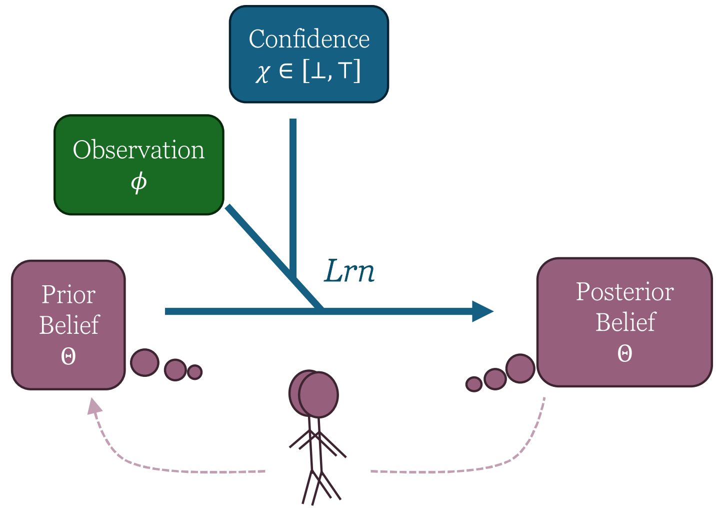
- How likely do I find it? DEGREE OF BELIEF
- How much should it influence my beliefs? DEGREE OF TRUST


DEGREE OF TRUST
low
high
DEGREE OF BELIEF
low
high
Irrelevant Garbage
The Credible Challenge
Even a Broken Clock...
Authoritative Corroboration

DEGREE OF BELIEF



DEGREE OF TRUST



A Simple Example: Linear Interpolation

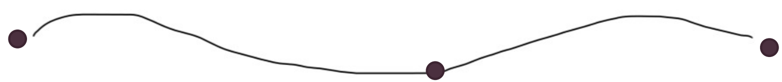
belief states $\mu \in \Theta = \Delta(W)$ are probability measures;

statements $A \in \Phi \subseteq W$ are events;

confidence $\chi \in [0, 1]$ in the unit interval;

Notes:
- no obvious probabilistic interpretation of χ?
- full-confidence update is a projection

$$Lrn(A, \chi, \mu) = (1 - \chi)\mu + \chi(\mu|A)$$



ignore @ no confidence

$$Lrn(A, \bot, \mu) = \mu$$

fully incorporate @ full confidence

$$Lrn(A, \top, \mu) = \mu|A$$

Unifying Existing Concepts

degree of belief in the statement $\chi \in [\bot, \top]$		Prior Belief State θ	Statement ϕ	Learner's Confidence (trust) $\chi \in [\bot, \top]$		Posterior Belief State $\mathrm{Lrn}(\phi, \chi, \theta)$
P(E)	←	probability measure P	event E	(full confidence)	→	conditioned measure
P(E)	←	probability measure P	event E	interpolation	→	partially conditioned measure
$\log P_\theta(y \mid x)$	←	NN params θ	example (x,y)	learning rate; number of epochs	→	trained model
$\mathcal{N}(z \mid x, \sigma^2)$	←	state (x, σ^2) of Kalman Filter	sensor reading z	Kalman Gain	→	tracked position
$Bel(E)$	←	Dempster-Shafer Belief func Bel	event E	weight of evidence	→	combined evidence
…		…	…	…		…

Confidence Domain

$\subset D \times D$

$: D \times D \to D$

$\in D$

$$[\perp, \top] = (D, \leq, \oplus, \top, \perp, \mathfrak{g})$$

preorder

no confidence

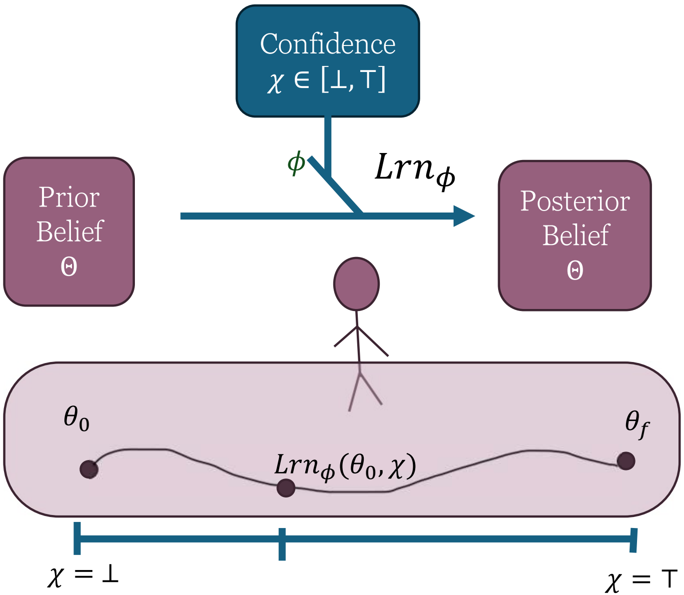
full confidence

geometry
(topology, diffble
structure on D)

independent
combination

$$\begin{aligned}
(\chi \oplus \chi') \oplus \chi'' &= \chi \oplus (\chi' \oplus \chi'') && \text{(associativity),} \\
\perp \oplus \chi &= \chi && \text{(that } \perp \text{ is neutral),} \\
\top \oplus \chi &= \top && \text{(and that } \top \text{ is absorbing).}
\end{aligned}$$

Axioms for Confidence



no confidence [L1] $Lrn_\phi(\bot, \theta) = \theta.$

full confidence [FC] $Lrn_\phi^\top \circ Lrn_\phi^\top = Lrn_\phi^\top.$

continuity [L2] $\chi \mapsto Lrn(\theta, \chi, \phi)$
is continuous, twice diffble

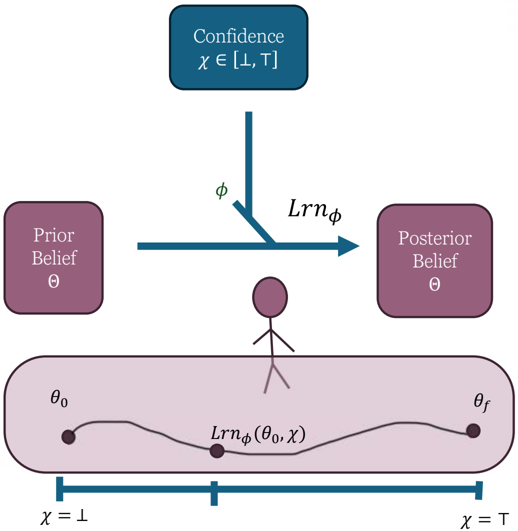
residuals [L3] $\chi < \chi' \implies$
$\exists \chi''.\ \bot < \chi'' \leq \chi'$
$Lrn_\phi^{\chi''} \circ Lrn_\phi^{\chi}(\theta) = Lrn_\phi^{\chi'}(\theta).$

acyclic [L4] If $\chi_0 \leq \chi \leq \chi_1$
and $Lrn_\phi(\chi_0, \theta) = Lrn_\phi(\chi_1, \theta)$,
then $Lrn_\phi(\chi, \theta) = Lrn_\phi(\chi_0, \theta)$.

combinative [L5] $Lrn_\phi(\chi, Lrn_\phi(\chi', \theta))$
$= Lrn_\phi(\chi \oplus \chi', \theta)$

An *action* of the confidence domain

$$\left(D,\ \perp,\ \top,\ \mathfrak{g},\ \leq\ \oplus\ \smile \right.$$



no confidence [L1] $Lrn_\phi(\perp, \theta) = \theta.$

full confidence [FC] $Lrn_\phi^\top \circ Lrn_\phi^\top = Lrn_\phi^\top.$

continuity [L2] $\chi \mapsto Lrn(\theta, \chi, \phi)$ is continuous, twice diffble

residuals [L3] $\chi < \chi' \implies \exists \chi''. \perp < \chi'' \leq \chi'$ $Lrn_\phi^{\chi''} \circ Lrn_\phi^{\chi}(\theta) = Lrn_\phi^{\chi'}(\theta).$

acyclic [L4] If $\chi_0 \leq \chi \leq \chi_1$ and $Lrn_\phi(\chi_0, \theta) = Lrn_\phi(\chi_1, \theta)$, then $Lrn_\phi(\chi, \theta) = Lrn_\phi(\chi_0, \theta).$

combinative [L5] $Lrn_\phi(\chi, Lrn_\phi(\chi', \theta)) = Lrn_\phi(\chi \oplus \chi', \theta)$

Canonical Representations of Confidence

Theorem (additive representation).
If Lrn satisfies [L1-5], then there is a translation $g(\chi,\theta)$ of confidence $\chi \in [\bot,\top]$ to the additive domain $[0,\infty]$ and a learner ${}^{+}Lrn$ such that

$$Lrn(\phi,\chi,\theta) = {}^{+}Lrn(\phi,g(\chi,\theta),\theta)$$

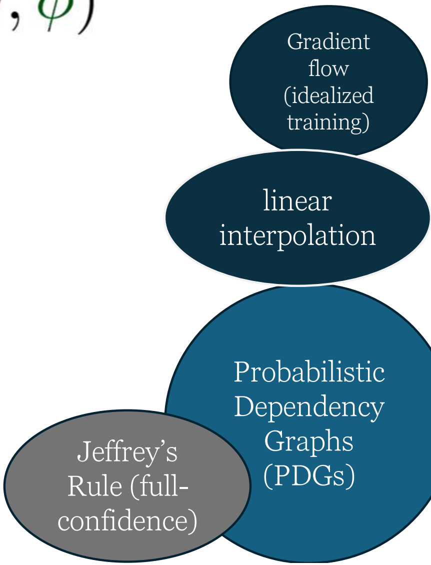
- This "flow form" implies a vector field representations of learners which can be very useful;

Optimizing Learners

[LB4] $$\frac{\partial}{\partial \chi} Lrn(\phi, \chi, \theta) = \nabla_\theta Bel(\theta, \phi)$$

learning is about locally increasing belief,
i.e., gradient descent to minimize loss.

Some examples using relative
entropy and log probability:



What about when learning objective is linear?

Defn (Loss-Linear Learner).
An optimizing learner with a linear objective, i.e., satisfying LB4 with $Bel(\theta, \phi) = \mathbb{E}_\theta\big[\, V_\phi \big]$, in the natural (Fisher) geometry.

Proposition. The additive form of a loss-linear learner is:

$$Boltz(P, \beta, \phi)(w) \propto P(w) \exp\Big(\, \beta\, V_\phi(w) \Big).$$

That is, the posterior is a Boltzman distribution with the prior as the base measure, the confidence as inverse temperature, and the value V_ϕ as the energy.

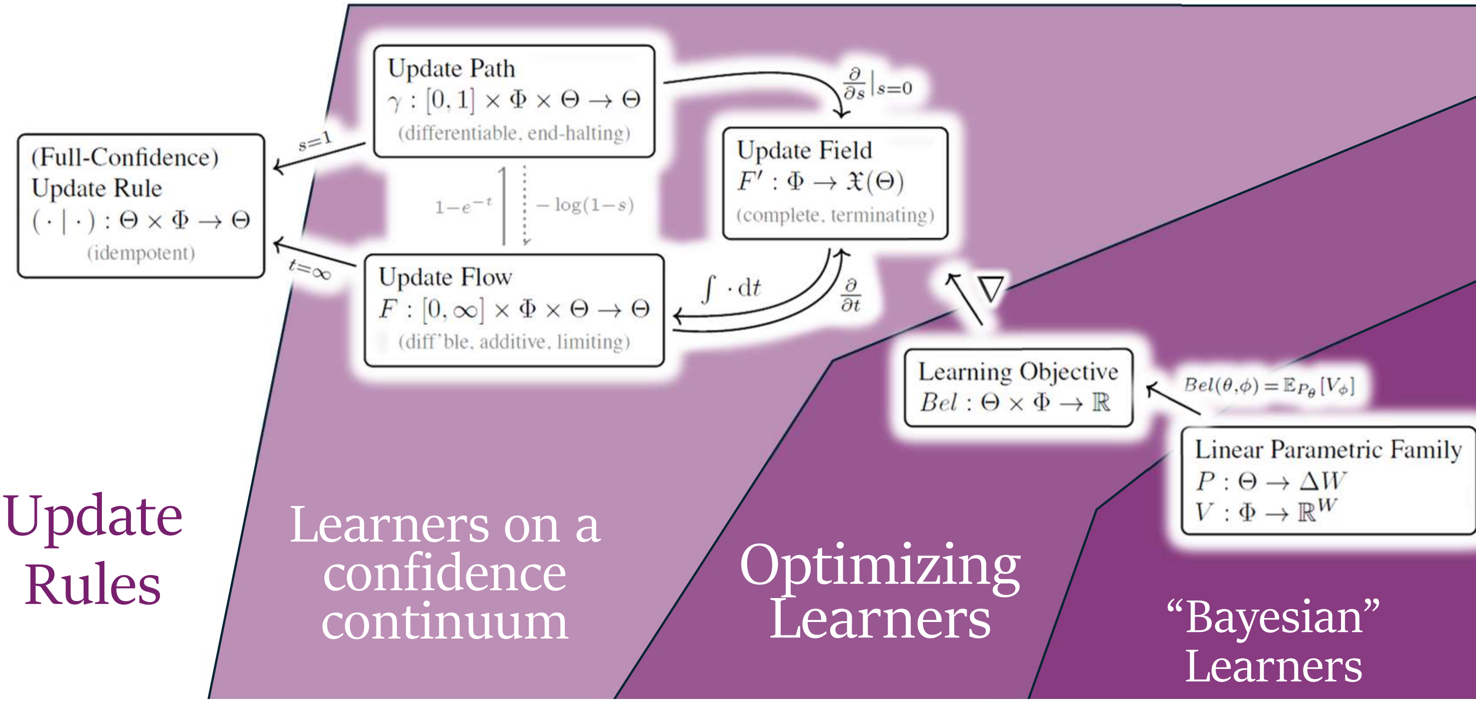
Defn (Bayesian Learner).

- Beliefs correspond to $P(H)$;
- H comes with likelihood $P(\,\phi \mid H\,)$;
- Updates by Bayes Rule: $\exists \star \in [\bot, \top]$.
  $Lrn\big(\phi, \star, P(H)\big) = P(\,H \mid \phi\,) \propto P(\,\phi \mid H\,) P(H)$

Proposition: A learner for probability distributions is Bayesian if and only if it is loss-linear, with

$$V_E(h) = \log P(E|h)$$

Representations of Confidence-based Learners



Conclusion

If certainty is about black and white,
then probability is about shades of gray,
learner's confidence is about transparency.

- Learner's confidence is distinct from likelihood;
- Unifies many concepts in the literature:
  - Sensor precision, Kalman gain, virtual evidence, weight of evidence, thermodynamic coldness, Boltzmann rationality constant β, learning rate, number of epochs, …
- Bayesian updates are a restrictive special case.