Learning with Confidence

Oliver Richardson
Mila / Université de Montréal

Formalism.

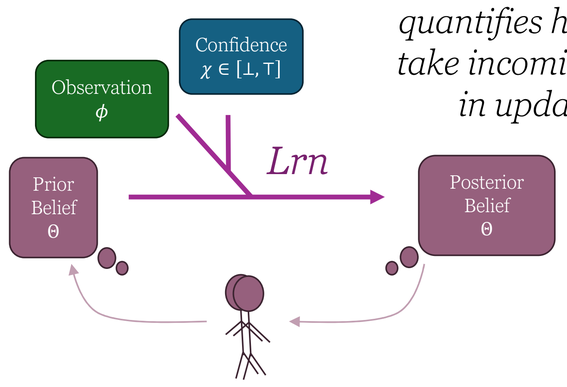
Θ : space of beliefs

Φ : language of observations

$[\bot,\top]$: confidence domain

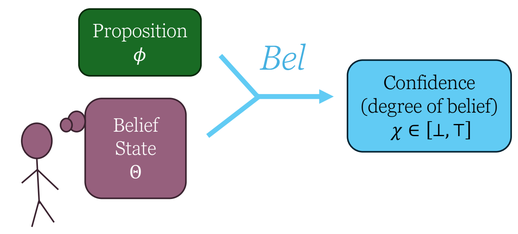
$= \big(D$: a set of confidence values,
$\bot \in D$ represents "no confidence",
$\top \in D$ represents "full confidence",
$\mathfrak{g}$ geometric information about D (e.g., topology and differentiable structure),
$\leq$ a pre-order on D,
$\oplus : D \times D \to D$ an operator that combines two independent degrees of confidence$\big)$

*Confidence as **trust** in incoming information;*
quantifies how seriously to take incoming information in updating beliefs



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

Contrast with the more common meaning of *confidence: as degree of belief* in some proposition



Axioms for (the role of confidence in) Learning

$Lrn : \Theta \times [\bot,\top] \times \Phi \to \Theta$

Lrn_ϕ should be an *action* of the confidence domain on the belief state

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

$\top$ *full confidence ⇒ projection* [FC] $Lrn^\top_\phi \circ Lrn^\top_\phi = Lrn^\top_\phi.$

$\mathfrak{g}$ *smoothness* [L2] $\chi \mapsto Lrn(\theta,\chi,\phi)$ continuous & differentiable; $\theta \mapsto Lrn(\phi,\chi,\theta)$ differentiable wherever continuous

$\leq$ *order and residuals* / *learning is acyclic* [L3] $\chi < \chi' \implies \exists \chi''.\ \bot < \chi'' \leq \chi'$ $Lrn^{\chi''}_\phi \circ Lrn^{\chi}_\phi(\theta) = Lrn^{\chi'}_\phi(\theta).$

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

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

(and Belief)

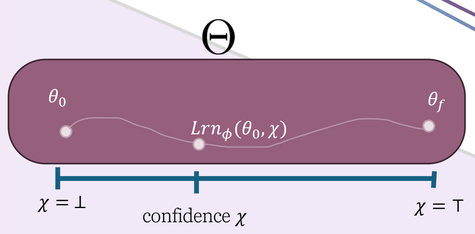
$Bel : \Theta \times \Phi \to [\bot,\top]$

[LB1] Learning with more confidence leads to more belief
$\chi \geq \chi' \implies Bel(\phi, Lrn(\phi,\chi,\theta)) \geq Bel(\phi, Lrn(\phi,\chi',\theta))$

[LB2] If you fully believe something, learning it has no effect
$Bel(\phi,\theta) = \top \implies Lrn(\phi,\chi,\theta) = \theta$

[LB3] After learning something with full confidence, you believe it
$Bel(\phi, Lrn(\phi,\top,\theta)) = \top$

Learners on a Confidence Continuum

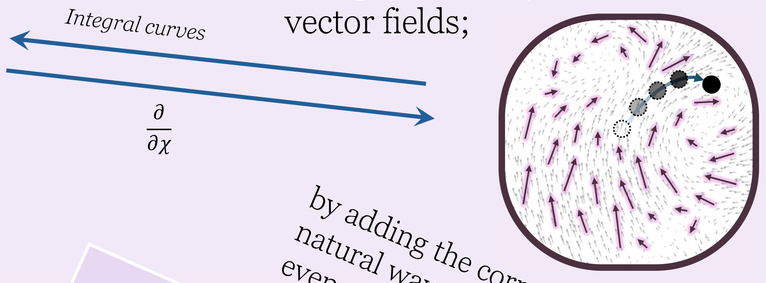


On a continuum (a connected totally ordered 1D manifold), we as well measure confidence additively.

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

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

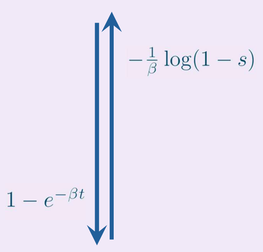
- Additive learners can be represented by vector fields;



by adding the corresponding vector fields, we get a natural way of making observations **simultaneously** even where order (ordinarily) matters.

additive domain

$[0,\infty]$ $\quad t \oplus t' := t + t'$



Proposition (Tempering).
These domains are isomorphic, and with isomorphisms naturally in correspondence with $\beta \in (0,\infty)$

$[0,1]$ $\quad s \oplus s' := s + s'(1-s) = s + s' - s \cdot s'$

fractional domain

Optimizing Learners

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

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

LB4 suffices to ensure simultaneous learning satisfies all axioms (including LB4)

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

Bayesian Learners

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

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

Examples

- Kalman Filtering (general case)
- Dempster's Rule of Combination
- Probabilistic Dependency Graphs (PDGs)
- Gradient flow (idealized training)
- Jeffrey's Rule (full-confidence)
- linear interpolation

optimizing learners for log likelihood and relative entropy

- "Approximate" and variational Bayesian methods
- DS rule of Combination with Simple Support Functions
- Kalman Filtering (1D, optimal gain)

Key Points

Metaphorically: if certainty is black and white, then probability represents shades of gray, and learner's confidence is transparency.

- High-confidence updates (like conditioning) are irreversible projections---but often simplify the posterior belief.
- Confidence is about more than accuracy. For example, when training a resume-screening classifier, one might have low confidence in past hiring decisions if they are discriminatory, even if they are accurate.
- Confidence allows us to be uncertain about observations, which is quite different from making observations that have uncertainty.

Three different kinds of (un)certainty:

1. learner's confidence = trust in incoming information;
2. internal/epistemic confidence = degree of belief;
3. statistical/aleatoric confidence (e.g., sensor precision).