On the Uncertainty Calibration of Equivariant Functions

Edward Berman

Northeastern University

Boston, MA, March 3, 2025

Collaborators


Figure: Jake Ginesin (left) and Robin Walters (right)

Outline

Motivation

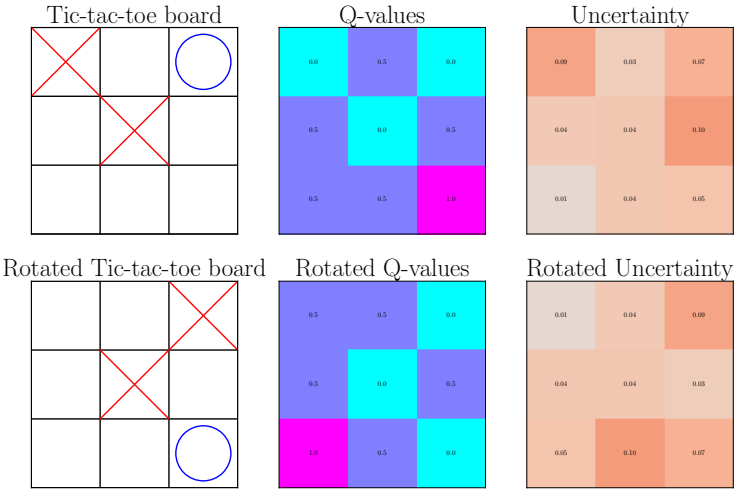


Figure: Equivariance with Uncertainty!

Motivation

Equivariance is a property of a function that allows for us to reason about group symmetries in neural networks and beyond.

- More sample efficient (auto-generalization across symmetry)
-
-

Motivation

Equivariance is a property of a function that allows for us to reason about group symmetries in neural networks and beyond.

- More sample efficient (auto-generalization across symmetry)
- More parameter efficient
-

Motivation

Equivariance is a property of a function that allows for us to reason about group symmetries in neural networks and beyond.

- More sample efficient (auto-generalization across symmetry)
- More parameter efficient
- Are they better **calibrated?**

Motivation

Preliminary works argue yes, but more work required

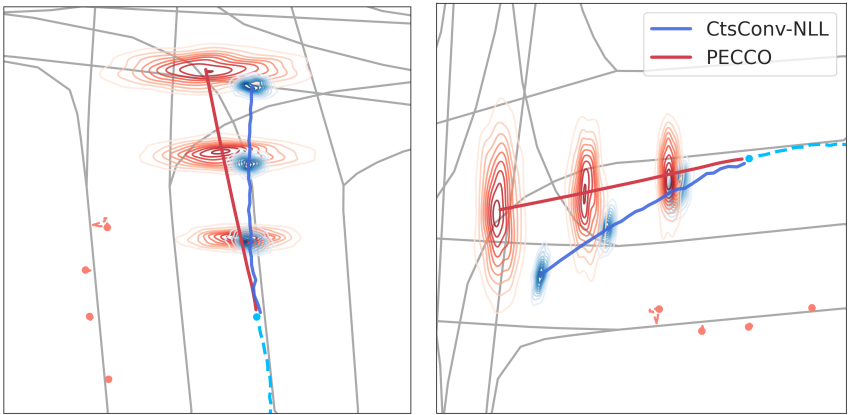


Figure: Equivariant Countour from `https://arxiv.org/pdf/2205.01927`

Outline

Equivariance

Equivariance is a property of an operator $\phi : \mathcal{X} \to \mathcal{Y}$ that maps between input and output vector spaces $\mathcal{X}$ and $\mathcal{Y}$. Given a group G and its representations $\rho^{\mathcal{X}}$ and $\rho^{\mathcal{Y}}$ which transform vectors in $\mathcal{X}$ and $\mathcal{Y}$ respectively, an *operator $\phi : \mathcal{X} \to \mathcal{Y}$ is said to be equivariant if it satisfies the following constraint*

$$\rho^{\mathcal{Y}}(g)[\phi(x)] = \phi(\rho^{\mathcal{X}}(g)[x]) \ , \ \text{for all } g \in G, x \in \mathcal{X} \ . \tag{1}$$

Invariance is a special case of equivariance in which $\rho^{\mathcal{Y}} = \mathcal{I}^{\mathcal{Y}}$ for all $g \in G$. I.e., *an operator $\phi : \mathcal{X} \to \mathcal{Y}$ is said to be invariant if it satisfies the following constraint*

$$\phi(x) = \phi(\rho^{\mathcal{X}}(g)[x]) \ , \ \text{for all } g \in G, x \in \mathcal{X} \ . \tag{2}$$

Equivariance

Examples

1. Galaxy Morphology Classification
2. Chemical Compound Spectral Lines
3. Pick and Place Robotics

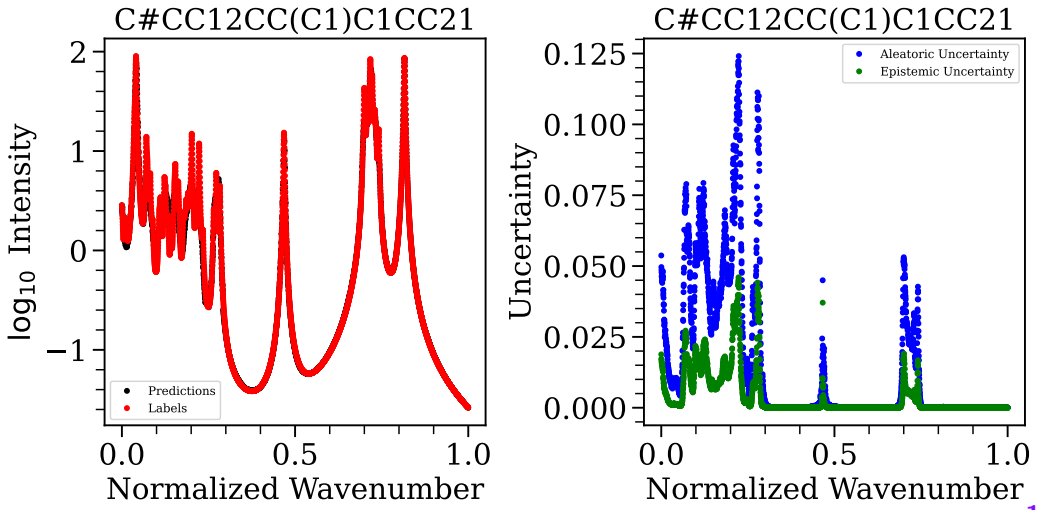
Equivariance

Galaxy Morphology Classification: $E(2)$ Classification Invariance

Figure: Galaxy Zoo Galaxies

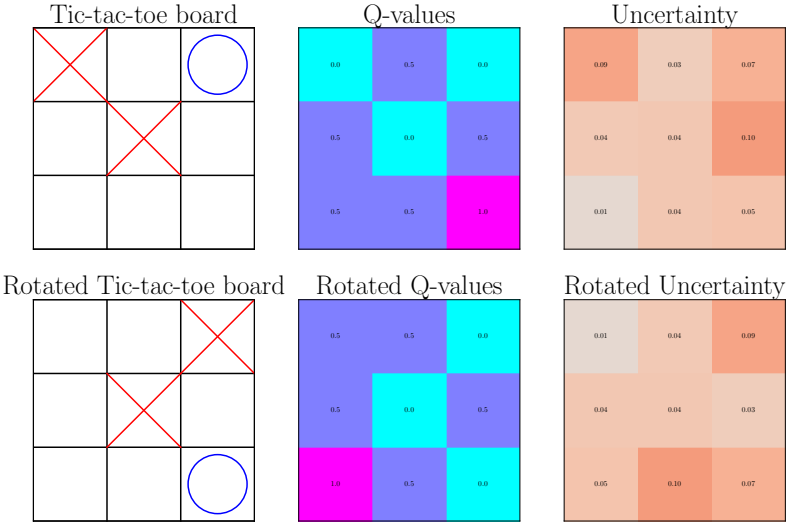
Equivariance

Chemical Compound Spectral Lines: $E(3)$ Regression Invariance



Equivariance

Pick and Place Robotics: $E(3)$ Regression Equivariance



Outline

Uncertainty

As outlined by our last slide, we seek to make rigorous the idea of calibration. We have the ground truth classification given by $f : X \to Y$ and a model class of function $\{h : X \to Y \times P\}$ used to approximate f.

$$\underset{\text{Classification}}{ECE}(h) = \mathbb{E}_{h_P}\left[\left| \mathbb{P}\left(f_Y = h_Y | h_P = p\right) - \underset{\text{Confidence}}{p} \right|\right], \tag{3}$$

we are well calibrated if a model's predicted **confidence** (as given by the logit) matches a model's **accuracy**.

- "If I am 80% confident, I should be correct 80% of the time"

Uncertainty

We can study regression analogously. Now, we have a model class $\{h : X \to \mu \times \sigma^2\}$ for μ, $\sigma^2 \in \mathbb{R}^n$ and $\mathbb{R}^n_+$ respectively. We are well calibrated if

$$\mathbb{E}[\|f - h_\mu\|_2^2 \mid h_{\sigma^2} = \sigma^2] = \sigma^2 \qquad \forall \sigma^2 \in [0, \infty). \tag{4}$$

Uncertainty

Equipped with these notions of uncertainty, we wish to understand how equivariance (or invariance) on the model class $\{h\}$ can effect our calibration.

Outline

Invariant Regression — A Case Study on Chemical Spectra

Our first goal is see if we can upper bound the calibration error for a class of functions $\{h : X \to \mu \times \sigma^2\}$ that are arbitrarily expressive except for being constrained to be invariant with respect to a group G.

Intuition: Let $e(x)$ be the error vector defined by taking the squared error between $f_Y(x)$ and $h_\mu(x)$ on each coordinate, $e(x)_i = (f_Y(x)_i - h_\mu(x)_i)^2$.

$$\mathbb{E}_{x,y}\left[\left\|\sigma^2 - e(x)\right\|_2^2\right] \leq \mathbb{E}_{x,y}\left[\left\|\sigma^2\right\|_2^2 + \left\|e(x)\right\|_2^2\right],$$

the miscalibration is in some sense upper bounded by the error.

Now, we consider the error on individual orbits in the input domain X. Specifically, by assumption of $G-$invariance, $h_{\sigma^2}(x) = \sigma^2 \implies h_{\sigma^2}(gx) = \sigma^2$.
In other words, every variance predicted by the model corresponds to (at least) one orbit in the domain X.

At a high level, we bound the calibration error by considering the regression error on individual orbits induced by σ^2.

Indeed, the upper bound comes out to be

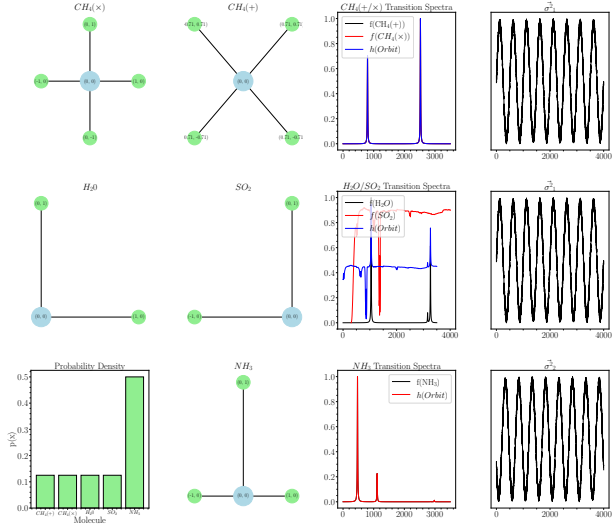
$$ENCE \leq 1 + \mathbb{E}\left[\frac{\mathrm{err}_{\mathrm{reg}}(h, \sigma^2)}{\|\sigma\|_2^2}\right],$$

the average regression error on orbits induced by σ^2!

Invariant Regression — A Case Study on Chemical Spectra

We now explore this on a fabricated example of chemical compounds. For the size and shapes of the compounds,..., hold a suspension of disbelief!

Invariant Regression — A Case Study on Chemical Spectra



Invariant Regression — A Case Study on Chemical Spectra

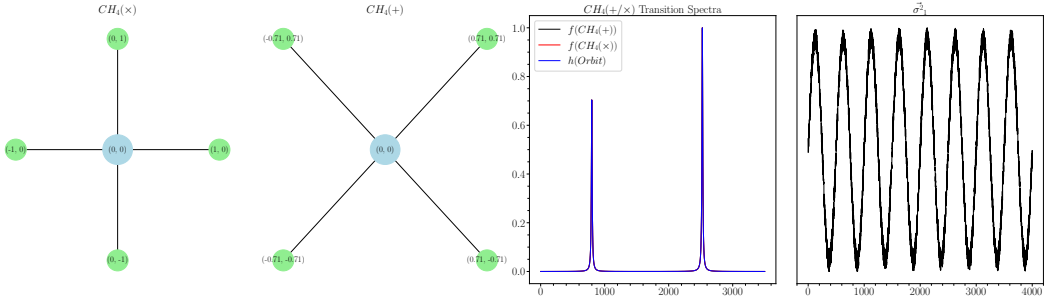


Figure: Methane Spectra

Invariant Regression — A Case Study on Chemical Spectra

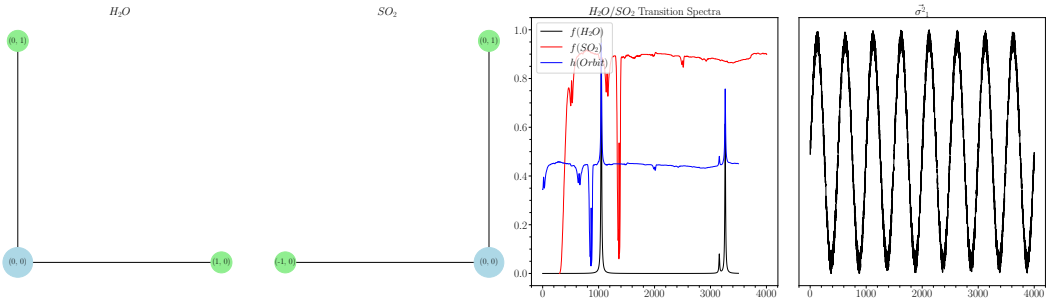


Figure: Polarized Molecule

Invariant Regression — A Case Study on Chemical Spectra

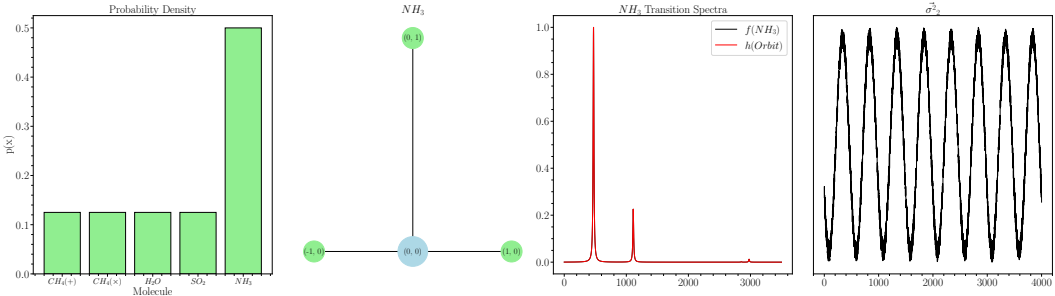


Figure: Ammonia Molecule

Invariant Regression — A Case Study on Chemical Spectra

From theory to practice! We train an $E(3)$ invariant model to predict both a spectral line and an uncertainty estimate on QM9s molecules.

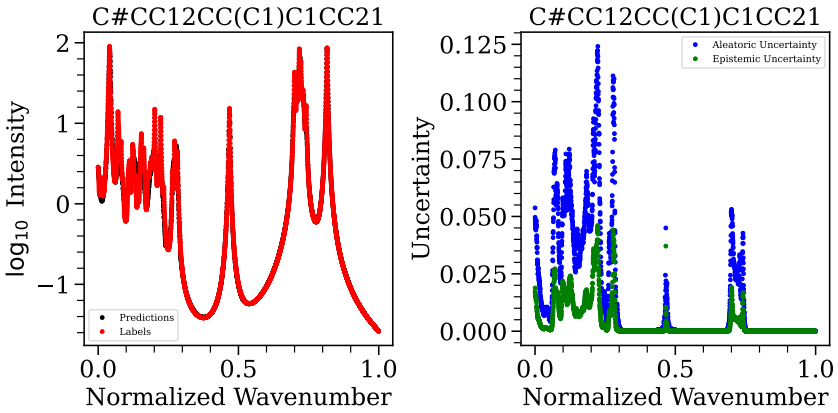


Figure: Spectral Lines

Due to binning approximations, it can be hard to estimate the true calibration error from data. However, regardless of binning approximation, we find that our bound is satisfied!

Invariant Regression — A Case Study on Chemical Spectra

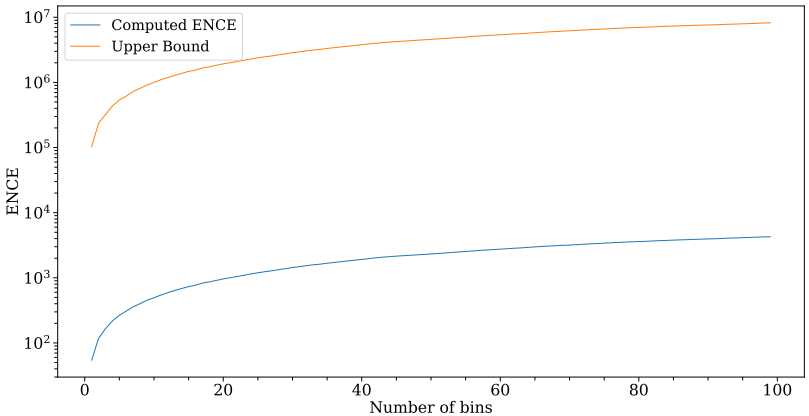
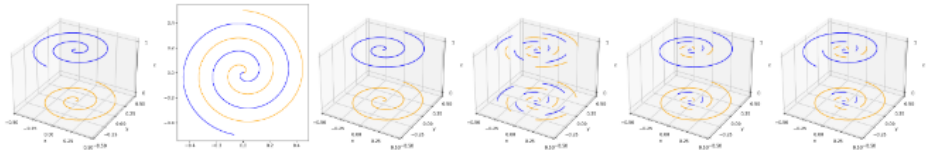


Figure: Upper Bound vs Computation

Outline

Equivariance Type on Swiss Roll Classification



(a) Swiss Roll Data | (b) C_2-Invariant View | (c) Incorrect | (d) Correct | (e) $c = .5, i = .5, e = 0$ | (f) $c = .5, i = 0, e = .5$

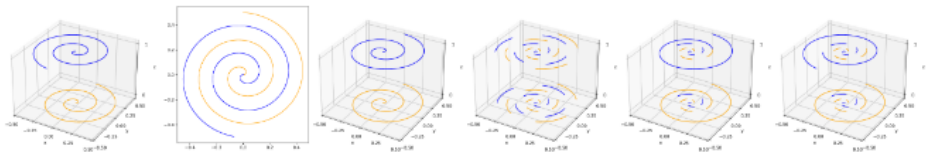
Figure 7: (a) (b) The Swiss Roll data distribution that leads to harmful extrinsic equivariance. (c) (d) The correct and incorrect data distribution in the Swiss Roll experiment. Here the spirals overlap with mismatched and matched labels respectively. (e) (f) Data distribution example with different correct ratio (c), incorrect ratio (i), and extrinsic ratio (e) values.

Figure: Swiss Roll Distributions from `https://arxiv.org/pdf/2303.04745`

Equivariance Type on Swiss Roll Classification

The Goal is to predict the color (blue/yellow) from the position (x,y,z). We sample in different proportions from the Correct and Incorrect Swiss Roll Distributions.

Equivariance Type on Swiss Roll Classification



(a) Swiss Roll Data | (b) C_2-Invariant View | (c) Incorrect | (d) Correct | (e) $c = .5, i = .5, e = 0$ | (f) $c = .5, i = 0, e = .5$

Figure 7: (a) (b) The Swiss Roll data distribution that leads to harmful extrinsic equivariance. (c) (d) The correct and incorrect data distribution in the Swiss Roll experiment. Here the spirals overlap with mismatched and matched labels respectively. (e) (f) Data distribution example with different correct ratio (c), incorrect ratio (i), and extrinsic ratio (e) values.

Figure: Swiss Roll Distributions from `https://arxiv.org/pdf/2303.04745`

Equivariance Type on Swiss Roll Classification

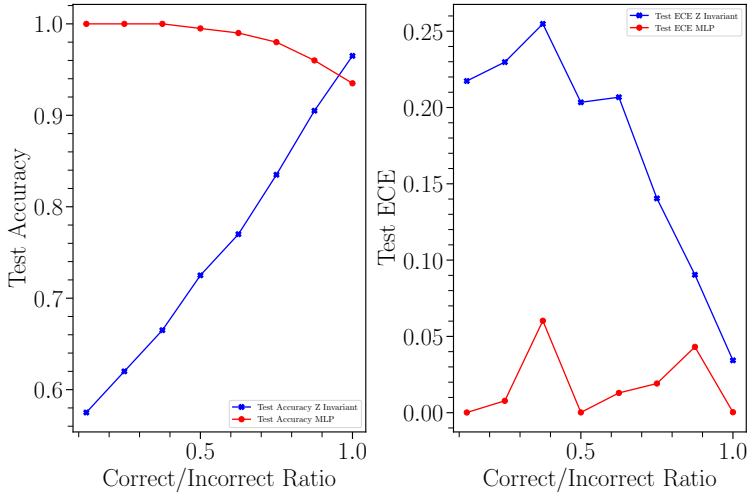


Figure: Accuracy and ECE as a function of correct equivariance in the dataset

Group Order on Galaxy Morphology Classification

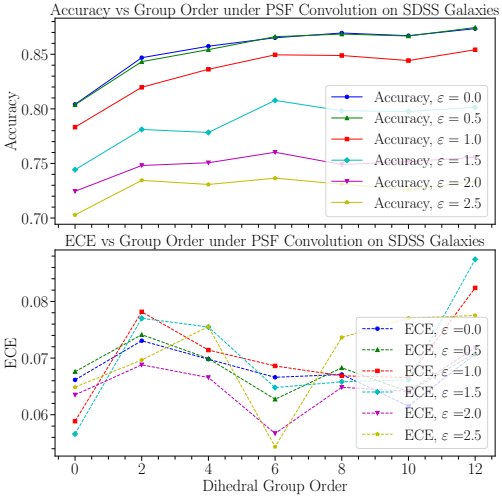


Figure: Galaxy Morphological Accuracy and Calibration vs. Group Order

Group Order on Galaxy Morphology Classification

- We consider accuracy and ECE under various levels of PSF convolution
- Study a model's ability to be accurate and well calibrated

Group Order on Galaxy Morphology Classification

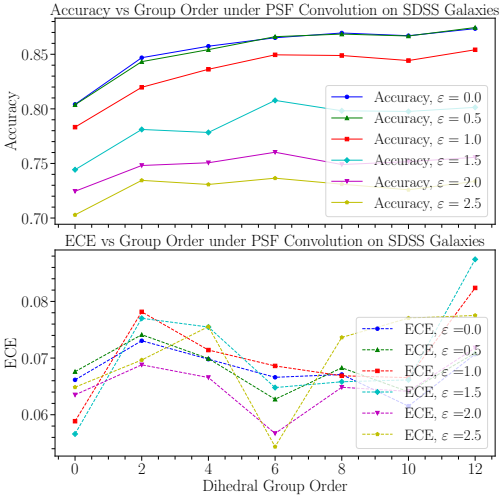


Figure: Galaxy Morphological Accuracy and Calibration vs. Group Order

Group Order on Galaxy Morphology Classification

- This tracks with accuracy having a lower bound related to the action of the group, but ECE for invariant classification does not!

Aleatoric Bleed

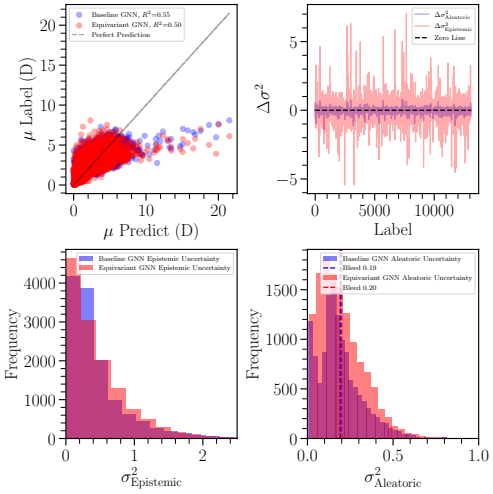


Figure: Aleatoric Bleed for Dipole Moment

Aleatoric Bleed

- We study if equivariant models can better disaggregate different types of uncertainty

Aleatoric Bleed

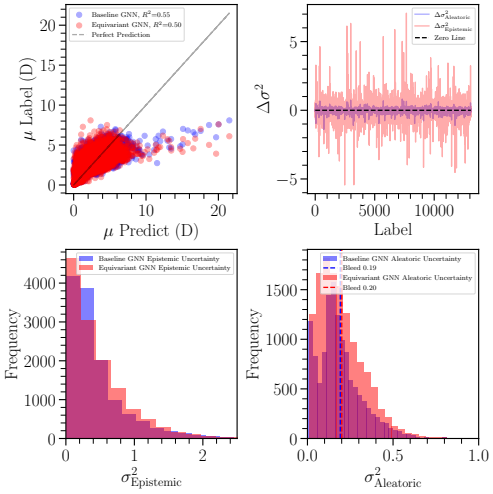


Figure: Aleatoric Bleed for Dipole Moment

Questions? Contact berman.ed@northeastern.edu or visit https://ebrmn.space/

Applying to PhD programs next fall :D

Outline

An uncertainty is... an error bar! Bigger error bars bad

An uncertainty is just an error bar. The easiest way to get an error bar is take the standard deviation of your data points. Consider n data points enumerated x_i, then

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{5}$$

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2} \tag{6}$$

This says nothing about *where* the error came from

The Model, The Data, and all that...

The Model

- Imagine you are fitting a function f with a model h, with parameters θ. We abbreivate h_θ.
- We update θ according to seen data!
- Given finite data, we can only approximate f so well!

Epistemic Uncertainty (Informal)

Uncertainty related to a model's inability to generalize from a finite data set

The Data

- Imagine you are fitting a function f with a model h, with parameters θ. We abbreivate h_θ.
- We update θ according to seen data!
- But the ground truth function f that characterizes the data is naturally probabilistic! e.g. $f = (1)x + (3) + \mathcal{N}(0, 1)$

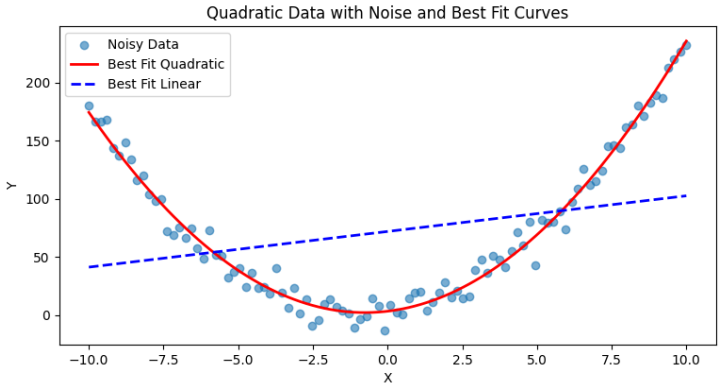
Aleatoric Uncertainty (Informal)

Uncertainty related to the data that can not be reconciled with more data

Your model is trash

Example Model Failure

Most of the literature on uncertainty quantification assumes your model is expressive enough to fit the underlying function with enough data, but this need not be the case!



Formalism

Assume your data points follow some underlying distribution. If I have n samples y_i, then

$$(y_1, \ldots, y_n) \sim \mathcal{N}(\mu, \sigma^2) \tag{7}$$

. We seek to characterize μ and σ^2. **We further impose evidential priors on** μ, σ^2:

$$\mu \sim \mathcal{N}(\gamma, \sigma^2 \nu^{-1}) \tag{8}$$
$$\sigma^2 \sim \Gamma^{-1}(\alpha, \beta) \tag{9}$$

where $\Gamma(\cdot)$ is the gamma function, $\gamma \in \mathbb{R}, \nu > 0, \alpha > 1, \beta > 0$.

The idea is that, putting priors on μ and σ^2 will give us enough parameters to define more rigorously epistemic and aleatoric uncertainties.

Formalism

Consider the posterior

$$q(\mu, \sigma^2) = p(\mu, \sigma^2 | y_1, \dots y_N) \tag{10}$$

With our choice of priors, we assumed $q(\mu, \sigma^2)$ can be factorized into $q(\mu)q(\sigma^2)$. Specifically, we consider the Normal-Inverse Gamma (NIG) distribution, a popular prior for a Gaussian with unknown mean and variance:

$$\begin{aligned} p(\mu, \sigma^2 | \gamma, \nu, \alpha, \beta) &= p(\mu)p(\sigma^2) && (11) \\ &= \mathcal{N}(\gamma, \sigma^2 \nu^{-1}) \Gamma^{-1}(\alpha, \beta) && (12) \\ &= \frac{\beta^\alpha \sqrt{\nu}}{\Gamma(\alpha)\sqrt{2\pi\sigma^2}} \left(\frac{1}{\sigma^2}\right)^{\alpha+1} \exp\left(-\frac{2\beta + \nu(\gamma - \mu)^2}{2\sigma^2}\right) && (13) \end{aligned}$$

Formalism

We now have all the tools we need to redefine the types of uncertainties!

$$\mathbb{E}[\mu] = \int_{\mu=-\infty}^{\infty} \mu p(\mu) d\mu = \gamma \qquad \text{(Prediction)} \tag{14}$$

$$\mathbb{E}[\sigma^2] = \int_{\sigma^2=0}^{\infty} \sigma^2 p(\sigma^2) d\sigma^2 \tag{15}$$

$$= \frac{\beta}{\alpha - 1} \quad \forall \alpha > 1, \qquad \text{(Aleatoric Uncertainty)} \tag{16}$$

$$\mathbb{V}[\mu] = \int_{\mu=-\infty}^{\infty} \mu^2 p(\mu) d\mu - (\mathbb{E}[\mu])^2 \tag{17}$$

$$= \frac{\beta}{\nu(\alpha - 1)} \quad \forall \alpha > 1 \qquad \text{(Epistemic Uncertainty)} \tag{18}$$

Formalism

But wait? How do we optimize for the right γ, ν, α, and β that will simultaneously give us accurate predictions and uncertainty estimates?

1. Solve for $p(y|\gamma, \nu, \alpha, \beta)$
2. Take the negative log likelihood of the distribution, and use it as a loss function to minimize!

See that

$$p(y_i|\gamma,\nu,\alpha,\beta) = \int_{\sigma^2=0}^{\infty}\int_{\mu=-\infty}^{\infty} p(y_i|\mu,\sigma^2)p(\mu,\sigma^2|\gamma,\nu,\alpha,\beta)d\mu\sigma^2 \tag{19}$$

which has the known solution

$$p(y_i|\gamma,\nu,\alpha,\beta) = St(y_i;\gamma,\frac{\beta(1+\nu)}{\nu\alpha},2\alpha) \tag{20}$$

where $St(y;\mu_{st},\sigma^2_{st},\nu_{st})$ is the Student-t distribution evaluated at y with location μ_{St}, scale σ^2_{St}, and ν_{St} degrees of freedom. The St distribution is given by

$$St(t;\mu,\sigma,\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\sigma\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{1}{\nu}\left(\frac{t-\mu}{\sigma}\right)^2\right)^{-(\nu+1)/2}. \tag{21}$$

Formalism

We now seek to maximize the likelihood of $St(y_i; \gamma, \frac{\beta(1+\nu)}{\nu\alpha} 2\alpha)$ by minimizing the negative log likelihood. Let $\Omega = 2\beta(1+\nu)$. We minimize

$$\mathcal{L}_i^{NLL}(w) = \frac{1}{2}\log\left(\frac{\pi}{\nu}\right) - \alpha\log(\Omega) + \left(\alpha + \frac{1}{2}\right)\log((y_i - \gamma)\nu + \Omega) + \log\left(\frac{\Gamma(\alpha)}{\Gamma(\alpha + \frac{1}{2})}\right). \tag{22}$$

And... Voilá!...?

Formalism

So far we have:

1. Solved for the closed for $p(\mu, \sigma^2|\gamma, \nu, \alpha, \beta)$ using NIG prior $\mu \sim \mathcal{N}(\gamma, \sigma^2\nu^{-1})$, $\sigma^2 \sim \Gamma^{-1}(\alpha, \beta)$.
2. Used the closed form to rigorously define the prediction, the aleatoric uncertainty, and the epistemic uncertainty B)
3. Solved for the closed form of $p(y_i|\gamma, \nu, \alpha, \beta)$
4. Took the negative log likelihood of this form for usage as a loss function, which allows us to solve for the $\gamma, \nu, \alpha, \beta$ that best fit the data B)

No.

No.

"Seen from this perspective, one may indeed wonder whether classical probability is the right paradigm for modeling the epistemic part. On the other side, measuring uncertainty and disaggregating total predictive uncertainty into its aleatoric and epistemic components do not necessarily become simpler for generalized formalisms [Hüllermeier et al., 2022]. Clearly, we are not yet at the end of the path toward a truly meaningful uncertainty representation and quantification." – Wimmer et al. 2023

No.

Due to

1. Nuances of training dynamics (iterative approximation algorithms)
2. Lack of access to the ground truth uncertainty in practice
3. Limitations of model expressability (c.f. fitting a linear model to a quadratic)
4. The relationship between aleatoric and epistemic uncertainties via ν

we often can not dissaggregate the two uncertainties very well in practice (Nevin et al. 2024, Valdenegro-Toro and Mori 2022, Jurgens et al. 2024, Wimmer et al. 2023).