AIAA 5047
Responsible AI
2025 Fall

Sihong Xie, AI Thrust, Information Hub

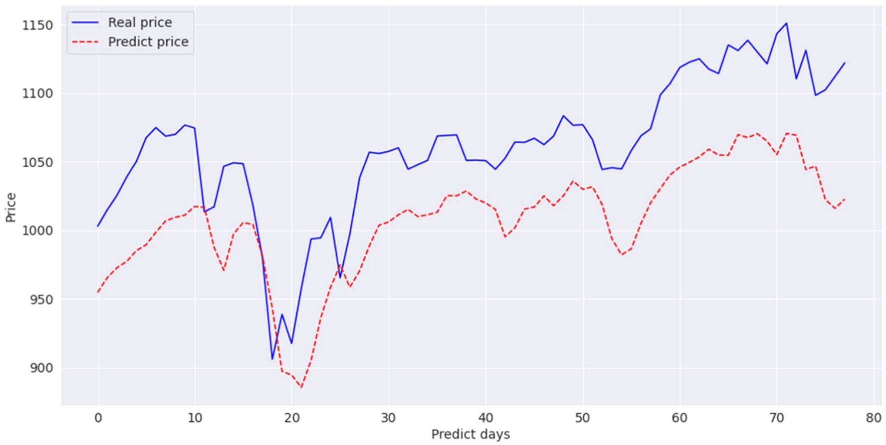
Lecture 8

W2 201, 9-11:50 AM F

Uncertainty in Machine Learning

Example: Fintech

For stock market forecasting, **regression** models analyze **numerical variables** to assist traders in predicting whether stock prices will rise or fall. However, traders prioritize sound investments over high-risk profit-seeking strategies, underscoring prediction uncertainty information.



Fintech: Stock Market Forecasting

Despite the model predicting a rise in stock prices tomorrow, I still want to sell the stocks that have already gone up today.

Beyond model predictions, uncertainty influences decision making.

Phuoc, T., Anh, P.T.K., Tam, P.H. et al. Applying machine learning algorithms to predict the stock price trend in the stock market – The case of Vietnam. Humanit Soc Sci Commun 11, 393 (2024). https://doi.org/10.1057/s41599-024-02807-x

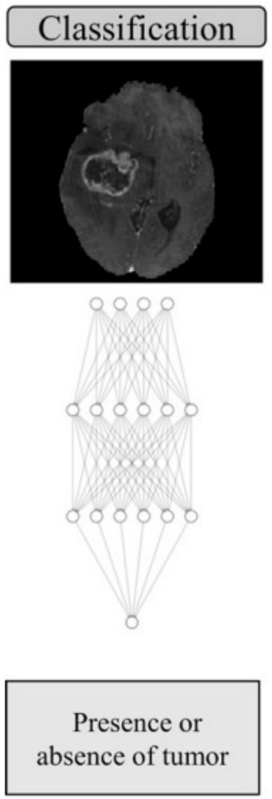
Uncertainty in Machine Learning

Example: Healthcare

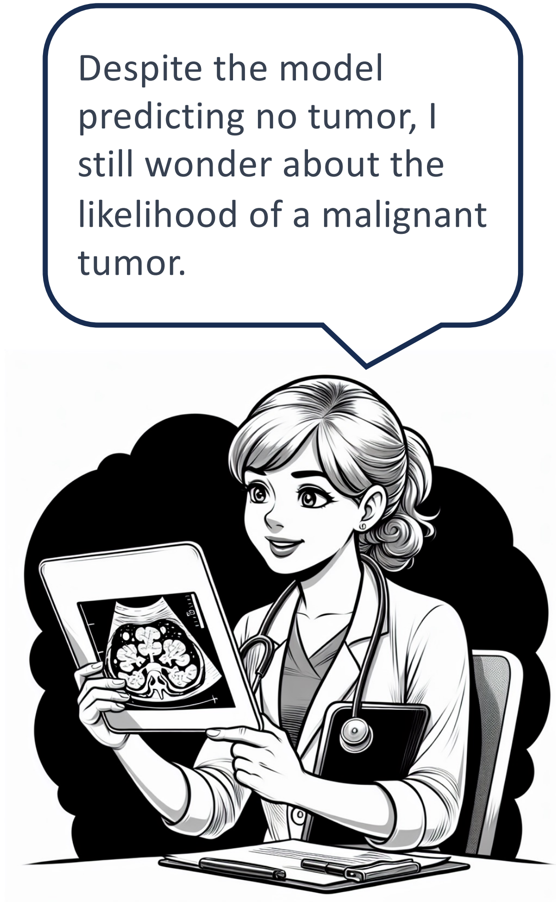
Doctors utilize **classification** models in healthcare to classify medical **images** and assist in diagnosing illnesses.

A binary prediction is insufficient and can mislead the decision making. Uncertainty has complementary information.

Doctors should consider the uncertainty of these predictions when deciding on treatment options for their patients.



Medicine Imagining

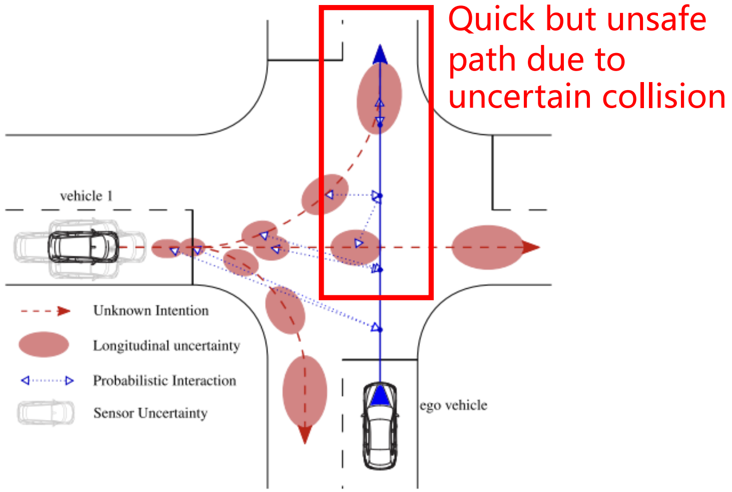


Faghani S, Moassefi M, Rouzrokh P, et al. Quantifying uncertainty in deep learning of radiologic images[J]. Radiology, 2023, 308(2): e222217.

Uncertainty in Machine Learning

Example: Auto-driving

Autonomous driving necessitates models to handle **multimodal inputs**, such as video and vehicle state data, to make driving decisions. The decisions should prioritize risk avoidance and rely on more certain predictions.



Auto-driving: Path Planning

Hubmann, Constantin, et al. "Automated driving in uncertain environments: Planning with interaction and uncertain maneuver prediction." IEEE transactions on intelligent vehicles 3.1 (2018): 5-17.

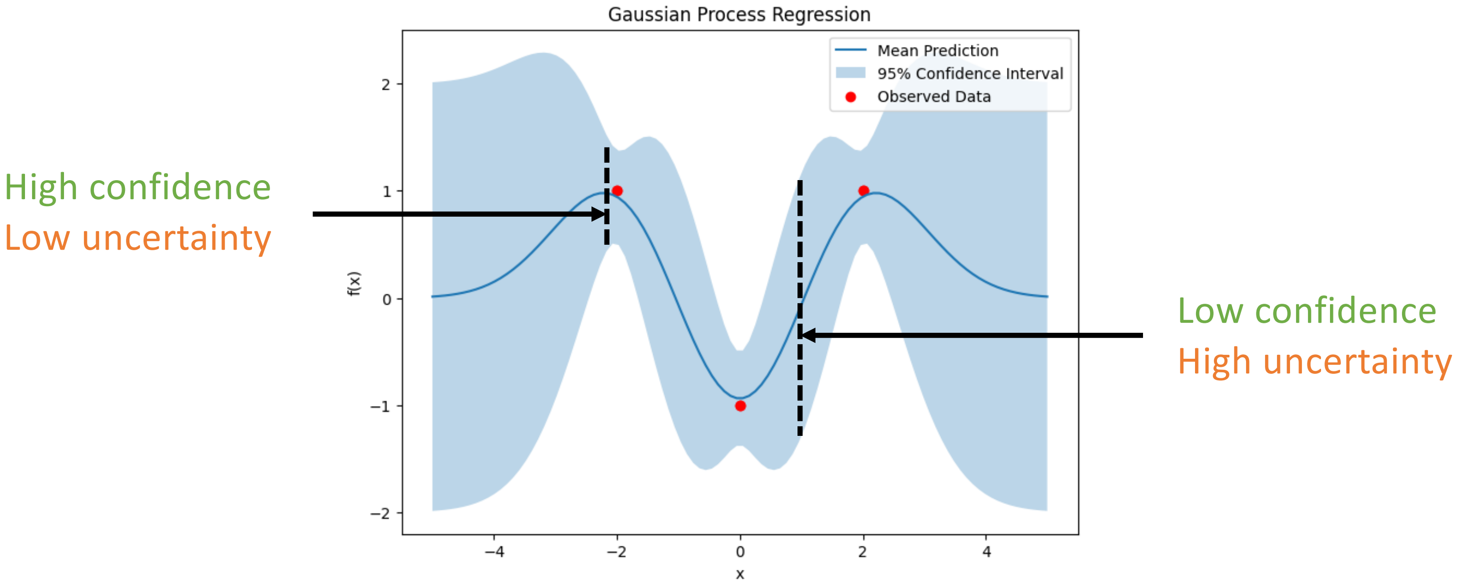
Uncertainty in Machine Learning

Motivation

In high-stakes fields like fintech, healthcare, and auto-driving, stakeholders should consider prediction confidence and uncertainty to make more informed decisions and mitigate potential risks.

Uncertainty is the general concept of how unsure a model is about its prediction. It quantifies the “lack of certainty.” Confidence is a specific measure of certainty (or the complement of uncertainty) for a particular prediction.

$$\text{Uncertainty} \approx 1 - \text{Confidence}$$

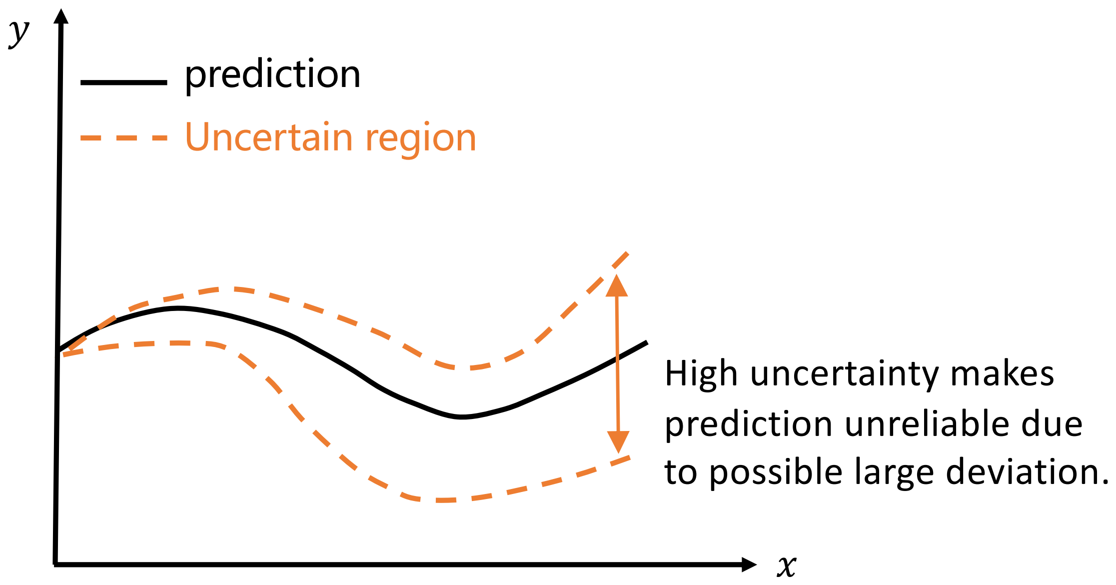


Jain, shashank. (2024, September 3). Understanding gaussian processes: From theory to stock price prediction. Medium. https://pub.aimind.so/understanding-gaussian-processes-from-theory-to-stock-price-prediction-336d6f2aaaab?gi=be5a1431ff81

Uncertainty in Machine Learning

Moreover, when uncertainty is too high, the prediction should be abstained in safety-critical applications.

For example, due to poor visibility in adverse weather conditions, it may be unsafe for the autonomous vehicle to make a decision with high uncertainty. In such cases, the system could be programmed to abstain from taking action. It should pull over to the side of the road until the uncertainty decreases to an acceptable level.



Autonomous driving under different condition

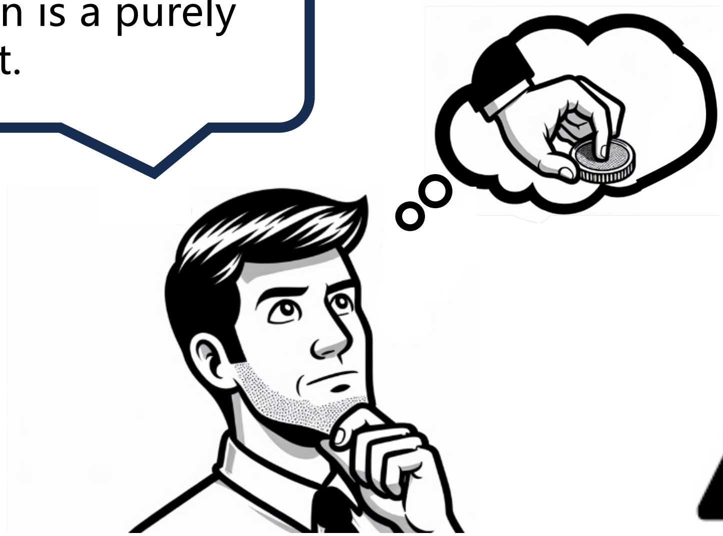
Lee, Unghui, et al. "EureCar turbo: A self-driving car that can handle adverse weather conditions." 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2016

Uncertainty in Machine Learning

Two Sources of Uncertainty: Aleatoric and Epistemic

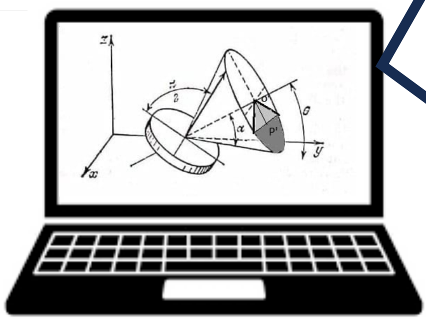
Aleatoric uncertainty arises from inherent data noise, whereas epistemic uncertainty stems from limited knowledge and is reducible with more data. However, **the two kinds can switch under different contexts.**

Flipping a coin is a purely random event.



Aleatoric Uncertainty

If I know flipping force, the coin mass, volume,initial position, and all the necessary physical information. I can simulate the process and predict whether it would be head or tail.



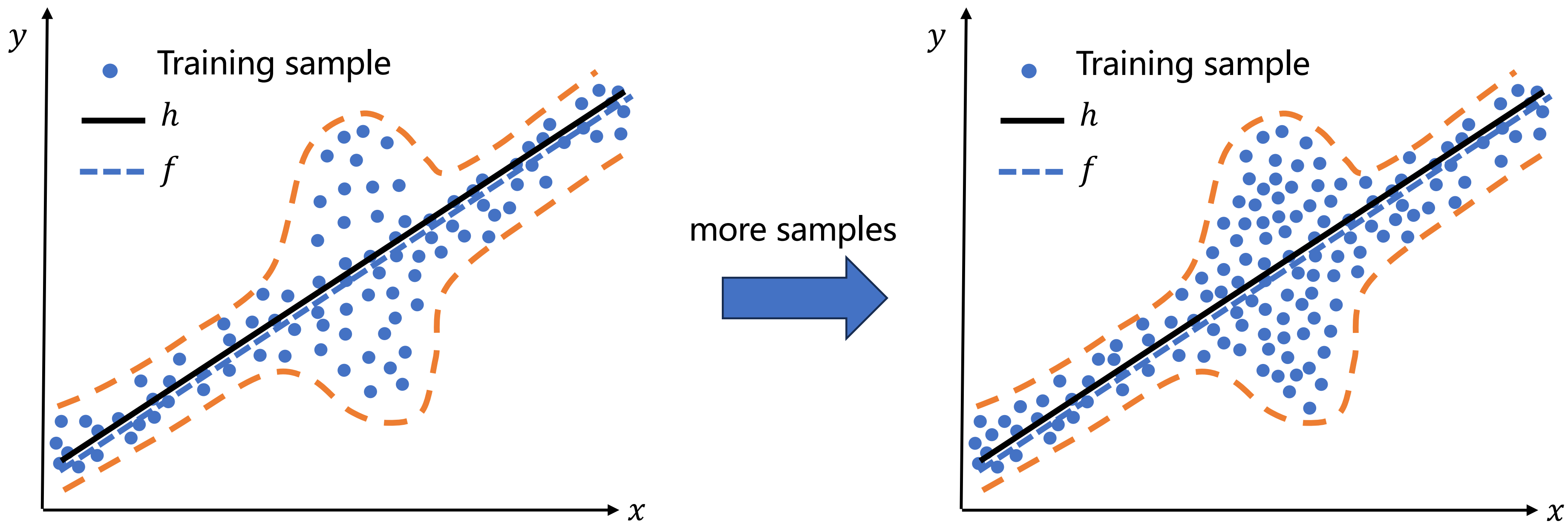
Epistemic Uncertainty

Physics of coin tossing and uncertainty: Cheenta probability series. cheenta academy. (n.d.).

Uncertainty in Machine Learning

Aleatoric uncertainty

Aleatoric uncertainty is the randomness inherent to the data generation process and is irreducible with increasing amount of data (i.e. more samples).



Aleatoric uncertainty is not reduced with more training samples

Gruber C, Schenk P O, Schierholz M, et al. Sources of Uncertainty in Machine Learning--A Statisticians' View[J]. arXiv preprint arXiv:2305.16703, 2023.

Uncertainty in Machine Learning

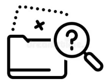
Four sources of aleatoric uncertainty:

$$\begin{bmatrix} X_1^{(1)} & X_1^{(2)} \\ X_2^{(1)} & X_2^{(2)} \\ X_3^{(1)} & X_3^{(2)} \\ X_4^{(1)} & X_4^{(2)} \\ X_5^{(1)} & X_5^{(2)} \\ X_6^{(1)} & X_6^{(2)} \\ X_7^{(1)} & X_7^{(2)} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \end{bmatrix}$$

Correct data

$$\begin{bmatrix} X_1^{(1)} & X_1^{(2)} \\ X_2^{(1)} & X_2^{(2)} \\ X_3^{(1)} & X_3^{(2)} \\ X_4^{(1)} & X_4^{(2)} \\ X_5^{(1)} & X_5^{(2)} \\ X_6^{(1)} & X_6^{(2)} \\ X_7^{(1)} & X_7^{(2)} \end{bmatrix}$$

Omitted Features
(unobservable data)



$$\begin{bmatrix} X_1^{(1)} & X_1^{(2)} \\ X_2^{(1)} & X_2^{(2)} \\ X_3^{(1)} & X_3^{(2)} \\ X_4^{(1)} & X_4^{(2)} \\ X_5^{(1)} & X_5^{(2)} \\ X_6^{(1)} & X_6^{(2)} \\ X_7^{(1)} & X_7^{(2)} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \end{bmatrix}$$

Missing Data
(incomplete datasets)

$$\begin{bmatrix} X_1^{(1)} + \delta_X & X_1^{(2)} + \delta_X \\ X_2^{(1)} + \delta_X & X_2^{(2)} + \delta_X \\ X_3^{(1)} + \delta_X & X_3^{(2)} + \delta_X \\ X_4^{(1)} + \delta_X & X_4^{(2)} + \delta_X \\ X_5^{(1)} + \delta_X & X_5^{(2)} + \delta_X \\ X_6^{(1)} + \delta_X & X_6^{(2)} + \delta_X \\ X_7^{(1)} + \delta_X & X_7^{(2)} + \delta_X \end{bmatrix}$$

Feature with random error δ_X
(low-resolution images)

$$\begin{bmatrix} Y_1 + \delta_Y \\ Y_2 + \delta_Y \\ Y_3 + \delta_Y \\ Y_4 + \delta_Y \\ Y_5 + \delta_Y \\ Y_6 + \delta_Y \\ Y_7 + \delta_Y \end{bmatrix}$$

Label with random error δ_Y
(biased labeling)

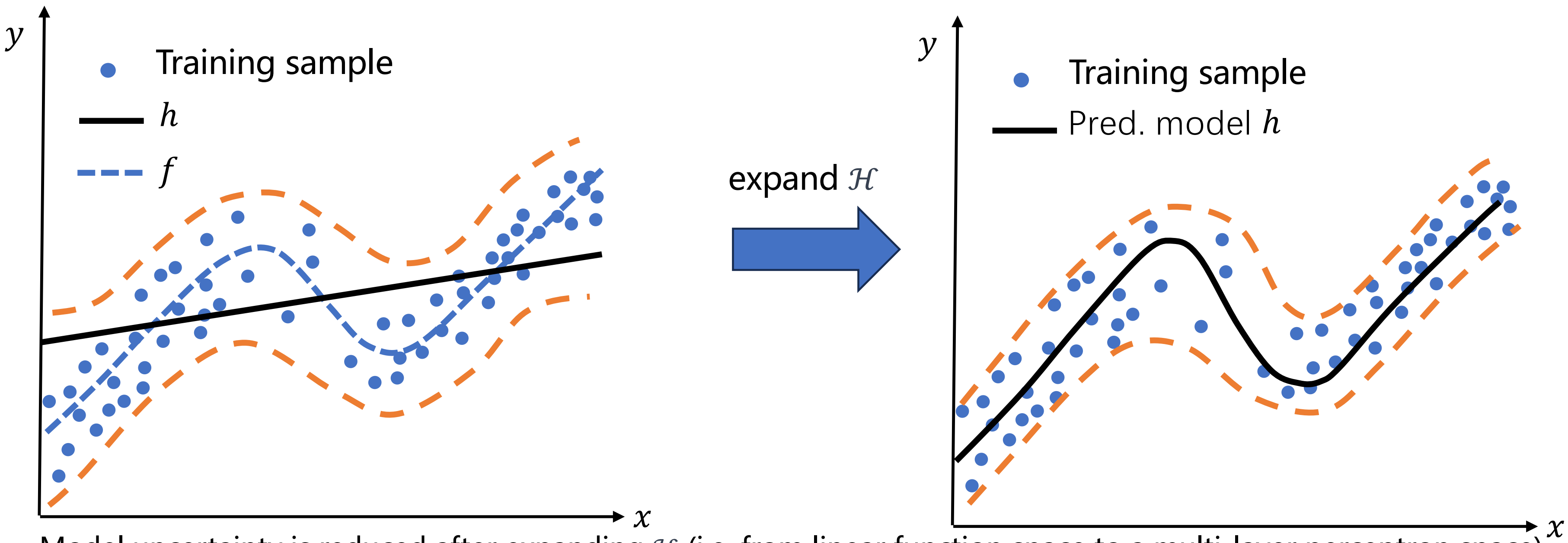
Gruber C, Schenk P O, Schierholz M, et al. Sources of Uncertainty in Machine Learning--A Statisticians' View[J]. arXiv preprint arXiv:2305.16703, 2023.

Uncertainty in Machine Learning

Epistemic uncertainty

Epistemic uncertainty is reducible uncertainty, rooted in the lack knowledge. It can be further decomposed into **model uncertainty** and **approximation uncertainty**.

Model uncertainty measures the difference between the ground truth function f and hypothesis $h \in \mathcal{H}$. It can be reduced by expanding function space $\mathcal{H}$. This is the "underfitting" problem in machine learning.

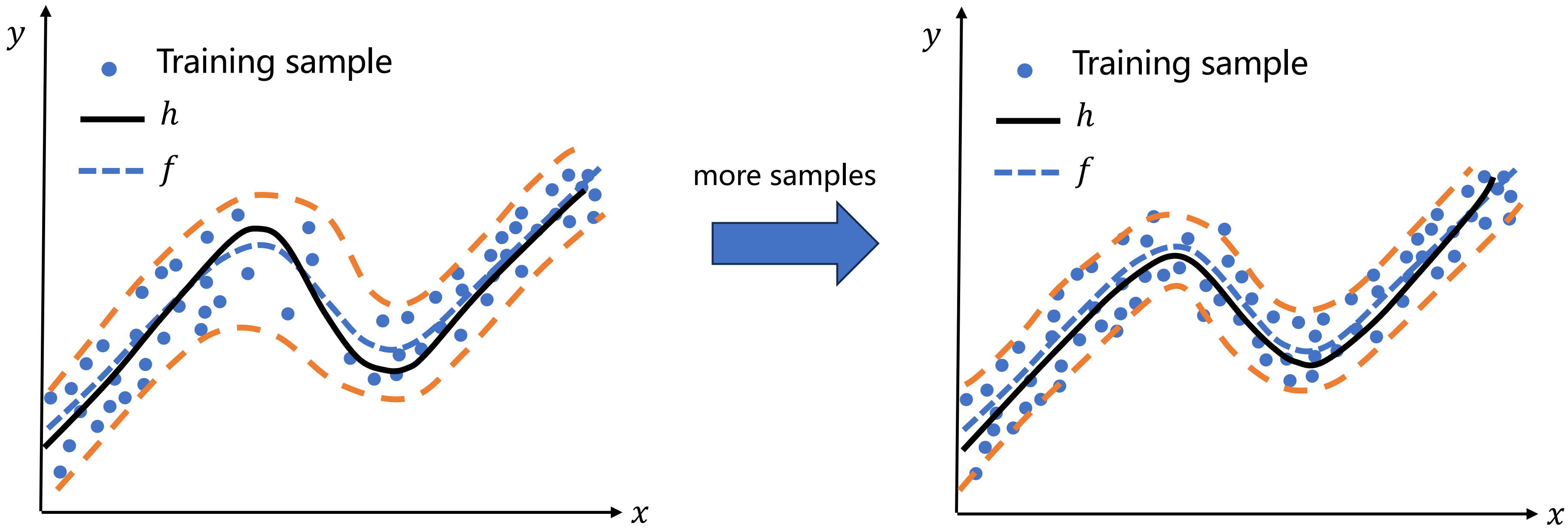


Model uncertainty is reduced after expanding $\mathcal{H}$ (i.e. from linear function space to a multi-layer perceptron space)

Gruber C, Schenk P O, Schierholz M, et al. Sources of Uncertainty in Machine Learning--A Statisticians' View[J]. arXiv preprint arXiv:2305.16703, 2023.

Uncertainty in Machine Learning

Approximation uncertainty is due to the limited number of training samples. It can be reduced with access to more training data.



Approximation uncertainty is reduced with more training samples

Gruber C, Schenk P O, Schierholz M, et al. Sources of Uncertainty in Machine Learning--A Statisticians' View[J]. arXiv preprint arXiv:2305.16703, 2023.

Uncertainty Quantification Methods

Bayesian Approach

Bayesian methods is a pivotal approach of representing uncertainty based on Bayes'Theorem, which updates the probability of Hypothesis with Evidence.

Example: Steve is a shy American. How likely is he to be a librarian?

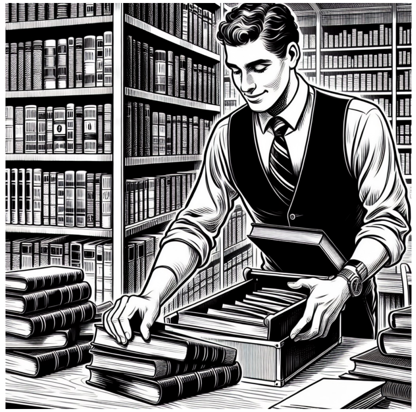
Hypothesis: Steve is a librarian.

Evidence: Steve is a shy.

You want: How likely is Steve to be a librarian?

$P(\text{Hypothesis given Evidence}) = P(H|E)$

Given: (a) the population ratio of librarian and non-librarian is 1:20. (b) 40% of librarians meet the description and only 10% of non-librarians would.



Is Steve a librarian?

Posterior $P(H|E)$

$$P(H|E) = \boxed{\frac{P(H)P(E|H)}{P(E)}} = \frac{P(H)P(E|H)}{P(H)P(E|H) + P(\neg H)P(E|\neg H)} = \frac{\underbrace{\frac{1}{20+1}}_{\text{Prior } P(H)}\times \underbrace{0.4}_{\text{Likelihood } P(E|H)}}{\underbrace{\frac{1}{20+1}}_{P(H)}\times\underbrace{0.4}_{P(E|H)}+\underbrace{\frac{20}{20+1}}_{P(\neg H)}\times\underbrace{0.1}_{P(E|\neg H)}}$$

Bayes' Theorem

Uncertainty Quantification Methods

Bayesian neural networks (BNNs)

BNNs include stochastic parameters θ with distribution $P(\theta)$.

In BNNs, the Hypothesis is θ and the Evidence is training data D, with inputs D_X and labels D_Y.

With Bayes' Theorem, we have $P(\theta|D) = \frac{P(\theta)P(D_Y|D_X,\theta)}{P(D_Y|D_X)}$.

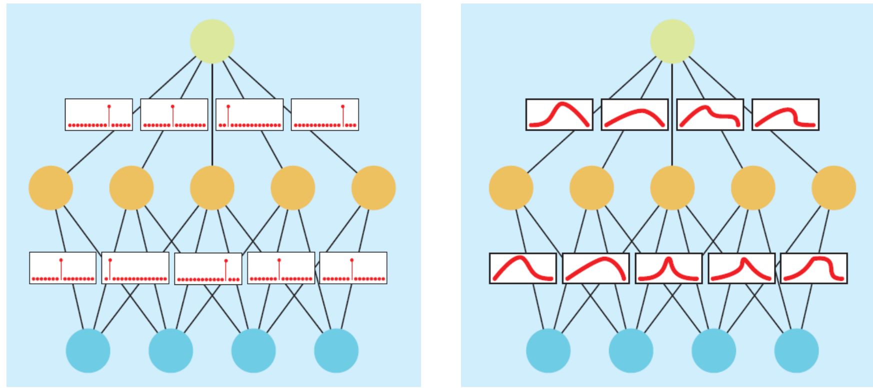
Training: To maximize the posterior $P(\theta|D)$, BNNs learn to maximize $P(\theta)P(D_Y|D_X,\theta)$.

Prediction: with input x, BNNs output

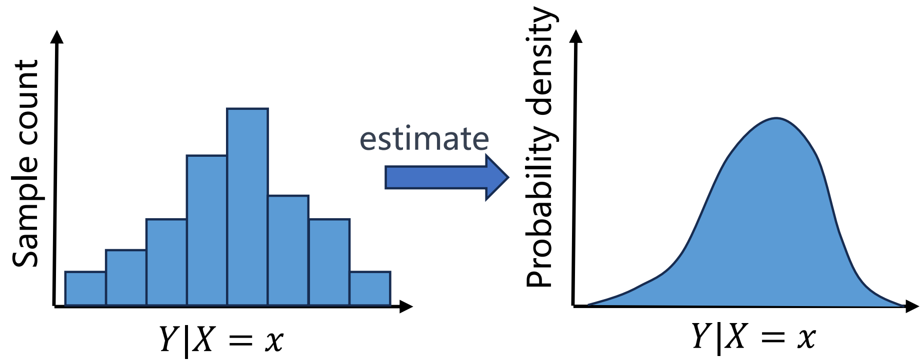
$$\hat{P}(Y|X=x) = \int_{\theta} \hat{P}(Y|X=x,\theta)P(\theta')d\theta'$$

to quantify prediction uncertainty, which is estimated by Monte Carlo sampling in practice.

Advantage: Flexibility with different choices of priors; Approximation of complex posteriors with simpler, tractable distributions.
Disadvantage: Computational complexity in posterior optimization.



Deterministic Parameters Stochastic Parameters



Monte Carlo Sampling for Prediction Uncertainty Estimation

Jospin L V, Laga H, Boussaid F, et al. Hands-on Bayesian neural networks—A tutorial for deep learning users[J]. IEEE Computational Intelligence Magazine, 2022, 17(2): 29-48.

Uncertainty Quantification Methods

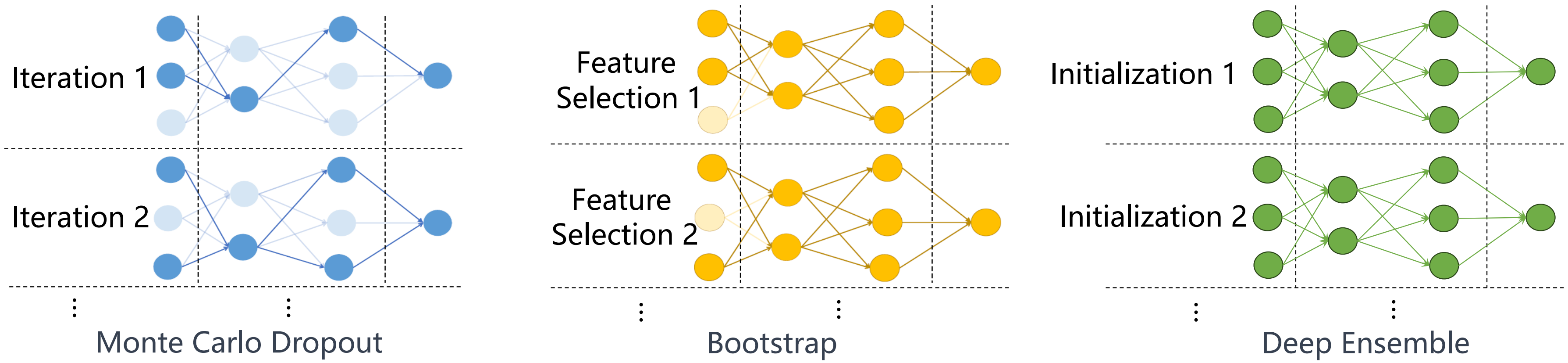
Monte Carlo Dropout

MC dropout is a convenient surrogate of Bayesian deep learning. It uses dropout layers to create variation in outputs.

Some other ensemble methods:

Bootstrap trains multiple models on different parts of the training data (i.e. feature selection) and then analyzes the variance in predictions across these models.

Deep ensemble trains multiple models with different random initializations independently on the same data and then aggregates their predictions.



Advantage: Simple implementation.

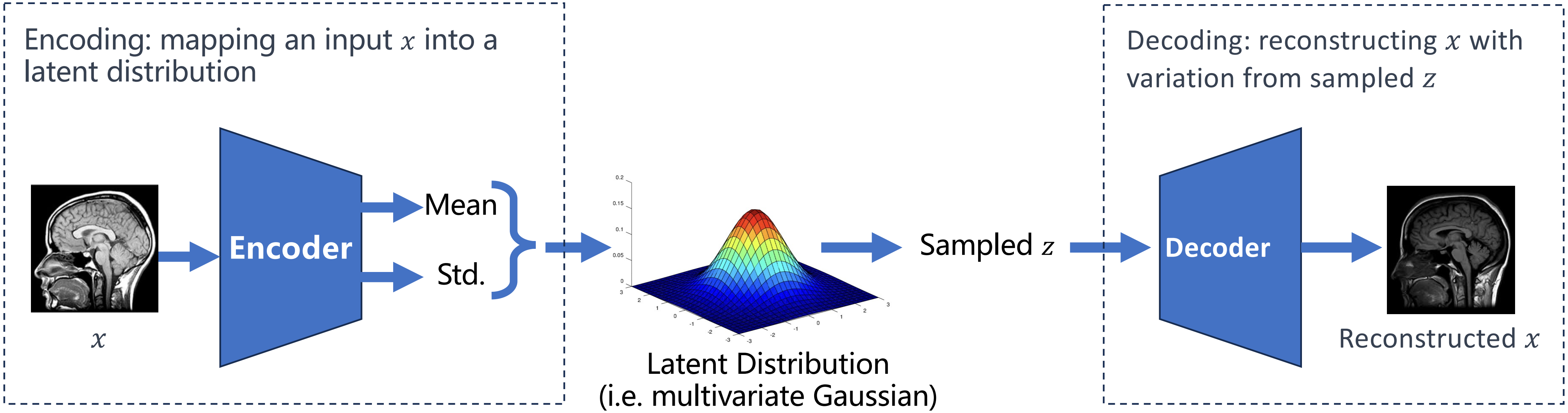
Disadvantage: Limited flexibility with single source of stochasticity. Lack of theoretical guarantees.

Abdar M, Pourpanah F, Hussain S, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges[J]. Information fusion, 2021, 76: 243-297

Uncertainty Quantification Methods

Variational Autoencoders (VAEs)

VAEs are generative models to learn the probability distribution of a given dataset.
A VAE model consists of an encoder and a decoder.



Advantage: Generative modeling for new samples.
Disadvantage: Lack of uncertainty estimation: VAEs do not provide a direct measure of uncertainty in predictions.

Abdar M, Pourpanah F, Hussain S, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges[J]. Information fusion, 2021, 76: 243-297

Uncertainty Quantification Methods

Closed-form UQ for Bayesian Regression Model

A scoring rule is an extended real-valued function S.

Denote the forecast $\hat{P}(Y|X = x)$ by $\hat{P}$ for simplicity with density $\hat{p}$ and CDF $F_{\hat{P}}$.

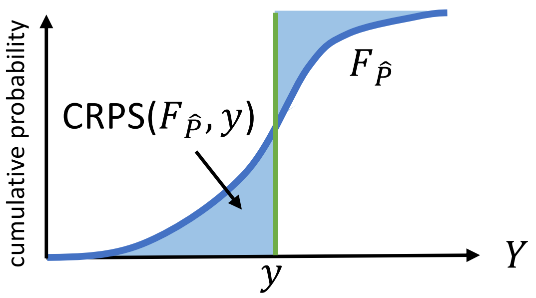
If true value is y, the forecaster's penalty is $S(\hat{P}, y)$. Different scoring rules $S(\hat{P}, y)$ are listed below.

Name	**Definition**
Continuous ranked probability score	$\mathrm{CRPS}(\widehat{P}, y) = \int_{\mathbb{R}} \big(F_{\widehat{P}}(t) - \mathbb{I}\{y \leq t\}\big)^2 \, dt$
Logarithmic score	$\mathrm{LS}(\widehat{P}, y) = -\log \widehat{p}(y)$
Quadratic score	$\mathrm{QS}(\widehat{P}, y) = -2\,\widehat{p}(y) + \int_{\mathbb{R}} \widehat{p}(t)^2 dt$
Squared error score	$\mathrm{SE}(\widehat{P}, y) = \big(y - \mathbb{E}_{Y \sim \widehat{P}}[Y]\big)^2$

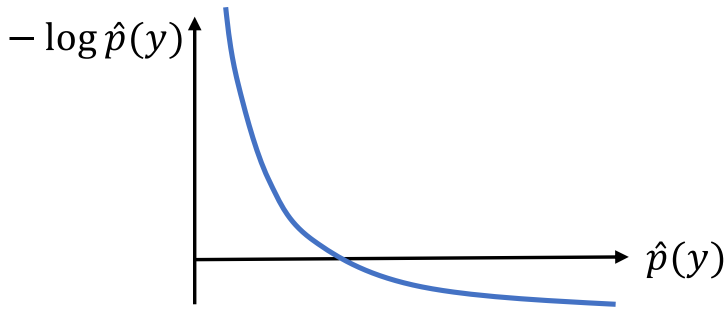
Fishkov, Alexander, et al. "Uncertainty Quantification for Regression using Proper Scoring Rules." arXiv preprint arXiv:2509.26610 (2025).

Uncertainty Quantification Methods

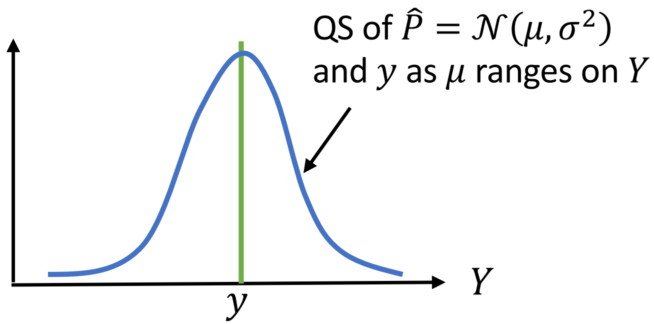
Closed-form UQ for Bayesian Regression Model



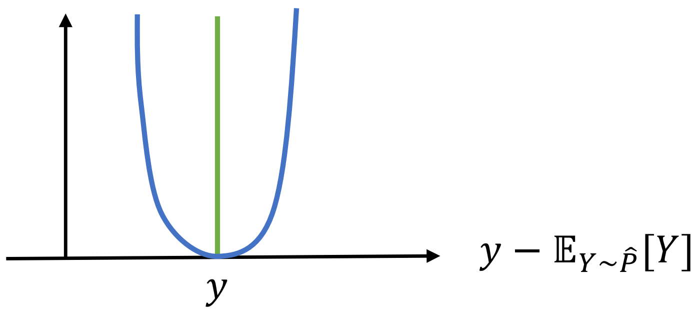
Continuous Ranked Probability Score (CRPS)



Logarithmic Score



Quadratic Score (QS)



Squared Error Score

Uncertainty Quantification Methods

Closed-form UQ for Bayesian Regression Model

We assess the predictive quality (**total uncertainty**) of the forecast $\hat{P}$ on average over possible values of y.

$$S(\widehat{P}, P) = \int S(\widehat{P}, y)\, dP(y)$$

where P is the true distribution of $P(Y|X = x)$.

Aleatoric uncertainty is the smallest possible error that one can achieve as

$$H(P) = S(P, P) = \int S(P, y)\, dP(y).$$

which is **entropy** if we define S as logarithmic scoring rule.

Epistemic uncertainty is due to the lack of knowledge about $\hat{P}$

$$d(\widehat{P}, P) = S(\widehat{P}, P) - S(P, P) = S(\widehat{P}, P) - H(P),$$

quantifying how much worse it is to predict with distribution $\hat{P}$ than with the true data distribution P.

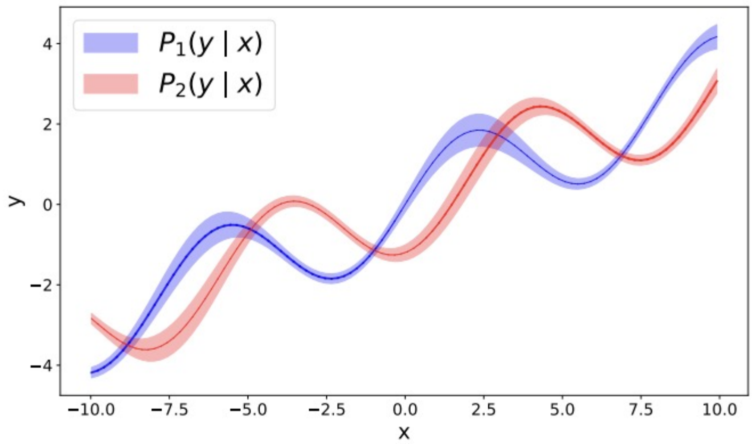
Fishkov, Alexander, et al. "Uncertainty Quantification for Regression using Proper Scoring Rules." arXiv preprint arXiv:2509.26610 (2025).

Uncertainty Quantification Methods

Closed-form UQ for Bayesian Regression Model

We show an example here. The ground-truth conditional is a heteroscedastic two-component mixture,

$$p^*(y \mid x) = \pi(x) P_1(y \mid x) + \big(1 - \pi(x)\big) P_2(y \mid x)$$

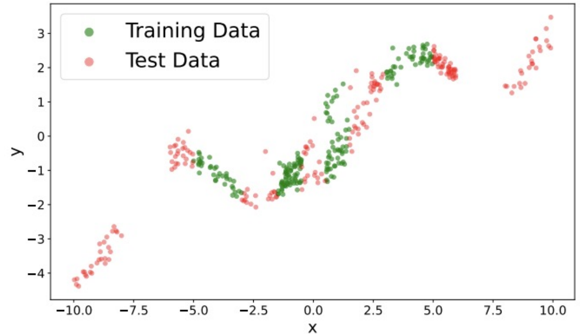


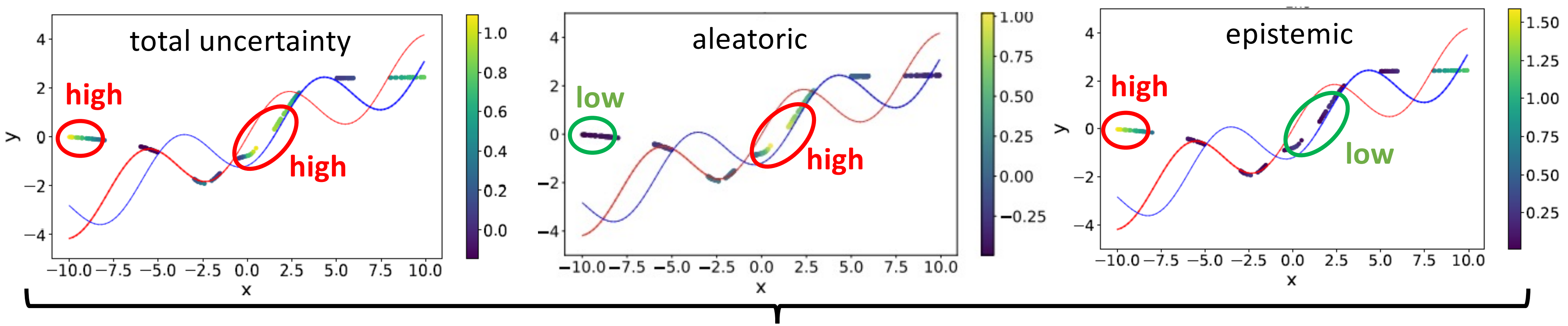
Fishkov, Alexander, et al. "Uncertainty Quantification for Regression using Proper Scoring Rules." arXiv preprint arXiv:2509.26610 (2025).

Uncertainty Quantification Methods

Closed-form UQ for Bayesian Regression Model

We draw training pairs from $p^*(y \mid x)$ and fit a regression network, and use test data to run inference.





test inputs at their predictive mean; color indicates the corresponding uncertainty

Fishkov, Alexander, et al. "Uncertainty Quantification for Regression using Proper Scoring Rules." arXiv preprint arXiv:2509.26610 (2025).

Uncertainty Quantification Methods

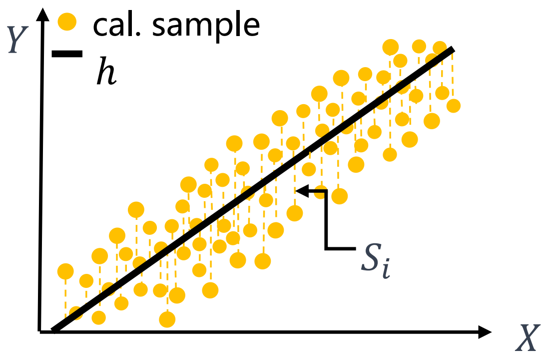
Conformal Prediction (CP)

CP uses calibration data to output $1-\alpha$ confidence set of the prediction of a test input.

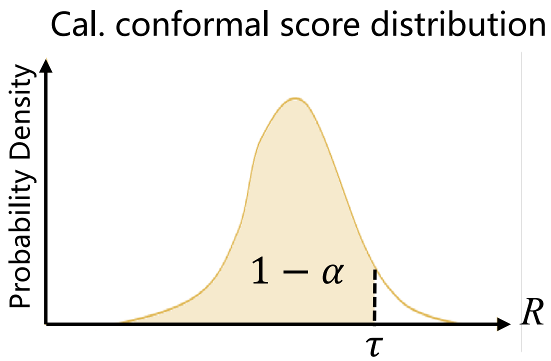
In regression tasks, with calibration samples $\{(X_i, Y_i)\}_{i=1}^{n}$ and a trained model h, we have the residuals $S_i = |h(X_i) - Y_i|$, namely conformal scores. More formally, we represent $S_i = s(X_i, Y_i)$, where $s(\cdot,\cdot)$ is score function.

τ is the empirical $1-\alpha$ quantile, which is the $\lceil(1-\alpha)n\rceil$-th largest element of $\{S_i\}_{i=1}^{n}$.

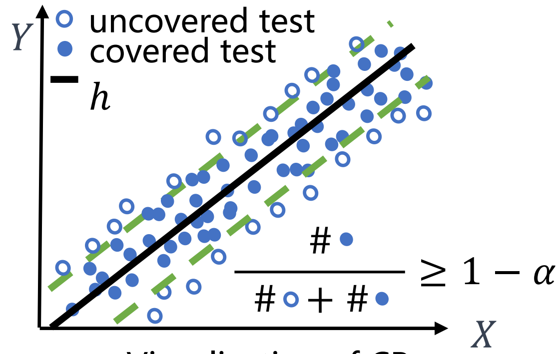
With an exchangeable (share the same distribution of calibration data) test sample (X_{n+1}, Y_{n+1}), prediction set is $C(X_{n+1}) = \{y: |h(X_{n+1}) - y| \le \tau\}$, and the **coverage guarantee** is $\Pr\left(Y_{n+1} \in C(X_{n+1})\right) \ge 1-\alpha$.



Conformal scores can be residuals



τ is empirical $1-\alpha$ quantile



Visualization of CP Prediction Interval

Advantage: CP is Post-hoc method with simple implementation; CP does not rely on specific assumptions about data distribution.

Disadvantage: Relying on exchangeable calibration and test samples; Lack of adaptiveness to different input values.

Angelopoulos A N, Bates S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification[J].arXiv preprint arXiv:2107.07511, 2021.

Uncertainty Quantification Methods

Proof of coverage guarantee

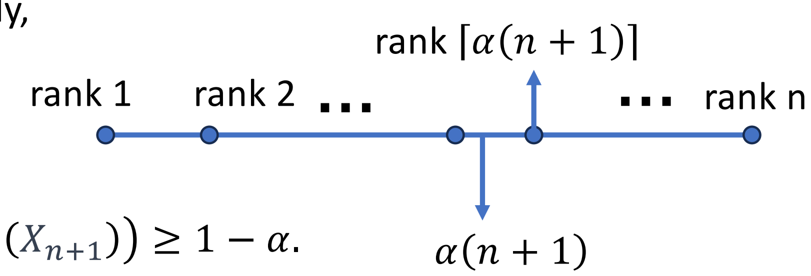
Conformal scores $S_1, \dots, S_n, S_{n+1}$ are exchangeable random variables.

Because of exchangeability, the rank of S_{n+1} among the $n+1$ values is uniformly distributed over $\{1,2, \dots, n+1\}$.

$$\Pr(\mathrm{rank}(S_{n+1}) = k) = \frac{1}{n+1}, \quad \forall k \in \{1, \dots, n+1\}.$$

Now, $S_{n+1} \leq \tau$ iff S_{n+1} is not among the largest $\lceil \alpha(n+1) \rceil$ scores. Equivalently,

$$\Pr(S_{n+1} \leq \tau) = \frac{\lceil (1-\alpha)(n+1) \rceil}{n+1}$$



Since $\frac{\lceil (1-\alpha)(n+1) \rceil}{n+1} \geq 1-\alpha,$ we have the guaranteed coverage $\Pr\big(Y_{n+1} \in C(X_{n+1})\big) \geq 1-\alpha.$

Intuition

Because the test score S_{n+1} is exchangeable with calibration scores, its rank is uniformly random. By taking the $(1-\alpha)$-quantile threshold τ, we ensure that at most an α-fraction of scores fall above it. Thus, the probability the new test score exceeds it is at most α.

Angelopoulos A N, Bates S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification[J].arXiv preprint arXiv:2107.07511, 2021.

Uncertainty Quantification Methods

Calibration

With label space $\mathcal{Y} = \{1, \ldots, K\}$, a classification models outputs a prediction as an probability vector $(v^1, \ldots, v^K)$.

The predicted label is $\hat{y} = \arg\max_{k \in \mathcal{Y}} v^k$.

The confidence of the prediction is $c = \max_{k \in \mathcal{Y}} v^k$.

A model is well-calibrated if the probability of correct prediction (i.e. reliability) equals the prediction confidence:

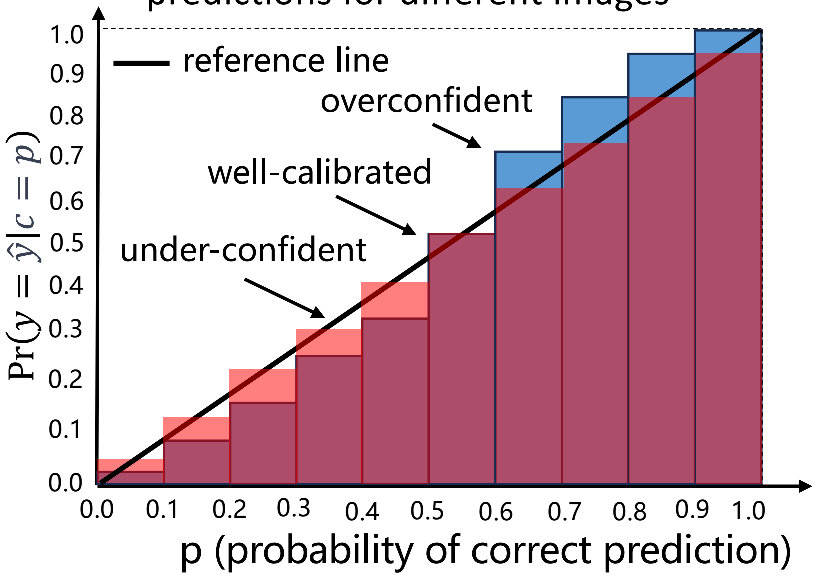
$$\Pr(y = \hat{y} | c = p) = p, \forall p \in [0,1].$$

Histogram binning: Grouping n calibration samples into different bins according to their confidences in predicted labels and plotting the bins. The plot informs if the model is well-calibrated.

Advantage: High Interpretability; Compatibility with various models.

Disadvantage: typically not applicable to regression tasks.

Distinct probability vector predictions for different images



Guo C, Pleiss G, Sun Y, et al. On calibration of modern neural networks[C]//International conference on machine learning. PMLR, 2017: 1321-1330.

Use of Uncertainty

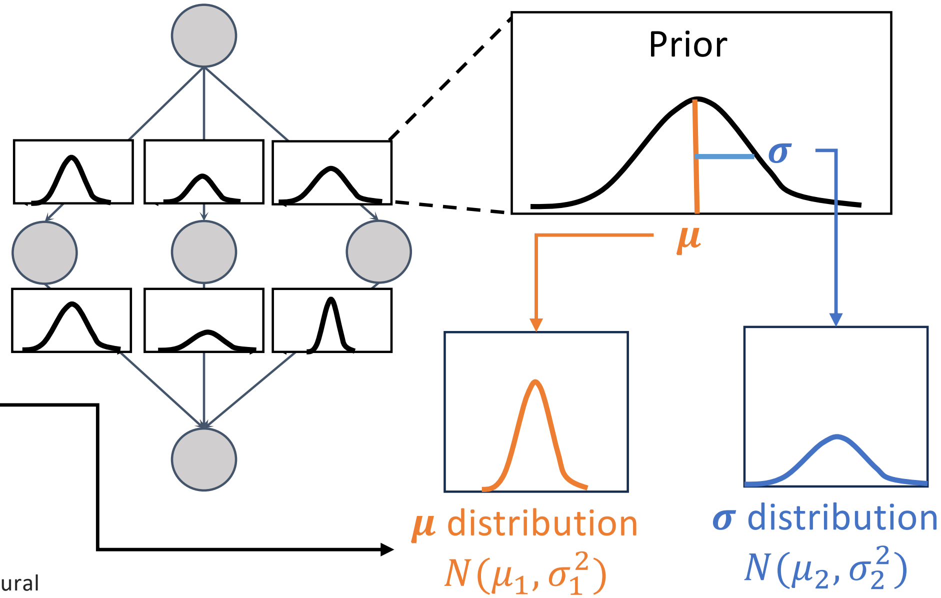
Adaptive Uncertainty Quantification

Adaptive Uncertainty Quantification aims to heteroscedasticity, which requires that prediction uncertainty is adaptive to different inputs.

Bayesian approaches' predictions can be more adaptive to inputs via the methods of increasing learning flexibility, such as architecture search and data augmentation.

More tailored strategies:
BNNs with hierarchical priors take the means and variances of priors are also random variables with their own distributions.

Hierarchical priors offer BNNs greater flexibility in fitting the training data with **more trainable parameters**, thereby improving their adaptability to various patterns within the dataset.



Karaletsos T, Bui T D. Hierarchical Gaussian process priors for Bayesian neural network weights[J]. Advances in Neural Information Processing Systems, 2020, 33: 17141-17152

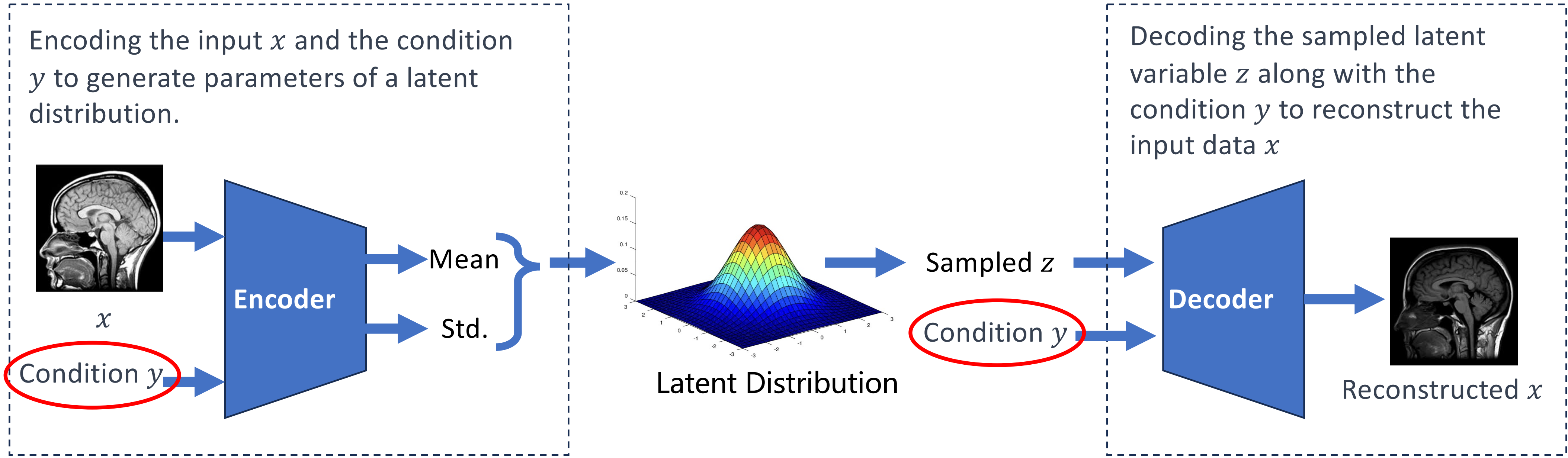
Use of Uncertainty

Adaptive Uncertainty Quantification

Conditional Variational Autoencoder (CVAE)

VAE can generate outputs by sampling from the encoder-estimated latent distribution.

The generation can be more context-aware by capturing the relationship between input x and condition y (i.e. label).



Zhao T, Zhao R, Eskenazi M. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders[J]. arXiv preprint arXiv:1703.10960, 2017

Use of Uncertainty

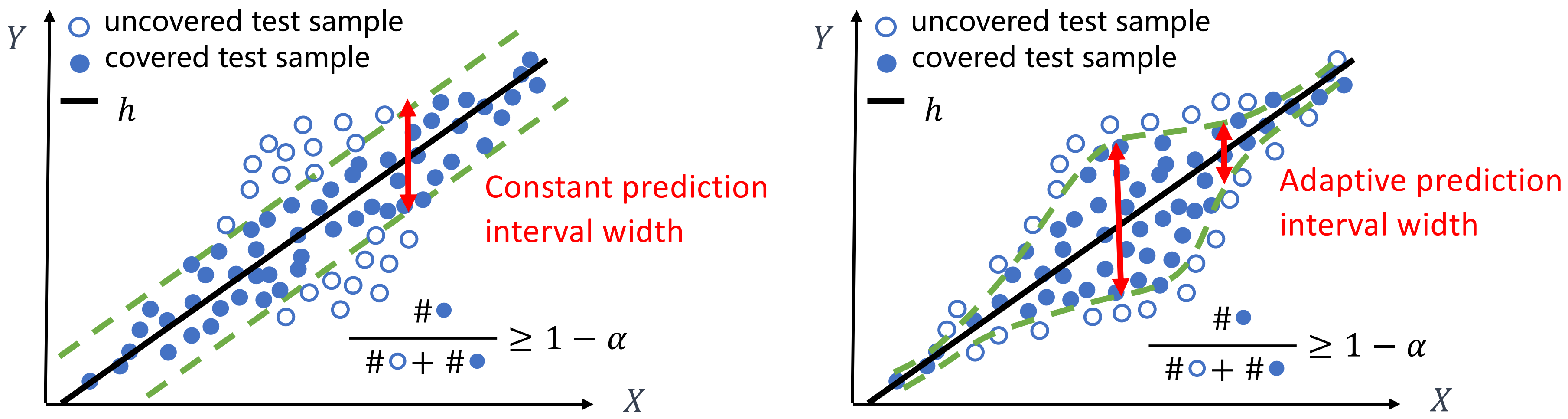
Adaptive Uncertainty Quantification

Localized Conformal Prediction weights calibration conformal scores, S_i, with distance-informed weights, such as

$$S_i(X_{n+1}) = e^{-5|X_{n+1}-X_i|} \times |h(X_i) - Y_i|$$

then the $1-\alpha$ quantile of the weighted conformal scores $\{S_i(X_{n+1})\}_{i=1}^{n}$ is a function of X_{n+1}, denoted by $\tau(X_{n+1})$.

As a result, the size of prediction set $C(X_{n+1}) = \{y: |h(X_{n+1}) - y| \leq \tau(X_{n+1})\}$ **is adaptive to** X_{n+1}.



Vanilla CP **Prediction Interval** without adaptiveness

Localized CP **Prediction Interval** with adaptiveness

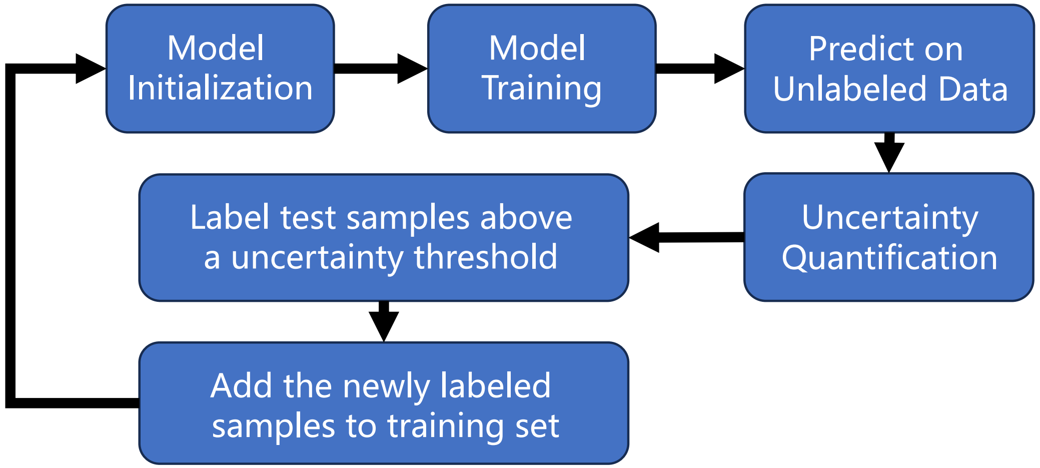
Guan, L. (2022). Localized conformal prediction: A generalized inference framework for conformal prediction. Biometrika, 110(1), 33–50. https://doi.org/10.1093/biomet/asac040

Use of Uncertainty

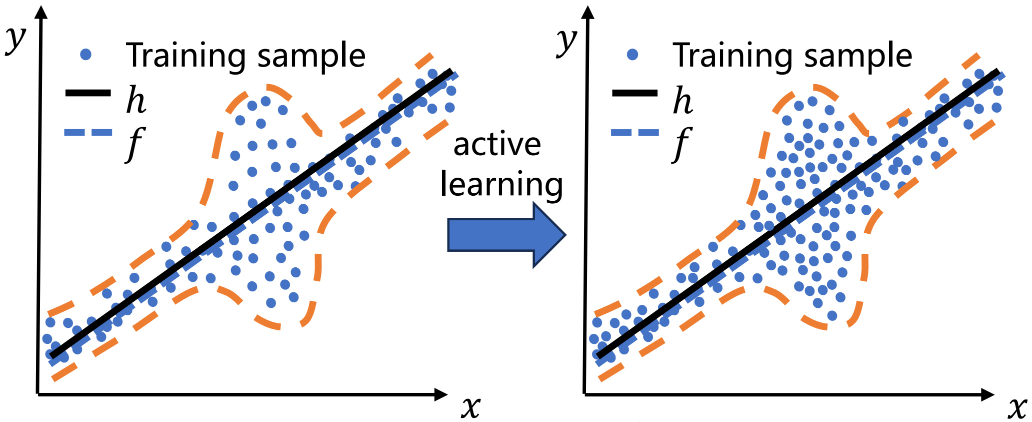
Active Learning in Uncertainty Quantification

Active Learning in Uncertainty Quantification is an essential approach for uncertainty reduction.

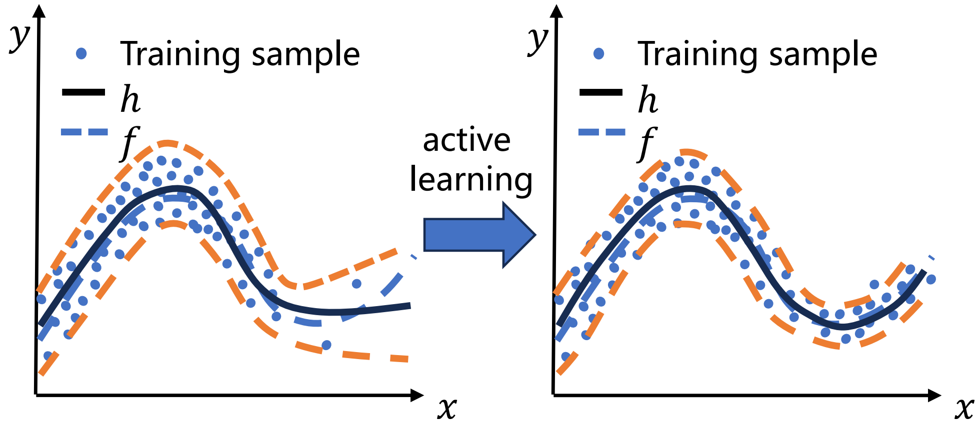
However, active learning can reduce epistemic uncertainty, while can not reduce aleatoric uncertainty.



Workflow of Active Learning in Uncertainty Quantification



Aleatoric uncertainty remains after active learning



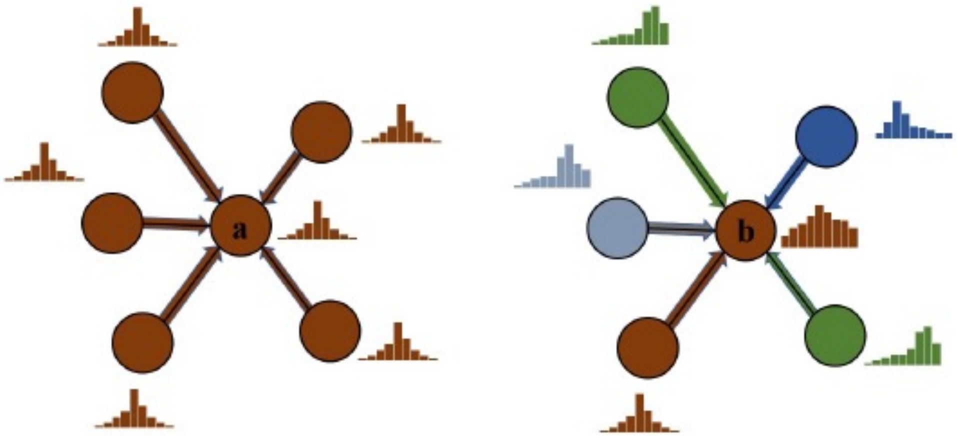
Epistemic uncertainty remains after active learning

Uncertainty Quantification for Structural Data

Uncertainty Quantification for Graph Data

Challenges: Most traditional uncertainty quantification methods assume that samples are independent, allowing them to statistically estimate uncertainty in predictions, such as coverage guarantees in CP.

However, in graph data, when a node is **homophilic with its neighbors**, neighbors provide **sufficient information for the prediction of the center node**, leading to a higher confidence. On the other hand, if a node is surrounded by **neighbors with different labels**, **we are not sure which community it belongs to**, so the confidence of the center node prediction should be lowered. This dependence make conventional methods ineffective.

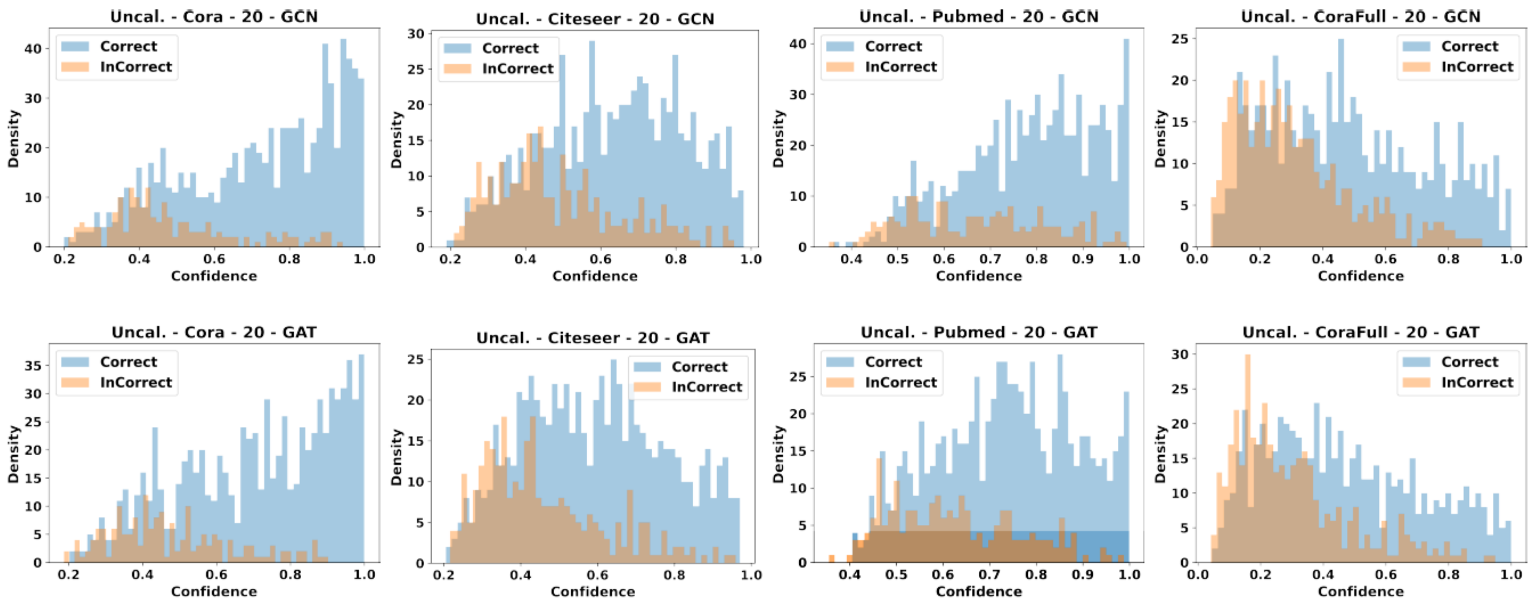


The ground-truth confidence distribution in a graph is interdependent

Wang X, Liu H, Shi C, et al. Be confident! towards trustworthy graph neural networks via confidence calibration[J]. Advances in Neural Information Processing Systems, 2021, 34: 23768-23779

Uncertainty Quantification for Structural Data

Base on confidence distributions of correct and incorrect prediction, we can see GCN tend to be under-confident in correct prediction.



Phenomenon: under-confident in correct predictions.

Wang X, Liu H, Shi C, et al. Be confident! towards trustworthy graph neural networks via confidence calibration[J]. Advances in Neural Information Processing Systems, 2021, 34: 23768-23779

Uncertainty Quantification for Structural Data

Work 1: CaGCN is a post-processed graph convolutional network (GCN) calibration method.

CaGCN is a model trained on a validation set D_{val} on top of a classification model. Denote the K-class one-hot label for node i as $(y_{i,1}, \ldots, y_{i,K})$ and prediction probability vector as $(z_{i,1}, \ldots, z_{i,K})$. CaGCN is trained with two purposes.

- Preserve the prediction accuracy of the classification model by NLL loss.

$$\mathcal{L}_{nll} = -\sum_{i=1}^{|D_{val}|} \sum_{k=1}^{K} y_{i,k} log(z_{i,k}).$$

- Increase the confidence of correct prediction and lower the confidence of incorrect prediction.

$$\mathcal{L}_{cal} = \frac{1}{n}\left(\sum_{i=1}^{|cor|} 1 - z_{i,m}^{(cor)} + z_{i,s}^{(cor)} + \sum_{i=1}^{|inc|} z_{i,m}^{(inc)} - z_{i,s}^{(inc)}\right)$$

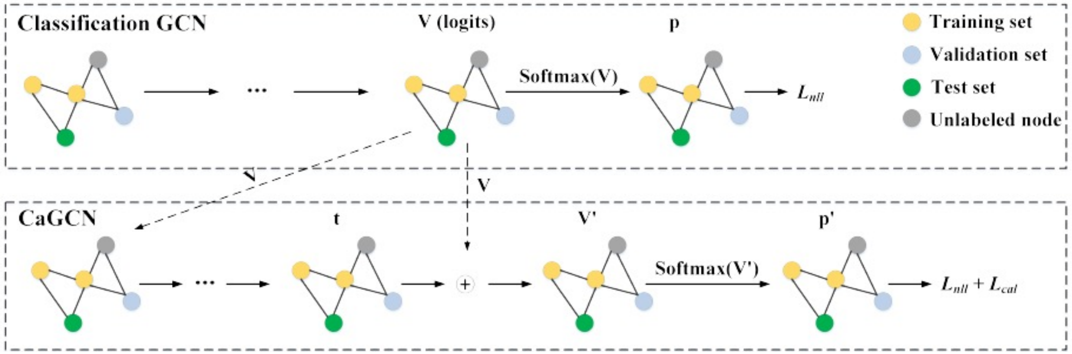
where $|cor|$ and $|inc|$ are the number of nodes correctly and incorrectly predicted, and $z_{i,m}$ and $z_{i,s}$ are the max and submax prediction probability. Intuitively, the confidence of incorrect predictions is decreased by reducing the gap between the max and the submax value of prediction probability and vice versa.

Wang X, Liu H, Shi C, et al. Be confident! towards trustworthy graph neural networks via confidence calibration[J]. Advances in Neural Information Processing Systems, 2021, 34: 23768-23779

Uncertainty Quantification for Structural Data

Classification Model for Prediction Accuracy

CaGCN to Calibrate for Prediction Confidence



To implement CaGCN, given output $\mathbf{V}$ of a classification GCN, CaGCN learns a temperature t_i for each node i.

$$\mathbf{t} = \sigma^{+}(\mathbf{A}\sigma(\cdots\mathbf{A}\sigma(\mathbf{A}\mathbf{V}\mathbf{W}^{(1)})\mathbf{W}^{(2)}\cdots)\mathbf{W}^{(l)}) = [t_1, \cdots, t_N]^{\top}(t_i > 0, \forall i \in \{1, \cdots, N\}),$$

Then, CaGCN get a calibrated logit $\mathbf{v}_i{}'$ by transforming its original logit $\mathbf{v}_i$ using t_i, and finally obtain calibrated confidence $\hat{p}_i$.

$$\mathbf{v}'_i = h(\mathbf{v}_i, t_i) = [v_{i,1}/t_i, \cdots, v_{i,K}/t_i]^{\top}, \mathbf{z}_i = [\sigma_{SM}(v'_{i,1}), \cdots, \sigma_{SM}(v'_{i,K})]^{\top}, \hat{p}_i = \max_k z_{i,k},$$

Pros: Topology-aware, Accuracy-preserving.
Cons: Requiring sufficient data to train the post-processed GCN model.

Wang X, Liu H, Shi C, et al. Be confident! towards trustworthy graph neural networks via confidence calibration[J]. Advances in Neural Information Processing Systems, 2021, 34: 23768-23779

Uncertainty Quantification for Structural Data

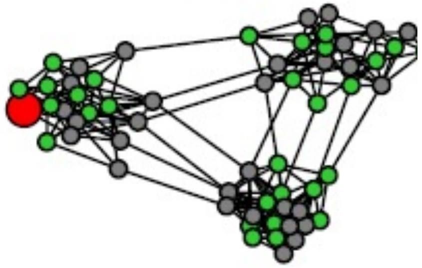
Work 2: Homophily: nodes in a neighborhood usually play similar roles in the network and are considered interchangeable on average. Therefore, nodes in a neighborhood better satisfying the exchangeability assumption.

Neighborhood Adaptive Conformal Prediction uses this insight by **taking the nodes surrounding the test node as calibration samples.**

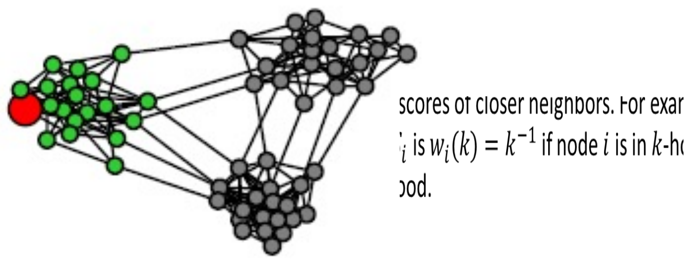
To do so, consider we have a predefined score function $s(\cdot)$. Let (X_i, Y_i) be the feature and label of node i, whose conformal score is $S_i = s(X_i, Y_i)$.

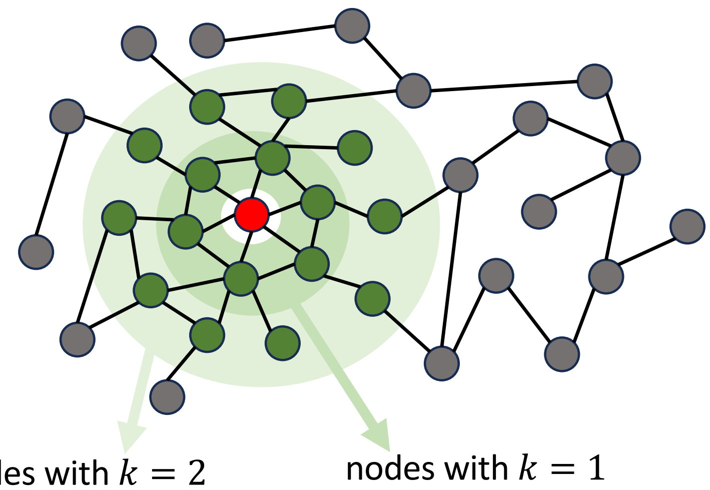
Given a test node with feature X_{test}, a scheme includes specified neighbors and assigns higher weights to conformal scores of closer neighbors. For example, the weight of S_i is $w_i(k) = k^{-1}$ if node i is in k-hop neighborhood.

Baseline



Neighborhood Adaptive Conformal Prediction





Nodes in k-hop neighborhood for $k = 1,2$

Clarkson, Jase. "Distribution free prediction sets for node classification." International Conference on Machine Learning. PMLR, 2023.

Uncertainty Quantification for Structural Data

We may consider including n nodes within 1-hop and 2-hop neighborhoods as calibration samples.

Assume $\{S_i\}_{i=1}^{n}$ is in increasing order and denote $\bar{w}_i$ is normalized weights of w_i.

We compute the quantile τ by

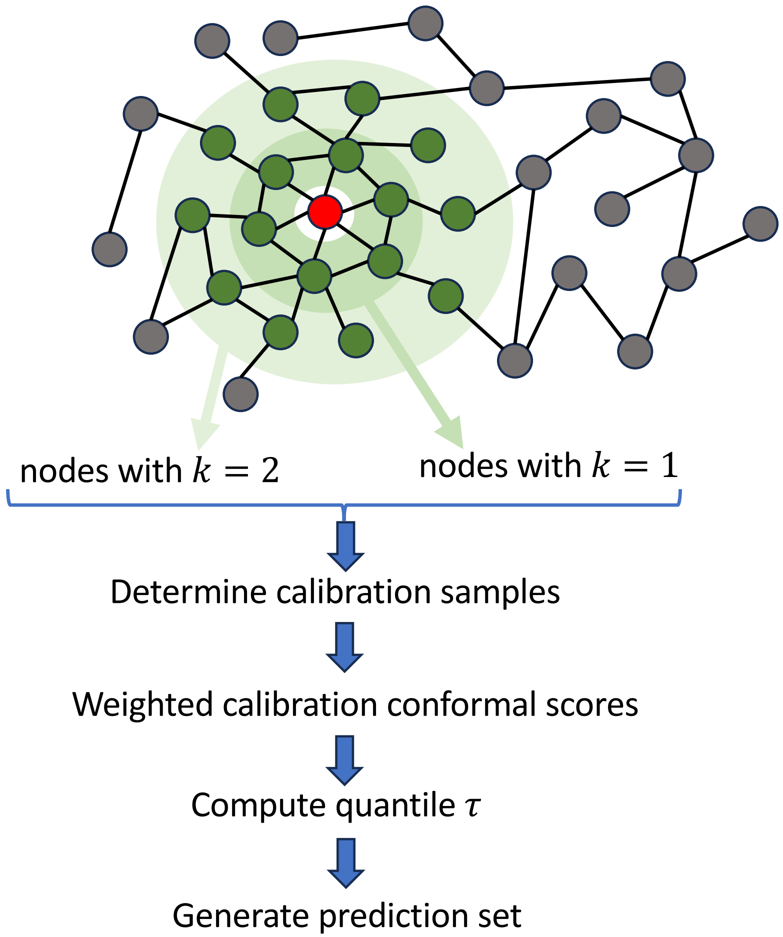
$$\tau = \inf\{S_q : \sum_{i=1}^{q} \bar{w}_i \geq 1 - \alpha\}.$$

Finally, the prediction set is

$$C(X_{test}) = \{y : s(X_{test}, y) \leq \tau\}.$$

In summary this work takes advantage of homophily to better satisfy the exchangeability assumption and thus make prediction set coverage closer to desired $1 - \alpha$.

Pros: Topology-aware; Adaptive to different test nodes.
Cons: Not directly applicable to both weighted and directed networks; Neighborhood size should be well-tuned; Rely on Homophily.



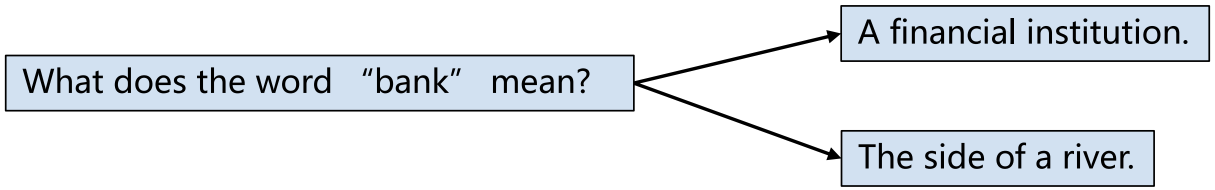
Clarkson, Jase. "Distribution free prediction sets for node classification." International Conference on Machine Learning. PMLR, 2023.

Uncertainty Quantification for Structural Data

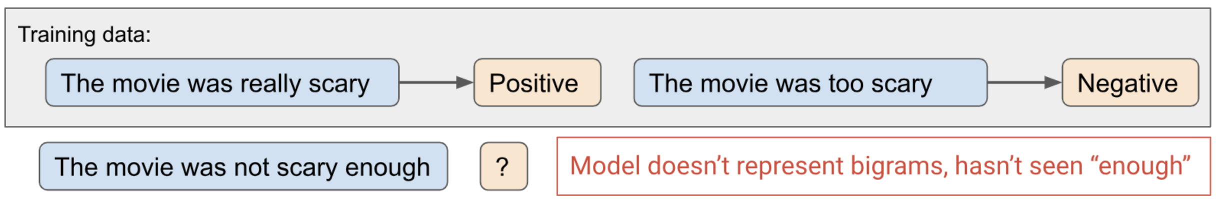
Uncertainty Quantification for Textual Data

Challenge: Aleatoric and epistemic uncertainties are prevalent in textual data.

Aleatoric uncertainty is inherent in the input task data.



Epistemic uncertainty exists due to lack of knowledge.



Fisch A, Jia R, Schuster T. Uncertainty estimation for natural language processing[C]. COLING, 2022.

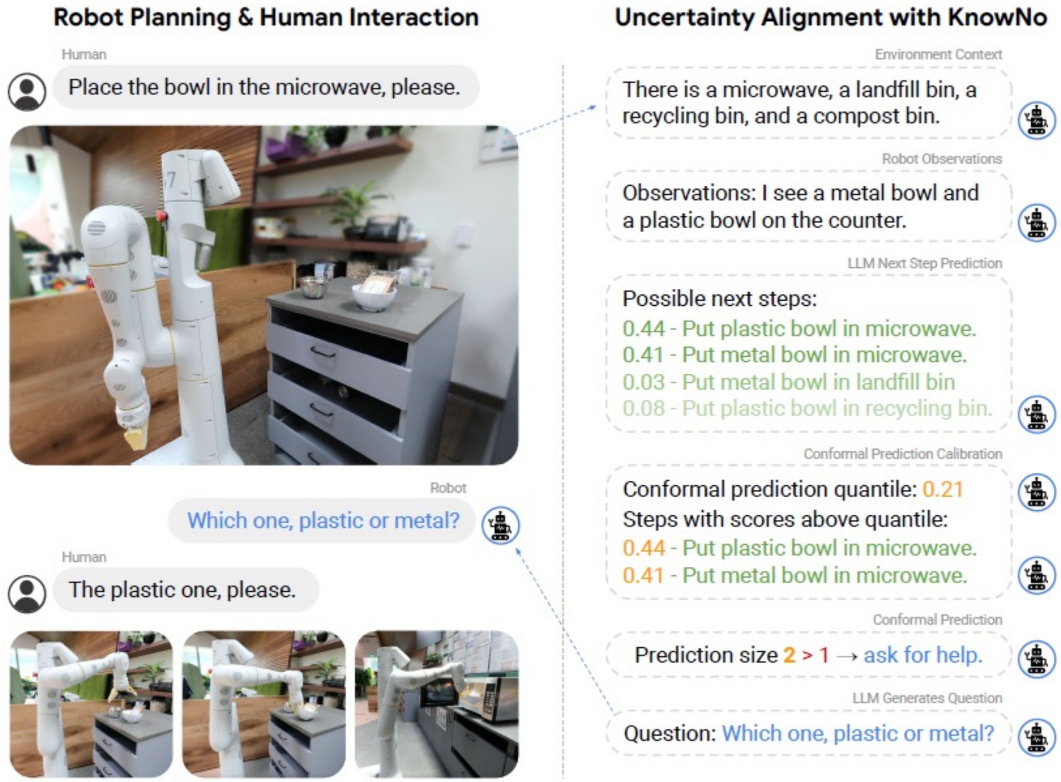
Uncertainty Quantification for Structural Data

Uncertainty Quantification for Textual data

Work: Large language models (LLM) can generate probabilistic outputs by autoregressive samplings where the model predicts the next word in a sequence based on the probabilities assigned to each word in its vocabulary.

Robots that ask for help is based on conformal prediction. It empowers LLM with visual context information to discern between high confidence in their knowledge and low confidence in what they don't know, enabling them to ask further questions for confirmation.

Pros: Addressing uncertainty alignment for language-instructed robots.
Cons: Environments must be fully grounded in the text input to the LLM.



Robots that ask for help

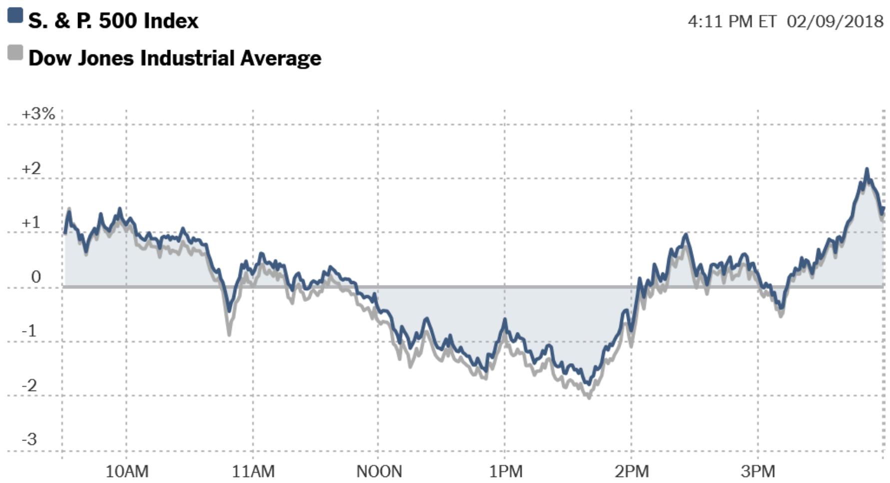
Ren A Z, Dixit A, Bodrova A, et al. Robots that ask for help: Uncertainty alignment for large language model planners[J]. arXiv preprint arXiv:2307.01928, 2023.

Uncertainty Quantification for Structural Data

Uncertainty Quantification for Time-series Data

Challenge: Time-series data poses unique challenges in uncertainty quantification due to the **sequential nature**.

For example, **temporal relationships and patterns** in the stock market data can lead to complex dependencies that need to be considered when forecasting prices and estimating uncertainties.



Stock market price data is typical time-series data

Flitter, E., & Stevenson, A. (2018, February 9). Tumultuous Week on wall street ends with a small rally. The New York Times.

Uncertainty Quantification for Structural Data

Uncertainty Quantification for Time-series Data

Work 1: The Bayesian Structural Time-series (BSTS) model contains stochastic parameters, introducing uncertainty into its predictions.
It decomposes the prediction of y_{n+1} of x_{n+1}, t_{n+1} with given training data $\{(x_i, t_i, y_i)\}_{i=1}^{n}$ into multiple components.

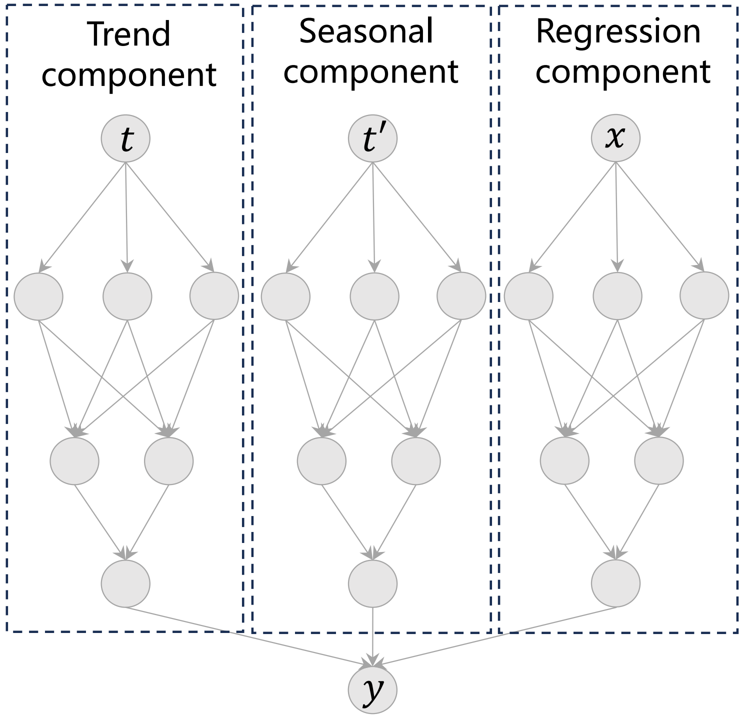
Trend and seasonal components model the temporal dependencies and capture the sequential structure inherent in time series data.

- Trend component captures the movement in the y over t.
- Seasonal component models the repeating patterns by studying y over t', where t' is the remainder of $t/7$ for weekly pattern.

Regression component includes the impact of external variables.

- Regression component relates the feature x with target y.

Bayesian Structural Time-series Model



Pros: Interpretable results; Disentangling the relation between external variables and time dependency
Cons: Require a sufficient amount of data; Choosing appropriate components and priors can be challenging

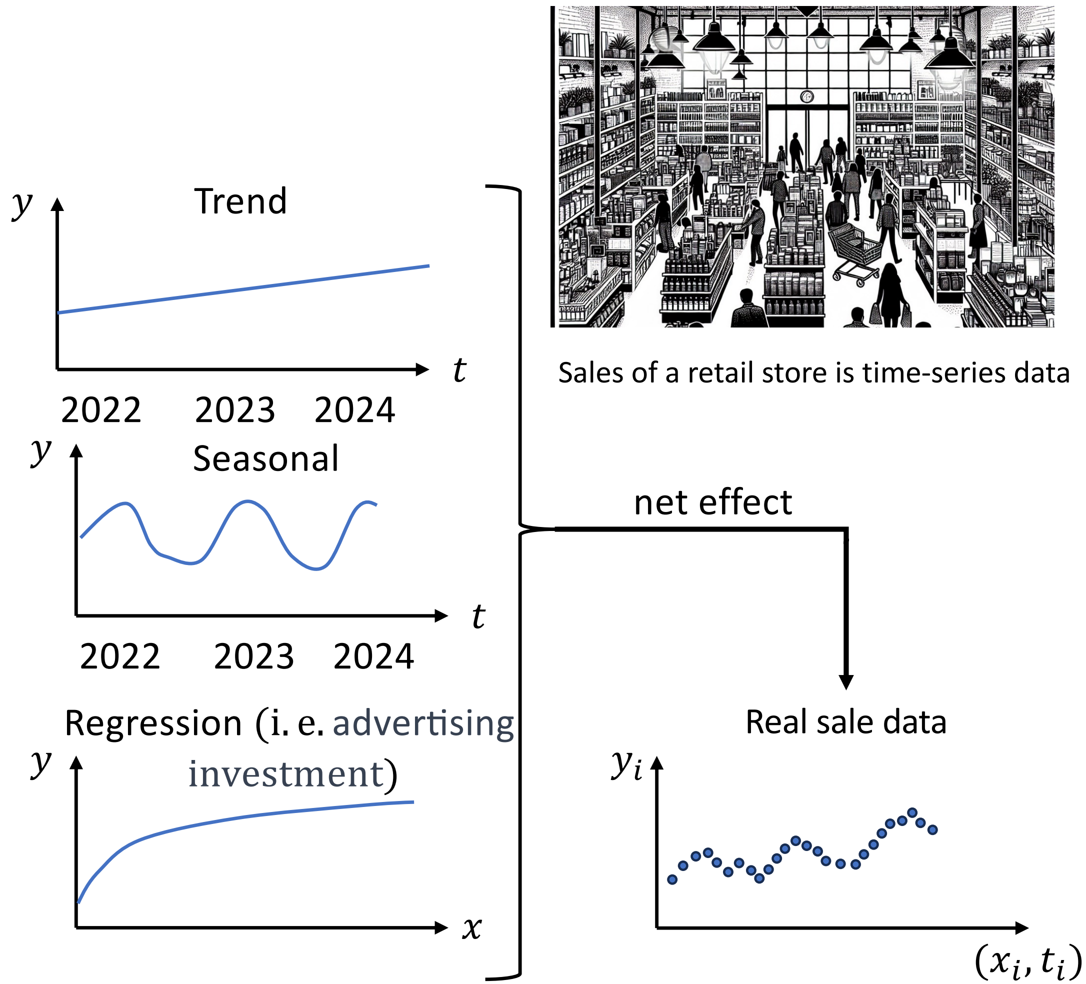
Scott S L, Varian H R. Predicting the present with Bayesian structural time series[J]. International Journal of Mathematical Modelling and Numerical Optimization, 2014, 5(1-2): 4-23

Uncertainty Quantification for Structural Data

Uncertainty Quantification for Time-series Data

Example: Sales of a Retail Store

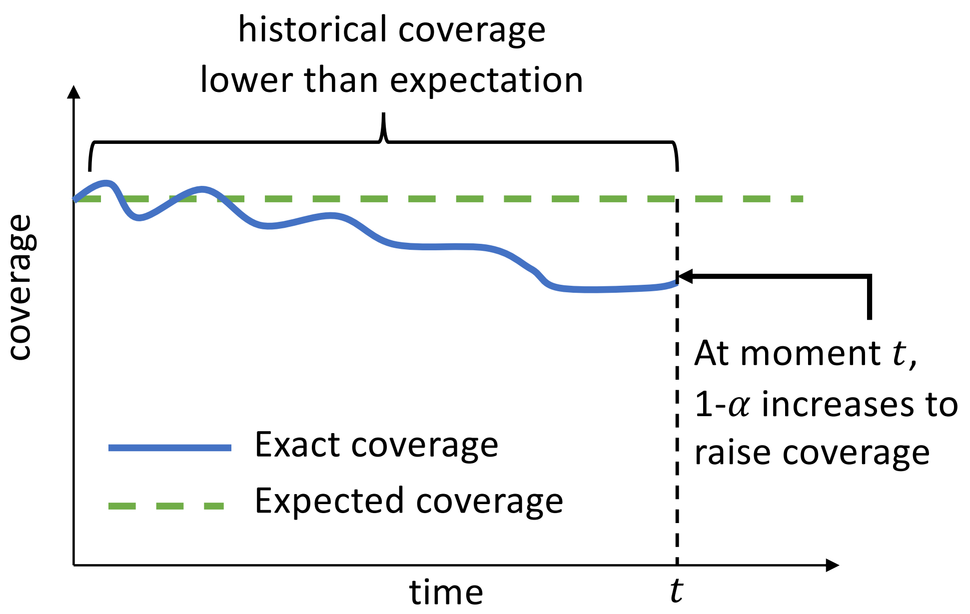
- Trend component: Suppose the store's sales have been steadily increasing by about 5% each year due to factors like improved marketing and increased customer base.
- Seasonal component: Sales might peak during the holiday season (November and December). This seasonal pattern could be modeled to account for these predictable variations in sales.
- Regression component: If an advertising campaign is launched in March, the model can assess its impact on sales during that month.

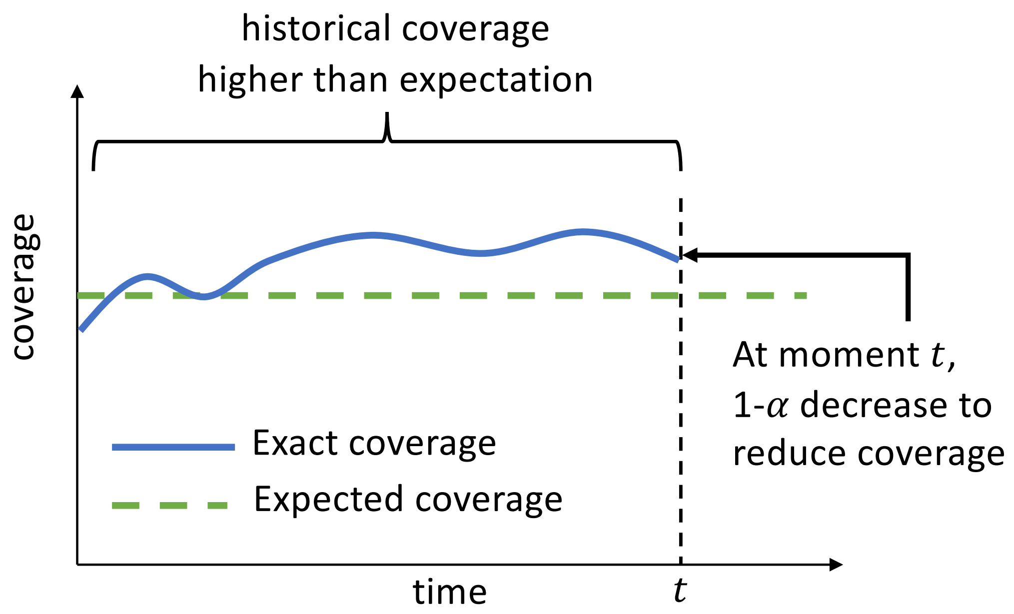


Uncertainty Quantification for Structural Data

Uncertainty Quantification for Time-series Data

Work 2: Time-series data breaks the exchangeability assumption of conformal prediction. We can dynamically adjust the parameter α based on the historical coverage of prediction sets. This approach ensures that prediction sets remain robust to distribution shifts over time.





Pros: adaptive to dynamic distribution shift.

Cons: Lack of theoretical coverage guarantee.

Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. Advances in Neural Information Processing Systems, 34:1660–1672, 2021.

Conclusion

In conclusion, uncertainty in machine learning can be quantified by various methods, particularly in structural data. It is crucial for making informed decisions and understanding the reliability of predictions. By incorporating uncertainty quantification techniques, practitioners can gain valuable insights into the robustness of their models, identify potential risks, and make more confident and reliable predictions.

- Aleatoric uncertainty
- Epistemic uncertainty

- Bayesian Approaches
- Conformal Prediction
- Calibration

- Graph Data
- Textual Data
- Time-series Data

Thank you

HKUST (GZ)