Single-Model Uncertainties for Deep Learning

Gardar Sigurdsson

ETH

2021

Outline

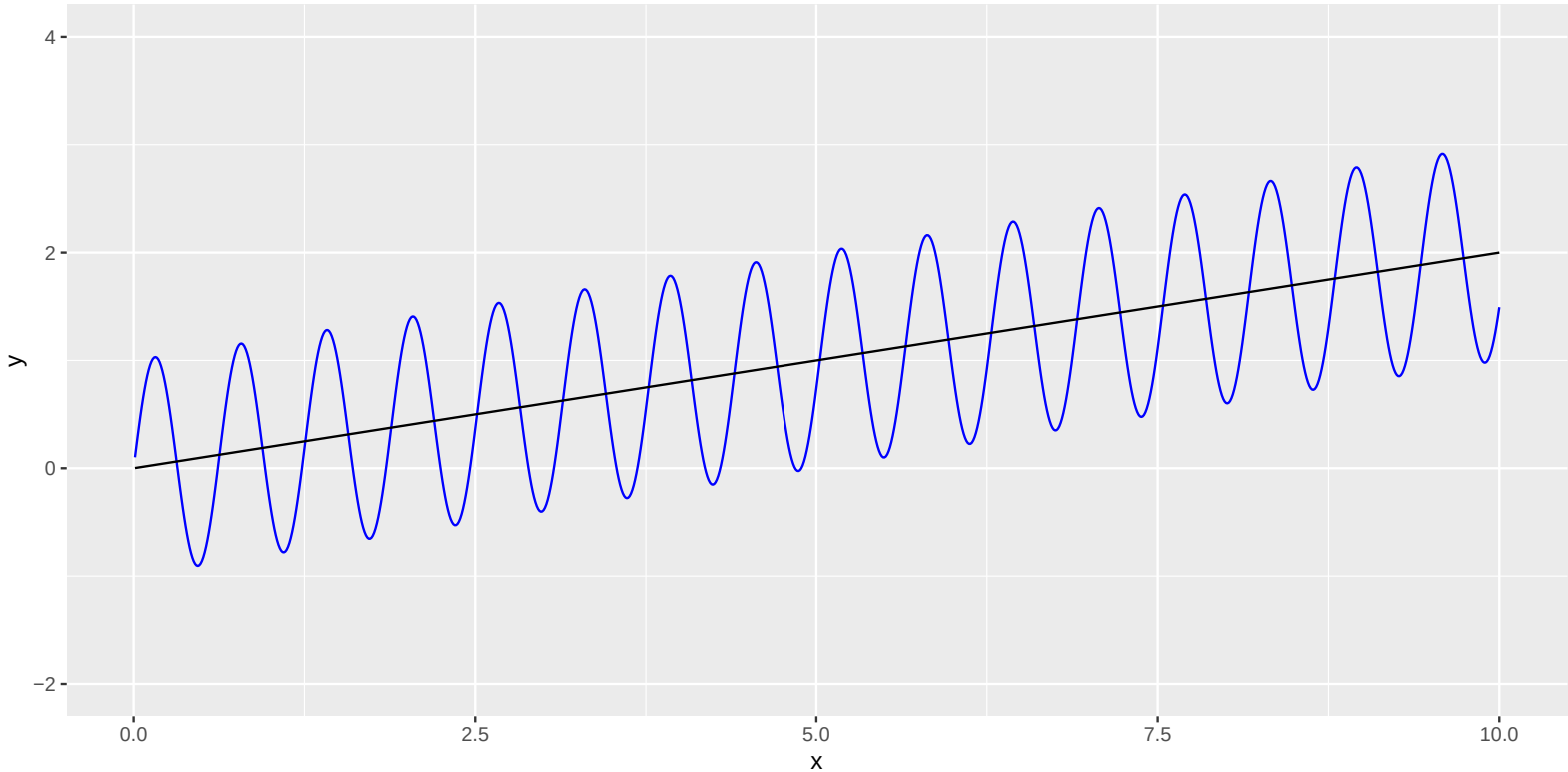
- Motivation
- Define three types of uncertainty
  - Approximation uncertainty
  - Aleatoric uncertainty
  - Epistemic uncertainty
- Simultaneous Quantile Regression
  - Measures aleatoric uncertainty
- Orthonormal Certificates
  - Measures epistemic uncertainty

Motivation

- Paper concerns itself with uncertainty
- Knowing the uncertainty of a model's prediction helps with:
  - **Abstention:** Hand-off high risk decisions to humans in cases of anomalies, adversarial attacks and out-of-distribution examples.
  - **Active learning:** What examples should be labeled to achieve the biggest benefit.
  - **Noise structure:** Prediction intervals, causal discovery.
  - **Interpretability**

Terms: Approximation Uncertainty

- Approximation uncertainty arises when the model is too simple



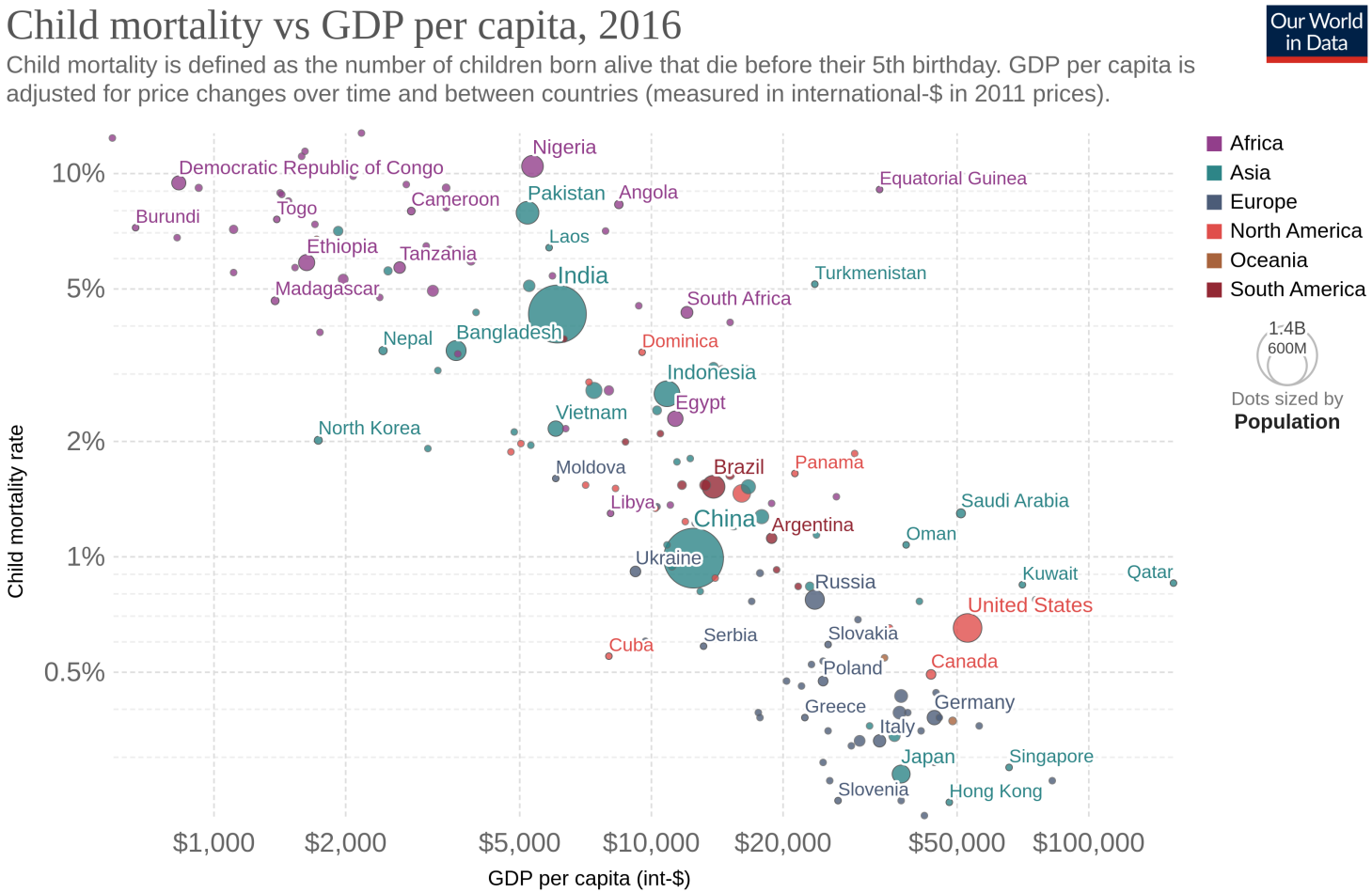
- Less of an issue when using over-parameterized NN models

Terms: Aleatoric Uncertainty

- Aleatoric stems from the Greek world 'alea' meaning rolling of the dice
- Describes the variance of $Y \mid X = x$
  - Unobserved variables
  - Measurement errors
  - Randomness

Terms: Aleatoric Uncertainty

Y = Child mortality, X = GDP per capita



Terms: Epistemic Uncertainty

- Errors associated with the lack of experience of our model
- Some parts of feature space are not present in our training data

Aleatoric Uncertainty: Previous Work

- ConditionalGaussian
  - $Y|X = x \sim N(\mu(x), \sigma(x)^2)$. Learn mean and variance for each x. The variance estimates the aleatoric uncertainty.
- Dropout
  - Enable dropout at test time to get multiple predictions. Then calculate the empirical quantiles to estimate uncertainty.
- QualityDriven
  - SOTA deep model estimating prediction intervals. Trained to minimize smooth surrogates of PICP/MPIW.
- QuantileForest
  - Random forest regression model trained with quantile regression.

Simultaneous Quantile Regression

- This method returns a $1 - \alpha$ prediction interval around the median of $Y \mid X$
  - Wide interval = High aleatoric uncertainty
  - Narrow interval = Low aleatoric uncertainty
- The interval is constructed by first estimating the quantiles of $Y \mid X$ and using those estimates to create the prediction interval

$$[\mathit{Quantile}(Y \mid X = x, \alpha/2), \mathit{Quantile}(Y \mid X = x, 1 - \alpha/2)]$$

Simultaneous Quantile Regression

- How do we estimate $\mathit{Quantile}(Y \mid X = x, \tau)$?
- Pinball loss

$$\ell_\tau(y, \hat{y}) = \begin{cases} \tau(y - \hat{y}) & \text{if } y - \hat{y} \geq 0 \\ (1 - \tau)(y - \hat{y}) & \text{else} \end{cases}$$

- It can be shown that

$$\frac{\partial E(\ell_\tau(Y, \hat{y}))}{\partial \hat{y}} = P(Y \leq \hat{y}) - \tau$$

i.e. the loss is minimized when $\hat{y}$ is the τ-th quantile.

Simultaneous Quantile Regression

- IID training dataset $\{(x_1, y_1), ..., (x_n, y_n)\}$
- Learning the τ-th quantile of $Y \mid X$:

$$\hat{f}_\tau \in \arg\min_f \frac{1}{n} \sum_{i=1}^{n} \ell_\tau(f(x_i), y_i)$$

- Simultaneous Quantile Regression:

$$\hat{f} \in \arg\min_f \frac{1}{n} \sum_{i=1}^{n} E_{\tau \sim U[0,1]}(\ell_\tau(f(x_i, \tau), y_i))$$

  - Train with stochastic gradient descent
  - Sample τ at each training step

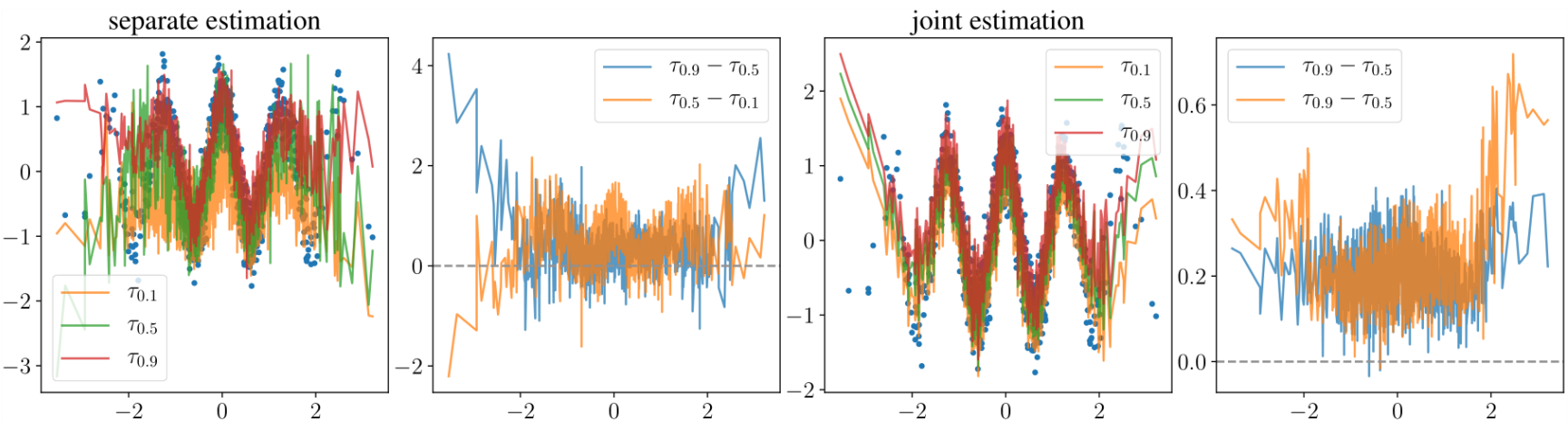
Simultaneous Quantile Regression

- Aleotoric uncertainty estimate for $Y \mid X = x$:

$$u_a(x) = \hat{f}(x, 1 - \alpha/2) - \hat{f}(x, \alpha/2)$$

Simultaneous Quantile Regression

- Crossing quantile issue: when the order of quantiles is incorrect
- Example: estimated 90th quantile is smaller than the 50th quantile



$$y = \cos(10x_1) + \mathcal{N}(0, \tfrac{1}{3})$$

Simultaneous Quantile Regression

- Comparison with other methods
  - Prediction Interval Coverage Probability (PICP, column 1): Fraction of samples that land within the prediction interval
  - Mean Prediction Interval Width (MPIW, column 2): Average width of prediction intervals

	ConditionalGaussian	Dropout	GradientBoostingQR
concrete	0.94 ± 0.03 (0.32 ± 0.09)	none	0.93 ± 0.00 (0.71 ± 0.00)
power	0.94 ± 0.01 (0.18 ± 0.00)	0.94 ± 0.00 (0.37 ± 0.00)	none
wine	0.94 ± 0.02 (0.49 ± 0.03)	none	none
yacht	0.93 ± 0.06 (0.03 ± 0.01)	0.97 ± 0.03 (0.10 ± 0.01)	0.95 ± 0.02 (0.79 ± 0.01)
naval	0.96 ± 0.01 (0.15 ± 0.25)	0.96 ± 0.01 (0.23 ± 0.00)	none
energy	0.94 ± 0.03 (0.12 ± 0.18)	0.91 ± 0.04 (0.17 ± 0.01)	none
boston	0.94 ± 0.03 (0.55 ± 0.20)	none	0.89 ± 0.00 (0.75 ± 0.00)
kin8nm	0.93 ± 0.01 (0.20 ± 0.01)	none	none

	QualityDriven	QuantileForest	**SQR (ours)**
concrete	none	0.96 ± 0.01 (0.37 ± 0.02)	0.94 ± 0.03 (0.31 ± 0.06)
power	0.93 ± 0.02 (0.34 ± 0.19)	0.94 ± 0.01 (0.18 ± 0.00)	0.93 ± 0.01 (0.18 ± 0.01)
wine	none	none	0.93 ± 0.03 (0.45 ± 0.04)
yacht	0.92 ± 0.05 (0.04 ± 0.01)	0.97 ± 0.04 (0.28 ± 0.11)	0.93 ± 0.06 (0.06 ± 0.04)
naval	0.94 ± 0.02 (0.21 ± 0.11)	0.92 ± 0.01 (0.22 ± 0.00)	0.95 ± 0.02 (0.12 ± 0.09)
energy	0.91 ± 0.04 (0.10 ± 0.05)	0.95 ± 0.02 (0.15 ± 0.01)	0.94 ± 0.03 (0.08 ± 0.03)
boston	none	0.95 ± 0.03 (0.37 ± 0.02)	0.92 ± 0.06 (0.36 ± 0.09)
kin8nm	0.96 ± 0.00 (0.84 ± 0.00)	none	0.93 ± 0.01 (0.23 ± 0.02)

Simultaneous Quantile Regression

Mean Prediction Interval Width: Average width of prediction intervals

Dataset	Conditional Gaussian	Dropout	Gradient BoostingQR	Quality Driven	Quantile Forest	SQR (ours)
concrete	0.32	NaN	0.71	NaN	0.37	**0.31**
power	**0.18**	0.37	NaN	0.34	**0.18**	**0.18**
wine	0.49	NaN	NaN	NaN	NaN	**0.45**
yacht	**0.03**	0.10	0.79	0.04	0.28	0.06
naval	0.15	0.23	NaN	0.21	0.22	**0.12**
energy	0.12	0.17	NaN	0.10	0.15	**0.08**
boston	0.55	NaN	0.75	NaN	0.37	**0.36**
kin8nm	**0.20**	NaN	NaN	0.84	NaN	0.23

Simultaneous Quantile Regression

- Benefits:
  - Can model complex distributions (doesn't assume $Y|X$ is gaussian)
  - Does not require ensambling, more efficient
  - Alleviates crossing quantiles
- Limitations:
  - Not model independent, τ needs to be added as a feature to the model

Orthonormal Certificates

- This method yields a measure $u_e(x)$ of the epistemic uncertainty at the point x.
  - High uncertainty = The model hasn't seen an example similar to this in the training set
  - Low uncertainty = The model has seen a similar during training
- Assumes deep model structure $f(\phi(x))$
  - $\phi(x) \in \mathbb{R}^h$ deep featurizer extracting high-level representation
  - f shallow classifier

Orthonormal Certificates

- Central idea: learn certificates $C_1, C_2, ..., C_k \in \mathbb{R}^h$ which minimize $\sum_{i=1}^{n} \ell_c(C_i^T \phi(x_i), 0)$
  - i.e. map the training set to 0
- To ensure diversity among the certificates, enforce orthogonality
- Orthonormal Certificates:

$$\hat{C} \in \arg \min_{C \in \mathbb{R}^{h \times k}} \frac{1}{n} \sum_{i=1}^{n} \ell_c(C^T \phi(x_i), 0) + \lambda \cdot ||C^T C - I_p||$$

Orthonormal Certificates

- Epistemic uncertainty measure:

$$u_e(x) = ||\hat{C}^T \phi(x)||$$

- Calculate a high quantile (e.g. 99th) of $||C^T \phi(x)||$ on the training set, use as threshold to descide if a new sample is in the training distribution

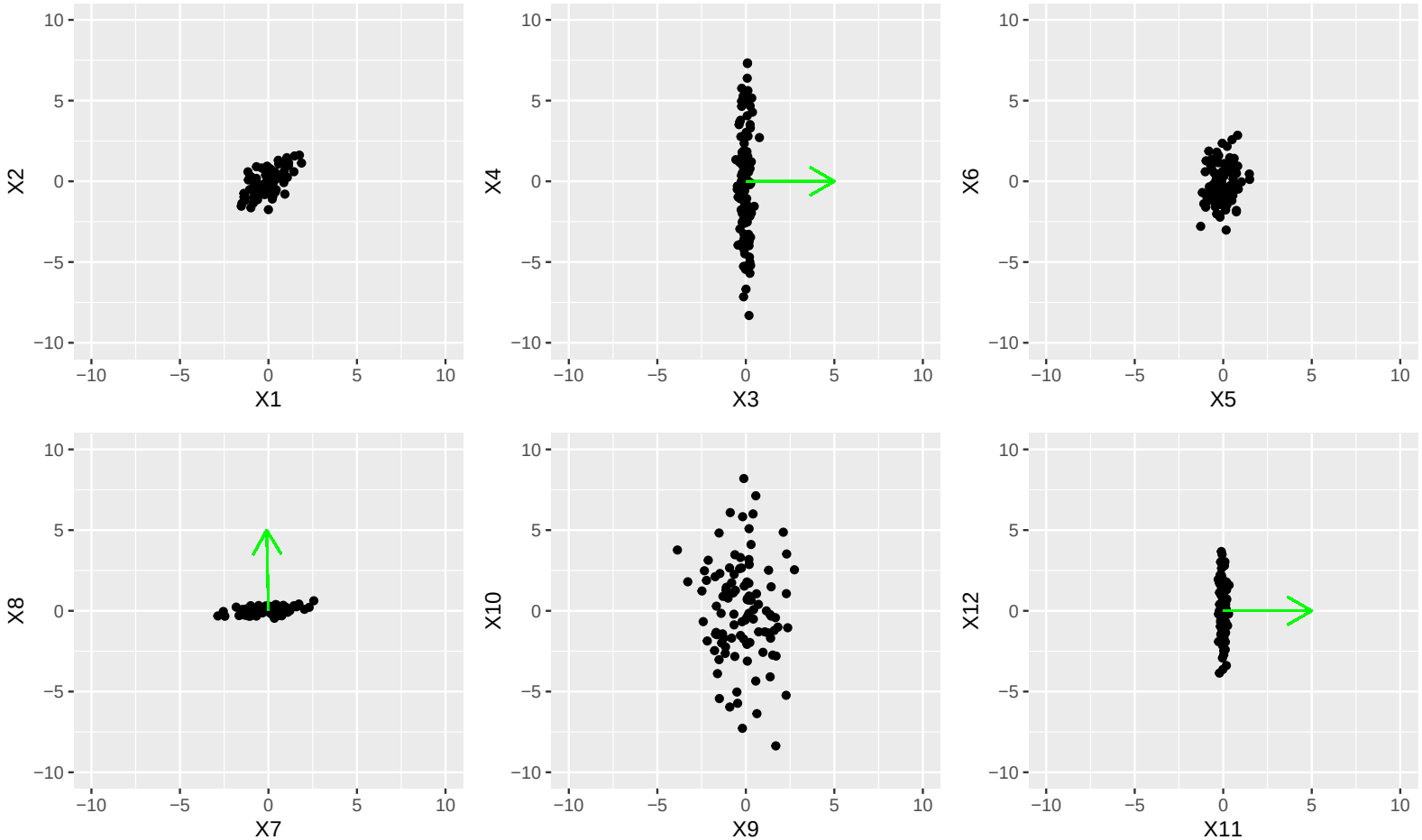
Orthonormal Certificates: PCA Interpretation

- Orthonormal Certificates:

$$\hat{C} \in \arg \min_{C \in \mathbb{R}^{h \times k}} \frac{1}{n} \sum_{i=1}^{n} \ell_c(C^T \phi(x_i), 0) + \lambda \cdot ||C^T C - I_p||$$

- When ℓ_c is the MSE loss, the certificates are the k orthonormal directions with the lowest variance

Orthonormal Certificates: PCA Interpretation



Orthonormal Certificates: Theorem

Theorem 1. *Let the in-domain data follow* $x \sim \mathcal{N}(0, \Sigma)$ *with* $\Sigma = V \Lambda V^\top$*, the out-domain data follow* $x' \sim \mathcal{N}(\mu', \Sigma')$ *with* $\Sigma' = V' \Lambda' V'^\top$*, and the certificates* $C \in \mathbb{R}^{d \times k}$ *be the bottom* k *eigenvectors of* Σ*, with associated eigenvalues* Γ*. Then,*

$$P\left(\|C^\top x\|^2 - \mathbb{E}[\|C^\top x\|^2] \geq t\right) \leq e^{-t^2/(2 \max_j \Gamma_j)},$$
$$P\left(\|C^\top x'\|^2 - \mathbb{E}[\|C^\top x'\|^2] \geq t\right) \leq e^{-t^2/(2 \max_j \Lambda'_j \|C_j V'_j\|^2)}.$$

Reminder: $u_e(x) = ||C^T \phi(x)||$

Orthonormal Certificates: Evaluation

- Evaluated on 4 classification datasets:
  - MNIST
  - CIFAR-10
  - Fashion-MNIST
  - SVHN
- Randomly split classes into in-domain and out-of-domain classes
- Use epistemic uncertainty measure to distinguish between the in- and out-domain classes

Orthonormal Certificates: Evaluation

- ROC AUC of different out-of-distribution classifiers

	cifar	fashion	mnist	svhn
covariance	0.64 ± 0.00	0.71 ± 0.13	0.81 ± 0.00	0.56 ± 0.00
distance	0.60 ± 0.11	0.73 ± 0.10	0.74 ± 0.10	0.64 ± 0.13
distillation	0.53 ± 0.01	0.62 ± 0.03	0.71 ± 0.05	0.56 ± 0.03
entropy	0.80 ± 0.01	0.86 ± 0.01	0.91 ± 0.01	0.93 ± 0.01
functional	0.79 ± 0.00	0.87 ± 0.02	0.92 ± 0.01	0.92 ± 0.00
geometrical	0.70 ± 0.11	0.66 ± 0.07	0.75 ± 0.10	0.77 ± 0.13
largest	0.78 ± 0.02	0.85 ± 0.02	0.89 ± 0.01	0.93 ± 0.01
ODIN	0.74 ± 0.09	0.84 ± 0.00	0.89 ± 0.00	0.88 ± 0.08
PCA	0.60 ± 0.09	0.57 ± 0.07	0.64 ± 0.06	0.55 ± 0.03
random	0.50 ± 0.00	0.51 ± 0.00	0.51 ± 0.00	0.50 ± 0.00
SVDD	0.52 ± 0.01	0.54 ± 0.03	0.59 ± 0.03	0.51 ± 0.01
BALD†	0.80 ± 0.04	$\mathbf{0.95 \pm 0.02}$	$\mathbf{0.95 \pm 0.02}$	0.90 ± 0.01
OCs (ours)	$\mathbf{0.83 \pm 0.01}$	0.92 ± 0.00	$\mathbf{0.95 \pm 0.00}$	$\mathbf{0.91 \pm 0.00}$
unregularized OCs	0.78 ± 0.00	0.87 ± 0.00	0.91 ± 0.00	0.88 ± 0.00
oracle	0.94 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.99 ± 0.00

Orthonormal Certificates

- Benefits:
  - Efficient method, no need to retrain the model
  - Performs well
- Limitations:
  - What if $\phi(x) \approx \phi(y)$ but y is out of distribution?
    - Might help to use a pre-trained model
  - Model dependent

Summary

- Simultaneous Quantile Regression: Method for measuring aleatoric uncertainty of a model
  - Perfoms well compared to other methods
  - Alleviates crossing quantiles
  - Requires a modification to the model
- Orthonormal Certificates: Method for measuring epistemic uncertainty
  - Perfoms well compared to other methods
  - Efficient method
  - Assumes a deep neural model i.e. model dependent

Thank you!

Q/A