Yandex Research

Ensembles Distribution Distillation and Uncertainty Estimation in Structured Prediction

Andrey Malinin

10 March 2020

Overview of the Talk

1. Motivation: Why do we need Uncertainty Estimation?
2. Sources of Uncertainty in Predictions
3. Ensemble Approaches
4. Ensemble Distribution Distillation
5. Uncertainty in Structured Prediction

Overview of the Talk

1. **Motivation: Why do we need Uncertainty Estimation?**
2. Sources of Uncertainty in Predictions
3. Ensemble Approaches
4. Ensemble Distribution Distillation
5. Uncertainty in Structured Prediction

Why is Uncertainty important?

- Philosophical $\rightarrow$ "Scio me nihil scire" - Socrates
  - Intelligent agents must know that they don't know $\rightarrow$
  - Agents must understand the limits of their knowledge
- Intelligent behaviour depends on detecting novel situations
  - Animals display fear or **curiosity**
  - Humans ask questions
- Uncertainty must affect **actions** of an intelligent agent

Why is Uncertainty important?

- Machine Learning (ML) systems are being deployed to many applications $\rightarrow$
  - Image Classification / Segmentation
  - Speech Recognition
  - Machine Translation
  - Etc...
- In some applications, the cost of a mistake is **high** or consequence **fatal**
  - Medical Applications
  - Financial Applications
  - Self-driving vehicles
- Obtaining measures of uncertainty in predictions helps avoid mistakes!
  - Increases safety and reliability of ML system

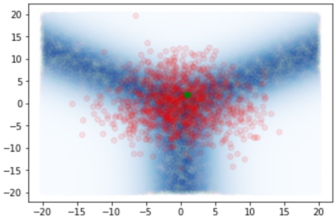
Scenario

- Given a deployed model and a test input $\boldsymbol{x}^*$ we wish to:
  - Obtain a prediction
  - Obtain a measure of **uncertainty in prediction**
- Take action based estimate of uncertainty
  - Reject prediction / stop decoding sentence
  - Modify policy / do exploration
  - Ask for human intervention
  - Use active learning

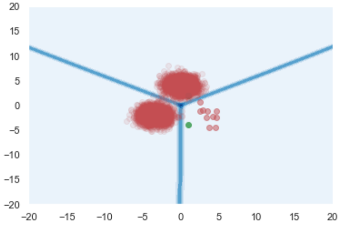
Overview of the Talk

1. Motivation: Why do we need Uncertainty Estimation?
2. **Sources of Uncertainty in Predictions**
3. Ensemble Approaches
4. Ensemble Distribution Distillation
5. Uncertainty in Structured Prediction

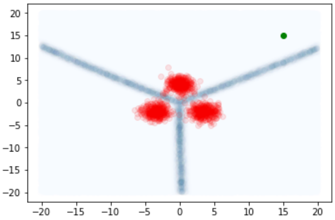
Sources of Uncertainty



(a) Data Uncertainty



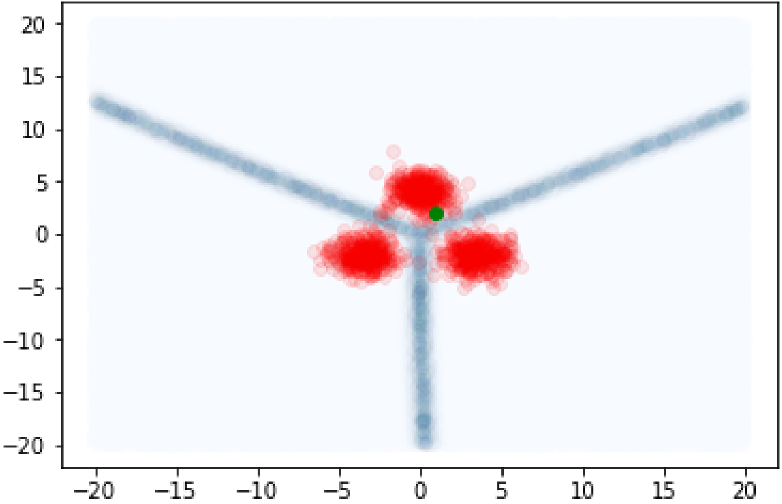
(b) Knowledge Uncertainty



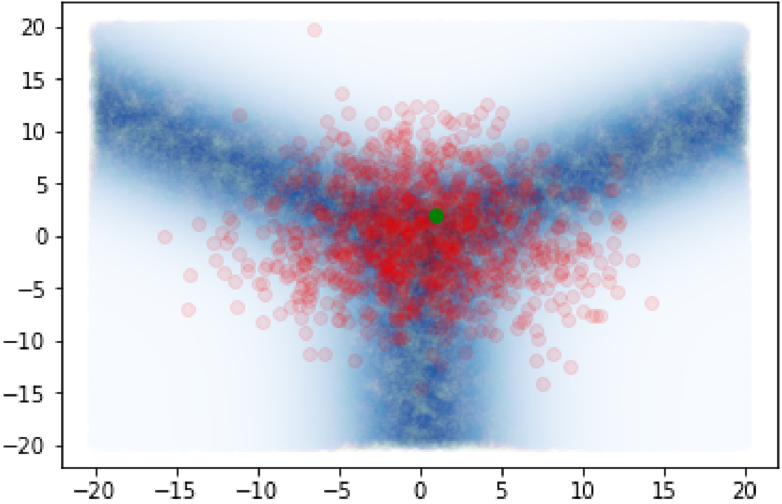
(c) Knowledge Uncertainty

- Knowledge (epistemic) uncertainty refers to both:
  - Data Sparsity and Knowledge Uncertainty

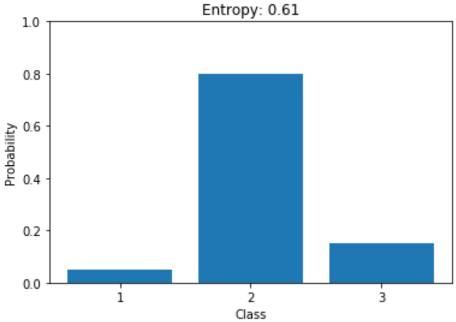
Data (Aleatoric) Uncertainty



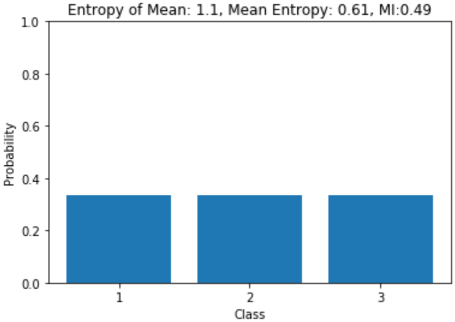
Data Uncertainty



Reminder - Entropy



(a) Low Entropy



(b) High Entropy

$$\mathcal{H}[\mathrm{P}_{\mathtt{tr}}(y|\boldsymbol{x}^*)] = -\sum_{c=1}^{K} \mathrm{P}_{\mathtt{tr}}(y=\omega_c|\boldsymbol{x}^*) \ln \mathrm{P}_{\mathtt{tr}}(y=\omega_c|\boldsymbol{x}^*)$$

Data (Aleatoric) Uncertainty

- Data Uncertainty is the *entropy* of the *true data distribution* $\rightarrow$

$$\mathcal{H}[\mathrm{P}_{\mathtt{tr}}(y|\boldsymbol{x}^*)] = -\sum_{c=1}^{K} \mathrm{P}_{\mathtt{tr}}(y=\omega_c|\boldsymbol{x}^*) \ln \mathrm{P}_{\mathtt{tr}}(y=\omega_c|\boldsymbol{x}^*)$$

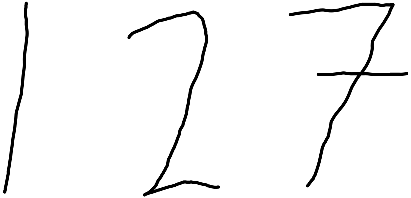
- Captured by the entropy of a model's posterior over classes $\rightarrow$

$$\mathcal{H}[\mathrm{P}(y|\boldsymbol{x}^*, \hat{\boldsymbol{\theta}})] = -\sum_{c=1}^{K} \mathrm{P}(y=\omega_c|\boldsymbol{x}^*, \hat{\boldsymbol{\theta}}) \ln \mathrm{P}(y=\omega_c|\boldsymbol{x}^*, \hat{\boldsymbol{\theta}})$$

- Data Uncertainty is captured as a consequence of Maximum Likelihood Estimation

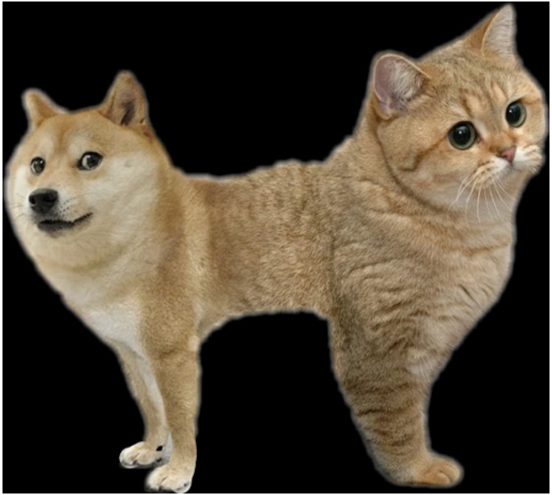
Data (Aleatoric) Uncertainty

- Distinct Classes

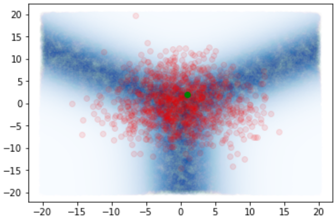


- Overlapping Classes

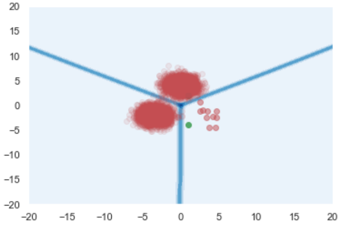
Data (Aleatoric) Uncertainty



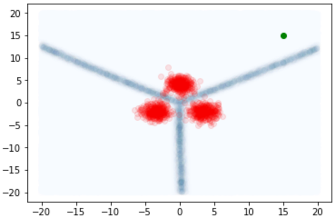
Sources of Uncertainty



(a) Data Uncertainty



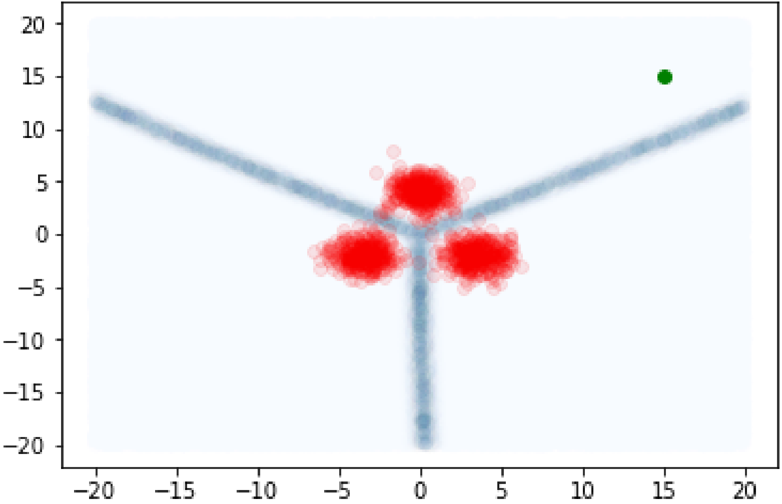
(b) Data Sparsity



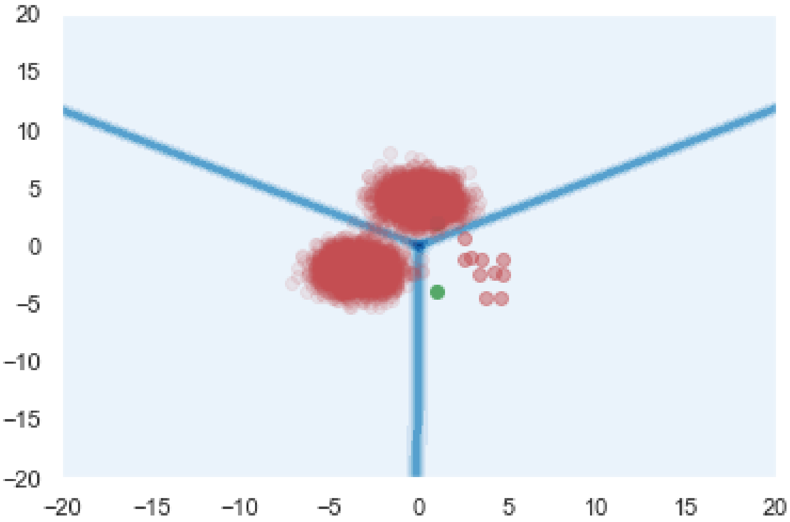
(c) Out-of-Distribution inputs

- Knowledge (epistemic) uncertainty refers to both:
  - Data Sparsity and Out-of-distribution inputs

Knowledge (Epistemic) Uncertainty

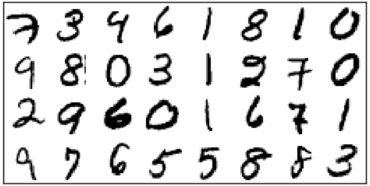


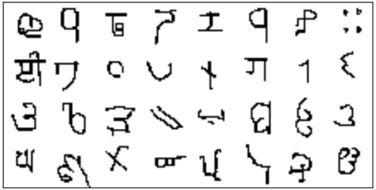
Knowledge (Epistemic) Uncertainty



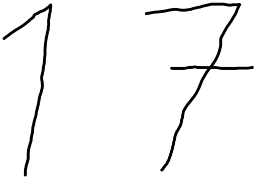
Knowledge (Epistemic) Uncertainty

- Unseen classes





- Unseen variations of seen classes



Knowledge (Epistemic) Uncertainty

Sources of Uncertainty

- Data Uncertainty → **Known-Unknown**
  - Class overlap (complexity of decision boundaries)
  - Homoscedastic and Heteroscedastic noise
- Knowledge Uncertainty → **Unknown-Unknown**
  - Test input in out-of-distribution region far from training data
  - Test input in out-of-distribution region of sparse training data
- Appropriate **action** depends on **source** of uncertainty
  - Separating sources of uncertainty requires **Ensemble approaches**

Overview

1. Motivation: Why do we need Uncertainty Estimation?
2. Sources of Uncertainty in Predictions
3. **Ensemble Approaches**
4. Ensemble Distribution Distillation
5. Uncertainty in Structured Prediction

- Uncertainty in $\boldsymbol{\theta}$ captured by model posterior $\mathtt{p}(\boldsymbol{\theta}|\mathcal{D}) \rightarrow$

$$\mathtt{p}(\boldsymbol{\theta}|\mathcal{D}) = \frac{\mathtt{p}(\mathcal{D}|\boldsymbol{\theta})\mathtt{p}(\boldsymbol{\theta})}{\mathtt{p}(\mathcal{D})}$$

- Can consider an ensemble of models $\rightarrow$

$$\{\mathtt{P}(y|\boldsymbol{x}^*, \boldsymbol{\theta}^{(m)})\}_{m=1}^{M}, \ \boldsymbol{\theta}^{(m)} \sim \mathtt{p}(\boldsymbol{\theta}|\mathcal{D})$$

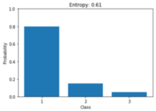
- Bayesian inference of $\mathtt{P}(y|\boldsymbol{x}^*, \boldsymbol{\theta}) \rightarrow$

$$\mathtt{P}(y|\boldsymbol{x}^*, \mathcal{D}) = \mathbb{E}_{\mathtt{p}(\boldsymbol{\theta}|\mathcal{D})}\big[\mathtt{P}(y|\boldsymbol{x}^*, \boldsymbol{\theta})\big] \approx \ \mathcal{H}\Big[\frac{1}{M}\sum_{m=1}^{M}\mathtt{P}(y|\boldsymbol{x}^*, \boldsymbol{\theta}^{(m)})\Big]$$

Ensemble for certain in-domain input

$$\mathrm{p}(\boldsymbol{\theta}) + \mathcal{D} \rightarrow \mathrm{p}(\boldsymbol{\theta}|\mathcal{D}) \longrightarrow$$

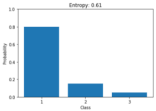
$\boldsymbol{\theta}^{(1)} \longrightarrow$



$\mathrm{P}(y|\boldsymbol{x}^*, \boldsymbol{\theta}^{(1)})$ $\rightarrow$

$\vdots$

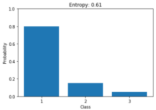
$\boldsymbol{\theta}^{(i)} \longrightarrow$



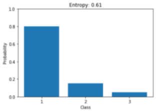
$\mathrm{P}(y|\boldsymbol{x}^*, \boldsymbol{\theta}^{(i)})$ $\rightarrow$

$\vdots$

$\boldsymbol{\theta}^{(N)} \longrightarrow$



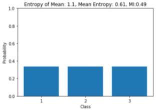
$\mathrm{P}(y|\boldsymbol{x}^*, \boldsymbol{\theta}^{(N)})$ $\rightarrow$



$\mathrm{P}(y|\boldsymbol{x}^*, \mathcal{D})$

Ensemble for uncertain in-domain input

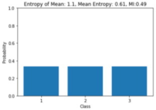
$\boldsymbol{\theta}^{(1)} \longrightarrow$



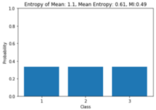
$\mathrm{P}(y|\boldsymbol{x}^*, \boldsymbol{\theta}^{(1)})$ $\rightarrow$

$\vdots$ $\vdots$

$\mathrm{p}(\boldsymbol{\theta}) + \mathcal{D} \rightarrow \mathrm{p}(\boldsymbol{\theta}|\mathcal{D}) \longrightarrow$ $\boldsymbol{\theta}^{(i)} \longrightarrow$



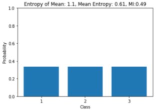
$\mathrm{P}(y|\boldsymbol{x}^*, \boldsymbol{\theta}^{(i)})$ $\rightarrow$



$\mathrm{P}(y|\boldsymbol{x}^*, \mathcal{D})$

$\vdots$ $\vdots$

$\boldsymbol{\theta}^{(N)} \longrightarrow$

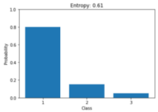


$\mathrm{P}(y|\boldsymbol{x}^*, \boldsymbol{\theta}^{(N)})$ $\rightarrow$

Ensemble for Out-of-Domain input

$$\mathrm{p}(\boldsymbol{\theta}) + \mathcal{D} \rightarrow \mathrm{p}(\boldsymbol{\theta}|\mathcal{D}) \longrightarrow$$

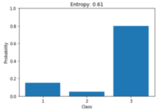
$$\boldsymbol{\theta}^{(1)} \longrightarrow$$



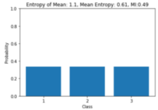
$\mathrm{P}(y|\boldsymbol{x}^*, \boldsymbol{\theta}^{(1)})$ $\rightarrow$

$\vdots$

$$\boldsymbol{\theta}^{(i)} \longrightarrow$$



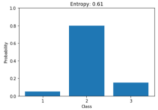
$\mathrm{P}(y|\boldsymbol{x}^*, \boldsymbol{\theta}^{(i)})$ $\rightarrow$



$\mathrm{P}(y|\boldsymbol{x}^*, \mathcal{D})$

$\vdots$

$$\boldsymbol{\theta}^{(N)} \longrightarrow$$



$\mathrm{P}(y|\boldsymbol{x}^*, \boldsymbol{\theta}^{(N)})$ $\rightarrow$

Measures of Uncertainty

- If the predictions from the models are consistent

$$\underbrace{\mathcal{H}\big[\mathbb{E}_{\mathrm{p}(\boldsymbol{\theta}|\mathcal{D})}[\mathrm{P}(y|\boldsymbol{x}^*,\boldsymbol{\theta})]\big]}_{\text{Total Uncertainty}} - \underbrace{\mathbb{E}_{\mathrm{p}(\boldsymbol{\theta}|\mathcal{D})}\big[\mathcal{H}[\mathrm{P}(y|\boldsymbol{x}^*,\boldsymbol{\theta})]\big]}_{\text{Expected Data Uncertainty}} = 0$$

- If the predictions from the models are diverse

$$\underbrace{\mathcal{H}\big[\mathbb{E}_{\mathrm{p}(\boldsymbol{\theta}|\mathcal{D})}[\mathrm{P}(y|\boldsymbol{x}^*,\boldsymbol{\theta})]\big]}_{\text{Total Uncertainty}} - \underbrace{\mathbb{E}_{\mathrm{p}(\boldsymbol{\theta}|\mathcal{D})}\big[\mathcal{H}[\mathrm{P}(y|\boldsymbol{x}^*,\boldsymbol{\theta})]\big]}_{\text{Expected Data Uncertainty}} > 0$$

- Difference of the two is a measure of knowledge uncertainty

$$\underbrace{\mathcal{I}[y,\boldsymbol{\theta}|\boldsymbol{x}^*,\mathcal{D}]}_{\text{Knowledge Uncertainty}} = \underbrace{\mathcal{H}\big[\mathbb{E}_{\mathrm{p}(\boldsymbol{\theta}|\mathcal{D})}[\mathrm{P}(y|\boldsymbol{x}^*,\boldsymbol{\theta})]\big]}_{\text{Total Uncertainty}} - \underbrace{\mathbb{E}_{\mathrm{p}(\boldsymbol{\theta}|\mathcal{D})}\big[\mathcal{H}[\mathrm{P}(y|\boldsymbol{x}^*,\boldsymbol{\theta})]\big]}_{\text{Expected Data Uncertainty}}$$

Approximate Inference

- Variational Inference:
  - Bayes by Backprop [Blundell et al., 2015]
  - Probabalistic Backpropagation [Hernández-Lobato and Adams, 2015]
- Monte-Carlo Methods:
  - Monte-Carlo Dropout [Gal, 2016, Gal and Ghahramani, 2016]
  - Stochastic Gradient Langevin Dynamics [Welling and Teh, 2011]
  - Fast-Ensembling via Mode Connectivity [Garipov et al., 2018]
  - Stochastic Weight Averaging Gaussian (SWAG) [Maddox et al., 2019]
- Non-Bayesian Ensembles:
  - Bootstrap DQN [Osband et al., 2016]
  - Deep Ensembles [Lakshminarayanan et al., 2017]

Limitations

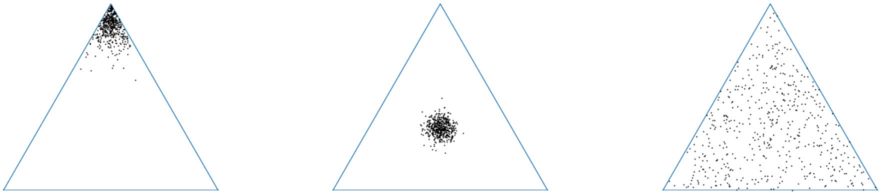
- Hard to guarantee diverse $\{\mathrm{P}(y|\boldsymbol{x}^*, \boldsymbol{\theta}^{(m)})\}_{m=1}^{M}$
- Diversity of ensemble depends on:
  - Selection of prior
  - Nature of approximations
  - Architecture of network
  - Properties and size of data
- Computationally expensive

Overview

1. Motivation: Why do we need Uncertainty Estimation?
2. Sources of Uncertainty in Predictions
3. Ensemble Approaches
4. **Ensemble Distribution Distillation**
5. Uncertainty in Structured Prediction

Distributions on a Simplex

- Ensemble $\{\mathtt{P}(y|\boldsymbol{x}^*, \boldsymbol{\theta}^{(m)})\}_{m=1}^{M}$ can be visualized on a simplex



(a) Confident (b) Data Uncertainty (c) Knowledge Uncertainty

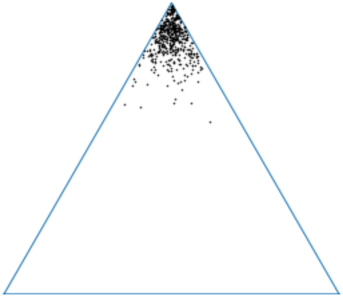
- Same as sampling from **implicit** Distribution over output Distributions

$$\mathtt{P}(y|\boldsymbol{x}^*, \boldsymbol{\theta}^{(m)}) \sim \mathtt{p}(\boldsymbol{\theta}|\mathcal{D}) \equiv \boldsymbol{\mu}^{(m)} \sim \mathtt{p}(\boldsymbol{\mu}|\boldsymbol{x}^*, \mathcal{D})$$

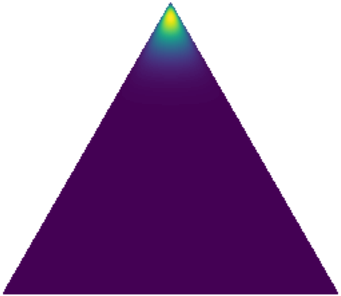
Distributions on a Simplex (cont)

- Expanding out $\boldsymbol{\mu}^{(m)} = \begin{bmatrix} \mathrm{P}(y = \omega_1) \\ \mathrm{P}(y = \omega_2) \\ \vdots \\ \mathrm{P}(y = \omega_K) \end{bmatrix}$, where each $\boldsymbol{\mu}^{(m)}$ is a point on a simplex.

Distribution over Distributions

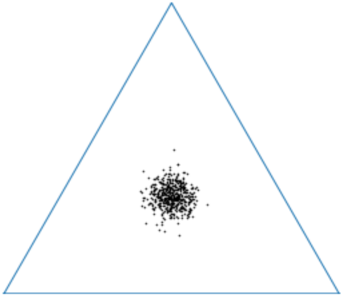


(a) $\{\boldsymbol{\mu}^{(m)}\}_{m=1}^{M}$

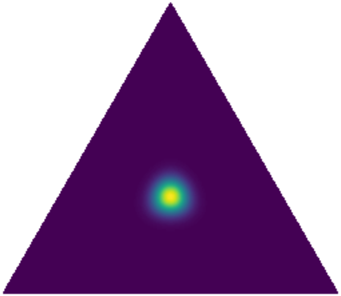


(b) $\mathtt{p}(\boldsymbol{\mu}|\boldsymbol{x}^*, \mathcal{D})$

Distribution over Distributions

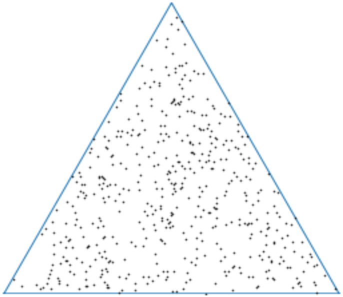


(a) $\{\boldsymbol{\mu}^{(m)}\}_{m=1}^{M}$

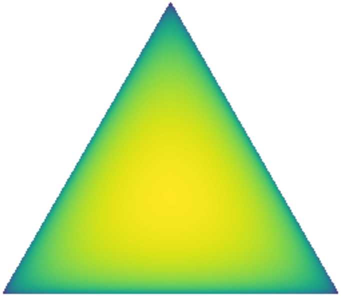


(b) $\mathrm{p}(\boldsymbol{\mu}|\boldsymbol{x}^*, \mathcal{D})$

Distribution over Distributions



(a) $\{\boldsymbol{\mu}^{(m)}\}_{m=1}^{M}$



(b) $\mathrm{p}(\boldsymbol{\mu}|\boldsymbol{x}^*, \mathcal{D})$

Prior Networks [Malinin and Gales, 2018]

- **Explicitly** model $\mathrm{p}(\boldsymbol{\mu}|\boldsymbol{x}^*, \mathcal{D})$ using a Prior Network $\mathrm{p}(\boldsymbol{\mu}|\boldsymbol{x}^*; \hat{\boldsymbol{\theta}})$

$$\mathrm{p}(\boldsymbol{\mu}|\boldsymbol{x}^*; \hat{\boldsymbol{\theta}}) \approx \mathrm{p}(\boldsymbol{\mu}|\boldsymbol{x}^*, \mathcal{D})$$

- Predictive posterior distribution is given by expected categorical

$$\mathrm{P}(y|\boldsymbol{x}^*; \hat{\boldsymbol{\theta}}) = \mathbb{E}_{\mathrm{p}(\boldsymbol{\mu}|\boldsymbol{x}^*; \hat{\boldsymbol{\theta}})}\big[\mathrm{p}(y|\boldsymbol{\mu})\big] = \hat{\boldsymbol{\mu}}$$

Uncertainty Measures for Prior Networks

- Ensemble uncertainty decomposition:

$$\underbrace{\mathcal{I}[y,\boldsymbol{\theta}|\boldsymbol{x}^*,\mathcal{D}]}_{\text{Knowledge Uncertainty}} = \underbrace{\mathcal{H}[\mathbb{E}_{\mathrm{p}(\boldsymbol{\theta}|\mathcal{D})}[\mathrm{P}(y|\boldsymbol{x}^*,\boldsymbol{\theta})]]}_{\text{Total Uncertainty}} - \underbrace{\mathbb{E}_{\mathrm{p}(\boldsymbol{\theta}|\mathcal{D})}[\mathcal{H}[\mathrm{P}(y|\boldsymbol{x}^*,\boldsymbol{\theta})]]}_{\text{Expected Data Uncertainty}}$$

- Prior Network uncertainty decomposition

$$\underbrace{\mathcal{I}[y,\boldsymbol{\mu}|\boldsymbol{x}^*;\hat{\boldsymbol{\theta}}]}_{\text{Knowledge Uncertainty}} = \underbrace{\mathcal{H}\big[\mathbb{E}_{\mathrm{p}(\boldsymbol{\mu}|\boldsymbol{x}^*;\hat{\boldsymbol{\theta}})}[\mathrm{P}(y|\boldsymbol{\mu})]\big]}_{\text{Total Uncertainty}} - \underbrace{\mathbb{E}_{\mathrm{p}(\boldsymbol{\mu}|\boldsymbol{x}^*;\hat{\boldsymbol{\theta}})}\big[\mathcal{H}[\mathrm{P}(y|\boldsymbol{\mu})]\big]}_{\text{Expected Data Uncertainty}}$$

Ensemble Distillation (EnD) [Hinton et al., 2015, Korattikara et al., 2015]

- Ensembles are computationally expensive
  - Distill an ensemble into a **single** model

$$\{\mathrm{P}(y|\boldsymbol{x}, \boldsymbol{\theta}^{(m)})\}_{m=1}^{M} \rightarrow \mathrm{P}(y|\boldsymbol{x}, \hat{\boldsymbol{\theta}})$$

- Minimize KL-divergence to mean of ensemble:

$$\mathcal{L}(\hat{\boldsymbol{\theta}}, \mathcal{D}) = \mathbb{E}_{\mathrm{p}(\boldsymbol{x})}\Big[\mathrm{KL}\big[\mathbb{E}_{\hat{\mathrm{p}}(\boldsymbol{\theta}|\mathcal{D})}[\mathrm{P}(y|\boldsymbol{x}, \boldsymbol{\theta})]||\mathrm{P}(y|\boldsymbol{x}, \hat{\boldsymbol{\theta}})\big]\Big]$$

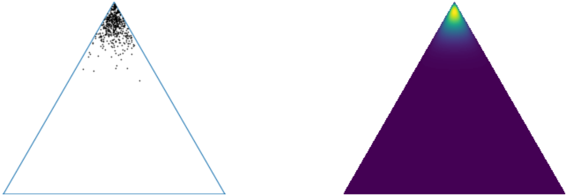
- Computational Performance gain
- Robustness to Adversarial Attack (Defensive Distillation)

Ensemble Distillation (EnD)

- EnD $\rightarrow$ model captures only *mean* of ensemble
- Diversity of ensemble is lost $\rightarrow$
  - Cannot separate measures of uncertainty
- Solution $\rightarrow$ Ensemble Distribution Distillation

Ensemble Distribution Distillation (End^2) [Malinin et al., 2019]

- Distill an ensemble into a **single** Prior Network



$$\{\mathrm{P}(y|\boldsymbol{x},\boldsymbol{\theta}^{(m)})\}_{m=1}^{M} \quad \rightarrow \quad \mathrm{p}(\boldsymbol{\mu}|\boldsymbol{x};\hat{\boldsymbol{\theta}})$$

- Goal $\rightarrow$ Maximum information extraction from ensemble.

Ensemble Distribution Distillation (End^2)

- Parameterize a Dirichlet distribution using Neural Network:

$$\mathrm{p}(\boldsymbol{\mu}|\boldsymbol{x};\boldsymbol{\theta}) = \mathtt{Dir}(\boldsymbol{\mu};\boldsymbol{\alpha}), \quad \boldsymbol{\alpha} = \ \boldsymbol{f}(\boldsymbol{x};\boldsymbol{\theta}), \quad \alpha_c > 0$$

- Training data are ensemble predictions for every input:

$$\mathcal{D} = \left\{\left\{p(y|\boldsymbol{x}^{(i)};\boldsymbol{\theta}^{(j)}), \boldsymbol{x}^{(i)}\right\}_{j=1}^{N}\right\}_{i=1}^{M} \sim \hat{\mathrm{p}}(\boldsymbol{\mu},\boldsymbol{x})$$

- Train via Maximum Likelihood:

$$\mathcal{L}(\boldsymbol{\theta},\mathcal{D}) = \ -\mathbb{E}_{\hat{\mathrm{p}}(\boldsymbol{x})}\Big[\mathbb{E}_{\hat{\mathrm{p}}(\boldsymbol{\mu}|\boldsymbol{x})}[\ln \mathrm{p}(\boldsymbol{\mu}|\boldsymbol{x};\boldsymbol{\theta})]\Big]$$

Ensemble Distribution Distillation: Image Classification

Dataset	Individual	Ensemble	EnD	EnD^2
CIFAR-10	8.0	**6.2**	6.7	6.9
CIFAR-100	30.4	**26.3**	28.2	28.0
TinyImageNet	41.8	**36.6**	38.5	37.3

Table: Classification Performance (% Error).

Ensemble Distribution Distillation: Misclassification Detection

Dataset	Individual	Ensemble	EnD	EnD^2
CIFAR-10	84.6	**86.8**	85.1	85.7
CIFAR-100	72.5	**75.0**	74.0	74.0
TinyImageNet	70.8	**73.8**	72.6	72.7

Table: Misclassification detection performance (% PRR).

Ensemble Distribution Distillation: OOD Detection

Test OOD Dataset	Model	CIFAR-10		CIFAR-100	
		Total Unc.	Knowledge Unc.	Total Unc.	Knowledge Unc.
LSUN	Individual	91.3	-	75.6	-
	EnD	89.0	-	76.5	-
	EnD^2	94.4	**95.3**	83.5	86.9
	Ensemble	94.5	94.4	82.4	**88.4**
TIM	Individual	88.9	-	70.5	-
	EnD	86.9	-	70.0	-
	EnD^2	91.3	**91.8**	76.4	79.3
	Ensemble	**91.8**	91.4	76.6	**81.7**

Table: OOD detection performance (% AUC-ROC) for CIFAR-10 and CIFAR-100 models.

Overview

1. Motivation: Why do we need Uncertainty Estimation?
2. Sources of Uncertainty in Predictions
3. Ensemble Approaches
4. Ensemble Distribution Distillation
5. **Uncertainty in Structured Prediction**

Auto-regressive structured prediction models

- Consider auto-regressive model mapping $\boldsymbol{X} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T\} \rightarrow \boldsymbol{y} = \{y_1, \cdots, y_L\}$:

$$\mathrm{P}(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta}) = \prod_{l=1}^{L} \mathrm{P}(y_l|\boldsymbol{y}_{<l}, \boldsymbol{X}; \boldsymbol{\theta})$$

- Suchs models are commonly applied for
  - Neural Machine Translation (NMT)
  - End-to-End Automatic Speech Recognition (ASR)
- Can ensemble methods to be applied to obtain uncertainty estimates?
- Can we consider uncertainties are multiple levels?
  - Token-level uncertainties
  - Sequence-level uncertainties

Ensemble combination for auto-regressive models

- Consider an ensemble of auto-regressive models

$$\left\{\mathrm{P}(y_l|\boldsymbol{y}_{<l}, \boldsymbol{X}, \boldsymbol{\theta}^{(m)})\right\}_{m=1}^{M}, \quad \boldsymbol{\theta}^{(m)} \sim \mathrm{p}(\boldsymbol{\theta}|\mathcal{D})$$

- We can combine the predictive posterior as an *expectation-of-products* (ExPr)...

$$\mathrm{P}(\boldsymbol{y}|\boldsymbol{X}, \mathcal{D}) = \mathbb{E}_{\mathrm{p}(\boldsymbol{\theta}|\mathcal{D})}\Big[\mathrm{P}(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta})\Big] = \mathbb{E}_{\mathrm{p}(\boldsymbol{\theta}|\mathcal{D})}\Big[\prod_{l=1}^{L}\mathrm{P}(y_l|\boldsymbol{y}_{<l}, \boldsymbol{X}, \boldsymbol{\theta})\Big]$$

- ...or as a *product-of-expectations* (PrEx):

$$\mathrm{P}(\boldsymbol{y}|\boldsymbol{X}, \mathcal{D}) = \prod_{l=1}^{L}\mathrm{P}(y_l|\boldsymbol{y}_{<l}, \boldsymbol{X}, \mathcal{D}) = \prod_{l=1}^{L}\mathbb{E}_{\mathrm{p}(\boldsymbol{\theta}|\mathcal{D})}\Big[\mathrm{P}(y_l|\boldsymbol{y}_{<l}, \boldsymbol{X}, \boldsymbol{\theta})\Big]$$

- Both are **valid** ways of combining an ensemble of models.

Ensemble combination for auto-regressive models

Model	NMT BLEU		ASR WER	
	EN-DE	EN-FR	Libr-TC	Libr-TO
Single	28.8 ± 0.2	45.4 $\pm$ 0.4	5.6 ± 0.2	14.7 ± 0.5
PrEx	**30.0**	**46.3**	**4.3**	**11.3**
ExPr	29.6	46.2	4.5	12.5

Table: Beam-Search decoding BLEU/WER on newstest14/LibriSpeech.

Ensemble combination for auto-regressive models

Model	NMT		ASR	
	EN-DE	EN-FR	Libr-TC	Libr-TO
PrEx	**1.352**	**1.043**	**0.209**	**0.501**
ExPr	1.381	1.052	0.236	0.606

Table: Teacher-forcing NLL on newstest14 and LibriSpeech.

Token-level measures of uncertainty

- Token-level measures of uncertainty are a direct application of ensemble methods:

$$\underbrace{\mathcal{I}[y_l, \boldsymbol{\theta}|\boldsymbol{y}_{<l}, \boldsymbol{X}, \mathcal{D}]}_{\text{Knowledge Uncertainty}} = \underbrace{\mathcal{H}[\mathrm{P}(y_l|\boldsymbol{y}_{<l}, \boldsymbol{X}, \mathcal{D})]}_{\text{Total Uncertainty}} - \underbrace{\mathbb{E}_{\mathrm{p}(\boldsymbol{\theta}|\mathcal{D})}[\mathcal{H}[\mathrm{P}(y_l|\boldsymbol{y}_{<l}, \boldsymbol{X}, \boldsymbol{\theta})]]}_{\text{Expected Data Uncertainty}}$$

- Uncertainty in prediction given an input $\boldsymbol{X}$ and context $\boldsymbol{y}_{<l}$

Sequence-level measures of uncertainty

- Sequence-level uncertainty estimates are more challenging to obtain:

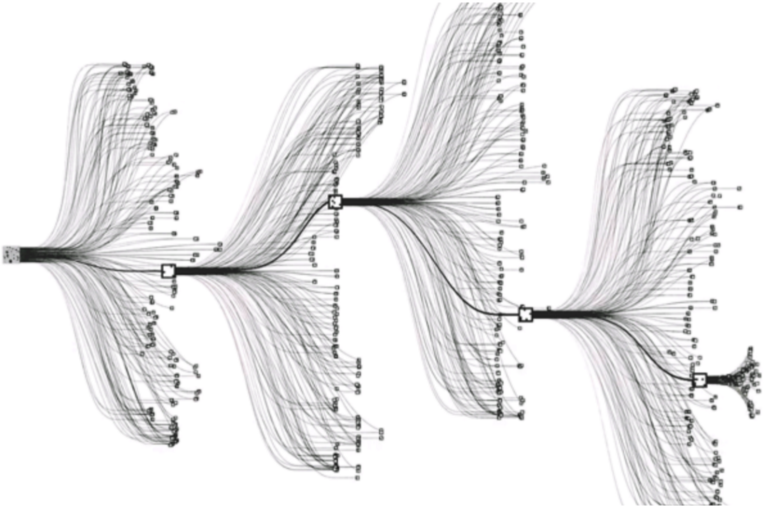
$$\underbrace{\mathcal{I}[\boldsymbol{y},\boldsymbol{\theta}|\boldsymbol{X},\mathcal{D}]}_{\text{Know. Uncertainty}} = \underbrace{\mathcal{H}[\mathrm{P}(\boldsymbol{y}|\boldsymbol{X},\mathcal{D})]}_{\text{Total Uncertainty}} - \underbrace{\mathbb{E}_{\mathrm{p}(\boldsymbol{\theta}|\mathcal{D})}[\mathcal{H}[\mathrm{P}(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta})]]}_{\text{Expected Data Uncertainty}}$$

- Consider the expression for entropy:

$$\begin{aligned}\mathcal{H}[\mathrm{P}(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta})] &= -\sum_{\boldsymbol{y}} \mathrm{P}(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta}) \ln \mathrm{P}(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta}) \\ &= -\sum_{y_1} \cdots \sum_{y_L} \mathrm{P}(y_1,\cdots,y_L|\boldsymbol{X},\boldsymbol{\theta}) \ln \mathrm{P}(y_1,\cdots,y_L|\boldsymbol{X},\boldsymbol{\theta})\end{aligned}$$

- Intractable to compute!

Sequence-level measures of uncertainty



Sequence-level measures of uncertainty

- However, we can approximate sequence-level uncertainty as follows:

$$\mathcal{H}[\mathrm{P}(\boldsymbol{y}|\boldsymbol{X},\mathcal{D})] \approx \sum_{l=1}^{L} \mathcal{H}\big[\mathbb{E}_{\mathrm{q}(\boldsymbol{\theta})}[\mathrm{P}(y_l|\boldsymbol{y}_{<l},\boldsymbol{X},\boldsymbol{\theta})]\big]$$

$$\mathbb{E}_{\mathrm{q}(\boldsymbol{\theta})}\big[\mathcal{H}[\mathrm{P}(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta})]\big] \approx \sum_{l=1}^{L} \mathbb{E}_{\mathrm{q}(\boldsymbol{\theta})}\big[\mathcal{H}[\mathrm{P}(y_l|\boldsymbol{y}_{<l},\boldsymbol{X},\boldsymbol{\theta})]\big]$$

$$\mathcal{I}\big[\boldsymbol{y},\boldsymbol{\theta}|\boldsymbol{X},\mathcal{D}\big] \approx \sum_{l=1}^{L} \mathcal{I}\big[y_l,\boldsymbol{\theta}|\boldsymbol{y}_{<l},\boldsymbol{X},\mathcal{D}\big]$$

- Approximation are accurate if:
  - We combine ensemble as a *Product-of-Expectations*
  - Distributions are independent of context

Sequence-level Error Detection

Task	Test set	Total Uncertainty		Data Uncertainty	Knowledge Uncertainty	
		TU	SCR-PE		MI	EPKL
ASR	Libr-TC	**67.0**	66.6	66.6	64.2	62.3
	Libr-TO	**73.3**	72.3	71.7	70.9	67.4
NMT EN-DE	newstest14	37.5	**45.7**	36.4	30.1	28.5
NMT EN-FR		32.6	**37.9**	37.8	32.8	31.9

Table: Sequence-level Error Detection % PRR in Beam-Search decoding regime.

Token-level Error Detection

Task	Test Data	Total Uncertainty TU	Total Uncertainty SCR-PE	Data Uncertainty	Knowledge Uncertainty MI	Knowledge Uncertainty EPKL	% Error
ASR	Libr-TC	36.8	**37.2**	35.1	33.9	29.6	3.9
	Libr-TO	**44.1**	43.5	42.6	41.9	37.8	10.3

Table: Token-level Error Detection %AUPR for LibriSpeech in Beam-Search Decoding regime.

OOD Detection for Speech Recognition

ID Data	OOD Data	Total Uncertainty	Data Uncertainty	Knowledge Unc. MI	Knowledge Unc. EPKL
Libr-TC	Libr-TO	**77.4**	76.4	76.7	77.1
Libr-TC	AMI-EVL	97.1	**97.2**	95.7	95.4
Libr-TO	AMI-EVL	90.9	**91.0**	88.0	86.7
Libr-TC	LNG-FR	**100.0**	**100.0**	99.9	99.9
	LNG-RU	**100.0**	**100.0**	99.9	99.9

Table: OOD Detection % ROC-AUC in Beam-Search decoding regime for ASR.

OOD Detection for Translation

ID Data	OOD Data	Total Uncertainty	Data Uncertainty	Knowledge Unc. MI	Knowledge Unc. EPKL
newstst14	MED	52.2	50.8	64.9	**65.2**
	Libr-TC	74.4	72.9	**77.1**	76.5
	Libr-TO	72.0	70.6	**76.2**	75.9
	PERM	83.9	80.3	97.0	**97.3**
	LNG-DE	33.5	29.7	73.2	**78.2**
	LNG-FR	20.4	19.1	57.7	**64.7**

Table: OOD Detection % ROC-AUC in Beam-Search decoding regime for NMT.

Take away points

- Uncertainty is important $\rightarrow$
  - Philosophically and practically necessary
- Sources of Uncertainty $\rightarrow$
  - Data Uncertainty and Knowledge Uncertainty
- Uncertainty Estimation via Ensembles $\rightarrow$
  - Theoretically motivated separation of uncertainty sources
  - Can reduce complexity via Ensemble Distribution Distillation
- Uncertainty for Structured Prediction $\rightarrow$
  - Can use ensembles of auto-regressive models
  - Successfully applied for ASR/NMT error detection and OOD detection
  - Cool new effects!

Thank You!

Any questions?

References I

[Blundell et al., 2015] Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015).
Weight uncertainty in neural networks.
arXiv preprint arXiv:1505.05424.

[Gal, 2016] Gal, Y. (2016).
Uncertainty in Deep Learning.
PhD thesis, University of Cambridge.

[Gal and Ghahramani, 2016] Gal, Y. and Ghahramani, Z. (2016).
Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning.
In *Proc. 33rd International Conference on Machine Learning (ICML-16)*.

References II

[Garipov et al., 2018] Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D. P., and Wilson, A. G. (2018).
Loss surfaces, mode connectivity, and fast ensembling of dnns.
In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 8789–8798. Curran Associates, Inc.

[Hernández-Lobato and Adams, 2015] Hernández-Lobato, J. M. and Adams, R. (2015).
Probabilistic backpropagation for scalable learning of bayesian neural networks.
In *International Conference on Machine Learning*, pages 1861–1869.

References III

[Hinton et al., 2015] Hinton, G., Vinyals, O., and Dean, J. (2015).
Distilling the knowledge in a neural network.
In *NIPS Deep Learning and Representation Learning Workshop*.

[Korattikara et al., 2015] Korattikara, A., Rathod, V., Murphy, K. P., and Welling, M. (2015).
Bayesian dark knowledge.
In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 3438–3446. Curran Associates, Inc.

References IV

[Lakshminarayanan et al., 2017] Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017).
Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles.
In *Proc. Conference on Neural Information Processing Systems (NIPS)*.

[Maddox et al., 2019] Maddox, W., Garipov, T., Izmailov, P., Vetrov, D. P., and Wilson, A. G. (2019).
A simple baseline for bayesian uncertainty in deep learning.
CoRR, abs/1902.02476.

[Malinin and Gales, 2018] Malinin, A. and Gales, M. (2018).
Predictive uncertainty estimation via prior networks.
In *Advances in Neural Information Processing Systems*, pages 7047–7058.

[Malinin et al., 2019] Malinin, A., Mlodozeniec, B., and Gales, M. (2019).
Ensemble distribution distillation.
arXiv preprint arXiv:1905.00076.

[Osband et al., 2016] Osband, I., Blundell, C., Pritzel, A., and Van Roy, B. (2016).
Deep exploration via bootstrapped dqn.
In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 4026–4034. Curran Associates, Inc.

[Welling and Teh, 2011] Welling, M. and Teh, Y. W. (2011).
Bayesian Learning via Stochastic Gradient Langevin Dynamics.
In *Proc. International Conference on Machine Learning (ICML)*.