Deep Classifiers with Label Noise Modeling and Distance Awareness

2022.12.23 (Fri.)

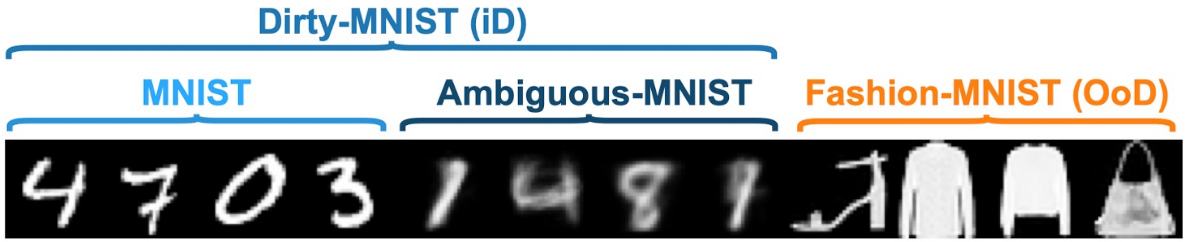
Superb AI Machine Learning Team

Presenter : Kyeongryeol, Go

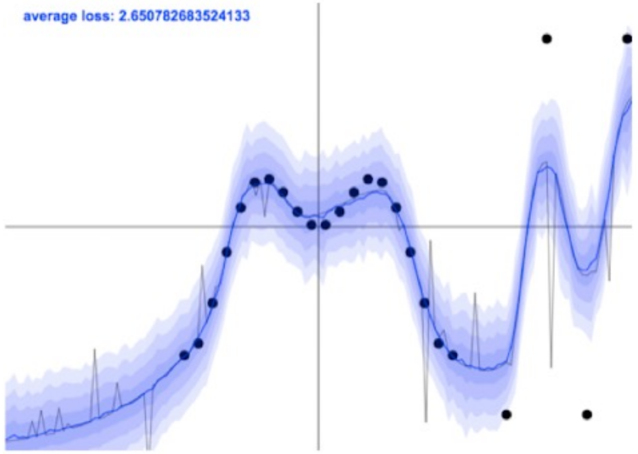
Two types of uncertainty

1. Epistemic (model) uncertainty
   - Lack of knowledge about data generating mechanism
   - model mis-specification (structural), <u>parameter estimation</u> (parametric)
   - reducible (effective in small data regime)
   - out-of-distribution detection, active learning

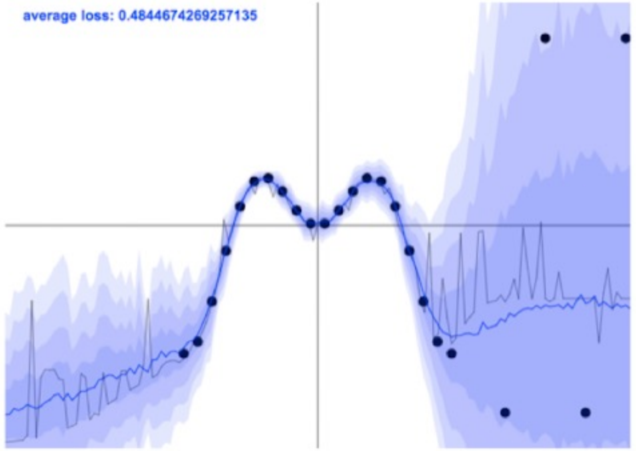
2. Aleatoric (data) uncertainty
   - Stochastic variability inherent in data generating process
   - measurement noise (regression), <u>labeling error</u> (classification)
   - irreducible (effective in big data regime)
   - in-distribution calibration, mis-label detection



aleatoric ↑ epistemic ↑



Homoscedastic



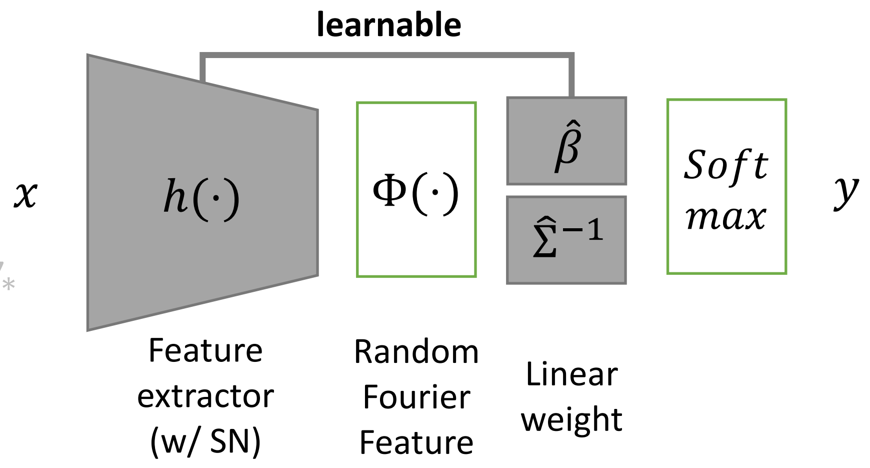
Heteroscedastic

Kendall, Alex, and Yarin Gal. "What uncertainties do we need in bayesian deep learning for computer vision?." *Advances in neural information processing systems* 30 (2017).

Epistemic : From SNGP..

1. Make the “feature extractor” input distance-preserving
   - apply spectral normalization (SN) with residual connection

2. Make “the classifier” feature distance-aware
   - use gaussian process to feature outputs (not scalable)
     - $p(D) = \mathcal{N}\big(0, K(h,h)\big) \rightarrow p(y|x,D) = K(h_*,h)K(h,h)^{-1}y_*$
   - use random fourier feature and a linear weight β
     - $K(h,h) \approx \Phi(h)\Phi(\mathrm{h})^{\mathrm{T}} \Rightarrow \Phi(h)^T\beta \quad where \quad \beta \sim \mathcal{N}(0,I)$
   - use laplace approximation to estimate $p(\beta|D)$
     - $p(y|x,D) = \mathbb{E}_{\beta\sim p(\beta|D)}[softmax(\Phi(h)^T\beta)]$



Epistemic uncertainty : $H(y|x,D) = -\int p(y|x,D)\log p(y|x,D)$

Liu, Jeremiah, et al. "Simple and principled uncertainty estimation with deterministic deep learning via distance awareness." *Advances in Neural Information Processing Systems* 33 (2020): 7498-7512.

Aleatoric : From econometrics literature..

Latent utility : $u^{(c)} = l^{(c)} + \epsilon^{(c)}$

$$p^{(c)} = p(y = c | x, D) = p\left(u^{(c)} > u^{(k)}, \forall k \neq c\right) = p\left(\arg\max_k u^{(k)} = c\right)$$

$$\Rightarrow \mathbb{E}_{\epsilon \sim G(0,1)}\left[1\left\{\arg\max_k u^{(k)} = c\right\}\right] = \exp\left(u^{(c)}\right) / \sum_{k=1}^{K} \exp\left(u^{(k)}\right) \qquad \text{(homoscedastic, i.i.d)}$$

$$\Rightarrow \mathbb{E}_{\epsilon \sim \mathcal{N}(0,\sigma(x;w))}\left[1\left\{\arg\max_k u^{(k)} = c\right\}\right] = \mathbb{E}\left[\lim_{\tau \to 0} \frac{\exp\left(u^{(c)}/\tau\right)}{\sum_{k=1}^{K} \exp\left(u^{(k)}/\tau\right)}\right] \qquad \text{(heteroscedastic, i.}\cancel{\text{i}}\text{,d)}$$

$$\approx \mathbb{E}\left[\frac{\exp\left(u^{(c)}/\tau\right)}{\sum_{k=1}^{K} \exp\left(u^{(k)}/\tau\right)}\right], \qquad \tau > 0$$

Bias-variance trade-off with temperature : $\tau \to 0 \Rightarrow$ bias $\downarrow$, variance $\uparrow$

Collier, Mark, et al. "A simple probabilistic method for deep classification under input-dependent label noise." *arXiv preprint arXiv:2003.06778* (2020).

Aleatoric : Inter-class correlation

Feature $\boldsymbol{h(x)}$ — **Output $\boldsymbol{l(x), \epsilon(x)}$** — **Sample $\boldsymbol{u(x)}$**

$h(x) \in \mathbb{R}^D$

- logit: $\mathrm{l}(x) = W_l h(x) + b_l \in \mathbb{R}^K$
- diagonal covariance: $\sigma(x) = \exp(W_\sigma h(x) + b_\sigma) \in \mathbb{R}^K \quad \Rightarrow \quad l(x) + \sigma(x) \odot \epsilon_K$
- full covariance
  - low-rank approximation: $V(x) = W_V h(x) + b_V \in \mathbb{R}^{K\times R} \quad \Rightarrow \quad l(x) + V(x)\cdot\epsilon_R + d(x) \odot \epsilon'_K$
  - further parameter-efficient: $v(x) = W_v h(x) + b_v \in \mathbb{R}^K, V \in \mathbb{R}^{K\times R} \quad \Rightarrow \quad l(x) + v(x) \odot (V\cdot\epsilon_R) + d(x) \odot {\epsilon_K}'$

$$\Rightarrow \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\Sigma(x;w))}\left[1\left\{\arg\max_k u^{(k)} = c\right\}\right] = \mathbb{E}\left[\lim_{\tau\to 0}\frac{\exp\left(u^{(c)}/\tau\right)}{\sum_{k=1}^K \exp(u^{(k)}/\tau)}\right]$$

(heteroscedastic, ~~i.i~~,d)

$$\approx \mathbb{E}\left[\frac{\exp\left(u^{(c)}/\tau\right)}{\sum_{k=1}^K \exp(u^{(k)}/\tau)}\right], \qquad \tau > 0$$

Covariance b/t commonly confused class pairs are strengthened during training

$$\Sigma(x;w) = V(x)V(x)^T + d^2(x)I \in \mathbb{R}^{K\times K} \quad where \quad V(x) \in R^{K\times R} \quad and \quad R \ll K$$

(More parameter-efficient version)

$$\Rightarrow \mathrm{V(x)} = \mathrm{v(x)}\mathbf{1}_R^T \odot V \in \mathbb{R}^{K\times R} \quad where \quad v(x) \in \mathbb{R}^K \quad and \quad V \in \mathbb{R}^{K\times R}$$

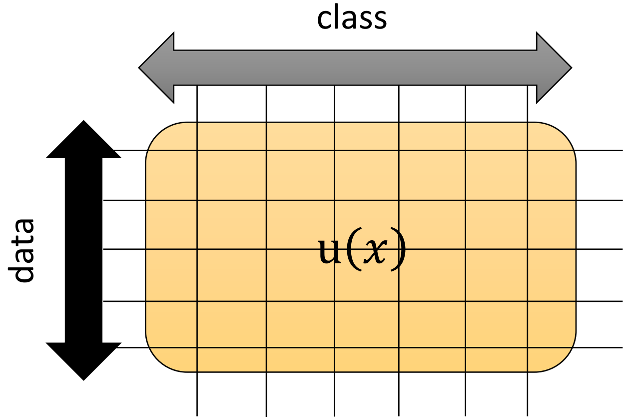
Collier, Mark, et al. "Correlated input-dependent label noise in large-scale image classification." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

Combining Epistemic (SNGP) and Aleatoric (Inter-class labeling noise)

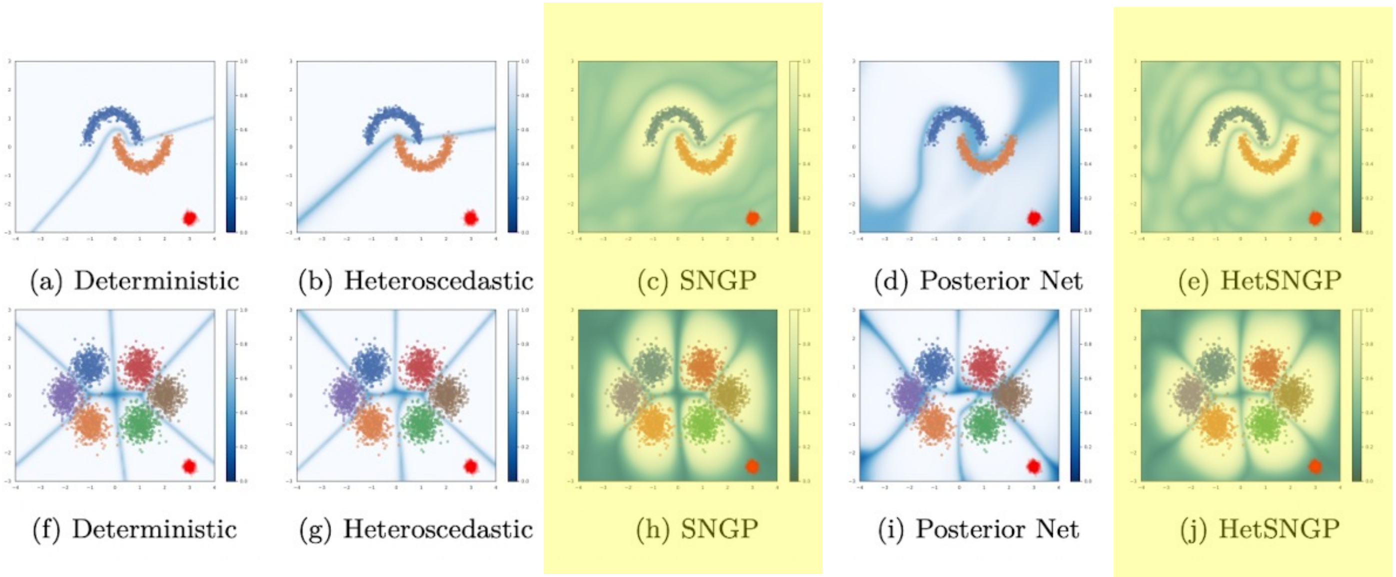
Model name : Heteroskedastic SNGP (Het-SNGP)

Latent utility : $u(x) = \boxed{\mathbb{E}_{\beta\sim p(\beta|D)}[\Phi(h(x))^T\beta]} + \boxed{v(x) \odot (V \cdot \epsilon_R) + d(x) \odot \epsilon'_K}$

<u>SNGP</u> <u>Inter-class labeling noise</u>

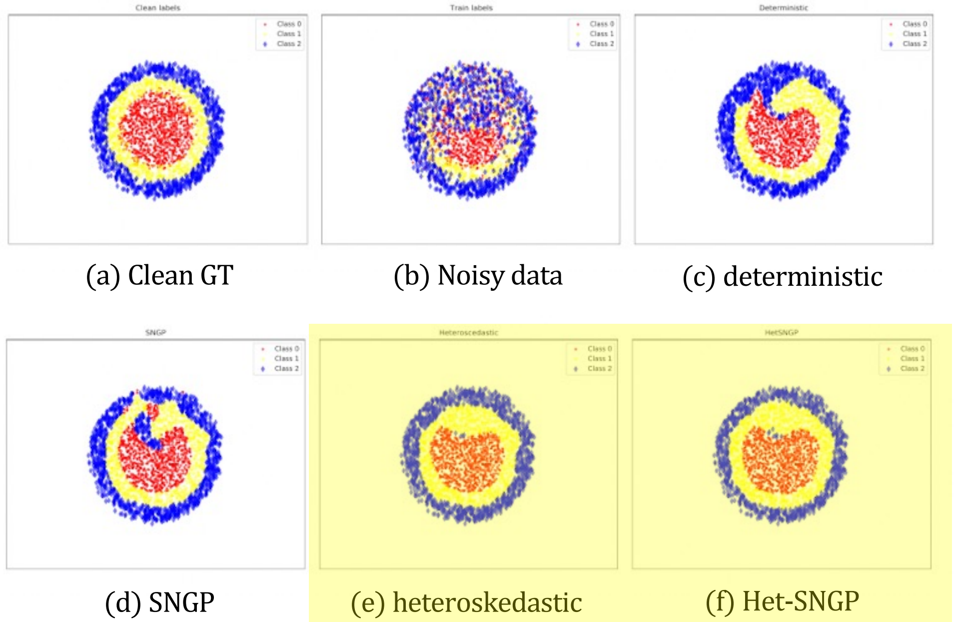


<u>"Whether combining the two can demonstrate the complementary benefits of the two methods"</u>



(a) Deterministic (b) Heteroscedastic (c) SNGP (d) Posterior Net (e) HetSNGP

(f) Deterministic (g) Heteroscedastic (h) SNGP (i) Posterior Net (j) HetSNGP



(a) Clean GT (b) Noisy data (c) deterministic

(d) SNGP (e) heteroskedastic (f) Het-SNGP

Fortuin, Vincent, et al. "Deep classifiers with label noise modeling and distance awareness." *arXiv preprint arXiv:2110.02609* (2021).

Combining Epistemic (SNGP) and Aleatoric (Inter-class labeling noise)

In-distribution (ID)

Out-of-distribution (OOD)

Method	↑ID Acc	↓ID NLL	↓ID ECE	↑ImC Acc	↓ImC NLL	↓ImC ECE	↑ImA Acc	↓ImA NLL	↓ImA ECE
Det.	0.759 ± 0.000	0.952 ± 0.001	0.033 ± 0.000	0.419 ± 0.001	3.078 ± 0.007	0.096 ± 0.002	0.006 ± 0.000	8.098 ± 0.018	0.421 ± 0.001
Het.	**0.771** ± 0.000	**0.912** ± 0.001	0.033 ± 0.000	0.424 ± 0.002	3.200 ± 0.014	0.111 ± 0.001	0.010 ± 0.000	7.941 ± 0.014	0.436 ± 0.001
SNGP	0.757 ± 0.000	0.947 ± 0.001	**0.014** ± 0.000	0.420 ± 0.001	**2.970** ± 0.007	**0.046** ± 0.001	0.007 ± 0.000	7.184 ± 0.009	**0.356** ± 0.000
HetSNGP (ours)	0.769 ± 0.001	0.927 ± 0.002	0.033 ± 0.000	**0.428** ± 0.001	2.997 ± 0.009	0.085 ± 0.001	**0.016** ± 0.001	**7.113** ± 0.018	0.401 ± 0.001

Method	↑ImR Acc	↓ImR NLL	↓ImR ECE	↑ImV2 Acc	↓ImV2 NLL	↓ImV2 ECE
Det.	0.229 ± 0.001	5.907 ± 0.014	0.239 ± 0.001	0.638 ± 0.001	1.598 ± 0.003	0.077 ± 0.001
Het.	**0.235** ± 0.001	5.761 ± 0.010	0.251 ± 0.001	**0.648** ± 0.001	1.581 ± 0.002	0.084 ± 0.001
SNGP	0.230 ± 0.001	**5.344** ± 0.009	**0.175** ± 0.001	0.637 ± 0.001	**1.552** ± 0.001	**0.041** ± 0.001
HetSNGP (ours)	0.232 ± 0.001	5.452 ± 0.011	0.225 ± 0.002	**0.647** ± 0.001	1.564 ± 0.003	0.080 ± 0.001

Method	↑ID Acc	↓ID NLL	↓ID ECE	↑ImC Acc	↓ImC NLL	↓ImC ECE
Det Ensemble	0.779	0.857	0.017	0.449	2.82	0.047
Het Ensemble	0.795	**0.790**	**0.015**	0.449	2.93	0.048
SNGP Ensemble	0.781	0.851	0.039	0.449	2.77	0.050
HetSNGP Ensemble (ours)	**0.797**	0.798	0.028	**0.458**	**2.75**	**0.044**

Further consider uncertainty over model parameter

Fortuin, Vincent, et al. "Deep classifiers with label noise modeling and distance awareness." *arXiv preprint arXiv:2110.02609* (2021).

E.O.D