# Feature Space Particle Inference for Neural Network Ensembles

2023.02.21 (Tue.)

Superb AI Machine Learning Team

Presenter : Kyeongryeol, Go

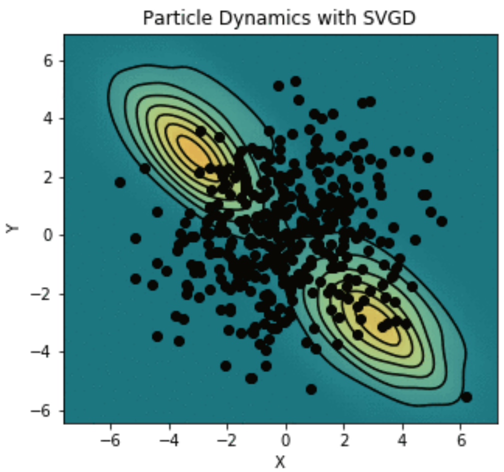
# Approximate Posterior Inference

Goal : Find $q(w)$ that approximates $p(w|D)$

- (Parametric) Variational Inference
  - $q(w)$ has certain parametric form (e.g. $\mathcal{N}(w; \mu, \sigma)$ )
  - $\min_{\{\mu,\sigma\}} D(q(w) \| p(w|D))$

- Markov Chain Monte Carlo (MCMC)
  - $w^{t+1} \sim m(w|w^t)$
  - Accept prob : $\min\left(1, \frac{p\left(w^{t+1}\middle|D\right)}{p\left(w^t\middle|D\right)}\right) = \min\left(1, \frac{p(w^{t+1})p\left(D\middle|w^{t+1}\right)}{p(w^t)p\left(D\middle|w^t\right)}\right)$
  - $q(w) \approx \frac{1}{T}\sum_{t=1}^{T} \delta(w - w^t)$

# Particle-based Variational Inference (PVI)




Goal : Find $q(w)$ that approximates $p(w|D)$

- $\{w_i^0\}_{i=1}^n \sim q_0(\mathrm{w}),\ \ \{w_i^1\}_{i=1}^n \sim \mathcal{A}q_0(w)$
- $\mathcal{A}\ :\ w_i^{t+1} \leftarrow w_i^t + \epsilon \cdot v\left(w_i^t\right)\quad \forall i,\ for\quad t = 0, \dots, T-1$

- $v = \arg\max_{v \in \mathcal{F}} \left\{ -\frac{d}{d\epsilon}\ D(\mathcal{A}q\ \|p)|_{\epsilon=0}\ s.t.\ \ \|v\|_{\mathcal{F}} \leq 1 \right\}$
- Steepest decreasing direction of the distance b/t $q$ and $p$

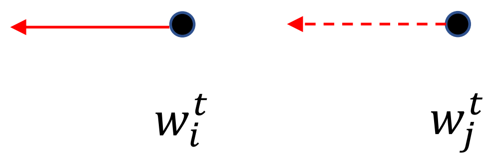
# Particle-based Variational Inference (PVI)

SVGD : $\mathcal{F} = \mathcal{H}$ s.t. $k(x, x') =< k(\cdot, x), k(\cdot, x') >_{\mathcal{H}}$,

- $v\left(w_i^t\right) = \sum_{j=1}^{n} k\left(w_i^t, w_j^t\right) \nabla_{w_j} \log p(w_j^t | D) + \nabla_{w_j} k\left(w_i^t, w_j^t\right)$

Driving term

Repulsive term




Compare : Langevin dynamics

- $v\left(w_i^t\right) = \nabla_{w_i} \log p(w_i^t | D) + \frac{2}{\sqrt{\epsilon}} z_i^t \quad where \quad z_i^t \sim \mathcal{N}(0,1)$
- Stochastic, computationally efficient

# Particle-based Variational Inference (PVI)

WGD : $\mathcal{F} = \mathcal{W}$,

- $v(w_i^t) = \nabla_{w_i} \log p(w_i^t | D) - \nabla_{w_i} \log q(w_i^t)$

Kernel density estimation (KDE) → vulnerable to curse of dimensionality

- $q(w_i^t) \propto \sum_{j=1}^{n} k(w_i^t, w_j^t)$
- $\nabla_{w_i} \log q(w_i^t) = \sum_{j=1}^{n} \nabla_{w_i} k(w_i^t, w_j^t) \,/\, \sum_{j=1}^{n} k(w_i^t, w_j^t)$

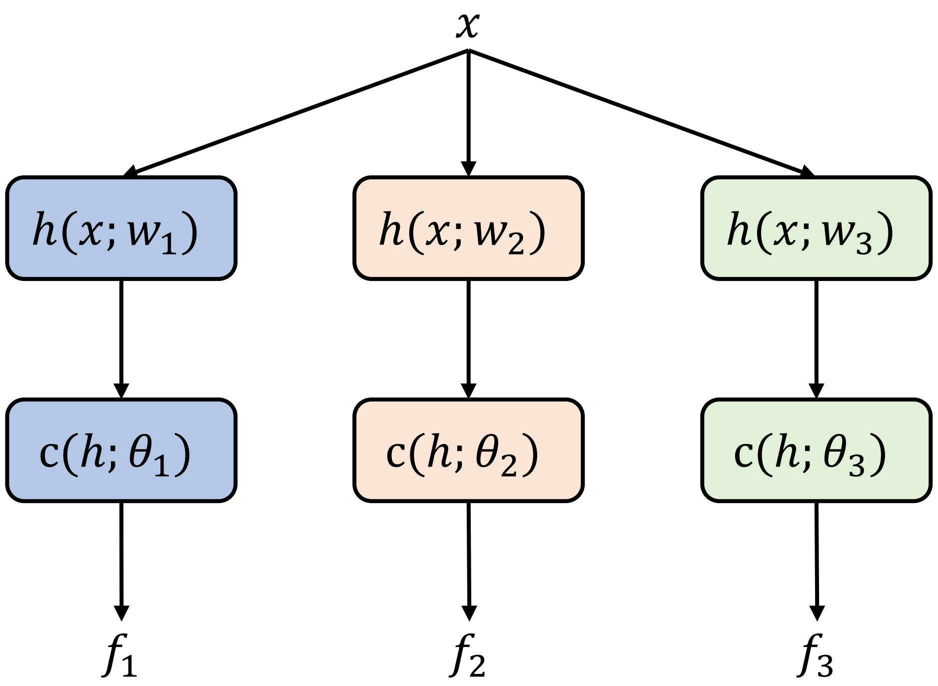
In summary

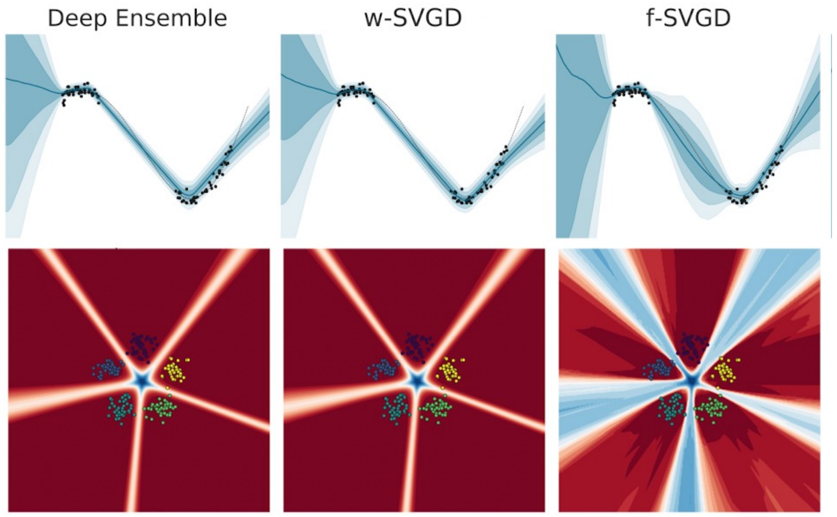
- greater flexibility than parametric VI
- greater sampling efficiency than MCMC
- lower redundancy than deep ensemble

# Where to apply PVI - previous

1. weight-WGD : $\phi = \{w, \theta\}$
   - $p(\phi|D) = p(\phi)\prod_{(x,y)\in D} p(y|x,\phi)$
   - $\phi_i^{t+1} \leftarrow \phi_i^t + \epsilon \cdot v(\phi_i^t)$
   - Overparameterized nature

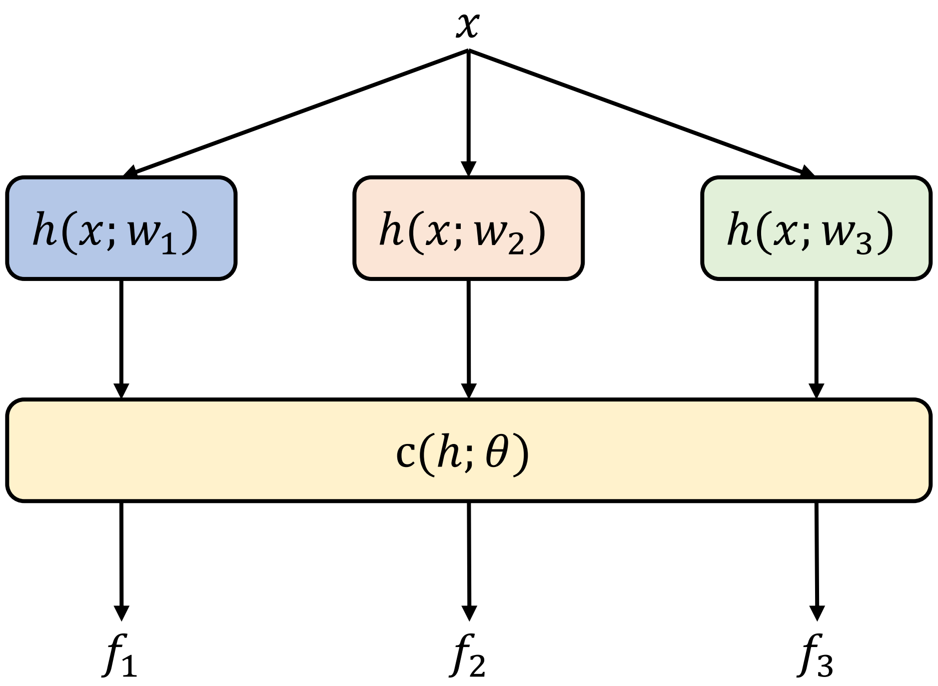
2. function-WGD : $f$
   - $p(f|D) = p(f)\prod_{(x,y)\in D} p(y|x,f)$
   - $\phi_i^{t+1} \leftarrow \phi_i^t + \epsilon \cdot \left(\frac{df_i^t}{d\phi_i^t}\right)^T v(f_i^t)$
   - Severe underfitting



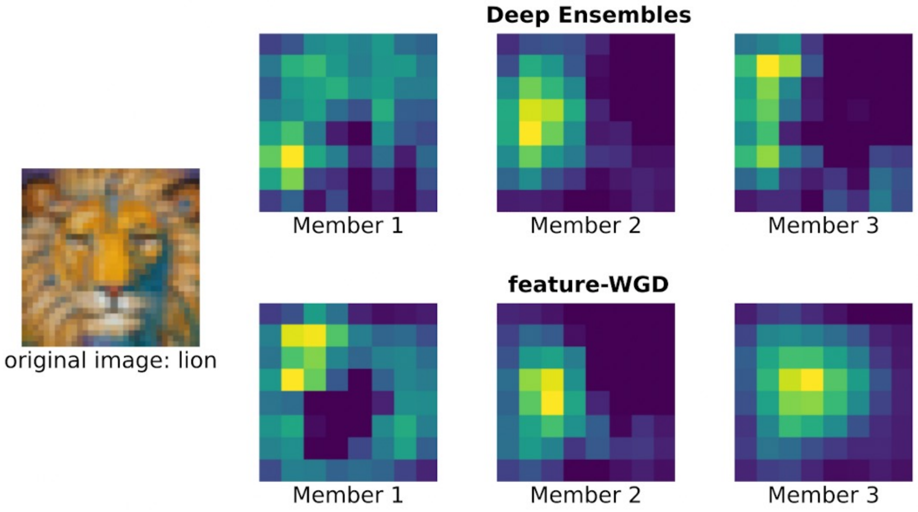



# Where to apply PVI - proposed

- feature-WGD : $h$
  - $\mathrm{p}(h|D) = p(h) \prod_{(x,y)\in D} p(y|x,h)$
  - $w_i^{t+1} \leftarrow w_i^t + \epsilon \cdot \left(\frac{dh_i^t}{dw_i^t}\right)^T v(h_i^t)$
  - $\theta_i^{t+1} \leftarrow \theta_i^t + \epsilon \cdot \frac{1}{n}\sum_{j=1}^n \log p(D|\theta_j^t)$
  - semantically shared feature space
  - multi-view structured data




** Curse of dimensionality of KDE **
Find subspace where likelihood change substantially
- (FxF) : $H^t = \frac{1}{n}\sum_{j=1}^n \nabla_h \log(D|h_j^t)\left(\nabla_h \log(D|h_j^t)\right)^T$
- (rxF) : $\Phi$ (r dominant eigenvectors)
- $q(h_i^t) \propto \sum_{j=1}^n k(h_i, h_j) \rightarrow \sum_{j=1}^n k(\Phi h_i, \Phi h_j)$



# Experiment

*Table 1.* Results for Wide ResNet-16-4 on CIFAR-10 with an ensemble size of 10, evaluated over 5 seeds.

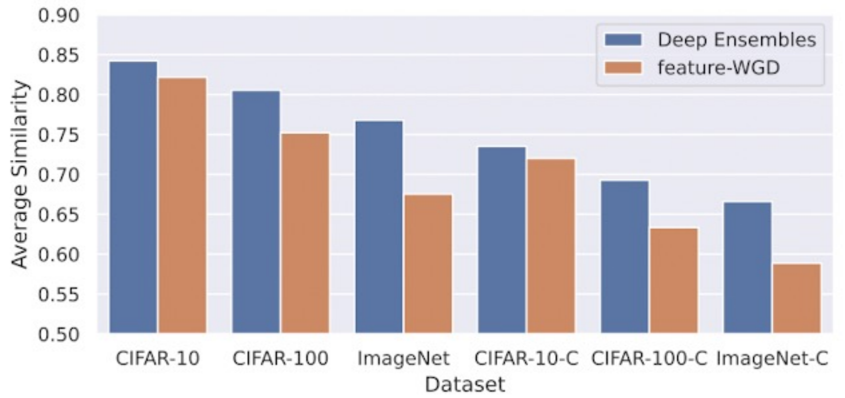
| METHOD | ACCURACY(↑) | NLL(↓) | BRIER(↓) | ECE(↓) | cA / cNLL / cBRIER / cECE |
|---|---|---|---|---|---|
| SINGLE | 95.4 ± 0.2 | 0.145 ± 0.006 | 0.069 ± 0.003 | 0.007 ± 0.000 | 73.7 / 0.796 / 0.349 / **0.020** |
| DEEP ENSEMBLES | 96.4 ± 0.1 | 0.110 ± 0.001 | 0.054 ± 0.001 | 0.007 ± 0.000 | 76.7 / 0.698 / 0.310 / 0.025 |
| WEIGHT-WGD | 96.4 ± 0.1 | 0.111 ± 0.002 | 0.054 ± 0.001 | 0.007 ± 0.001 | 76.7 / 0.702 / 0.312 / 0.026 |
| FUNCTION-WGD | 96.1 ± 0.1 | 0.124 ± 0.001 | 0.059 ± 0.001 | 0.007 ± 0.001 | 75.7 / 0.736 / 0.322 / 0.024 |
| FEATURE-WGD | **96.5 ± 0.1** | **0.107 ± 0.001** | **0.052 ± 0.001** | **0.006 ± 0.001** | **77.3 / 0.681 / 0.302 / 0.020** |

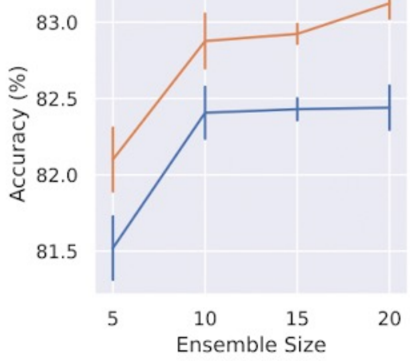
*Table 2.* Results for Wide ResNet-16-4 on CIFAR-100 with an ensemble size of 10, evaluated over 5 seeds.

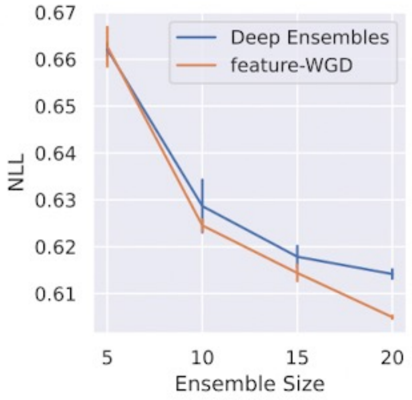
| METHOD | ACCURACY(↑) | NLL(↓) | BRIER(↓) | ECE(↓) | cA / cNLL / cBRIER / cECE |
|---|---|---|---|---|---|
| SINGLE | 77.4 ± 0.3 | 0.835 ± 0.007 | 0.316 ± 0.003 | 0.030 ± 0.003 | 46.7 / 2.279 / 0.658 / 0.035 |
| DEEP ENSEMBLES | 82.3 ± 0.2 | 0.632 ± 0.004 | 0.249 ± 0.001 | 0.020 ± 0.001 | 52.9 / 1.971 / 0.590 / 0.032 |
| WEIGHT-WGD | 82.3 ± 0.1 | 0.633 ± 0.002 | 0.250 ± 0.001 | 0.021 ± 0.001 | 52.8 / 1.967 / 0.589 / 0.031 |
| FUNCTION-WGD | 79.0 ± 0.1 | 0.715 ± 0.003 | 0.286 ± 0.001 | 0.018 ± 0.002 | 49.5 / 2.133 / 0.623 / 0.034 |
| FEATURE-WGD | **82.9 ± 0.2** | **0.624 ± 0.002** | **0.243 ± 0.001** | **0.017 ± 0.001** | **53.5 / 1.955 / 0.584 / 0.029** |

*Table 3.* Results for ResNet-50 on ImageNet with an ensemble size of 5. Note that we only evaluate 1 run due to the computational cost.

| METHOD | ACCURACY(↑) | NLL(↓) | BRIER(↓) | ECE(↓) | cA / cNLL / cBRIER / cECE |
|---|---|---|---|---|---|
| SINGLE | 75.7 | 0.954 | 0.338 | 0.018 | 37.7 / 3.235 / 0.738 / 0.021 |
| DEEP ENSEMBLES | **78.0** | **0.853** | **0.309** | 0.019 | 40.9 / 3.011 / 0.706 / **0.015** |
| FEATURE-WGD | **78.0** | 0.859 | **0.309** | **0.015** | **42.4 / 2.923 / 0.693** / 0.018 |



E.O.D