NEURAL INFORMATION PROCESSING SYSTEMS

NeurIPS 2025

Knowledge Distillation of Uncertainty using Deep Latent Factor Model

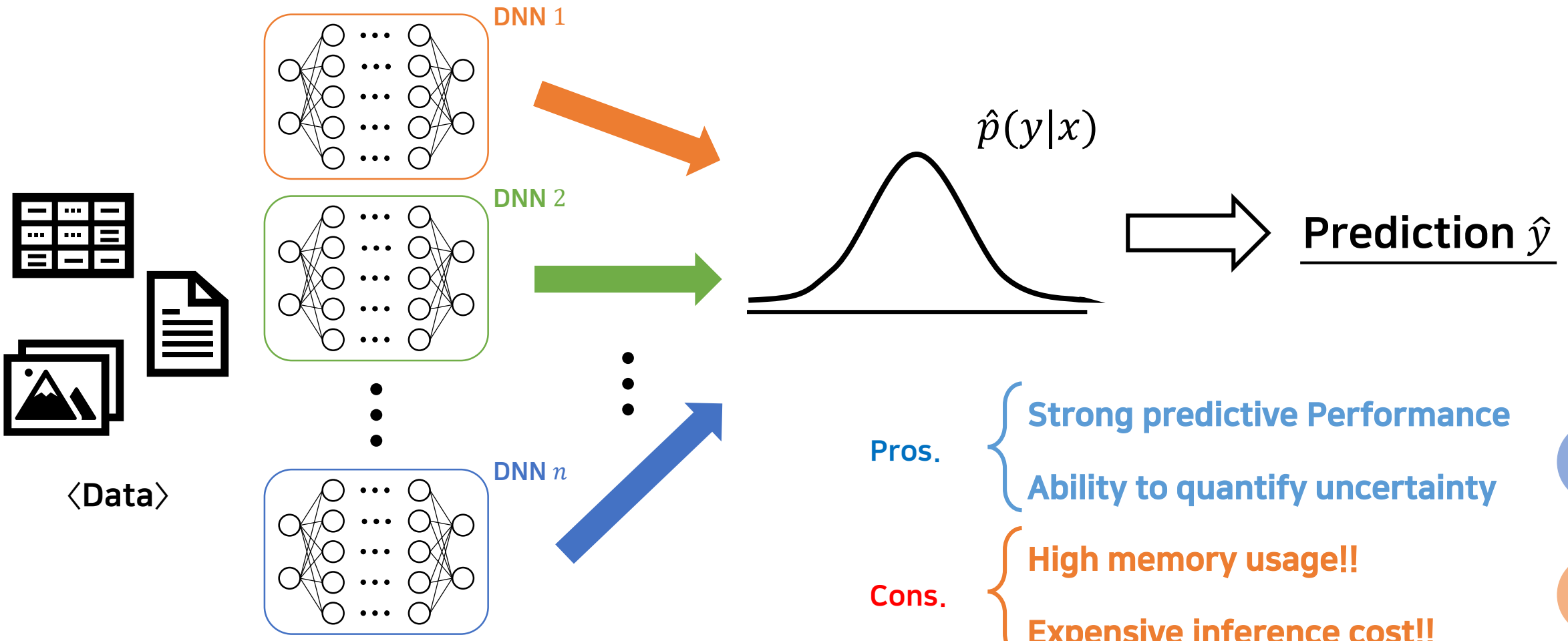
Sehyun Park[1], Jongjin Lee[2], Yunseop Shin[1], Ilsang Ohn[3] and Yongdai Kim[1]*

1 Department of Statistics, Seoul National University
2 Samsung Research
3 Department of Statistics, Inha University
* Corresponding author

SAMSUNG

Introduction

- **Deep Ensembles** combine multiple independently trained neural networks to improve prediction accuracy and estimate uncertainty.



Pros.
- Strong predictive Performance
- Ability to quantify uncertainty

Cons.
- High memory usage!!
- Expensive inference cost!!

Introduction

- **Knowledge Distillation (KD)** enables model compression while preserving the teacher's performance.

- We view each (teacher) ensemble member as a realization of a stochastic process, particularly a **Gaussian Process (GP)**.

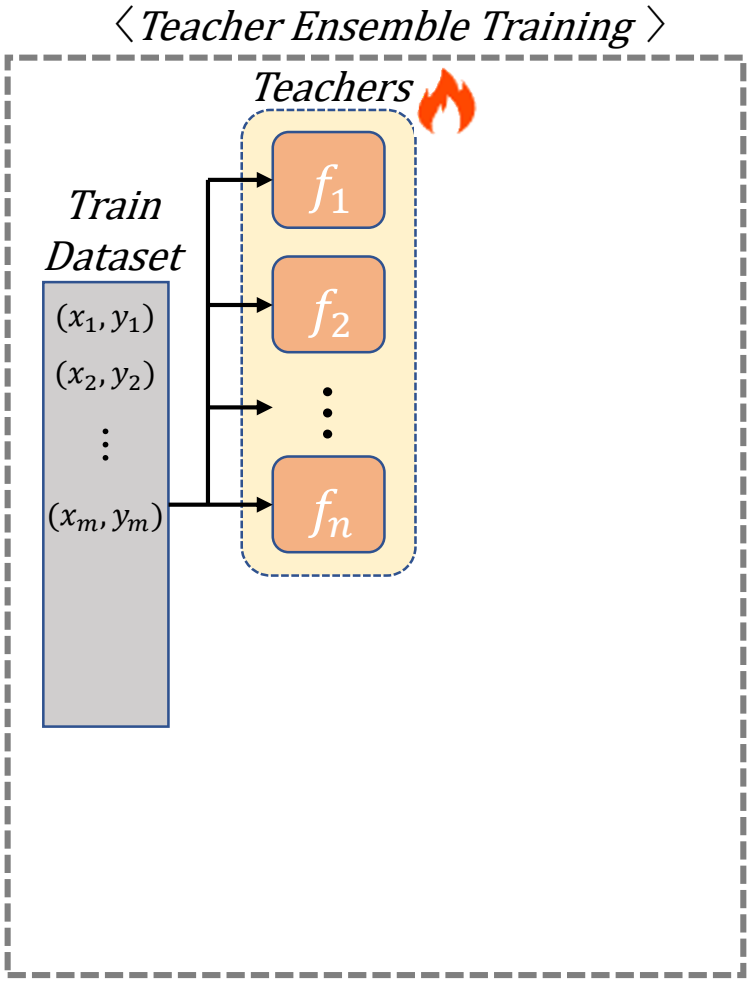
$$f(x) = \mu_\theta(x) + \Phi_\theta(x)^T Z, \qquad Z \sim \mathcal{N}_q\big(0, I_q\big).$$

$$f(\cdot) \sim \mathcal{GP}\big(\mu_\theta(\cdot), \Sigma_\theta(\cdot,\cdot)\big), \qquad \Sigma_\theta(x, x') = \Phi_\theta(x)^T \Phi_\theta(x')$$

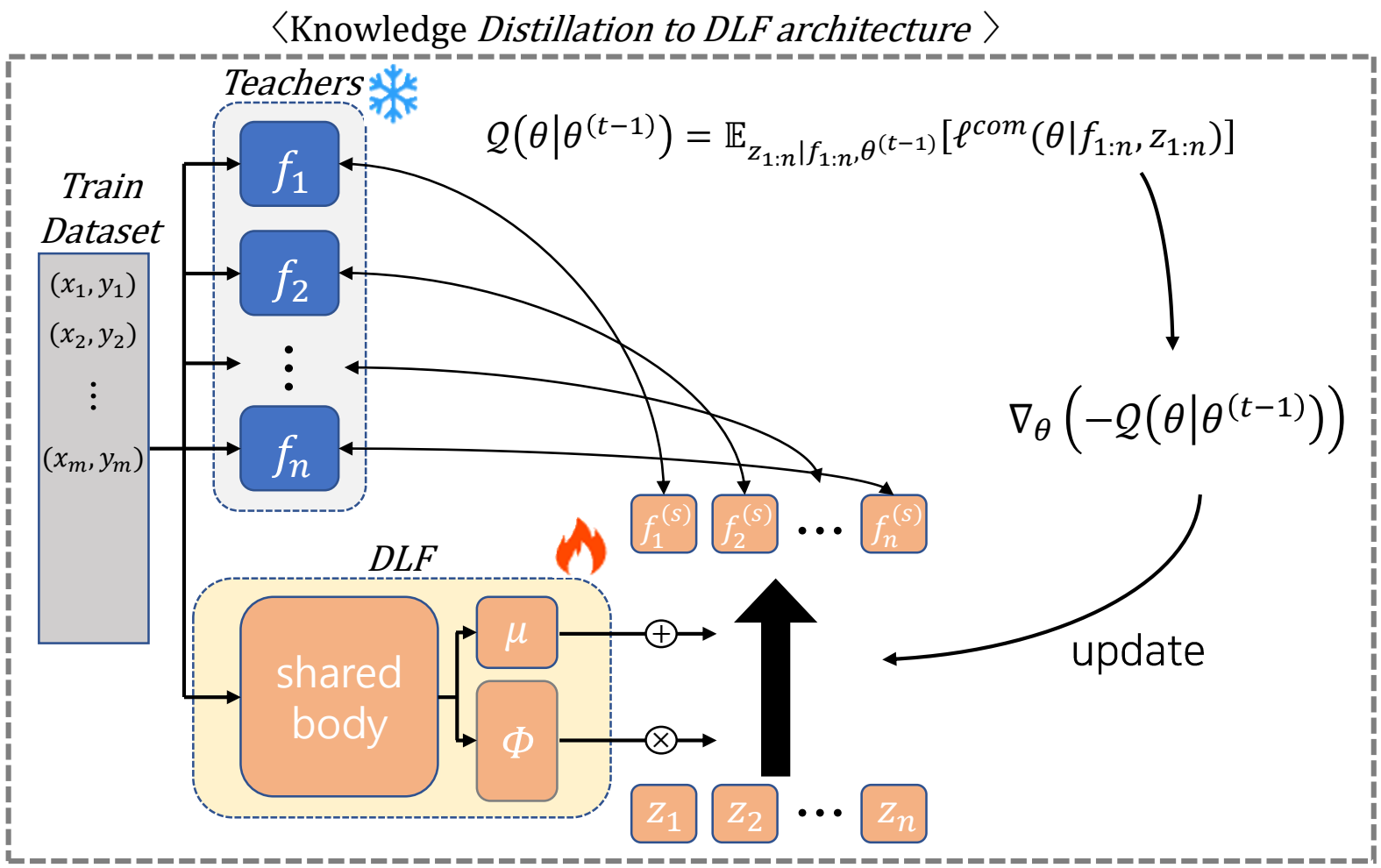
〈 **Deep Latent Factor Model** 〉

- By estimating the mean and covariance functions of this GP, we aim to distill the ensemble into a single, efficient student model.

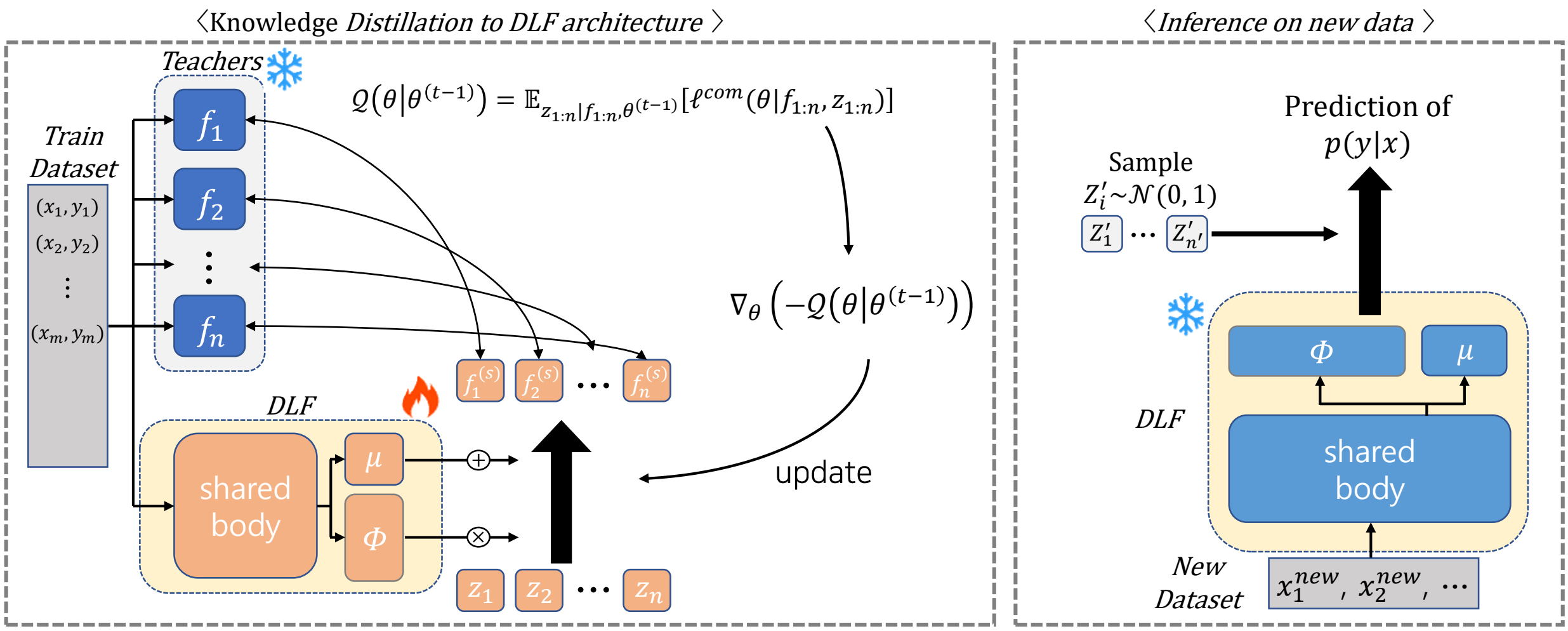
Overall Process



Overall Process



Overall Process



EM-algorithm

- **EM algorithm** :
  - ▶ E-step: Compute the expected complete log-likelihood

$$\mathcal{Q}\left(\theta\middle|\theta^{(t-1)}\right) = \mathbb{E}_{z_{1:n}|f_{1:n},\theta^{(t-1)}}[\ell^{com}(\theta|f_{1:n}, z_{1:n})]$$

    (It can be obtained in closed-form.)
  - ▶ M-step: Update $\theta^{(t)}$ by maximizing $\mathcal{Q}\left(\theta\middle|\theta^{(t-1)}\right)$, implemented via stochastic gradient descent.

Algorithm 1: EM algorithm for the univariate DLF model

Require : Teacher ensemble members $f_1(\cdot), \ldots, f_n(\cdot)$, learning rate $\eta > 0$.

Input : Design points $\mathcal{D}^{\text{design}} = \{\boldsymbol{x}_1^{(d)}, \ldots, \boldsymbol{x}_m^{(d)}\}$,

Initialize parameter: $\theta^{(0)}$

for $t = 1, 2 \ldots T$ **do**

  Shuffle the dataset $\mathcal{D}^{\text{design}}$ and divide into mini-batches $\{\mathcal{B}_1, \ldots, \mathcal{B}_M\}$

  for $l = 1, 2, \ldots, M$ **do**

    Calculate prediction values of teacher model $\boldsymbol{f}_{1:n,\mathcal{B}_l}$, where $\boldsymbol{f}_{1:n,\mathcal{B}_l} = (\boldsymbol{f}_{1:n}(\boldsymbol{x}_j^{\text{design}}), \boldsymbol{x}_j^{\text{design}} \in \mathcal{B}_l)$.

    Calculate $\mathbb{E}[\boldsymbol{z}|\theta^{(t-1)}]$ and $\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top|\theta^{(t-1)}]$ using $\boldsymbol{f}_{1:n,\mathcal{B}_l}$

    $Q(\theta|\theta^{(t-1)}) := \mathrm{E}_{\mathbf{z}_{1:n}|\boldsymbol{f}_{1:n,\mathcal{B}_l},\theta^{(t-1)}}[\ell^{com}(\theta; \boldsymbol{f}_{1:n,\mathcal{B}_l}, \mathbf{z}_{1:n})]$

    Calculate gradient $g^{(t)} := \nabla_\theta \left(-Q(\theta|\theta^{(t-1)})\right)$

    Update $\theta^{(t)} \leftarrow \theta^{(t-1)} - \eta \cdot g^{(t)}$

Output : Estimated parameter θ

Results

- Regression task is evaluated on six benchmark datasets from the UCI repository [1] including Boston housing, Concrete, Energy, Wine, Power and Kin8nm dataset.

Table 1: Results on UCI benchmark datasets. ($*$: closer to the coverage probability of a teacher ensemble is better)

Dataset	Method	Metric			
		RMSE ↓	NLL ↓	CRPS ↓	95% Coverage Probability $*$
Concrete	Teachers	5.6191	3.1134	2.9926	0.9515
	small-Ens	5.6952 (0.0494)	3.1672 (0.0167)	**2.9953 (0.0304)**	**0.9431 (0.0041)**
	Hydra	6.0558 (0.1366)	3.2586 (0.0293)	3.3084 (0.0639)	0.9097 (0.0115)
	BE	5.9777 (0.1475)	3.2241 (0.0332)	3.2320 (0.0894)	0.9282 (0.0099)
	DLF (Ours)	**5.6047 (0.1771)**	**3.1584 (0.0463)**	3.0622 (0.0954)	0.9291 (0.0125)
Wine	Teachers	0.5497	0.7980	0.2962	0.9750
	small-Ens	0.6002 (0.0083)	0.8861 (0.0160)	0.3307 (0.0067)	0.9569 (0.0042)
	Hydra	0.5689 (0.0114)	0.8322 (0.0166)	0.3075 (0.0040)	0.9594 (0.0053)
	BE	0.5560 (0.0187)	**0.8065 (0.0138)**	0.3020 (0.0059)	0.9612 (0.0092)
	DLF (Ours)	**0.5506 (0.0104)**	0.8230 (0.0199)	**0.2980 (0.0043)**	**0.9681 (0.0055)**

[1] Arthur Asuncion, David Newman, et al. Uci machine learning repository, 2007.
[2] TRAN, Linh, et al. Hydra: Preserving ensemble diversity for model distillation. arXiv preprint arXiv:2001.04694, 2020.
[3] NAM, Giung, et al. Diversity matters when learning from ensembles. Advances in neural information processing systems, 2021.

Results

Table 2: Results on CIFAR-10 and CIFAR-100

dataset	method	Acc(%) ↑	NLL ↓	ECE(%) ↓
CIFAR-10	Teachers	94.24	0.1539	0.9
	small-Ens	92.87 (0.35)	0.2377 (0.0019)	3.93 (0.14)
	Hydra	93.16 (0.06)	0.2660 (0.0063)	4.15 (0.01)
	LBE	93.25 (0.26)	0.2480 (0.0053)	4.11 (0.10)
	Proxy-EnD^2	90.92 (0.28)	0.2861 (0.0027)	2.08 (0.19)
	EDFM	90.62 (0.23)	0.2858 (0.0025)	2.78 (0.17)
	DLF (Ours)	**93.40 (0.14)**	**0.2246 (0.0023)**	**2.79 (0.20)**
CIFAR-100	Teachers	81.36	0.7167	1.41
	small-Ens	79.29 (0.26)	1.0413 (0.0145)	12.90 (0.24)
	Hydra	77.42 (0.15)	1.2912 (0.0272)	12.70 (0.37)
	LBE	79.58 (0.40)	1.0110 (0.0087)	13.42 (0.42)
	Proxy-EnD^2	67.62 (0.22)	1.2355 (0.0151)	**7.35 (0.25)**
	EDFM	64.17 (0.32)	1.6741 (0.0242)	11.35 0.47)
	DLF (Ours)	**79.68 (0.23)**	**0.8974 (0.0042)**	9.45 (0.31)

Table 3: Results on SuperGLUE benchmark datasets

dataset	method	Acc (%) ↑	NLL ↓	ECE (%) ↓
RTE	Teachers	75.09	0.8401	17.70
	small-Ens	67.15 (0.0057)	0.6739 (0.0124)	13.09 (0.0148)
	Hydra	62.82 (0.1650)	0.9034 (0.1733)	23.31 (0.4907)
	LBE	65.97 (0.1406)	0.9235 (0.0664)	25.63 (0.1909)
	DLF (Ours)	**67.06 (0.1040)**	**0.6658 (0.0762)**	**9.74 (0.5050)**
WiC	Teachers	68.03	0.6395	9.14
	small-Ens	65.02 (0.0062)	**0.7674 (0.0206)**	16.69 (0.0116)
	Hydra	65.05 (0.0210)	1.1628 (0.0355)	26.26 (0.0191)
	LBE	65.36 (0.1209)	0.8809 (0.0950)	20.28 (0.0300)
	DLF (Ours)	**66.18 (0.0146)**	0.7706 (0.0543)	**15.99 (0.0198)**

- We also evaluate classification task on CIFAR-10 and CIFAR-100 [4], and analyze two SuperGLUE [5] sub-tasks: RTE, and WiC.

[4] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.2009.
[5] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, FelixHill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purposelanguage understanding systems. Advances in neural information processing systems, 32, 2019.
[6] RYABININ, Max; MALININ, Andrey; GALES, Mark. Scaling ensemble distribution distillation to many classes with proxy targets. Advances in Neural Information Processing Systems, 2021.
[7] PARK, Jonggeon, et al. Ensemble Distribution Distillation via Flow Matching. In: Forty-second International Conference on Machine Learning, 2025.

Thank you