ICLR

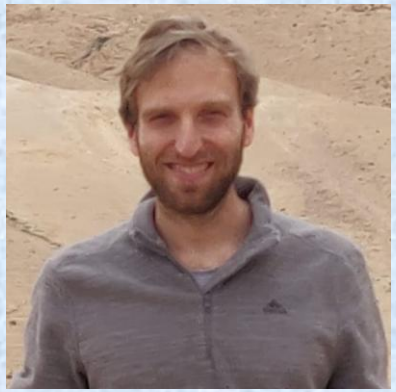
SGD-Based Knowledge Distillation with Bayesian Teachers: Theory and Guidelines

Itai Morad

Ph.D. student
Ben-Gurion University
of the Negev, Israel

Nir Shlezinger

Professor
Ben-Gurion University
of the Negev, Israel



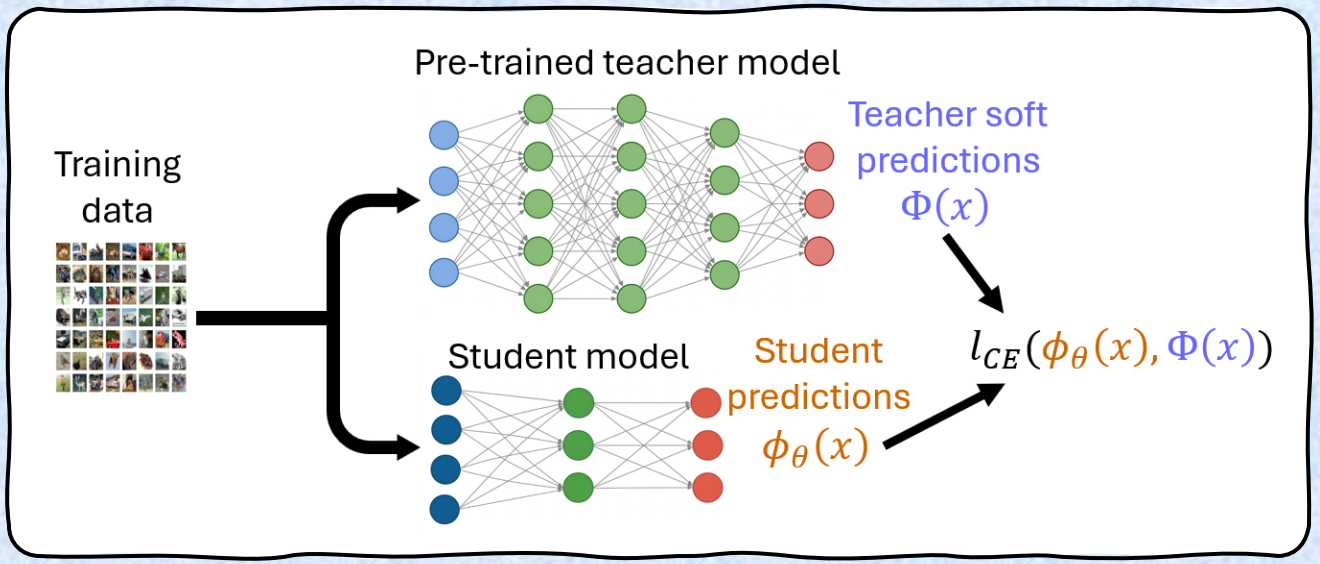
Yonina C. Eldar

Professor
Weizmann Institute
of Science, Israel

Knowledge Distillation

The process of transferring knowledge from a large model to a smaller one

- Model compression (memory)
- Efficiency (speed)
- Knowledge transfer
- Performance maintenance



Teacher soft predictions are viewed as approximations of the Bayes conditional probabilities (BCPs)

Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," International Journal of Computer Vision, vol. 129, no. 6, pp. 1789–1819, 2021.

How do BCPs affect the learning dynamics and performance of students trained with SGD?

BCP-Based Risk

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} f_{\mathcal{P}}(\boldsymbol{\theta}) \coloneqq \mathbb{E}_{\boldsymbol{x}\sim\mathcal{P}}[l(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\boldsymbol{x}), \mathcal{P}(\boldsymbol{y}|\boldsymbol{x}))]$$

BCP instead of One-hot

- ✓ Same f^* and $\boldsymbol{\theta}^*$ as standard (one-hot) risk.
- ✓ Satisfies the **interpolation** property.
- ✓ $\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\phi}_{\boldsymbol{\theta}^*}(\boldsymbol{x}), \mathcal{P}(\boldsymbol{y}|\boldsymbol{x})) = 0 \ \forall \boldsymbol{x}$.
- ✓ $\boldsymbol{\phi}_{\boldsymbol{\theta}^*}$ also minimizes the **empirical risk** for $\mathcal{D} \subset \mathrm{supp}(\mathcal{P})$.

$$\boldsymbol{\phi}_{\theta^*}(\boldsymbol{x}) = [p(y_1|\boldsymbol{x}), \ldots, p(y_K|\boldsymbol{x})]$$

$$f_{\mathcal{P}}{}^* = H(\boldsymbol{y}|\boldsymbol{x})$$

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^d} f_{\mathcal{D}}(\boldsymbol{\theta}) \coloneqq \frac{1}{N}\sum_{n=1}^{N} l(\boldsymbol{\phi}_{\boldsymbol{\theta}}(\boldsymbol{x}), \mathcal{P}(\boldsymbol{y}|\boldsymbol{x}))$$

Menon, Aditya K., et al. "A statistical perspective on distillation." International Conference on Machine Learning. PMLR, 2021.
Garrigos, Guillaume, and Robert M. Gower. "Handbook of convergence theorems for (stochastic) gradient methods." arXiv preprint arXiv:2301.11235 (2023).

SGD Convergence Analysis

SGD iterates: $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \alpha \nabla f_{\zeta_t}(\boldsymbol{\theta}^t)$

Theorem. Under μ-strong convexity and $\mathcal{L}$-smoothness, for any $\alpha \leq \frac{1}{\mathcal{L}}$, the model parameters $\boldsymbol{\theta}$ converge as

$$\mathbb{E}[\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^*\|^2] \leq \underbrace{(1-\alpha\mu)^t \|\boldsymbol{\theta}^0 - \boldsymbol{\theta}^*\|^2}_{\text{Convergence speed term}} + \underbrace{\frac{2\alpha}{\mu}\sigma_f^*}_{\text{Neighborhood term}}$$

Theorem. Under μ-PL condition and $\mathcal{L}$-smoothness, for any $\alpha \leq \frac{\mu}{2L\mathcal{L}}$, the **risk** f converges as

$$\mathbb{E}[f_{\mathcal{P}}(\boldsymbol{\theta}^t) - f_{\mathcal{P}}^*] \leq \underbrace{(1-\alpha\mu)^t (f_{\mathcal{P}}^0 - f_{\mathcal{P}}^*)}_{\text{Convergence speed term}} + \underbrace{\frac{L\alpha}{\mu}\sigma_f^*}_{\text{Neighborhood term}}$$

Garrigos, Guillaume, and Robert M. Gower. "Handbook of convergence theorems for (stochastic) gradient methods." arXiv preprint arXiv:2301.11235 (2023).

Perfect BCP SGD Analysis

SGD iterates: $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \alpha \nabla f_{\zeta_t}(\boldsymbol{\theta}^t)$

Theorem. Under μ-strong convexity and $\mathcal{L}$-smoothness, for any $\alpha \leq \underbrace{\frac{1}{\mathcal{L}}}_{\text{2x larger}}$, the model parameters $\boldsymbol{\theta}$ converge as

$$\mathbb{E}[\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^*\|^2] \leq (1 - \alpha\mu)^t \|\boldsymbol{\theta}^0 - \boldsymbol{\theta}^*\|^2 + \underbrace{\frac{2\alpha}{\mu}\sigma_f^*}_{=0}$$

Theorem. Under μ-PL condition and $\mathcal{L}$-smoothness, for any $\alpha \leq \underbrace{\frac{\mu}{2L\mathcal{L}}}_{\text{2x larger}}$, the **risk** f converges as

$$\mathbb{E}[f_{\mathcal{P}}(\boldsymbol{\theta}^t) - f_{\mathcal{P}}^*] \leq (1 - \alpha\mu)^t (f_{\mathcal{P}}^0 - f_{\mathcal{P}}^*) + \underbrace{\frac{L\alpha}{\mu}\sigma_f^*}_{=0}$$

Noisy BCP SGD Analysis

Smaller gradient noise

One-hot labels:

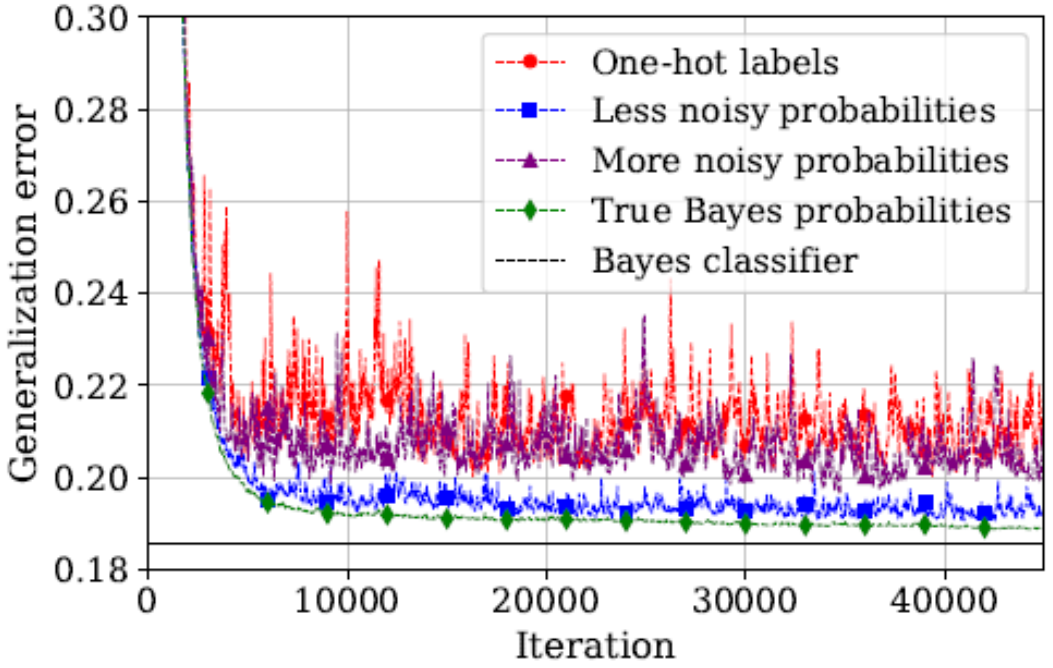
$$\sigma_f^* = \mathbb{E}[\sum_{k=1}^{K} \frac{\left\| J_{\boldsymbol{\theta},k}[\boldsymbol{\phi}_{\boldsymbol{\theta}^*}(\boldsymbol{x})] \right\|^2}{\mathcal{P}(y_k|\boldsymbol{x})}]$$

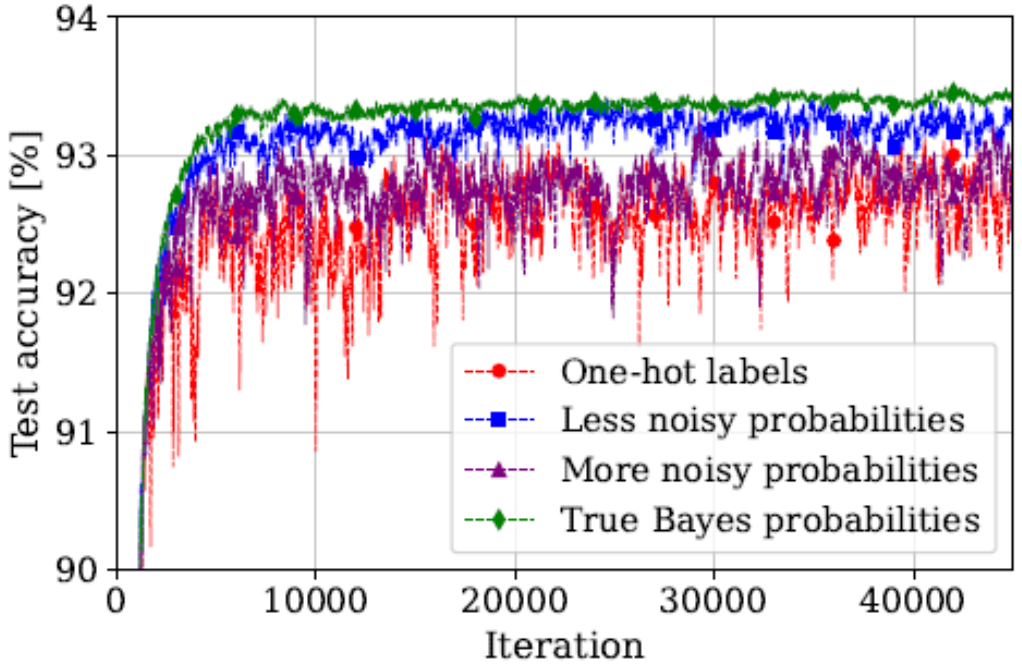
Noisy BCPs:

$$\sigma_f^* = \nu \cdot \mathbb{E}[\sum_{k=1}^{K} \frac{\left\| J_{\boldsymbol{\theta},k}[\boldsymbol{\phi}_{\boldsymbol{\theta}^*}(\boldsymbol{x})] \right\|^2}{[\mathcal{P}(y_k|\boldsymbol{x})]^2}]$$

Learning from BCPs leads to more **stable convergence** and guarantees convergence with larger learning rates.

Synthetic Study





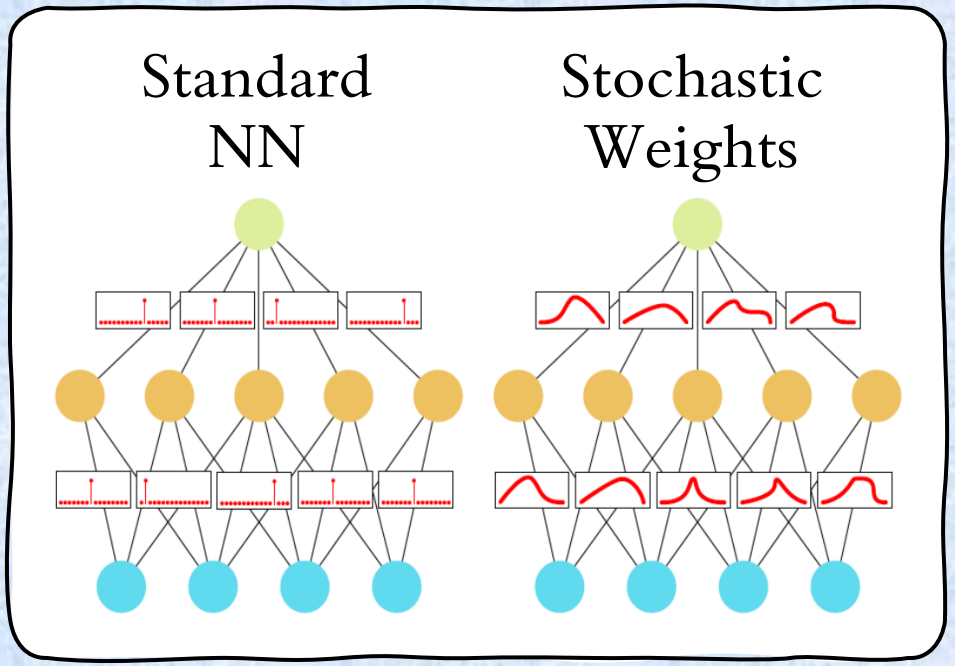
Teacher quality is directly proportional to error floor and learning curve fluctuations (learning stability).

Bayesian NN Teachers

How do we translate the observation that good teachers are ones that are best calibrated to the true BCPs?

We utilize **BNNs** as **teachers** in KD

- ✓ BNNs trained with variational inference.
- ✓ Bayesianize pretrained NNs via Laplace approximation.



Jospin, Laurent Valentin, et al. "Hands-on Bayesian neural networks—A tutorial for deep learning users." IEEE Computational Intelligence Magazine 17.2 (2022): 29-48.

Results

Higher student **accuracy** and improved **learning stability**

Highlight example: ResNet50→WRN-16-2

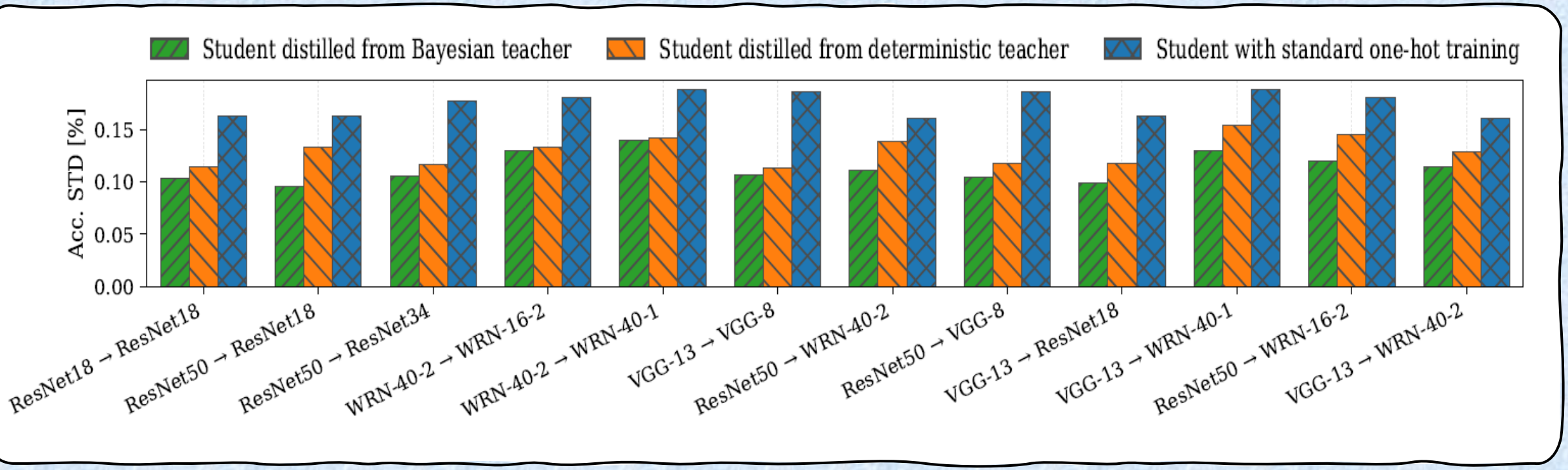
Teacher Kind	Teacher Accuracy [%]	Student Accuracy [%]
One-hot labels (none)	-	67.70
Deterministic	75.41	69.36
BNN Laplace Approximation	74.83	72.52
BNN Variational Inference	75.67	73.63

+0.26% VI BNN teacher → **+4.27%** student

-0.58% LA BNN teacher → **+3.16%** student

Results

Higher student **accuracy** and improved **learning stability**



$$\sigma_{BNN} < \sigma_{DNN} < \sigma_{one-hot}$$

Results

Higher student **accuracy** and improved **learning stability**

- Additional datasets
- Additional response-based distillation methods (DKD, DIST, WTTM)
- Few-shot classification
- Temperature scaling effects
- Ablation study (BNN parameters)