Knowledge Distillation Performs Partial Variance Reduction

Mher Safaryan

Marie Skłodowska-Curie Fellow

Institute of Science and Technology Austria

NEURAL INFORMATION PROCESSING SYSTEMS

Alexandra Peste

PhD student / Postdoctoral Researcher

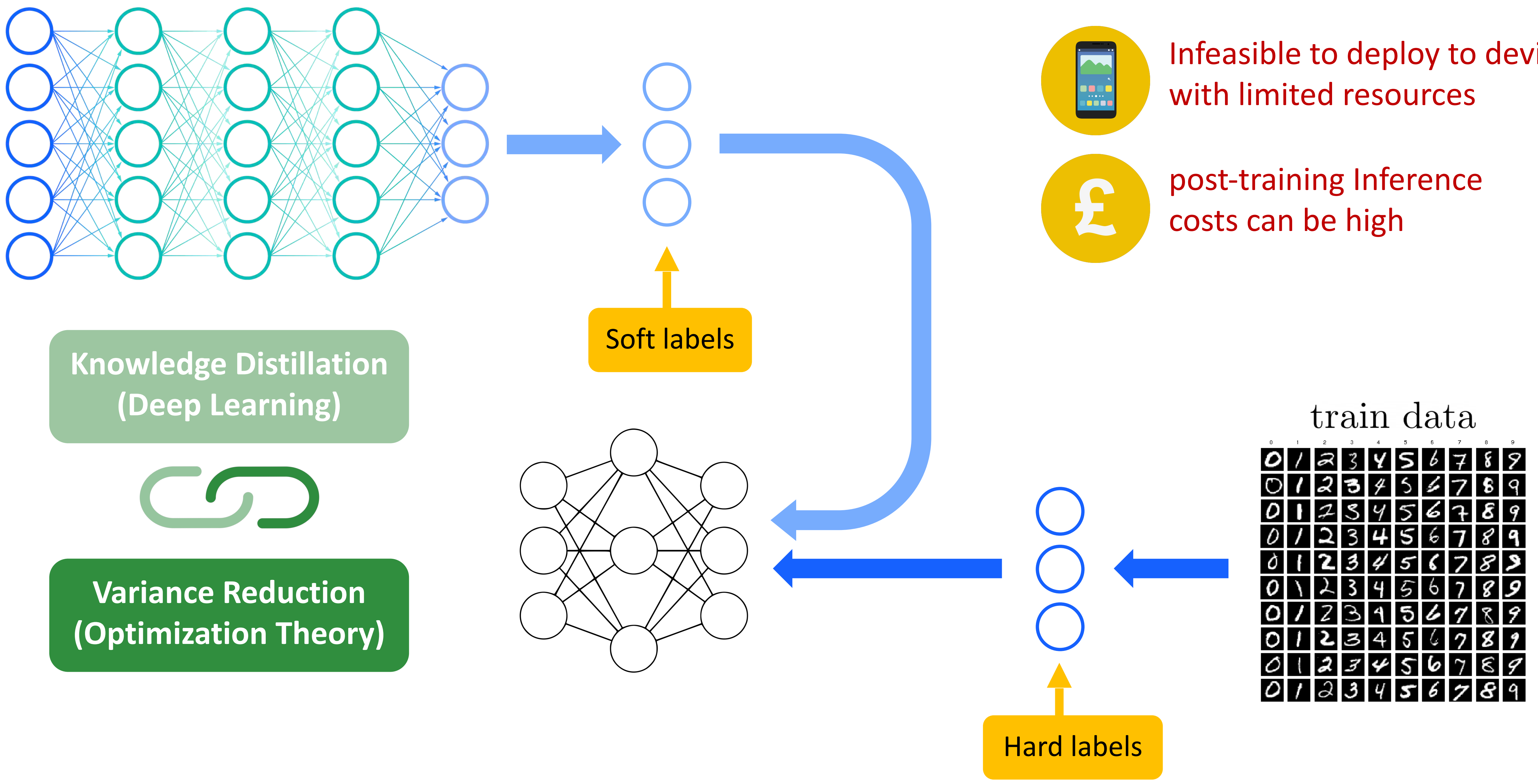
Institute of Science and Technology Austria

Dan Alistarh

Professor

Institute of Science and Technology Austria

Overview



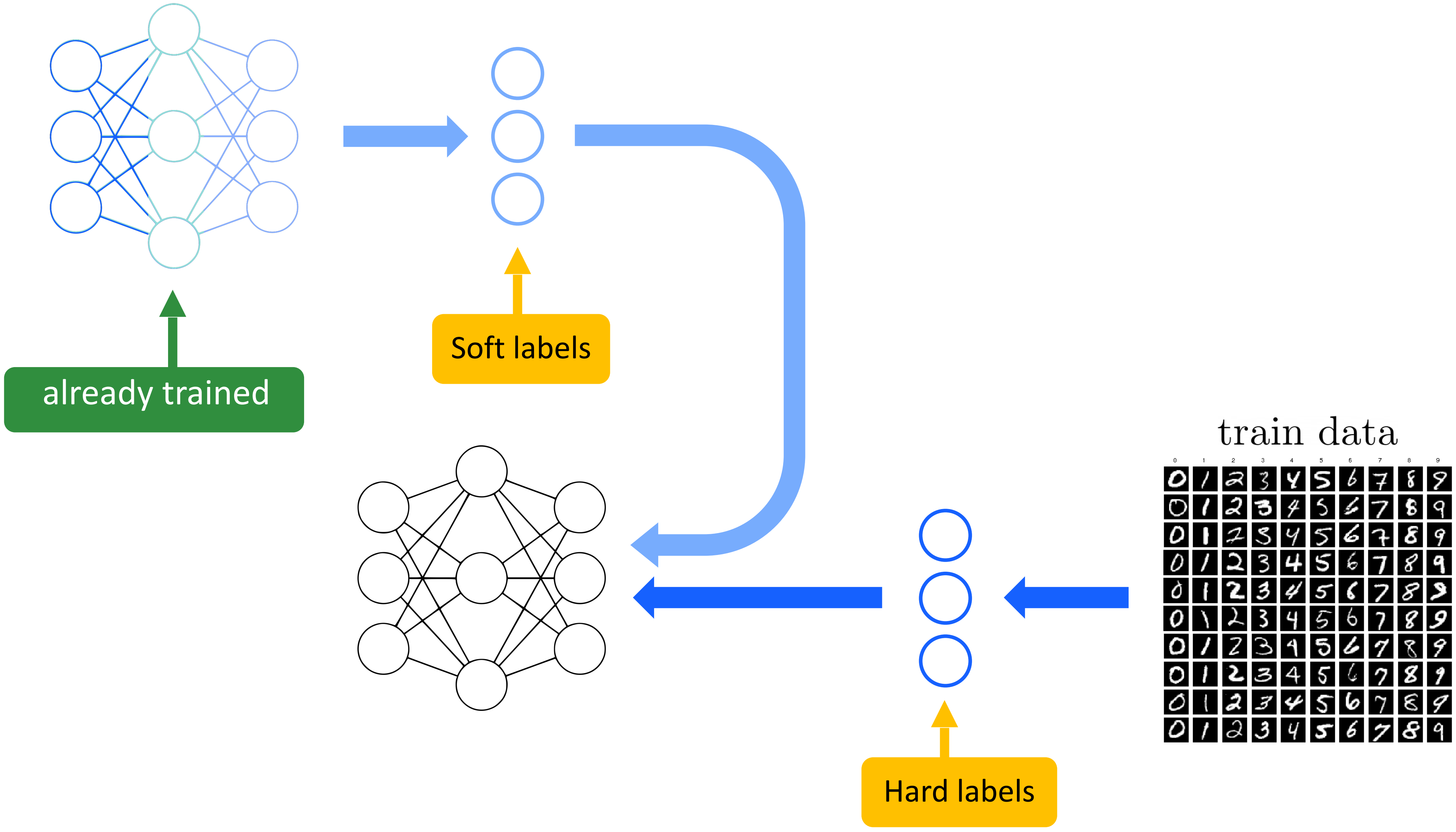
Infeasible to deploy to devices with limited resources

post-training Inference costs can be high

Knowledge Distillation (Deep Learning)

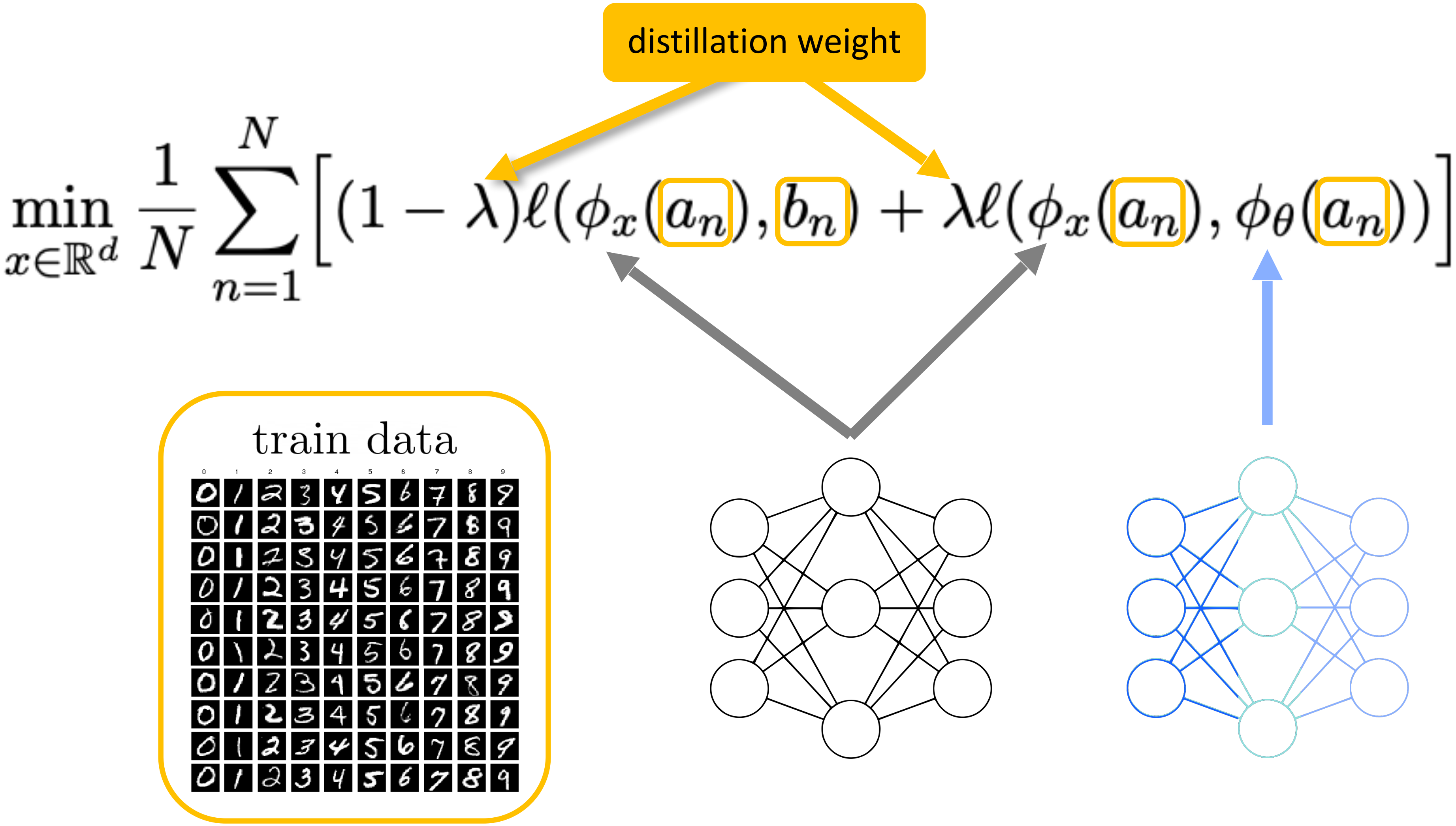
Variance Reduction (Optimization Theory)

Self-distillation



Self-distillation

$$\min_{x \in \mathbb{R}^d} \frac{1}{N} \sum_{n=1}^{N} \left[(1-\lambda)\ell(\phi_x(a_n), b_n) + \lambda \ell(\phi_x(a_n), \phi_\theta(a_n)) \right]$$



Self-distillation

standard loss with respect to train data

distillation loss with respect to teacher model

$$\min_{x \in \mathbb{R}^d} \frac{1}{N} \sum_{n=1}^{N} \Big[(1-\lambda) \ell(\phi_x(a_n), b_n) + \lambda \ell(\phi_x(a_n), \phi_\theta(a_n)) \Big]$$

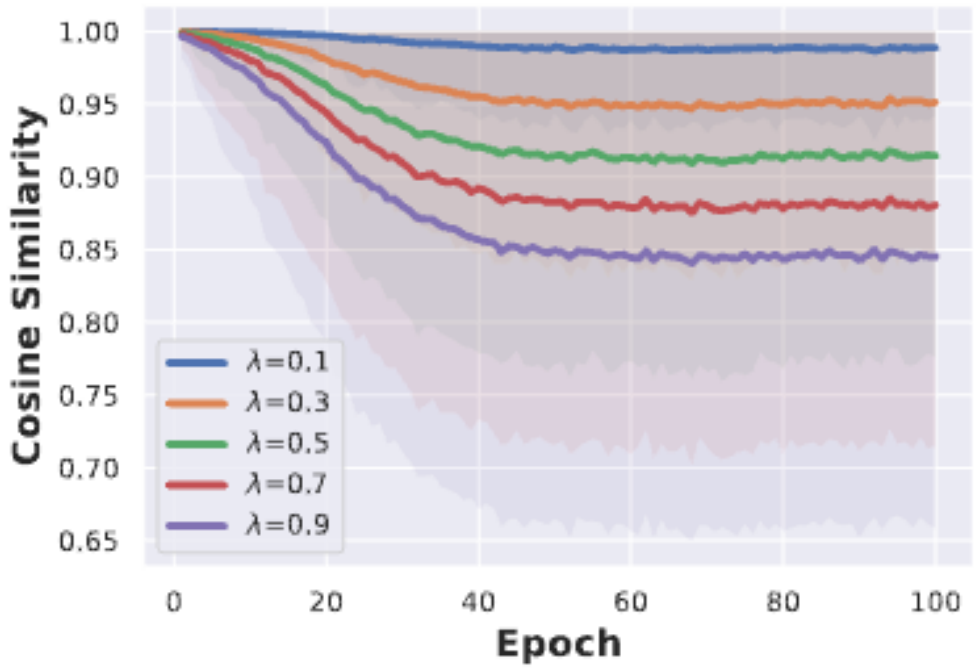
$f_n(x)$

$f_n(x \mid \theta, \lambda)$

Proposition (Distillation Gradient)

$$\nabla f_n(x \mid \theta, \lambda) \simeq \nabla f_n(x) - \lambda \nabla f_n(\theta)$$

- linear regression
- (deep) linear network
- non-linear network



Self-distillation

strong convexity

smoothness

Theorem 1 (See Appendix C.2). *Let Assumptions 1 and 3 hold. For any $\gamma \leq \frac{1}{8\mathcal{L}}$ and properly chosen distillation weight λ, the iterates (7) of SGD with self-distillation using teacher's parameters θ converge as*

$$\mathbb{E}\left[\|x^t - x^*\|^2\right] \leq (1-\gamma\mu)^t \|x^0 - x^*\|^2 + \frac{2\sigma_*^2}{\mu} \min(\gamma, \mathcal{O}(f(\theta) - f^*)),$$

where $\sigma_*^2 := \mathbb{E}[\|\nabla f_\xi(x^*)\|^2]$ is the stochastic noise at the optimum.

PL condition

smoothness

Theorem 2 (See Appendix C.3). *Let Assumptions 2 and 3 hold. For any $\gamma \leq \frac{1}{4\mathcal{L}} \frac{\mu}{L}$ and properly chosen distillation weight λ, the iterates (7) of SGD with self-distillation using teacher's parameters θ converge as*

$$\mathbb{E}\left[f(x^t) - f^*\right] \leq (1-\gamma\mu)^t \left(f(x^0) - f^*\right) + \frac{L\sigma_*^2}{\mu} \min(\gamma, \mathcal{O}(f(\theta) - f^*)),$$

- **Optimal distillation weight**
- **Unbiased knowledge distillation**
- **Distillation of compressed model**

Thank you