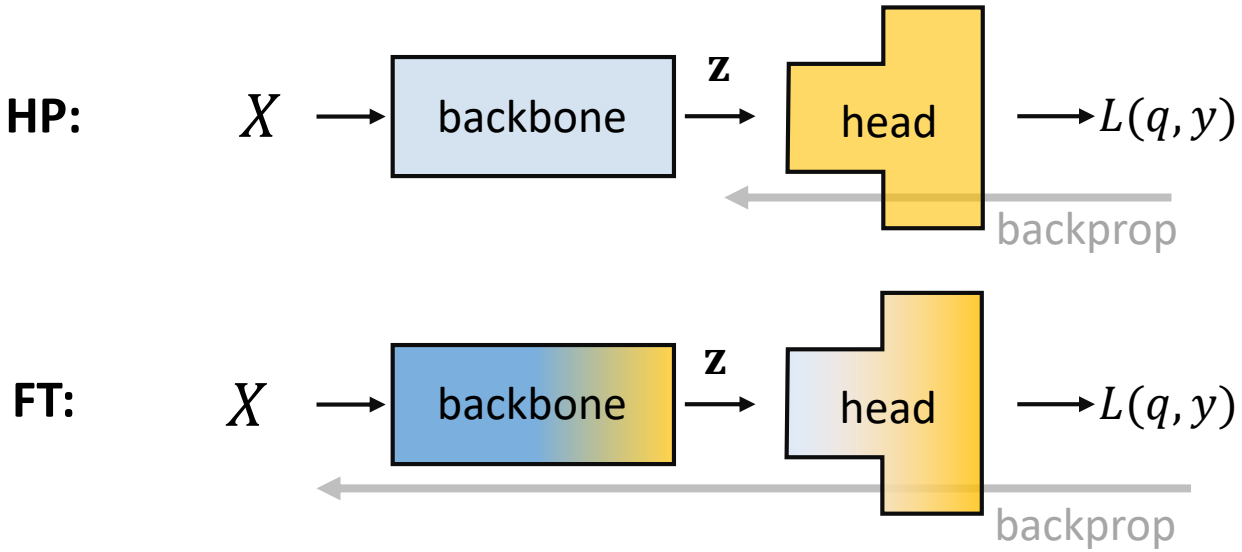
How to Prepare your Task-head for Finetuning

$$z_0 \rightarrow z_T$$

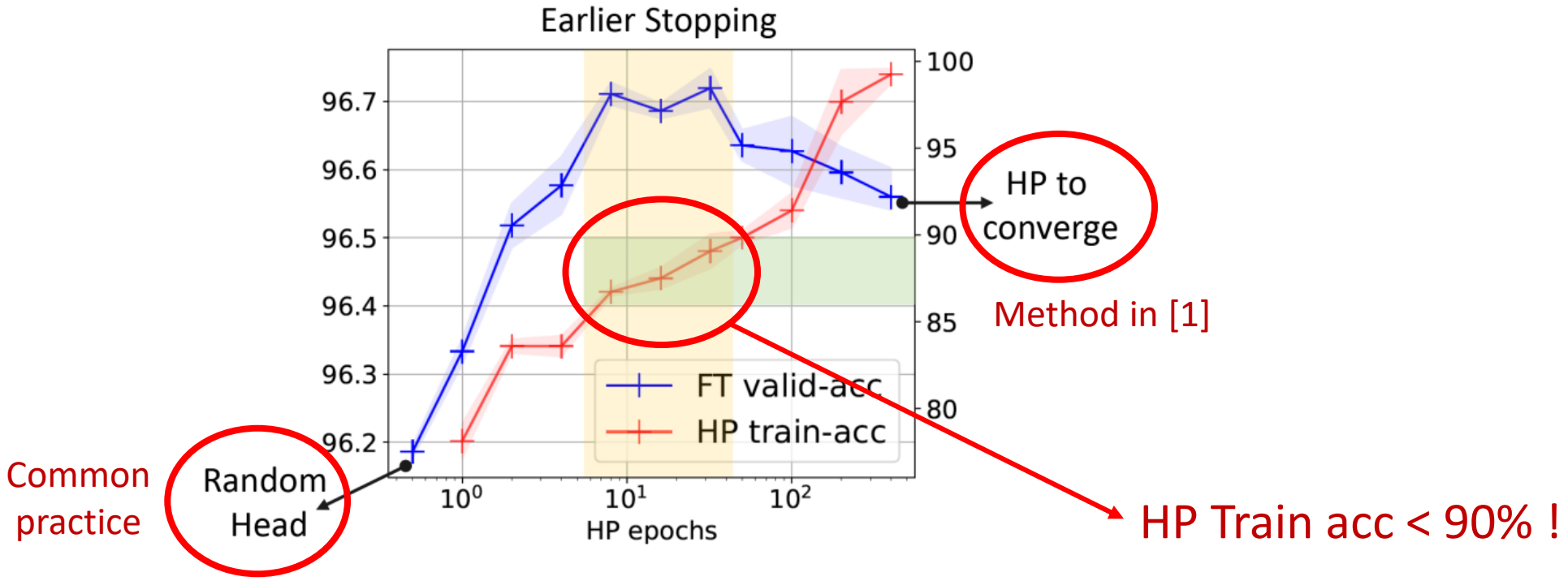
Yi Ren, Shangmin Guo, Wonho Bae, Danica J. Sutherland

- Popular flow: pretrain, re-initialize a random head, then finetune together
- But Kummar et.al claim hasty finetune will **distort** learned features.
- They propose a two-phase finetuning:
  - **Head-probing**: only update task-head $g(\mathbf{z})$, to **converge**
  - **Fine-tuning**: update head $g(\mathbf{z})$ and backbone $\mathbf{z} = f(\mathbf{x})$ together



[1] Kumar, Ananya, et al. "Fine-tuning can distort pretrained features and underperform out-of-distribution." ICLR 2022

- Following Kummar et. al, the gradient on $f(\mathbf{x})$ is **zero**, hence $\mathbf{z}$ is **unchanged**
- But we usually **require** $\mathbf{z}$ adapt to new downstream tasks
- Hence neither common practice nor solution in [1] is optimal

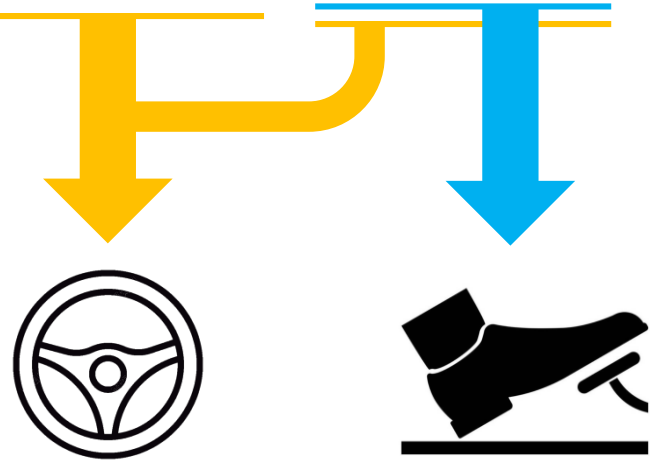


- We need to analyze the learning dynamics of $\mathbf{z}$:

$$z_{t+1}(\tilde{x}) - z_t(\tilde{x}) = \frac{\eta}{N} \sum_{i=1}^{N} \underbrace{\text{eNTK}^{f}_{\mathbf{w}_t}(\tilde{x}, x_i)}_{①} \underbrace{(\nabla_z q_t(x_i))^{\top}}_{②} \underbrace{(e_{y_i} - q_t(x_i))}_{③} + \mathcal{O}(\eta^2)$$

① eNTK of the backbone, **change slow** during finetune.
[For real model, refer to Figure 6 in Appendix A]
[For toy model, this is invariant]

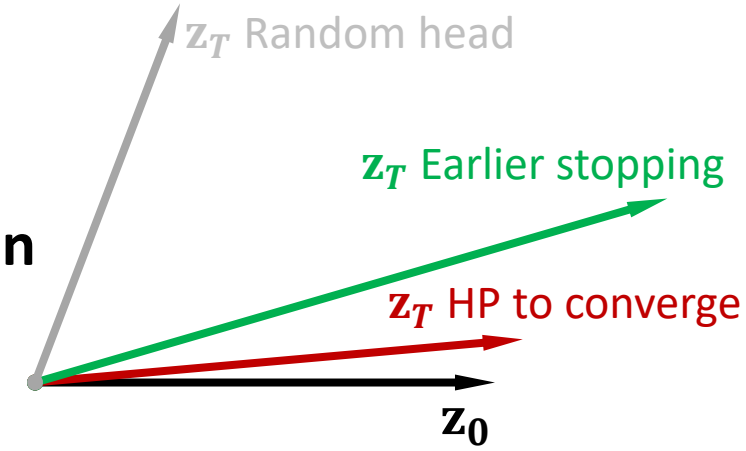
② **Direction** is determined by task-head at time t
[For real model, NTK approximation is usually good]
[For toy model, it should be just the head vector $\mathbf{v}$]

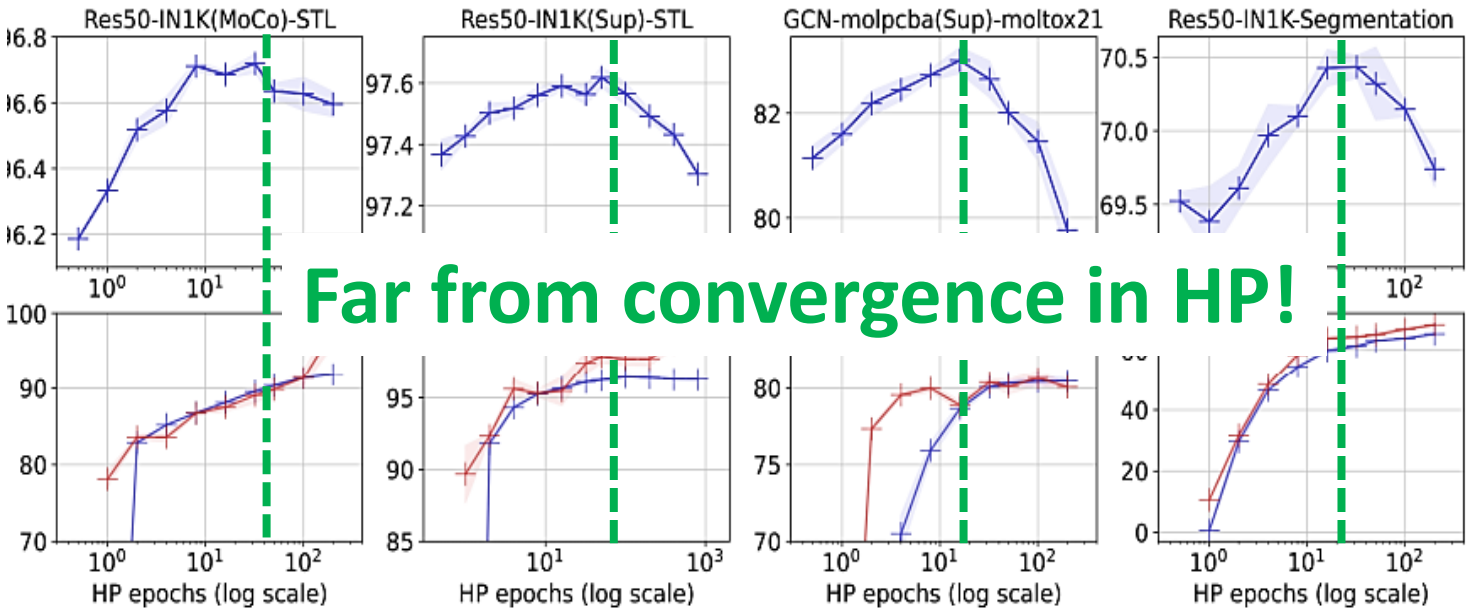


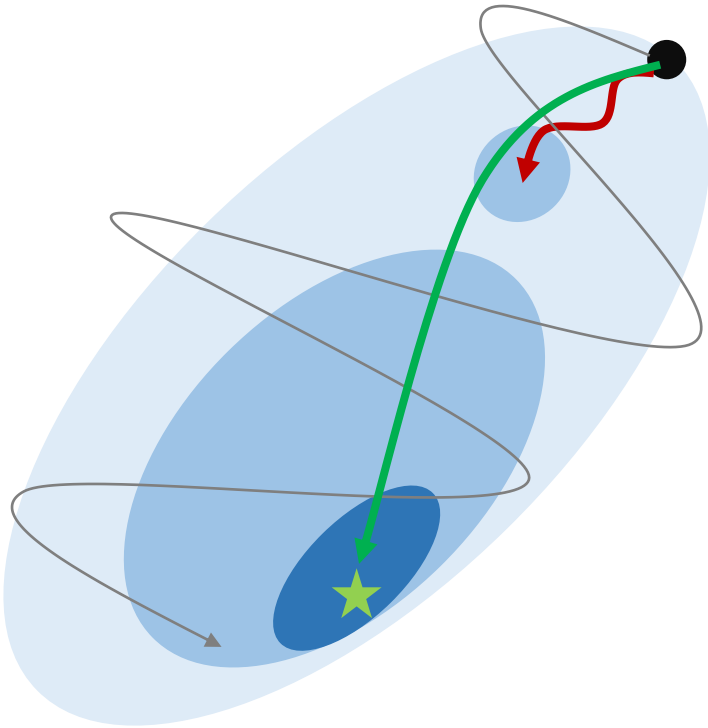
③ **Energy-vector** is determined by model's prediction at the beginning of finetune
Upper bound of adaptation: $\mathbb{E}\|\mathbf{z}_T - \mathbf{z}_0\|_2 \leq c \cdot \mathbb{E}\|\mathbf{e}_\mathbf{y} - \mathbf{q}_0\|_2$
[For toy model, we can analytically analyze $\mathbf{z}_T$]

- How to ensure $\mathbf{z}$ **sufficiently** adapt to **appropriate directions** for new tasks

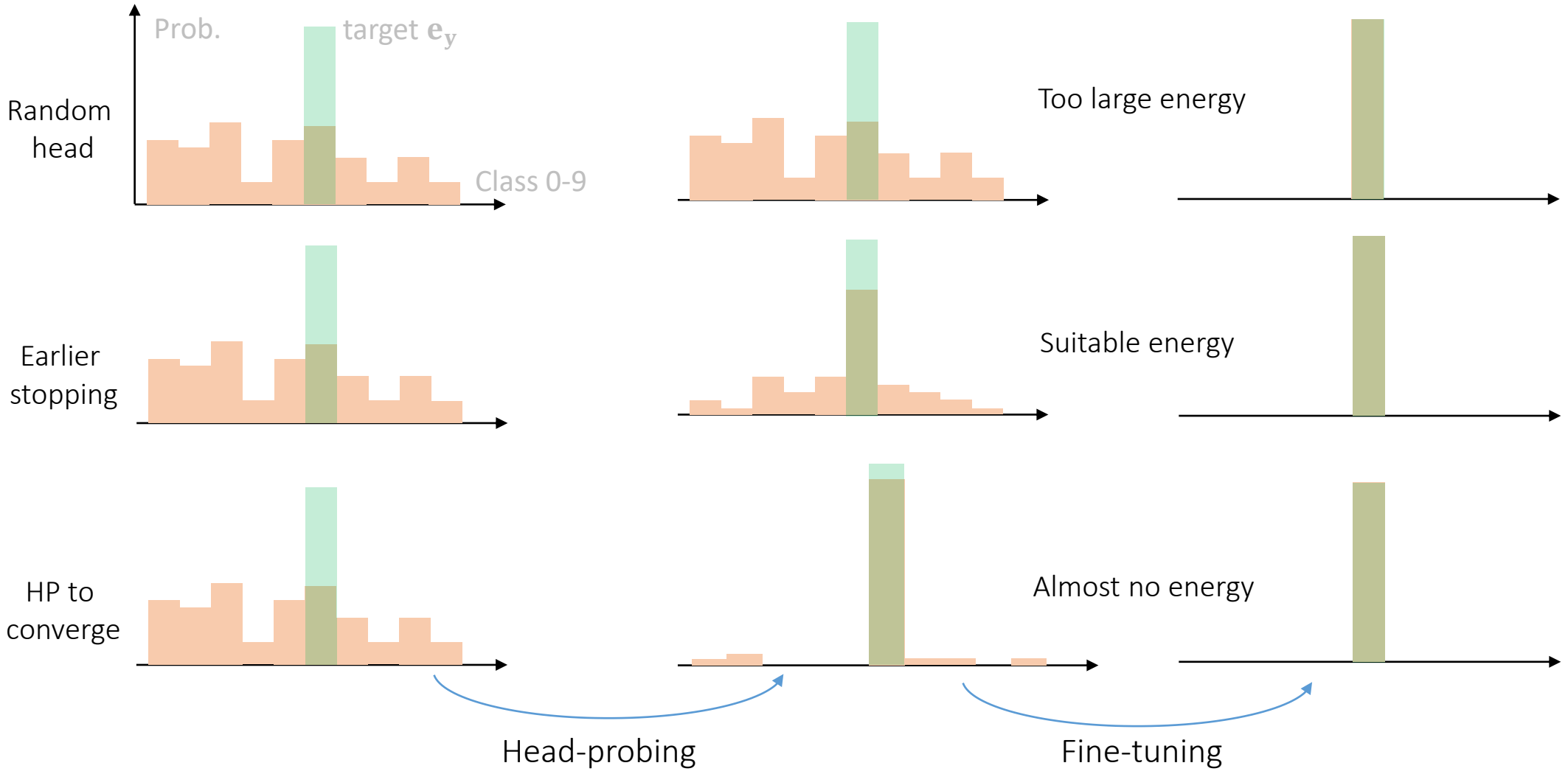
- Using the dynamic to describe how $\mathbf{z}$ adapts
- Random head: too big energy and **inconsistant direction**
- HP to converge: too small energy, almost no adaptation
- Earlier stopping: enough energy and stable direction





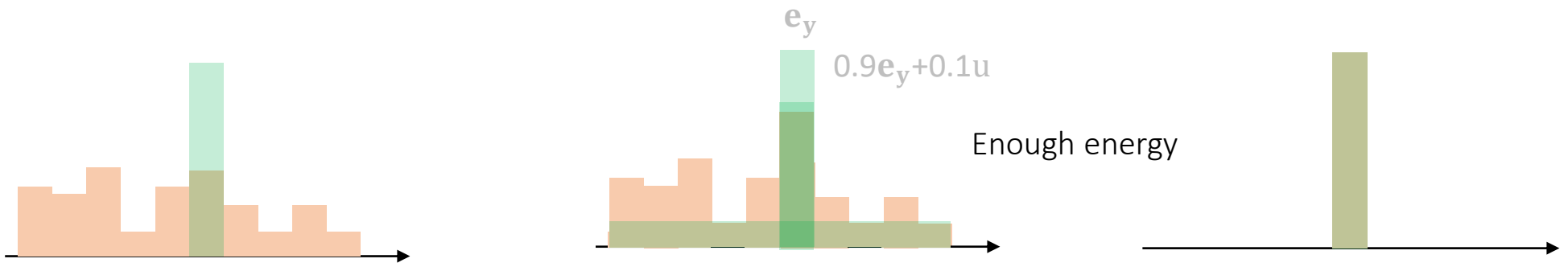


- Understanding the how $\mathbf{e_y} - \mathbf{q_0}$ controls of energy

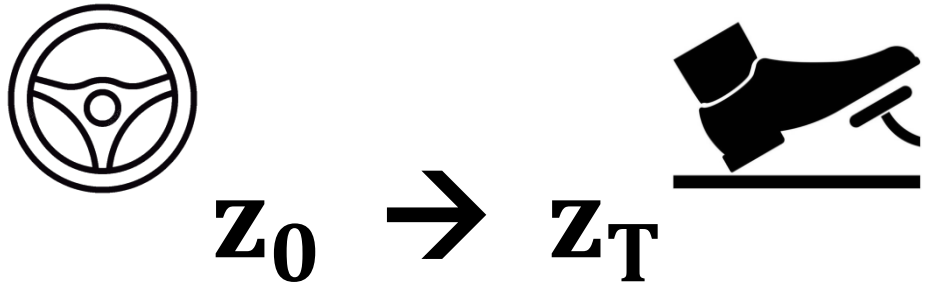


- If the energy still too big after long head-probing
  - Use more complex head to increase HP-accuracy
  - Only copy lower layers from pretrained model

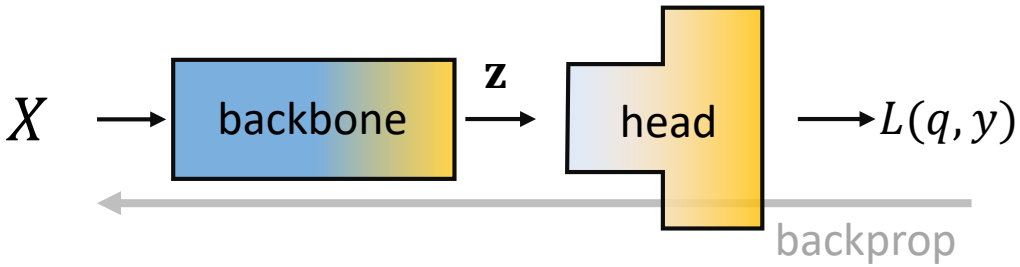
- If HP-accuracy converge too fast
  - Use **smoothed label** during HP, hence reserve some energy
    Now it is safe to HP to converge!

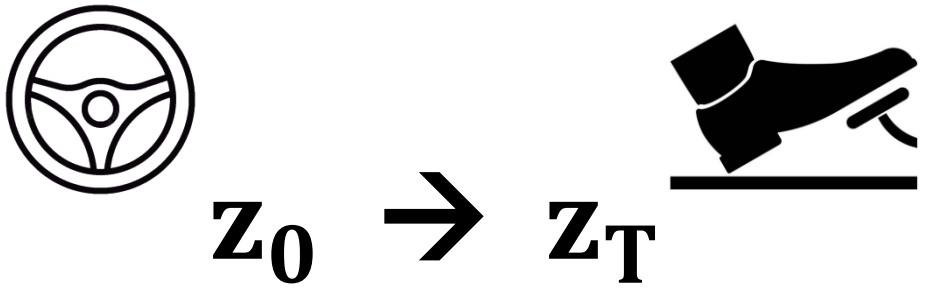


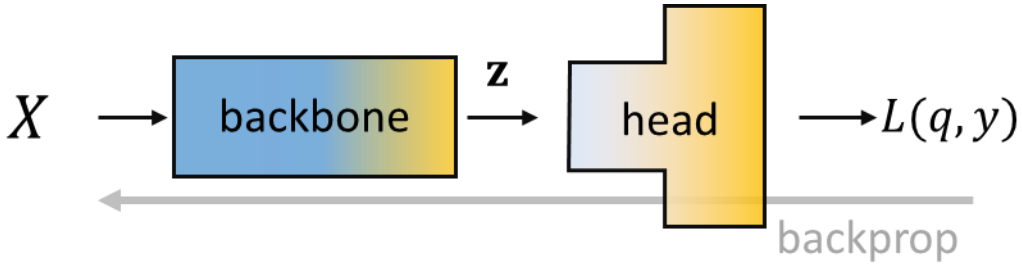
Thanks for your attention!

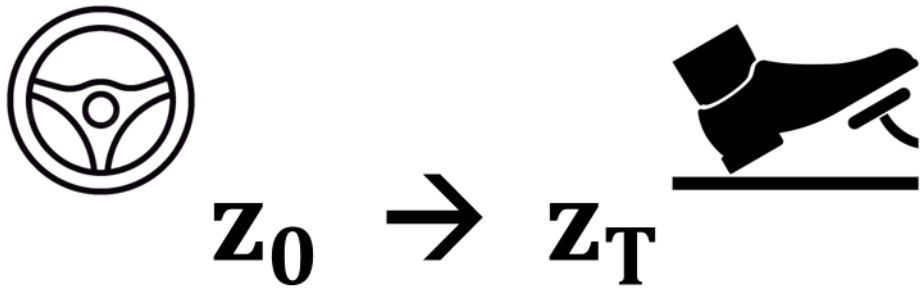


- Stopping earlier in head probing
- Try MLP head or partial copy
- Try label smoothing during HP


X
backbone
z
head
$L(q, y)$
backprop


$\mathbf{z_0}$
$\mathbf{z_T}$


X
backbone
z
head
$L(q, y)$
backprop


$\mathbf{z_0}$
$\mathbf{z_T}$