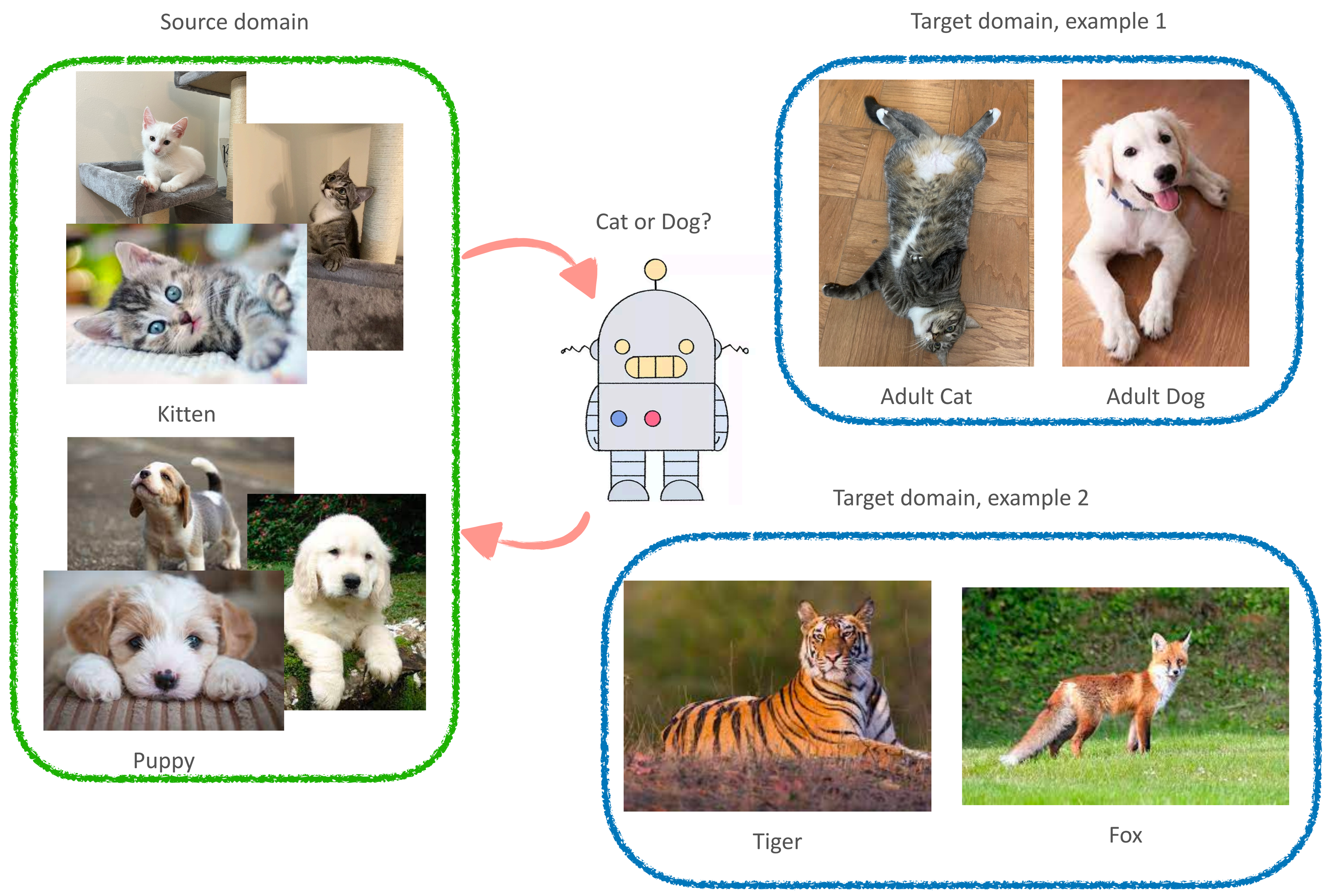
The Power and Limitation of Pretraining-Finetuning for Linear Regression under Covariate Shift

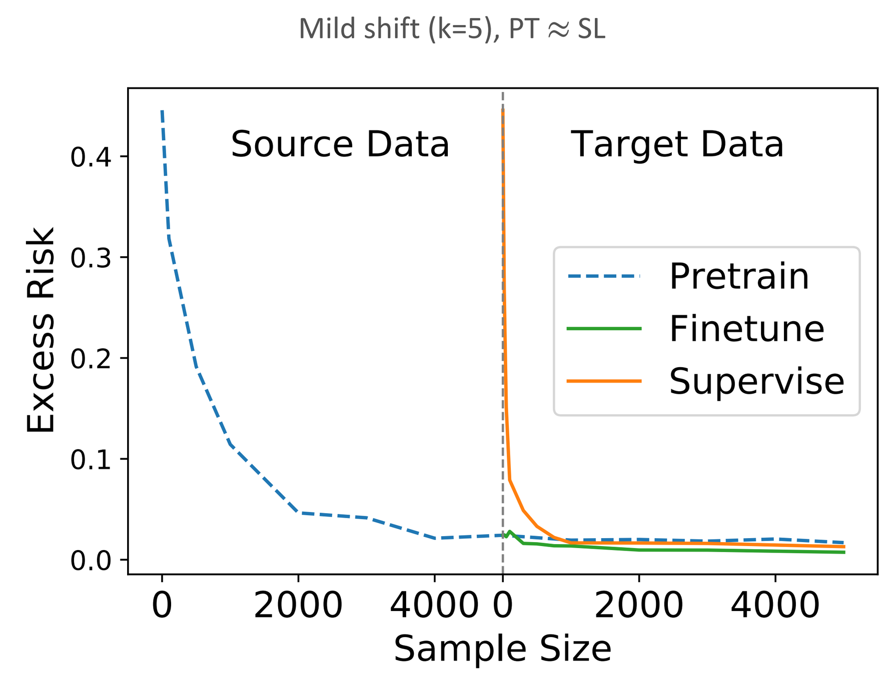
Jingfeng Wu*[1], Difan Zou*[2], Vladimir Braverman[1], Quanquan Gu[3], Sham M. Kakade[4]
[1]Johns Hopkins University, [2]The University of Hong Kong, [3]UCLA, [4]Harvard University

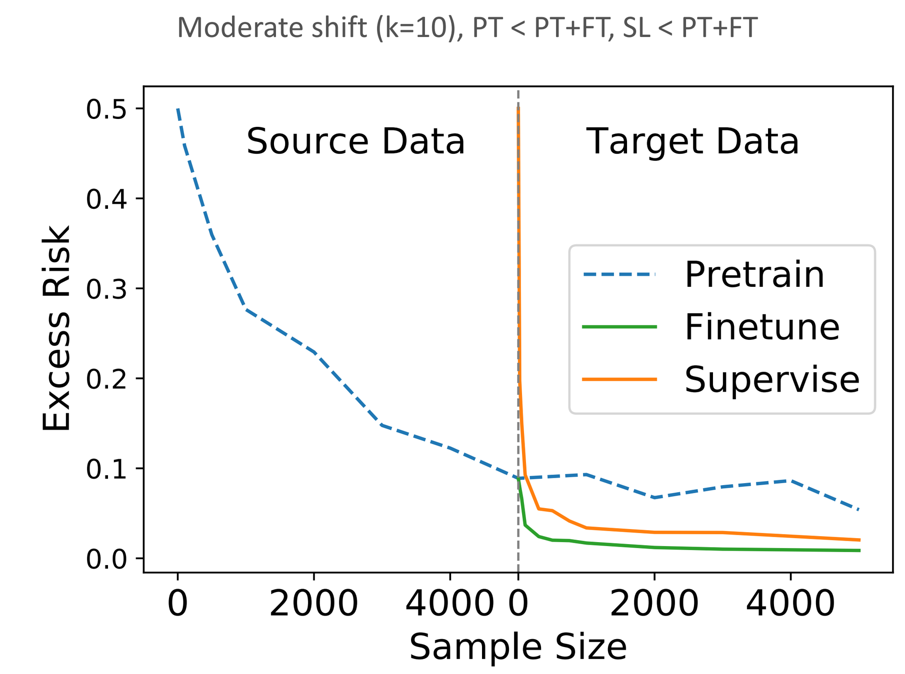
NEURAL INFORMATION PROCESSING SYSTEMS

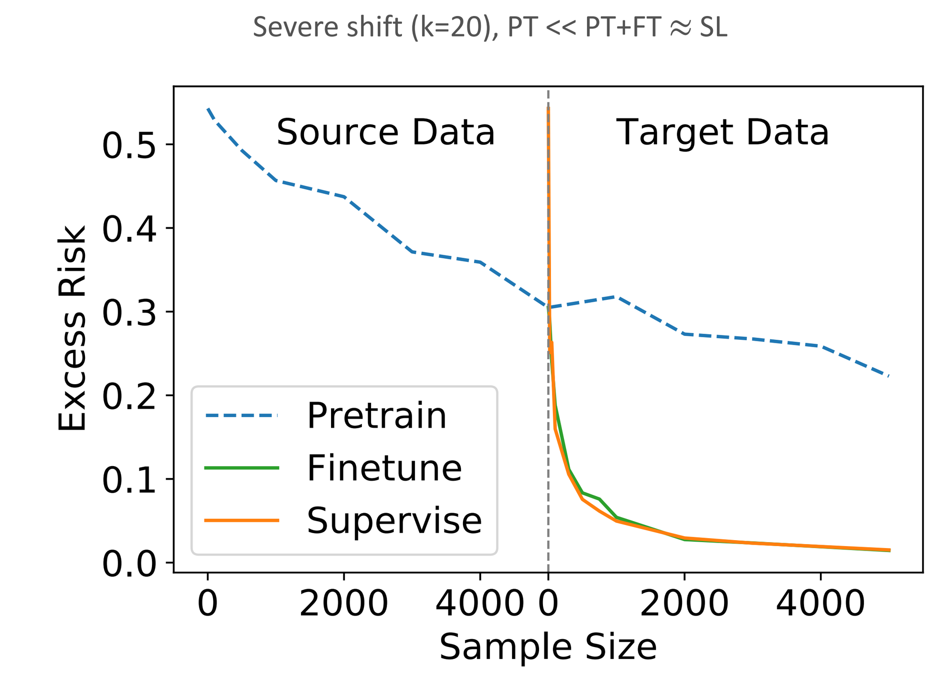
Covariate Shift

$\mathbb{P}_{\mathtt{source}}(x) \neq \mathbb{P}_{\mathtt{target}}(x)$ but $\mathbb{P}_{\mathtt{source}}(y\,|\,x) = \mathbb{P}_{\mathtt{target}}(y\,|\,x)$









Problem Formulation

Linear Regression under Covariate Shift

- Shared Labeling Function $y = \mathbf{x}^\top \mathbf{w}^* + \mathcal{N}(0,\sigma^2)$
- Source/Target Covariance Matrix $\mathbf{G} := \mathbb{E}_{\mathtt{source}}[\mathbf{x}\mathbf{x}^\top]$ $\mathbf{H} := \mathbb{E}_{\mathtt{target}}[\mathbf{x}\mathbf{x}^\top]$
- Target Risk $\mathcal{L}(\mathbf{w}) := \mathbb{E}_{\mathtt{target}}(y - \mathbf{x}^\top\mathbf{w})^2$
- Target Excess Risk $\Delta(\mathbf{w}) := \mathcal{L}(\mathbf{w}) - \mathcal{L}(\mathbf{w}^*) = (\mathbf{w}-\mathbf{w}^*)^\top \mathbf{H}(\mathbf{w}-\mathbf{w}^*)$

Pretraining-Finetuning via Online SGD

Input

- M source data $(\mathbf{x}_t, y_t)_{t=1}^{M} \in \mathbb{R}^{d\times 1}$
- N target data $(\mathbf{x}_{M+t}, y_{M+t})_{t=1}^{N} \in \mathbb{R}^{d\times 1}$
- Initial stepsize η_0 for pretraining, initial stepsize η_M for finetuning

Output $:= \mathbf{w}_{M+N}$, given by

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \eta_{t-t} \cdot (y_t - \mathbf{x}_t^\top \mathbf{w}_{t-1}) \cdot \mathbf{x}_t$$

$$\eta_t = \begin{cases} \eta_0/2^{\ell}, & 0 \le t < M, \ell = \lfloor t/\log(M) \rfloor \\ \eta_M/2^{\ell}, & M \le t < N, \ell = \lfloor (t-M)/\log(N) \rfloor \end{cases}$$

Main Result

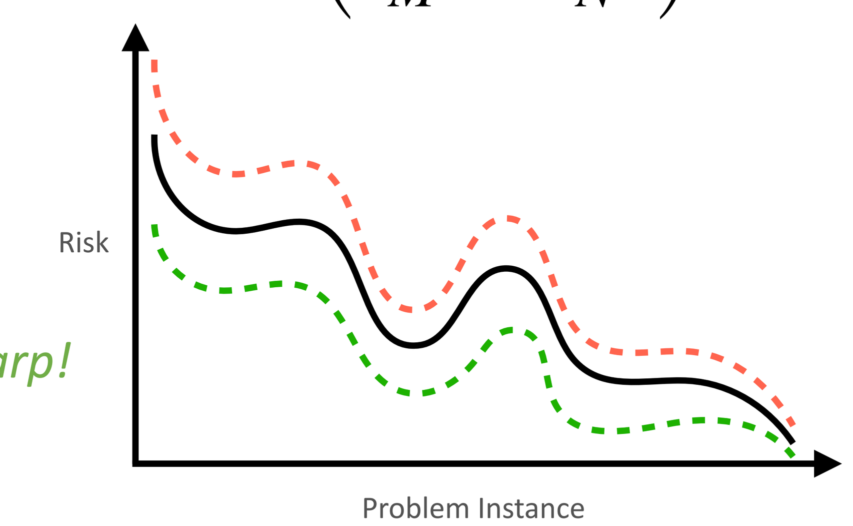
Overview

$$\sigma^2 \cdot \left(\frac{D_{\mathrm{eff}}^{\mathrm{ft}}}{M} + \frac{D_{\mathrm{eff}}}{N} \right) \lesssim \mathbb{E}[\Delta(\mathbf{w}_{M+N})] \lesssim (1+\mathrm{SNR}) \cdot \sigma^2 \cdot \left(\frac{D_{\mathrm{eff}}^{\mathrm{ft}}}{M} + \frac{D_{\mathrm{eff}}}{N} \right)$$

Effective Dimensions

- $D_{\mathrm{eff}}^{\mathrm{ft}}$ is a function of $M, N, \eta_0, \eta_M, \mathbf{G}, \mathbf{H}$
- D_{eff} is a function of $N, \eta_M, \mathbf{H}$

The bound is point-wisely sharp!



A Formal Upper Bound

[Hypercontractivity] Suppose for each $\mathbf{v}$,

$$\mathbb{E}_{\mathtt{source}}\langle \mathbf{v}, \mathbf{x}\rangle^4 \le \alpha\big(\mathbb{E}_{\mathtt{source}}\langle \mathbf{v}, \mathbf{x}\rangle^2\big)^2,\ \mathbb{E}_{\mathtt{target}}\langle \mathbf{v}, \mathbf{x}\rangle^4 \le \alpha\big(\mathbb{E}_{\mathtt{target}}\langle \mathbf{v}, \mathbf{x}\rangle^2\big)^2.$$

Suppose $\eta_0, \eta_M < \min\{1/(4\alpha\mathrm{tr}(\mathbf{G})),\ 1/(4\alpha\mathrm{tr}(\mathbf{H}))\}$. Then

$$\mathbb{E}[\Delta(\mathbf{w}_{M+N})] \lesssim \left\| \prod_{t=M}^{M+N-1} (\mathbf{I} - \eta_t \mathbf{H}) \prod_{t=0}^{M-1} (\mathbf{I} - \eta_t \mathbf{G})(\mathbf{w}_0 - \mathbf{w}^*) \right\|_{\mathbf{H}}^2 + (1+\mathrm{SNR}) \cdot \sigma^2 \cdot \left(\frac{D_{\mathrm{eff}}^{\mathrm{ft}}}{M_{\mathrm{eff}}} + \frac{D_{\mathrm{eff}}}{N_{\mathrm{eff}}} \right)$$

where

- **Signal-to-noise ratio**

$$\mathrm{SNR} := \left(\|\mathbf{w}_0 - \mathbf{w}^*\|_{\mathbf{G}}^2 + \left\| \prod_{t=0}^{M-1} \big(\mathbf{I} - \eta_t \mathbf{G}\big)(\mathbf{w}_0 - \mathbf{w}^*) \right\|_{\mathbf{H}}^2 \right) / \sigma^2$$

- **Effective steps** $M_{\mathrm{eff}} := M/\log(M)$, $N_{\mathrm{eff}} := N/\log(N)$
- **Effective dimension**

$$D_{\mathrm{eff}}^{\mathrm{ft}} = \mathrm{tr}\left(\prod_{t=0}^{N-1} \big(\mathbf{I} - \eta_{M+t}\mathbf{H}\big)^2 \mathbf{H}\big(\mathbf{G}_{\mathbb{J}}^{-1} + M_{\mathrm{eff}}^2 \eta_0^2 \mathbf{G}_{\mathbb{J}^c}\big) \right)$$

$$D_{\mathrm{eff}} := |\mathbb{K}| + N_{\mathrm{eff}}^2 \eta_M^2 \sum_{i \notin \mathbb{K}} \lambda_i^2(\mathbf{H})$$

- **Learnable indexes**

$$\mathbb{J} := \big\{ j : \lambda_j(\mathbf{G}) > 1/(\eta_0 M_{\mathrm{eff}}) \big\},\ \mathbb{K} := \big\{ k : \lambda_k(\mathbf{H}) > 1/(\eta_M N_{\mathrm{eff}}) \big\}$$

Simulations

For k=5, 10 and 20, we consider the following problems:

$$\mathbf{w}^* = \Big(\underbrace{1,\ldots,1}_{k \text{ copies}}, \frac{1}{k+1}, \frac{1}{k+2}, \ldots \Big)^\top, \quad \sigma^2 = 1,$$

$$\mathbf{G} = \mathrm{diag}\left(\frac{1}{k^2}, \ldots, \frac{1}{2^2}, 1, \frac{1}{(k+1)^2}, \ldots \right), \quad \mathbf{H} = \mathrm{diag}\left(1, \frac{1}{2^{1.5}}, \ldots, \frac{1}{k^{1.5}}, \frac{1}{(k+1)^{1.5}}, \ldots \right)$$

Implications

Power of PT / FT

For every problem in $\mathscr{C}$,

$$\mathscr{C} := \big\{ \mathbf{w}^*, \mathbf{H}, \mathbf{G}, \sigma^2 : \|\mathbf{w}^*\|_{\mathbf{G}}^2 \lesssim \sigma^2,\ \mathbf{GH} = \mathbf{HG} \big\}$$

we have

$$\mathbb{E}\Delta(\mathbf{w}_{M+0}) \lesssim \mathbb{E}\Delta(\mathbf{w}_{0+N^{\mathrm{sl}}}) \ \Leftarrow\ \boxed{M \gtrsim (N^{\mathrm{sl}})^2} \cdot \frac{\boxed{\|\mathbf{H}_{\mathbb{K}}\|_{\mathbf{G}}}}{D_{\mathrm{eff}}^{\mathrm{sl}}}$$

$O(n^2)$ source data $\gtrsim n$ target data

- *≈ 1 when $\mathbf{G}$ weakly aligns with $\mathbf{H}$*
- *can be improved with finetune*

Limitation of PT vs. Power of FT

Fix a small $\epsilon > 0$. Consider the following covariate shift problem

$$\mathbf{w}^* = (1,1,0,0,\ldots)^\top, \quad \sigma^2 = 1,$$

$$\mathbf{G} = \mathrm{diag}(\epsilon^2, 1, 0, 0, \ldots), \quad \mathbf{H} = \mathrm{diag}(1, \underbrace{\epsilon^{0.5}, \ldots, \epsilon^{0.5}}_{2\epsilon^{-0.5} \text{ copies}}, 0, 0, \ldots)$$

- **Supervised Learning** $\mathbb{E}\Delta(\mathbf{w}_{0+N}) \lesssim \epsilon \ \Rightarrow\ N \gtrsim \epsilon^{-1.5}$
- **Pretraining** $\mathbb{E}\Delta(\mathbf{w}_{M+0}) \lesssim \epsilon \ \Rightarrow\ M \gtrsim \epsilon^{-2}$
- **Pretraining-Finetuning**

$$\mathbb{E}\Delta(\mathbf{w}_{M+N}) \lesssim \epsilon \ \Leftarrow\ M \eqsim \epsilon^{-1}\log\epsilon^{-1},\ N \eqsim \epsilon^{-1}\log^2\epsilon^{-1}$$

PT+FT could save poly samples than PT or SL alone