The Power and Limitation of Pretraining-Finetuning

for Linear Regression under Covariate Shift

Jingfeng Wu
Joint work with Difan Zou, Vladimir Braverman, Quanquan Gu, Sham M. Kakade

Covariate Shift

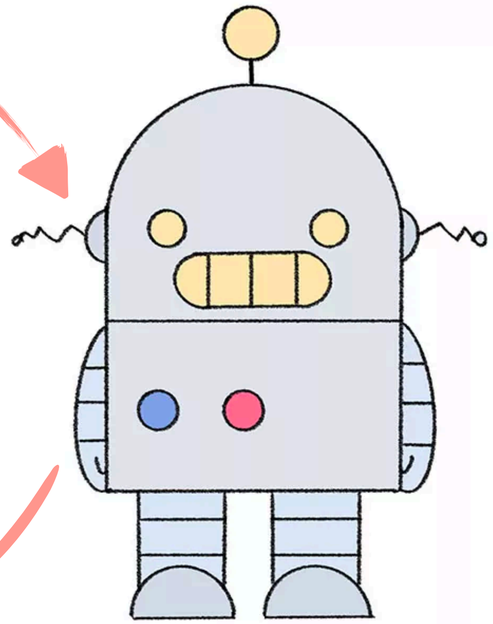
$\mathbb{P}_{\texttt{source}}(x) \neq \mathbb{P}_{\texttt{target}}(x)$ **but** $\mathbb{P}_{\texttt{source}}(y \mid x) = \mathbb{P}_{\texttt{target}}(y \mid x)$

Source domain

Kitten

Puppy

Cat or Dog?



How to generalize on the target domain?

Target domain, example 1

Adult Cat

Adult Dog

Target domain, example 2

Tiger

Fox

A Basic Yet Fundamental Approach

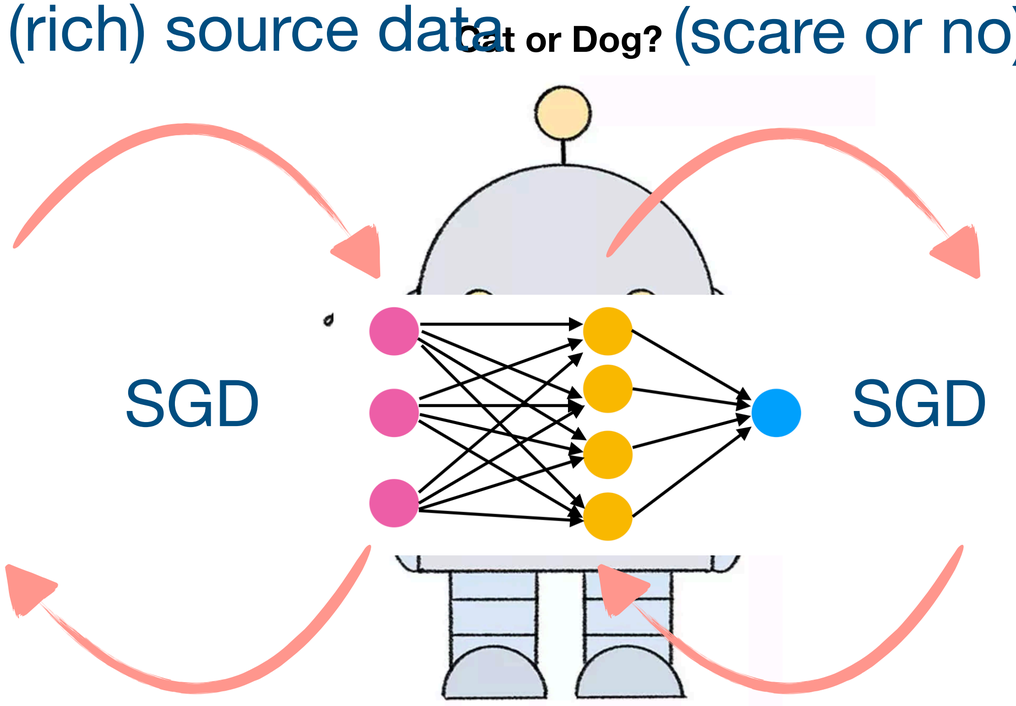
Source domain

Kitten

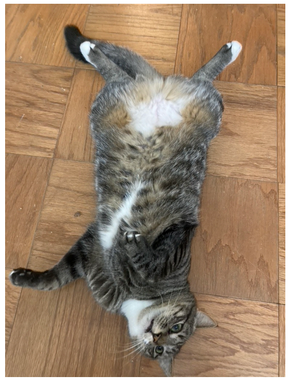
Puppy

Pretrain with
(rich) source data

Finetune with
(scare or no) target data



Target domain



Adult Cat

Adult Dog

How good is the simple pretraining-finetuning approach?

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 $\mathbf{G} := \mathbb{E}_{\texttt{source}}[\mathbf{x}\mathbf{x}^\top] \quad \mathbf{H} := \mathbb{E}_{\texttt{target}}[\mathbf{x}\mathbf{x}^\top]$
- Target Risk $\mathscr{L}(\mathbf{w}) := \mathbb{E}_{\texttt{target}}(y - \mathbf{x}^\top \mathbf{w})^2$
- Target Excess Risk

$$\Delta(\mathbf{w}) := \mathscr{L}(\mathbf{w}) - \mathscr{L}(\mathbf{w}^*) = (\mathbf{w} - \mathbf{w}^*)^\top \mathbf{H} (\mathbf{w} - \mathbf{w}^*)$$

Problem Formulation

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

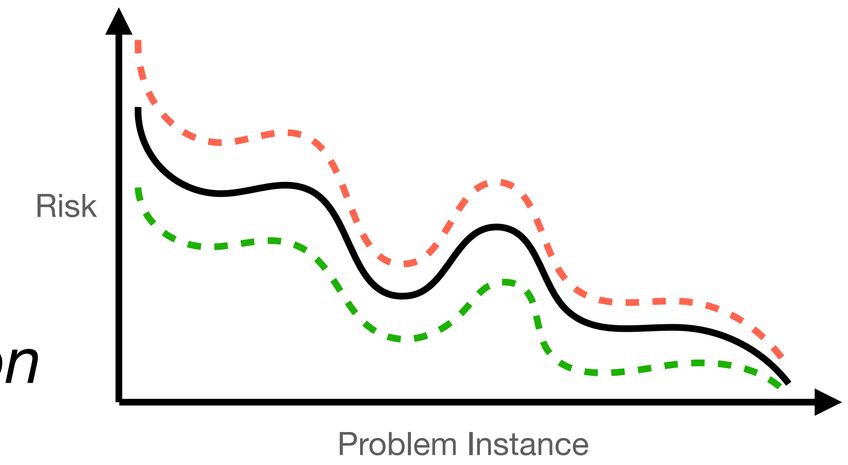
A Sharp Risk Bound for Pretraining-Finetuning

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

Effective Dimension

- $D_{\texttt{eff}}^{\texttt{ft}}$ is a function of M, N, η_0, η_M, $\mathbf{G}$, $\mathbf{H}$
- $D_{\texttt{eff}}$ is a function of N, η_M, $\mathbf{H}$

Our bound is point-wisely sharp



Assumption

- Source & target data: *hyper-contractivity condition*
- $\texttt{SNR} := \|\mathbf{w}_0 - \mathbf{w}^*\|_2^2 \,/\, \sigma^2 \lesssim 1$

Implication 1

Pretraining with M data vs. SL with $N^{\mathtt{sl}}$ data

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

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

we have

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

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

≈ 1 when $\mathbf{G}$ weakly aligns with $\mathbf{H}$

Implication 2

PT with M data vs. SL with N data vs. PT+FT with M+N data

For each $\epsilon > 0$, there exists a problem instance in $\mathscr{C}$, such that:

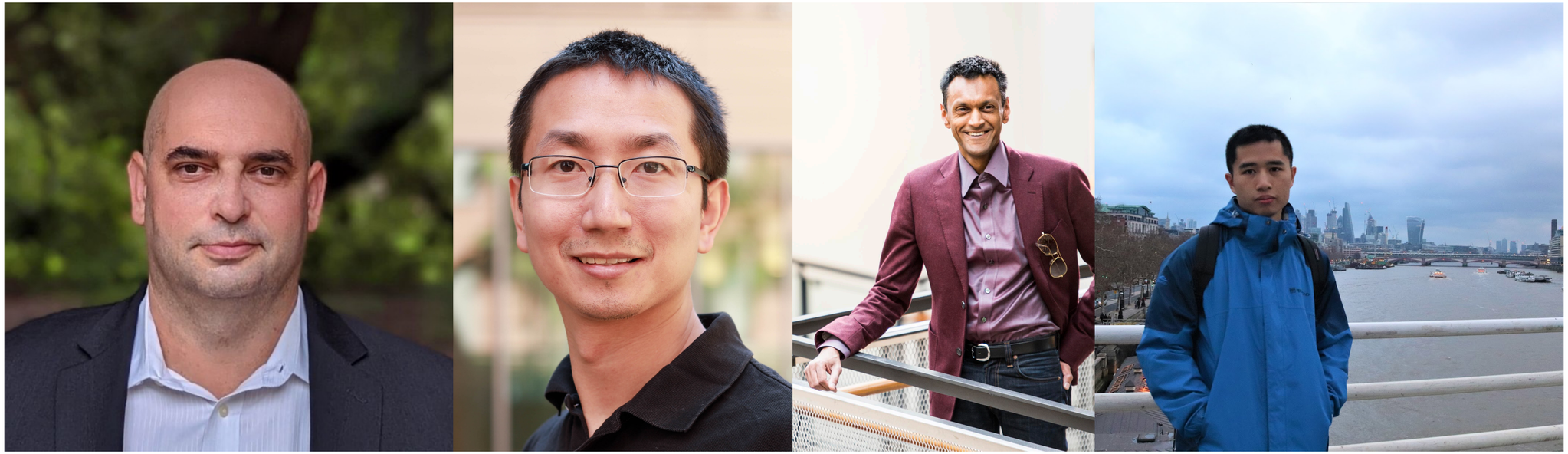
to achieve ϵ-excess risk:

- **Pretraining** $\Rightarrow \; M \gtrsim \epsilon^{-2}$
- **Supervised Learning** $\Rightarrow \; N \gtrsim \epsilon^{-1.5}$
- **Pretraining-Finetuning** $\Leftarrow \; M \eqsim \epsilon^{-1} \log \epsilon^{-1}, \; N \eqsim \epsilon^{-1} \log^2 \epsilon^{-1}$

PT+FT could save polynomially than PT or SL alone

Take Home

1. Point-wisely sharp bounds for linear regression under covariate shift
2. $O(n^2)$ source data $\gtrsim n$ target data, when H weakly aligns with G
3. Pretraining-finetuning could save *polynomially* than pretraining (or SL)



Vova Bravermen Quanquan Gu Sham Kakade Difan Zou