Guarantees for Nonlinear Representation Learning

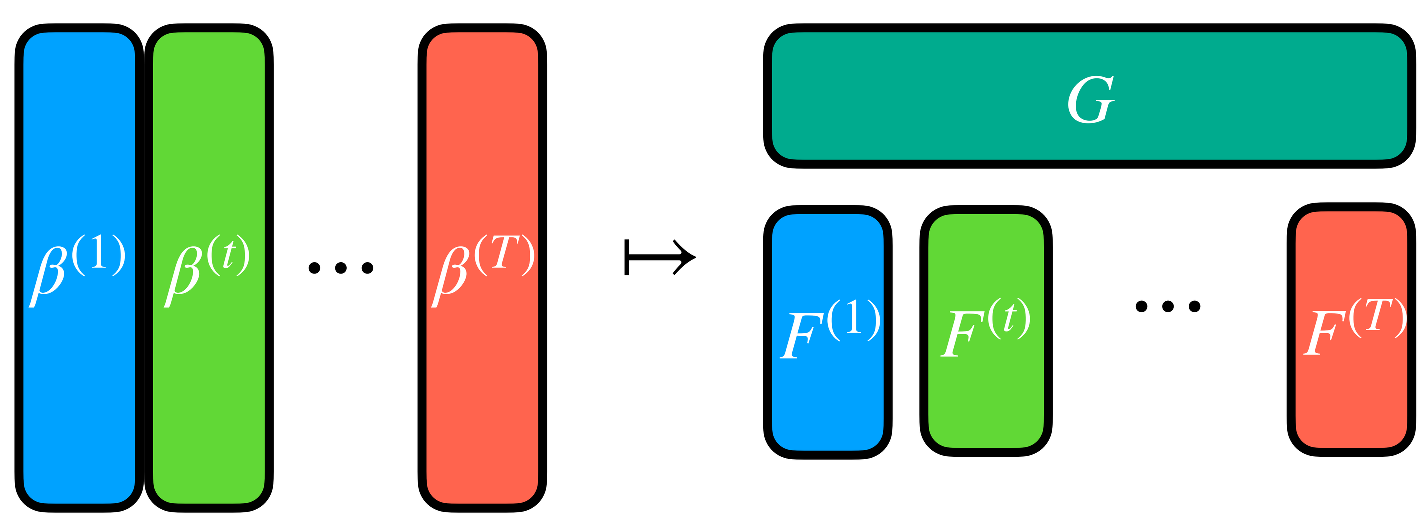
Non-Identical Covariates, Dependent Data, Fewer Samples

Thomas T.C.K. Zhang, Bruce D. Lee, Ingvar Ziemann, George J. Pappas, Nikolai Matni
Department of Electrical and Systems Engineering, University of Pennsylvania   Email: ttz2@seas.upenn.edu

Motivation: Multi-Task Learning

Modern deep learning is driven by ability to learn **meaningful representations from diverse data**.

Natural way to encourage learning performant representations from multi-task data is to enforce a **shared representation**. Task specification comes from small model trained on top of representation.



Cartoon of parameter-efficiency via multi-task rep. learning

From a theoretical perspective, want to formalize:

- Per-task sample-efficiency better than single-task setting.
- Data across all tasks should contribute to representation learning.
- Gains determined by some (tight) measure of **task diversity** or **task coverage**.

Problem Set-Up: Regression

Receive data from $t = 1,\ldots,T$ tasks of the form

$$y^{(t)} = h_\star^{(t)}(x^{(t)}) + w^{(t)}, \qquad w^{(t)} \sim \mathcal{D}(0, \sigma_w^2)$$

Shared representation: each task's predictor factorizes into task-specific linear heads $F_\star^{(t)} \in \mathbb{R}^{1\times r}$ and shared **nonlinear** rep $g_\star : \mathbb{R}^{d_x} \to \mathbb{R}^r$

$$h_\star^{(t)}(\cdot) = F_\star^{(t)} g_\star(\cdot).$$

Want to understand transfer risk onto downstream task $h_\star^{(0)} = F_\star^{(0)} g_\star(\cdot)$.

Setting Expectations

Consider Empirical Risk Minimizer (ERM):

$$\{\hat{F}^{(t)}\}, \hat{g} \in \operatorname*{argmin}_{\{F^{(t)}\}, g} \sum_{i,t} \left\| y_i^{(t)} - F^{(t)} g(x_i^{(t)}) \right\|^2$$

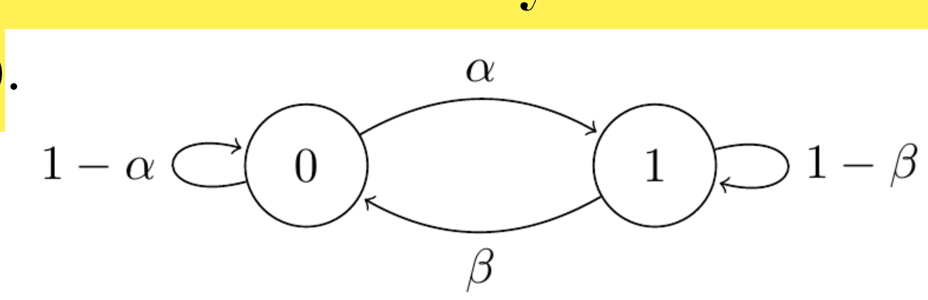
What kind of guarantees to expect?

1. When optimal rep $g_\star$ given, each task becomes r**-dim lin reg problem**: burn-in (sample requirement) is $\Omega(r)$ and gen bound scales $\mathcal{O}(r/N)$ for N points per task.
2. When tasks are **identical** $F_\star^{(0)} = F_\star^{(1)} = \cdots = F_\star^{(T)}$, $D_x^{(0)} = \cdots = D_x^{(T)}$, task coverage measure should be ideal regardless of structure of $F_\star^{(0)}, D_x^{(0)}$, e.g. $|\operatorname{supp}(F_\star^{(0)})| \ll r$.
3. **Beyond independent covariates:** by recent work, effect of (sequential) dependence in single-task regression only enters burn-in, not gen bound.

Deficiencies of Prior Work

Prior guarantees make the following key assumptions:

- Covariates are **independent** and **identically distributed** across all tasks. Precludes sequential settings—**non-identical stationary dists** induced by different $h_\star^{(t)}(\cdot)$.



- Large burn-in required per task. E.g. linear setting requires $\Omega(d_x) \gg r$ samples—**each task is already solvable from scratch**.
- Task coverage is assumed **uniform**, i.e. $\left[F_\star^{(1)\prime} \ \cdots \ F_\star^{(T)\prime}\right]$ has **full rank** $= r$ and **well-conditioned**. Implies rep dimension **cannot be overestimated!**

Generalization Guarantees

Given N datapoints per training task $t = 1,\ldots,T$ and transfer task (0):

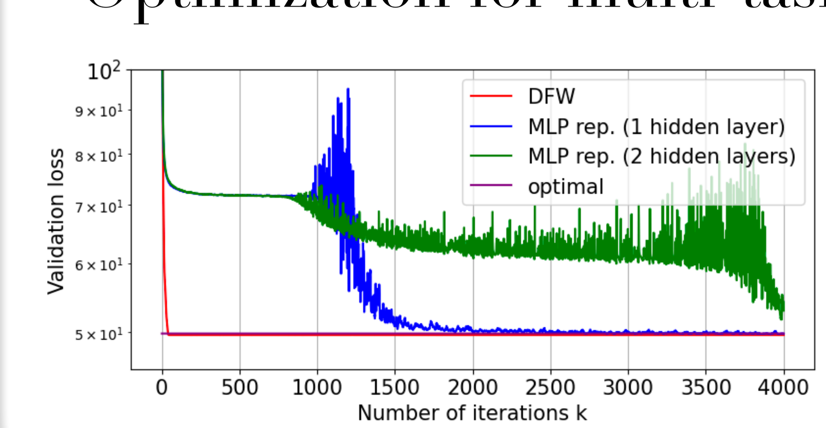
Key Theorem: as long as $N \gtrsim \tau_{\mathrm{mix}}(r + \mathrm{Comp}(G)/T)$, with high probability ERM satisfies

$$\mathrm{Risk}(\hat{F}^{(0)}, \hat{g}) \leq C_{\mathbf{X}} C_{\mathbf{F}} \cdot \sigma_w^2 \left(\frac{r}{N} + \frac{\mathrm{Comp}(G)}{NT} \right).$$

- τ_{mix}: effect of dependent data. **Generalization bound is unaffected!**
- $\mathrm{Comp}(G)$: complexity measure of rep. class $g \in G$. Effect of rep class G is **distributed across tasks!**
  - When T large, burn-in and rate approaches optimal $\Omega(r)$ and $\mathcal{O}(r/N)$ when $g_\star$ given.
- $C_{\mathbf{X}}$: **"overlap" of covariate distributions**.
  - $C_{\mathbf{X}} = 1$ when covariate dists. identical.
  - $C_{\mathbf{X}} = \infty$ when $\operatorname{supp}(D_x^{(0)}) \cap \operatorname{supp}(\{D_x^{(t)}\}) = \varnothing$.
- $C_{\mathbf{F}}$: **"overlap" of task-specific predictors**.
  - $C_{\mathbf{F}} = 1$ when $F_\star^{(0)} = F_\star^{(1)} = \cdots = F_\star^{(T)}$.
  - $C_{\mathbf{F}} = \infty$ when $F_\star^{(0)} \notin \operatorname{range}(F_\star^{(1)}, \ldots, F_\star^{(T)})$.

Discussion and Future Directions

- Guarantees for regression can be ported into various settings, e.g. stochastic contextual bandits.
- Existence result for in-context learning: $\exists$ algorithm (ERM) that benefits from multi-task data.
- Optimization for multi-task models is non-trivial!



See our concurrent work for more details