Transfer Learning: Statistical Perspectives

Yuyao Wang

Boston University

Reference: AC295 Lecture 5, “Intro to Transfer Learning” (Protopapas).

Why Transfer Learning? A Realistic Motivating Example

Scenario: Deploying a model under distribution shift

You trained a predictive model on **City A** (abundant historical data) but must deploy in **City B** (new sensors, limited labels). Examples: traffic speed forecasting, demand prediction, air-quality mapping, health-risk scoring, fraud detection.

What changes in practice (typical sources of shift)?

- **Population / behavior:** commuting patterns, purchasing habits, patient mix.
- **Measurement system:** sensor placements, missingness patterns, device calibration.
- **Environment & policy:** road network changes, new regulations, seasonality differences, interventions.

Transfer learning goal. Use *large source data* to learn reusable structure, then *adapt* using *limited target data* to reduce target error and avoid negative transfer.

Problem Setup and Guiding Questions

Setup. Training data come from a *source* distribution, but deployment occurs on a (possibly different) *target* distribution:

$$(X, Y) \sim P_{\mathrm{s}} \quad \text{(source)} \qquad \Longrightarrow \qquad (X, Y) \sim P_{\mathrm{t}} \quad \text{(target)}, \quad P_{\mathrm{s}} \neq P_{\mathrm{t}}.$$

Objective. Learn a predictor f that minimizes the *target risk*

$$R_{\mathrm{t}}(f) = \mathbb{E}_{(X,Y)\sim P_{\mathrm{t}}}\left[\ell(f(X), Y)\right],$$

potentially leveraging *source* data and limited *target* supervision, while avoiding *negative transfer*.

Guiding questions.

1. **What to transfer?** instances, representations, parameters, priors, or structure
2. **When to transfer?** quantify mismatch; select/weight sources; detect negative transfer
3. **How to transfer?** reweighting, alignment, adaptation, regularization, meta-learning, fine-tuning

What Changes from Source to Target? A Shift Taxonomy

Transfer learning is not a single problem: **which part of the data-generating process changes** determines what is feasible.

Common shift types (with canonical assumptions)

- **Covariate shift:** $P_s(X) \neq P_t(X)$ but $P_s(Y \mid X) = P_t(Y \mid X)$.
- **Label shift:** $P_s(Y) \neq P_t(Y)$ but $P_s(X \mid Y) = P_t(X \mid Y)$.
- **Conditional / concept shift:** $P_s(Y \mid X) \neq P_t(Y \mid X)$ (hardest; requires structure).

Target risk. Under covariate shift,

$$R_t(f) = \mathbb{E}_{(X,Y)\sim P_s}\Big[\underbrace{\tfrac{p_t(X)}{p_s(X)}}_{w(X)}\,\ell(f(X), Y)\Big],$$

which motivates *importance weighting* if $w(X)$ can be estimated and overlap holds.

Statistical Takeaway: What Must Be Answered Under Shift?

The key questions become:

1. **Identifiability:** What target quantities are *estimable* from available source/target data under the assumed shift?
2. **Robustness:** What happens if the assumed shift is *slightly wrong*? Can we control worst-case or sensitivity?
3. **Uncertainty:** How confident are we on the target (e.g., valid CIs / prediction intervals / calibrated uncertainty)?

Practical implication. A "good" transfer method should (i) state its shift assumptions clearly, (ii) degrade gracefully when assumptions fail, and (iii) quantify when transfer may be harmful (negative transfer).

Where Is Statistical Transfer Learning Focused Today?

Current focal points

- **Transportability / generalizability:** conditions under which target risk/effects are identifiable from biased or shifted data.
- **Shift-aware estimation:** importance weighting, doubly-robust estimators, and semiparametric efficiency under covariate/label shift.
- **Uncertainty quantification:** valid inference on the target (CIs, prediction intervals, calibration) under shift and limited target labels.
- **Robustness & safety:** distributionally robust objectives (e.g., f-divergence / Wasserstein balls) and negative-transfer control.
- **Multi-source borrowing:** source selection/weighting, hierarchical (empirical Bayes) shrinkage, and adaptive regularization.
- **Structured domains:** transfer on graphs, time series, and spatial data via interpretable inductive biases (smoothness, low-rank, graphon).

Deep Transfer Learning: Current Focus

Main idea: learn *general representations* from large data, then *adapt quickly* to new tasks/domains.

What the field is optimizing for

- **Pretraining $\rightarrow$ transfer:** self-supervised / foundation models + fine-tuning or prompting.
- **Cheap adaptation:** parameter-efficient tuning (adapters/LoRA) instead of full retraining.
- **Little/no target labels:** domain adaptation via alignment + pseudo-labeling/self-training.
- **Robustness to new domains:** OOD generalization and test-time adaptation.
- **Long-term deployment:** continual learning (avoid forgetting as domains shift over time).

Summary. *Scale representations, adapt efficiently, stay stable under shift.*

Statistical Transfer Learning: Current Focus

Main idea: transfer under *explicit shift assumptions*, with *validity* and *uncertainty* on the target.

What the field is optimizing for

- **Shift assumptions:** what changes from P_s to P_t?
- **Safe borrowing:** weighting, shrinkage/priors, multi-source selection.
- **Robustness:** avoid negative transfer under mismatch.
- **Uncertainty:** calibrated prediction and valid intervals on target.

Summary. *Make assumptions explicit, borrow cautiously, quantify uncertainty.*

When Does Negative Transfer Happen?

Negative transfer: using source information makes target worse:

$$R_{\mathrm{t}}(f_{\text{transfer}}) > R_{\mathrm{t}}(f_{\text{target-only}}).$$

- **Mismatch is large:** weak overlap / different regimes between source and target.
- **Wrong invariance:** assuming $P_{\mathrm{s}}(Y \mid X) = P_{\mathrm{t}}(Y \mid X)$ when it actually changes.
- **Shortcut features:** model transfers spurious correlations that do not hold in target.
- **Too much source pull:** strong priors/alignment blocks target-specific fitting.
- **Self-training errors:** noisy pseudo-labels reinforce mistakes (confirmation bias).

Practice: estimate mismatch, down-weight unreliable sources, and keep a target-only baseline.

My Research (I): Phase Transition in Transfer Learning

Focus. When is transfer *beneficial* versus *harmful* under distribution shift?

Key contribution

- **Phase transition:** characterize a threshold (in target sample size and/or source–target mismatch) where the optimal strategy switches from *borrowing source information* to *target-only fitting*.
- **Mechanism:** formalize the bias–variance tradeoff induced by source regularization/priors, and identify regimes that lead to **negative transfer**.

Method. Adaptive transfer strength (data-driven shrinkage):

$$\text{transfer weight} \uparrow \quad \text{when similarity/overlap is high,}$$
$$\text{transfer weight} \downarrow \quad \text{when mismatch is high.}$$

My Research (II): Graphon-Based Transfer for Networks

Focus. Transfer learning for *network structure*: estimate target edge probabilities with limited/noisy target graphs.

Key contribution

- **Graphon model:** represent each network by a latent edge-probability function, capturing shared structure across domains.
- **Transfer via alignment:** match latent positions/blocks between source and target (e.g., OT-based matching), then borrow information to improve target estimation.

Methodological shift. From black-box transfer to **model-based transfer**:

$$\text{latent structure learning \& alignment} \Rightarrow \text{transport of edge-probability information} \Rightarrow \text{improved target estimation.}$$

My Research (III): SCOT — Cross-City Transfer via Optimal Transport

Focus. Cross-city transfer for urban prediction (e.g., mobility / socio-economic signals) when target labels are scarce.

Key contribution

- **OT alignment of regions:** learn city-specific region embeddings and compute a Sinkhorn OT coupling to match regions across cities.
- **OT-guided transfer:** use the coupling to (i) align representations and (ii) weight contrastive objectives, improving cross-city generalization.
- **Structure preservation:** incorporate reconstruction/geometry terms so the alignment respects local neighborhood semantics.

Method. A principled pipeline for cross-city transfer:

$$\text{learn region embeddings} \Rightarrow \text{Sinkhorn OT coupling (region matching)}$$
$$\Rightarrow \text{OT-guided alignment} + \text{reconstruction}$$
$$\Rightarrow \text{improved target prediction.}$$

Transfer Learning Strategies (I): What to Transfer?

Four common transfer "objects"

- **Instance:** reuse (reweighted) source samples, $w(x) \approx \frac{p_t(x)}{p_s(x)}$.
- **Representation:** learn features $h(\cdot)$ that reduce source–target mismatch (alignment).
- **Parameter:** reuse/shrink parameters or priors, e.g. $\|\theta - \hat{\theta}_s\|^2$.
- **Structure:** transfer relational patterns (graph/space/time dynamics).

Rule of thumb: pick the transfer object that is most likely *invariant* across domains.

Transfer Learning Strategies (II): Which Setting Are You In?

Setting = domain/task relationship + target label availability.

Canonical settings

- **Inductive transfer learning: tasks differ** ($\mathcal{T}_s \neq \mathcal{T}_t$), target has (some) labels.
  Typical: pretrain $\rightarrow$ fine-tune; multitask learning; meta-learning.
- **Transductive transfer learning (domain adaptation): same task**, **different domains** ($\mathcal{D}_s \neq \mathcal{D}_t$), target labels scarce/none.
  Typical: feature alignment (MMD/adversarial/OT) + self-training/pseudo-labels.
- **Unsupervised transfer learning:** target task is unsupervised (e.g., clustering/representation); labels absent in target (often both).
  Typical: transferable embeddings, self-supervised objectives.

How to Think About Transfer Learning (A Checklist)

A practical way to formulate a TL problem

1. **Specify the shift:** what differs between source and target (data, labels, mechanism, structure)?
2. **Specify supervision:** how many target labels, and what is available at test time?
3. **Choose the transfer object:** instance / representation / parameter / structure.
4. **Pick the control knob:** how strongly to trust the source (e.g., λ in a transfer penalty).
5. **Define success & safety:** target risk + robustness (avoid negative transfer; add uncertainty/calibration if needed).

Template: $\min_{f} \ \widehat{R}_{t}(f) \ + \ \lambda\,\Omega(f; \text{source}, \text{target})$ (fit target) + (transfer / alignment / prior)

Where Are the Innovation Opportunities in Transfer Learning?

Most TL papers contribute by changing one (or more) of these pieces

- **Better shift modeling:** new assumptions, new invariances, causal/structural views of what transfers.
- **Better transfer mechanism Ω:** e.g., OT-based matching, graph/time-aware alignment.
- **Better source selection/weighting:** decide *which* sources to trust and *how much* (adaptive λ).
- **Better training protocol:** pretrain $\rightarrow$ adapt; meta-learning; test-time adaptation; continual adaptation.
- **Better safety tools:** negative-transfer detection, robustness objectives (DRO), calibrated uncertainty / conformal prediction.
- **Better evaluation:** realistic shifts, multi-source benchmarks, ablations for “what transfers” and “when it fails”.