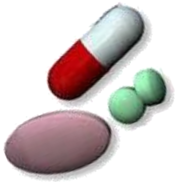
max planck institut
informatik

Multi-Task Learning for HIV Therapy Screening

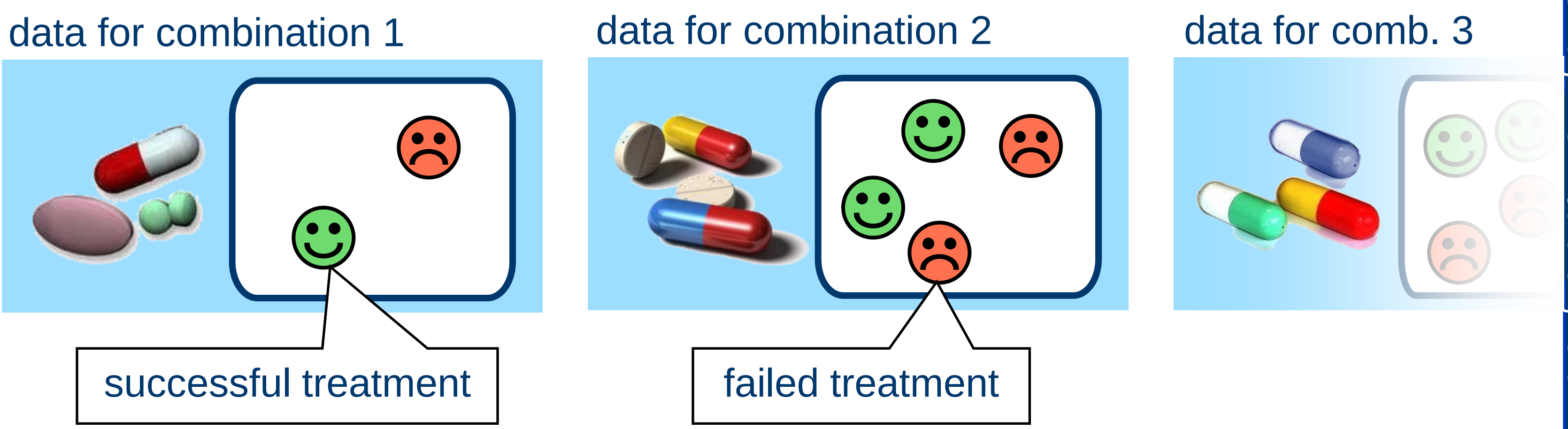
Steffen Bickel, Jasmina Bogojeska,
Thomas Lengauer, Tobias Scheffer




HIV Therapy Screening



- Usually combinations (3-6 drugs) out of around 17 antiretroviral drugs administered.
- Effect of combinations on virus similar but not identical.
- Scarce training data available from treatment records.



- Challenge: Prediction of therapy outcome from genotypic information.

Bickel, Bogojeska, Lengauer, Scheffer

Multi-Task Learning

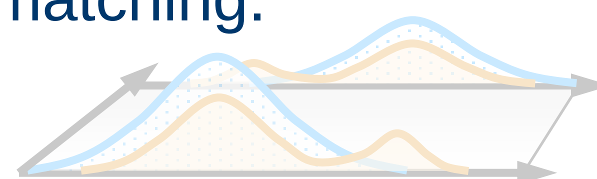
- Several related prediction problems (tasks).
  - Not necessarily identical conditional $p(y|\mathbf{x})$ of label given input.
  - Usually, some conditionals are similar.
- Challenge:
  - Use all available training data and account for the difference in distributions accross tasks.

- HIV therapy screening:
  - Can be modeled as multi-task learning problem.
  - Drug combinations (tasks) have similar but not identical effect on the virus.

Overview

- Motivation.
  - HIV therapy screening.
  - Multiple tasks with differing distributions.

- Multi-task learning by distribution matching.
  - Problem Setting.
  - Density ratio matches pool to target distribution.
  - Discriminative estimation of matching weights.



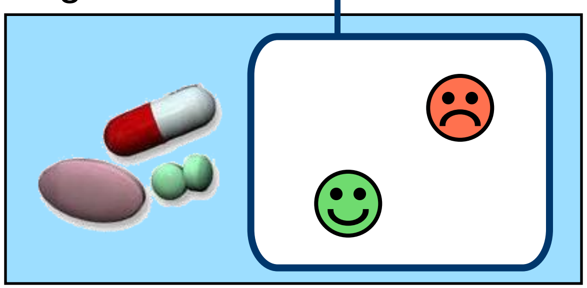
- Case study:
  - HIV therapy screening.

Multi-Task Learning – Problem Setting

Target distribution

$$p(\mathbf{x}, y | t)$$

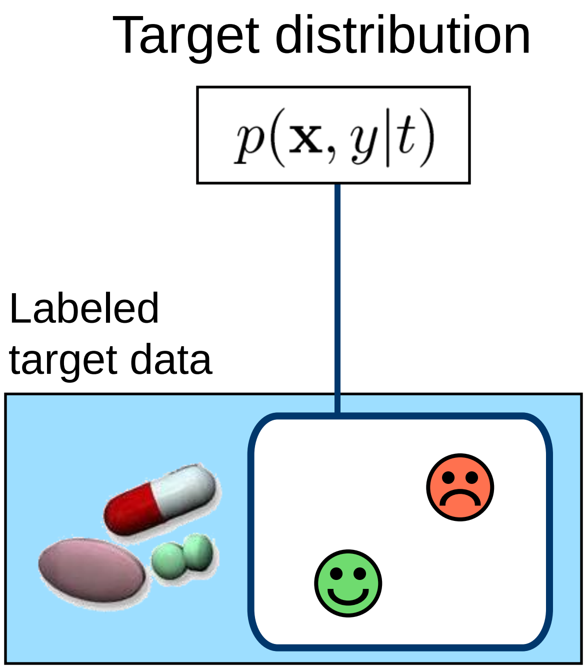
Labeled
target data



Bickel, Bogojeska, Lengauer, Scheffer

Multi-Task Learning – Problem Setting

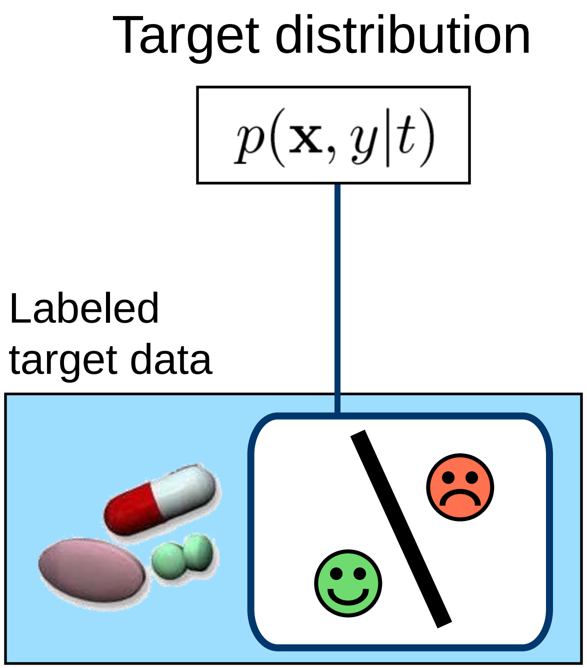
- Goal: Minimize loss under target distribution.
  - $\mathbf{E}_{(\mathbf{x},y)\sim p(\mathbf{x},y|t)}[\ell(f(\mathbf{x}),y)]$



Bickel, Bogojeska, Lengauer, Scheffer

Multi-Task Learning – Problem Setting

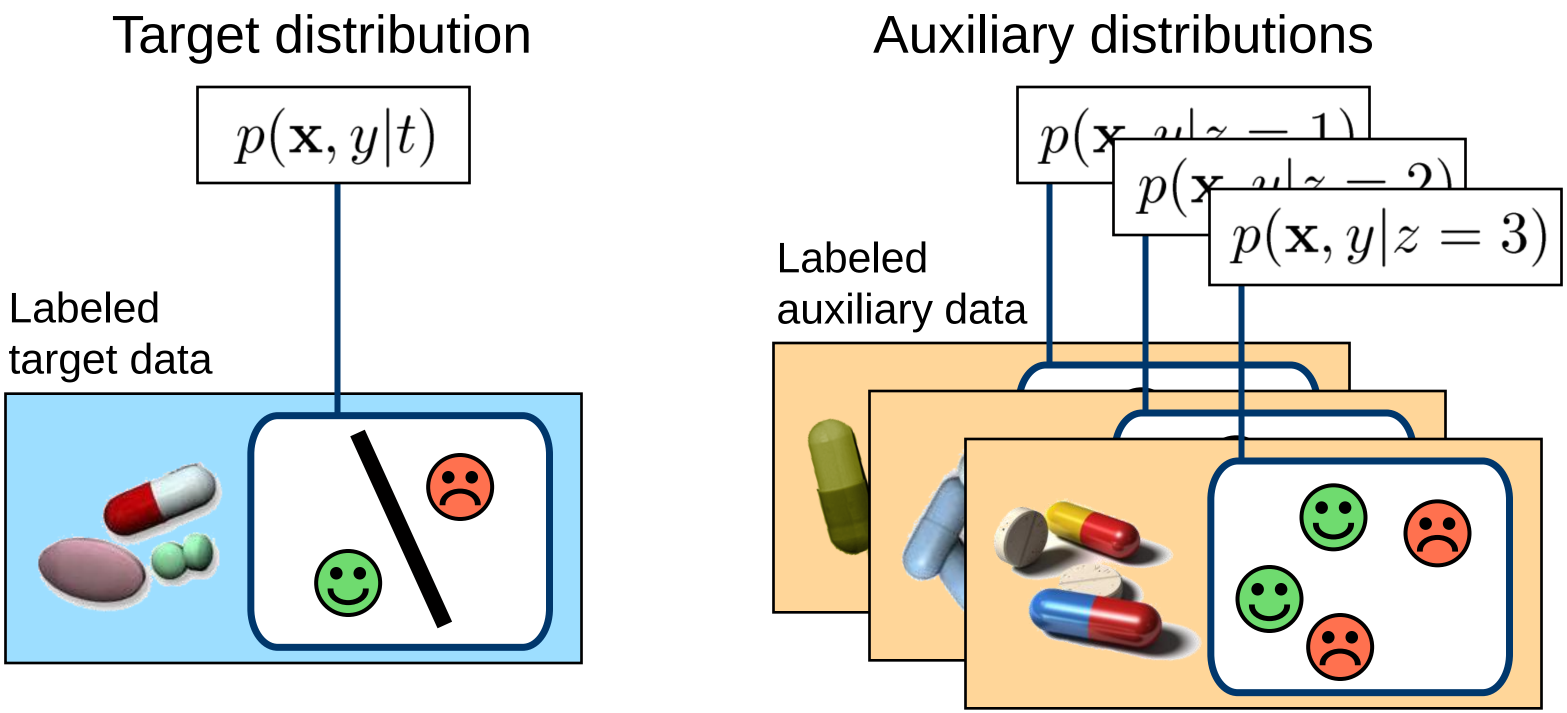
- Goal: Minimize loss under target distribution.
  - $\mathbf{E}_{(\mathbf{x},y)\sim p(\mathbf{x},y|t)}[\ell(f(\mathbf{x}),y)]$



Bickel, Bogojeska, Lengauer, Scheffer

Multi-Task Learning – Problem Setting

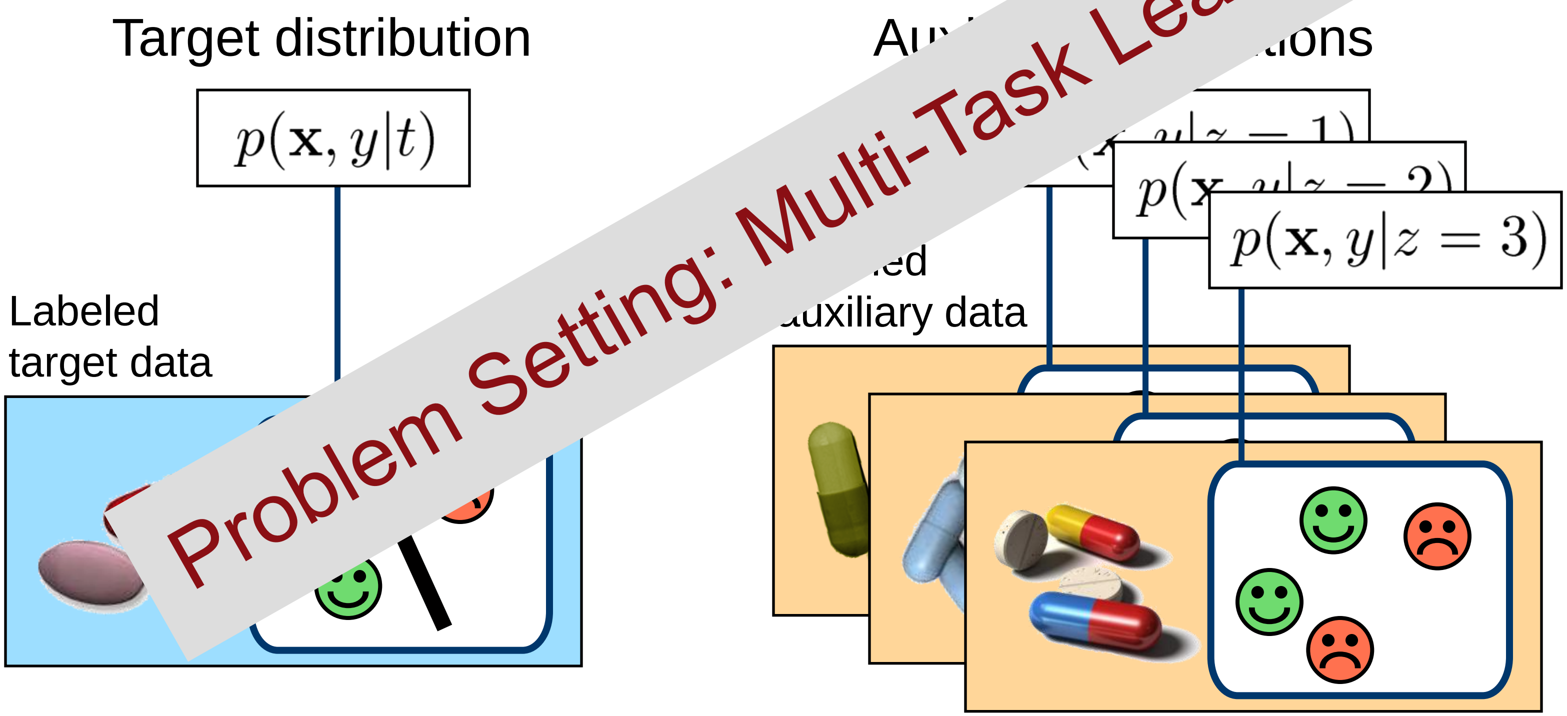
- Goal: Minimize loss under target distribution.
  - $\mathbf{E}_{(\mathbf{x},y)\sim p(\mathbf{x},y|t)}[\ell(f(\mathbf{x}),y)]$



Bickel, Bogojeska, Lengauer, Scheffer

Multi-Task Learning – Problem Setting

- Goal: Minimize loss under target distribution.
  - $\mathbf{E}_{(\mathbf{x},y)\sim p(\mathbf{x},y|t)}[\ell(f(\mathbf{x}),y)]$

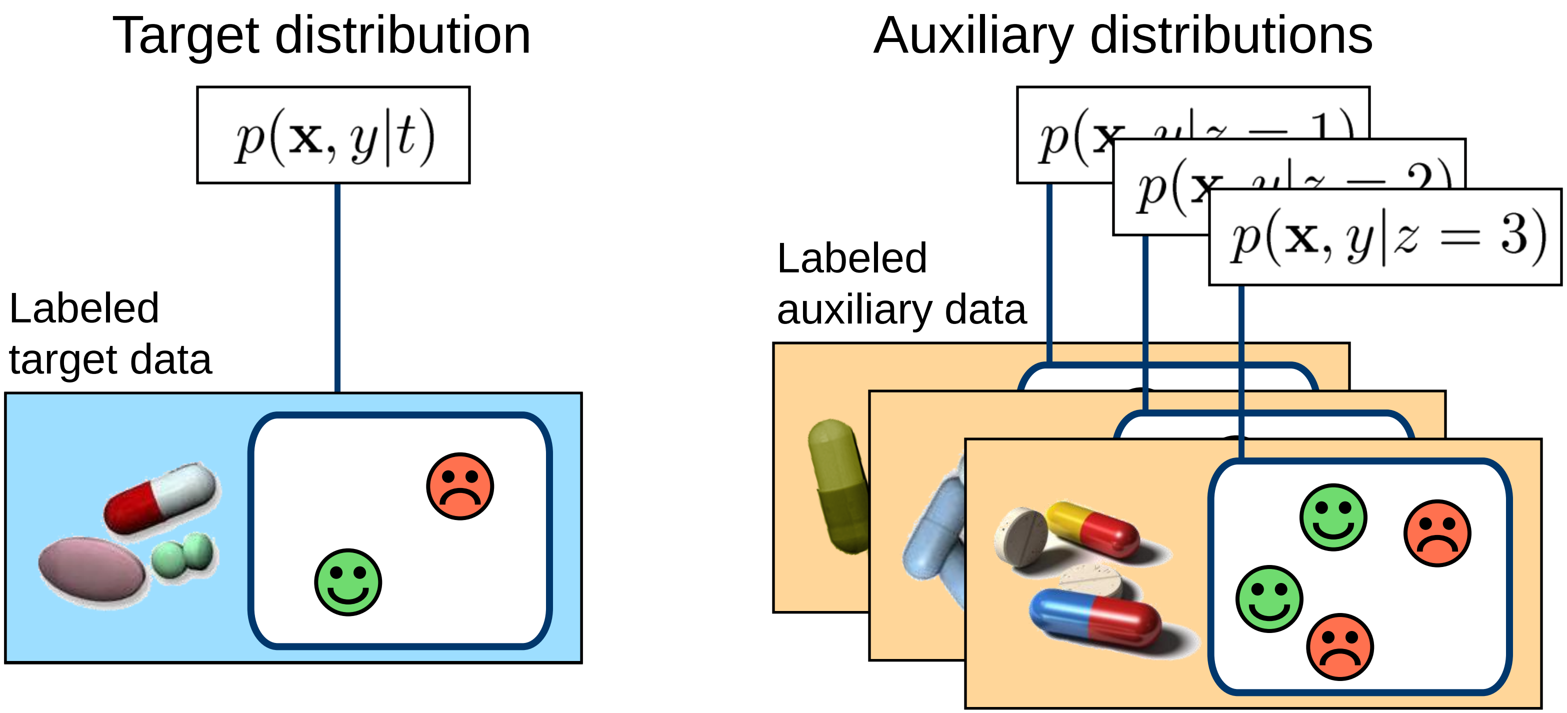


Problem Setting: Multi-Task Learning

Bickel, Bogojeska, Lengauer, Scheffer

Multi-Task Learning – Problem Setting

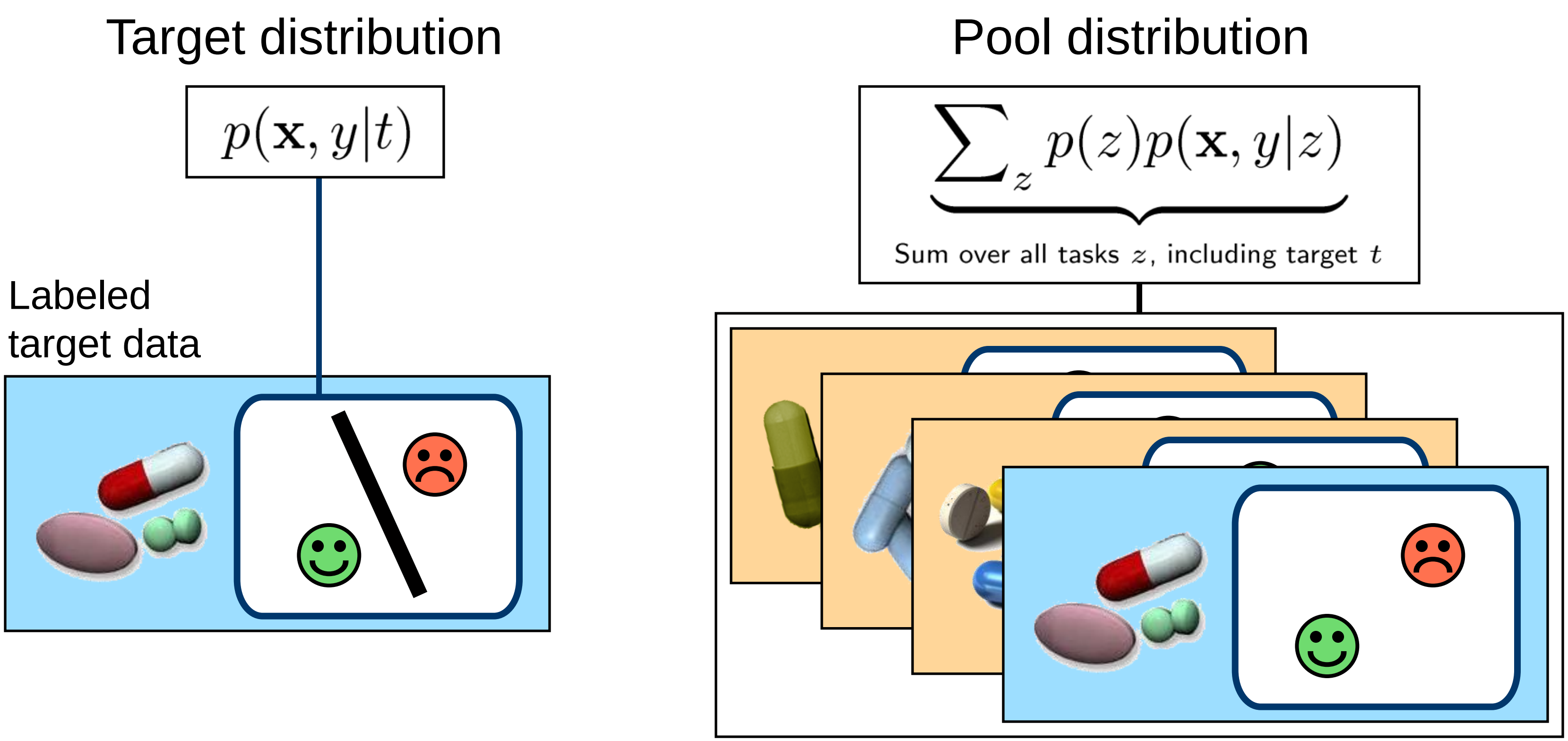
- Goal: Minimize loss under target distribution.
  - $\mathbf{E}_{(\mathbf{x},y)\sim p(\mathbf{x},y|t)}[\ell(f(\mathbf{x}),y)]$



Multi-Task Learning

- Goal: Minimize loss under target distribution.

$$\mathbf{E}_{(\mathbf{x},y)\sim\mathrm{Target}}[\ell(f(\mathbf{x}),y)] \neq \mathbf{E}_{(\mathbf{x},y)\sim\mathrm{Pool}}[\ell(f(\mathbf{x}),y)]$$

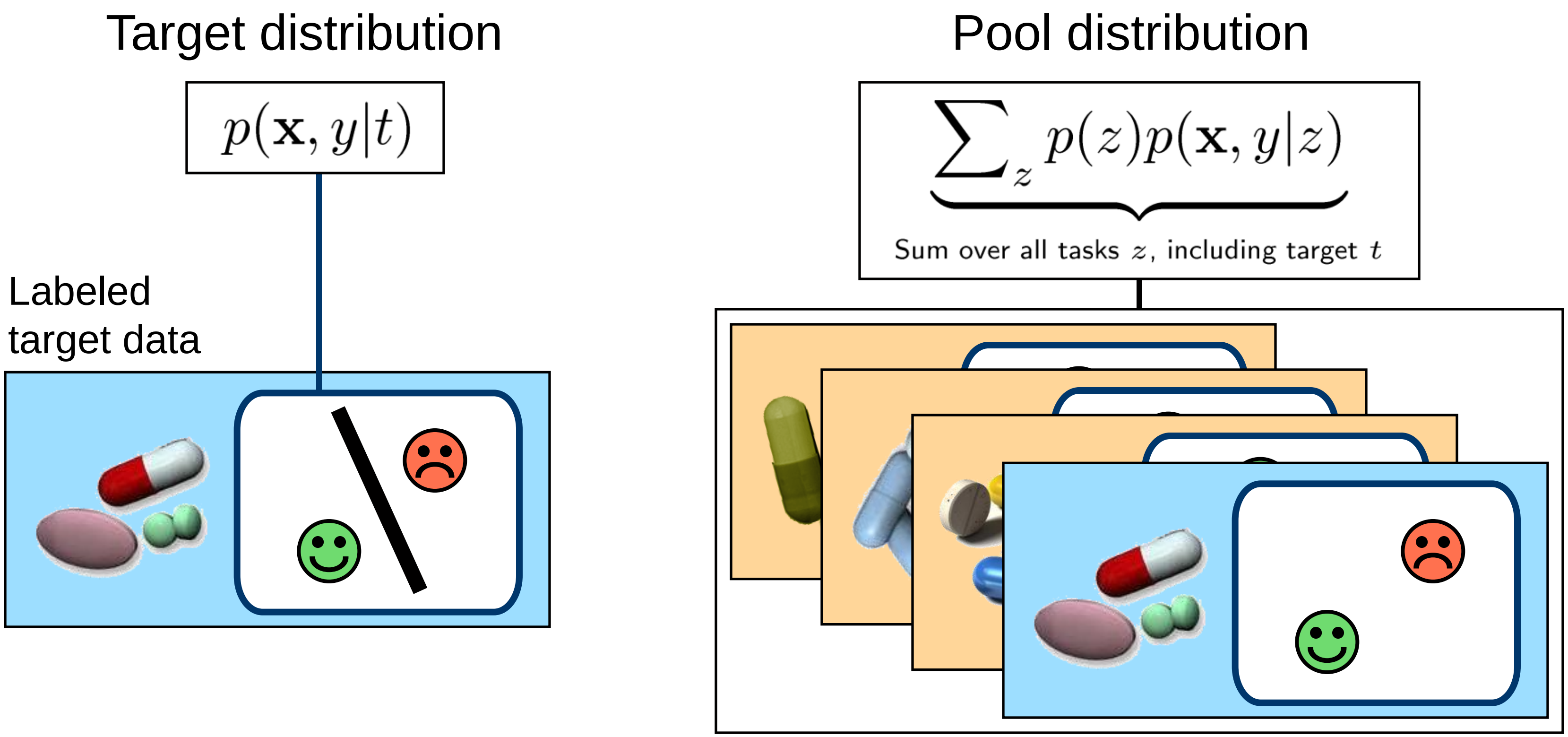


Bickel, Bogojeska, Lengauer, Scheffer

Distribution Matching

- Goal: Minimize loss under target distribution.

$$\mathbf{E}_{(\mathbf{x},y)\sim\text{Target}}[\ell(f(\mathbf{x}),y)] = \mathbf{E}_{(\mathbf{x},y)\sim\text{Pool}}[r_t(\mathbf{x},y)\ell(f(\mathbf{x}),y)]$$



Bickel, Bogojeska, Lengauer, Scheffer

Distribution Matching

- Goal: Minimize loss under target distribution.

$$\mathbf{E}_{(\mathbf{x},y)\sim\mathrm{Target}}[\ell(f(\mathbf{x}),y)] = \mathbf{E}_{(\mathbf{x},y)\sim\mathrm{Pool}}[r_t(\mathbf{x},y)\ell(f(\mathbf{x}),y)]$$

Expected loss under target distribution

Expectation over training pool

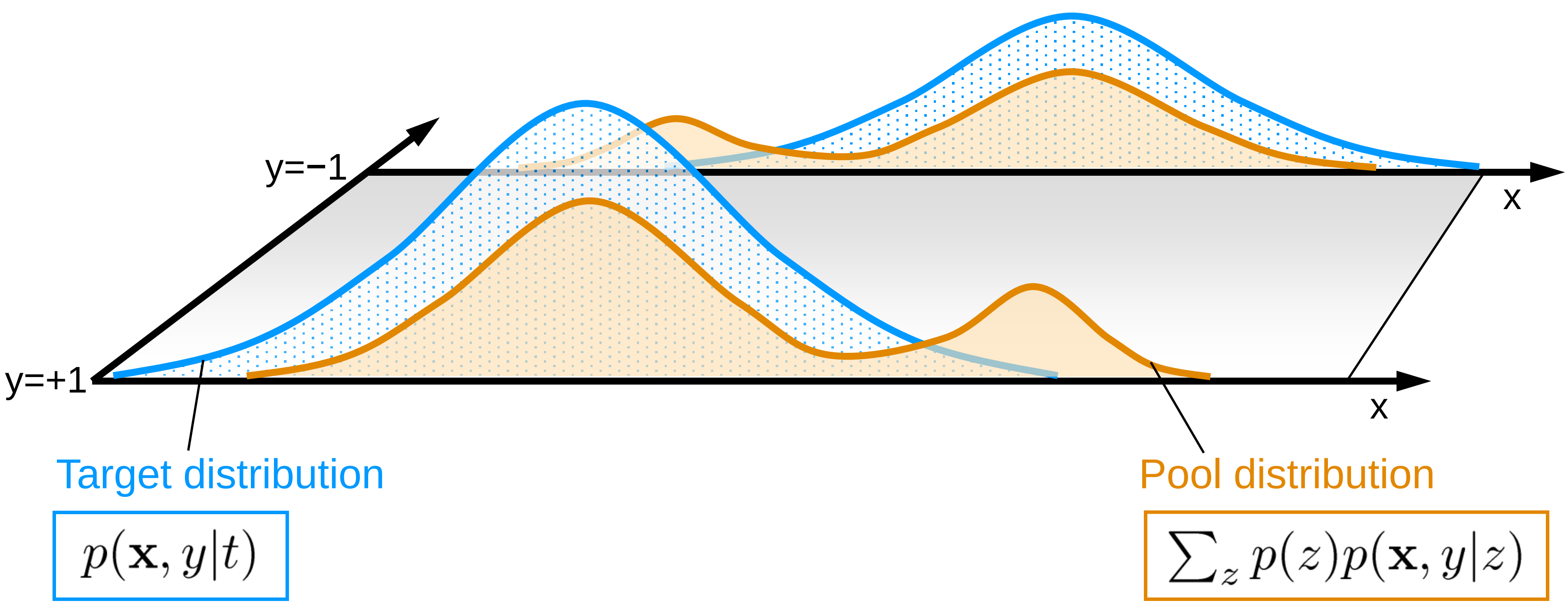
Rescale loss for each pool example

Bickel, Bogojeska, Lengauer, Scheffer

Distribution Matching

- Goal: Minimize loss under target distribution.

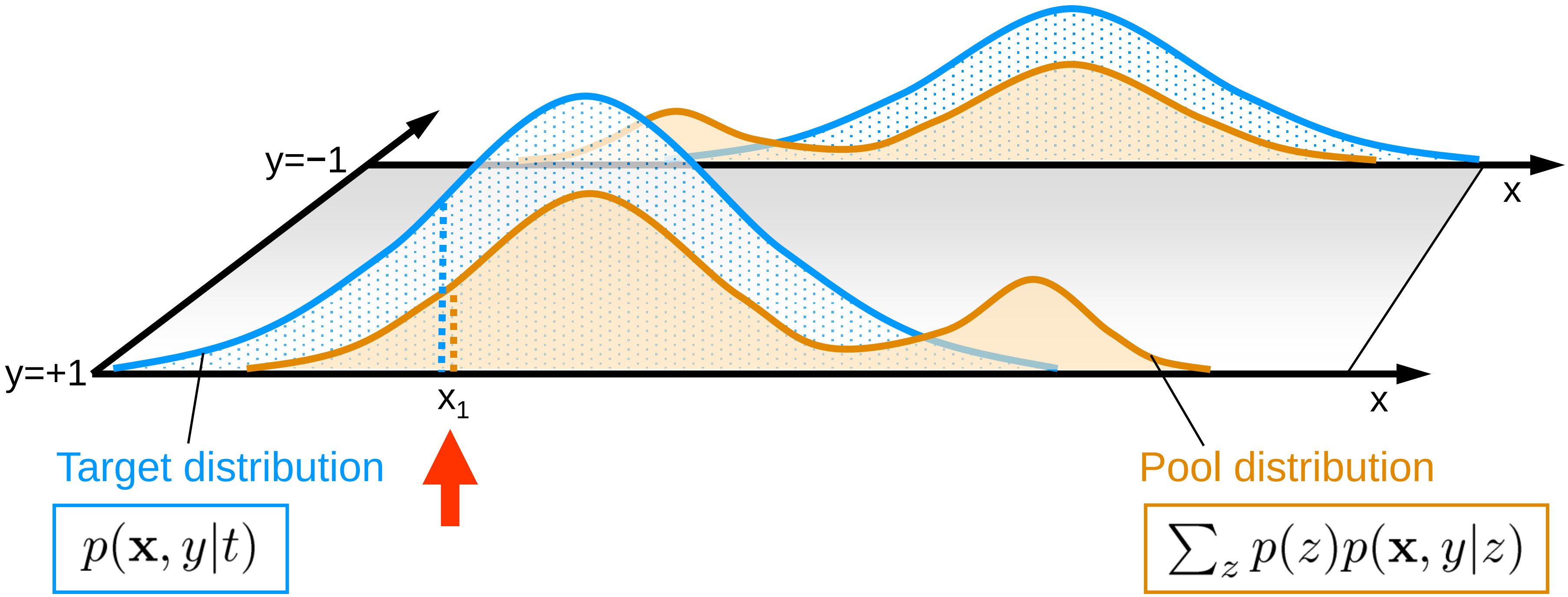
$$\mathbf{E}_{(\mathbf{x},y)\sim\text{Target}}[\ell(f(\mathbf{x}),y)] = \mathbf{E}_{(\mathbf{x},y)\sim\text{Pool}}\left[\quad r_t(\mathbf{x},y) \quad \ell(f(\mathbf{x}),y)\right]$$



Distribution Matching

- Goal: Minimize loss under target distribution.

$$\mathbf{E}_{(\mathbf{x},y)\sim\text{Target}}[\ell(f(\mathbf{x}),y)] = \mathbf{E}_{(\mathbf{x},y)\sim\text{Pool}}\left[\quad 2 \quad \ell(f(\mathbf{x}),y)\right]$$

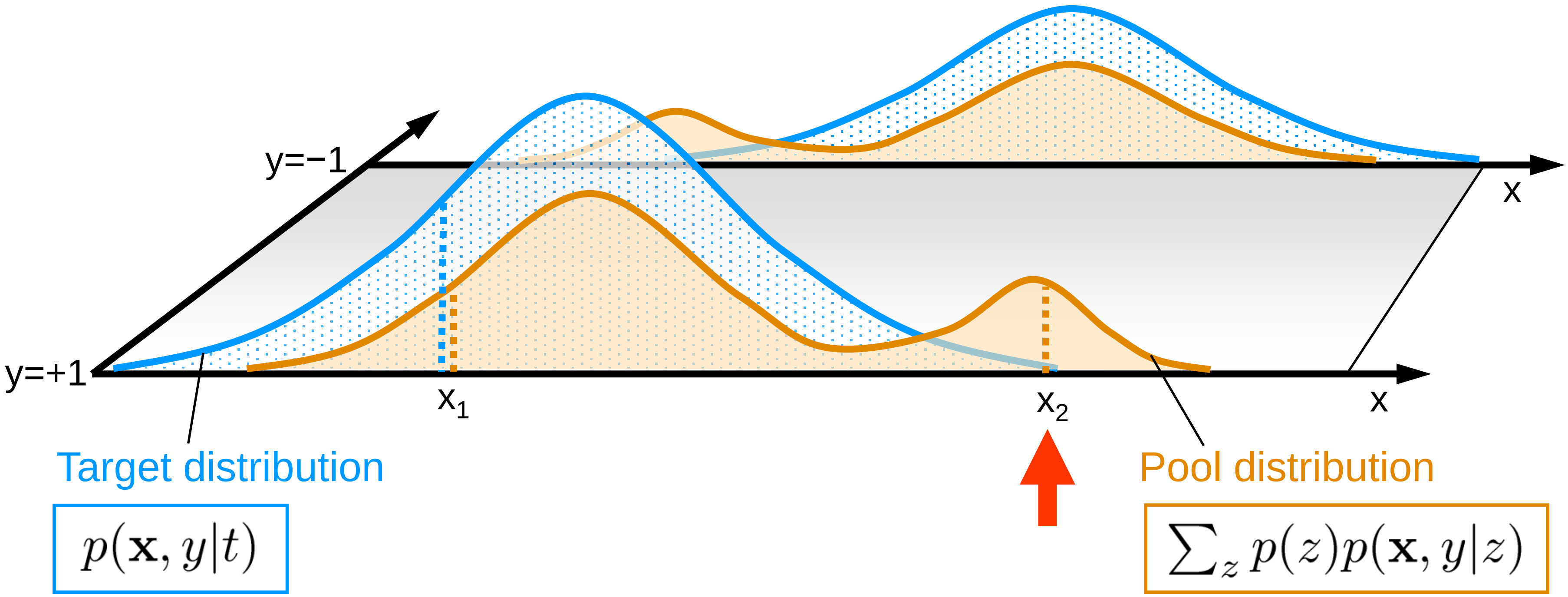


Bickel, Bogojeska, Lengauer, Scheffer

Distribution Matching

- Goal: Minimize loss under target distribution.

$$\mathbf{E}_{(\mathbf{x},y)\sim\mathrm{Target}}[\ell(f(\mathbf{x}),y)] = \mathbf{E}_{(\mathbf{x},y)\sim\mathrm{Pool}}\left[\quad \mathbf{0} \quad \ell(f(\mathbf{x}),y)\right]$$



Estimation of Density Ratio

- Goal: Minimize loss under target distribution.

$$\mathbf{E}_{(\mathbf{x},y)\sim\text{Target}}[\ell(f(\mathbf{x}),y)] = \mathbf{E}_{(\mathbf{x},y)\sim\text{Pool}}\left[\frac{p(\mathbf{x},y|t)}{\sum_z p(z)p(\mathbf{x},y|z)}\ell(f(\mathbf{x}),y)\right]$$

Estimation of Density Ratio

- Goal: Minimize loss under target distribution.

$$\mathbf{E}_{(\mathbf{x},y)\sim\text{Target}}[\ell(f(\mathbf{x}),y)] = \mathbf{E}_{(\mathbf{x},y)\sim\text{Pool}}\left[\frac{p(\mathbf{x},y|t)}{\sum_z p(z)p(\mathbf{x},y|z)}\ell(f(\mathbf{x}),y)\right]$$

- Theorem:

$$\frac{p(\mathbf{x},y|t)}{\sum_z p(z)p(\mathbf{x},y|z)} = \frac{p(t|\mathbf{x},y)}{p(t)}$$

Potentially high-dimensional densities

One binary conditional density

Bickel, Bogojeska, Lengauer, Scheffer

Estimation of Density Ratio

- Goal: Minimize loss under target distribution.

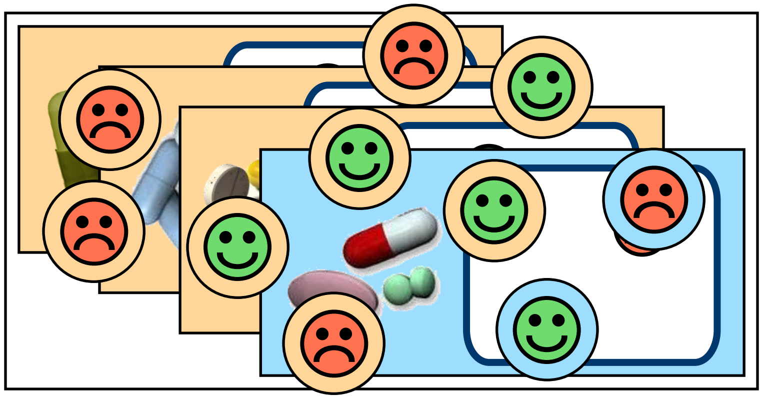
$$\mathbf{E}_{(\mathbf{x},y)\sim\mathrm{Target}}[\ell(f(\mathbf{x}),y)] = \mathbf{E}_{(\mathbf{x},y)\sim\mathrm{Pool}}\left[\frac{p(\mathbf{x},y|t)}{\sum_z p(z)p(\mathbf{x},y|z)}\ell(f(\mathbf{x}),y)\right]$$

- Theorem:

$$\frac{p(\mathbf{x},y|t)}{\sum_z p(z)p(\mathbf{x},y|z)} = \frac{p(t|\mathbf{x},y)}{p(t)}$$

- Intuition of $p(t|\mathbf{x},y)$: how much more likely is $(\mathbf{x},y)$ to be drawn from target than from auxiliary density.

Pool



Bickel, Bogojeska, Lengauer, Scheffer

Estimation of Density Ratio

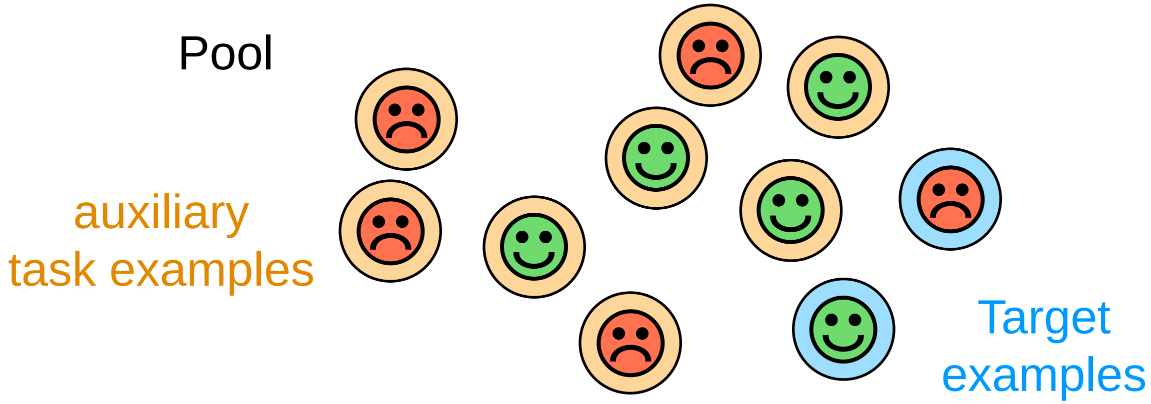
- Goal: Minimize loss under target distribution.

$$\mathbf{E}_{(\mathbf{x},y)\sim\text{Target}}[\ell(f(\mathbf{x}),y)] = \mathbf{E}_{(\mathbf{x},y)\sim\text{Pool}}\left[\frac{p(\mathbf{x},y|t)}{\sum_z p(z)p(\mathbf{x},y|z)}\ell(f(\mathbf{x}),y)\right]$$

- Theorem:

$$\frac{p(\mathbf{x},y|t)}{\sum_z p(z)p(\mathbf{x},y|z)} = \frac{p(t|\mathbf{x},y)}{p(t)}$$

- Intuition of $p(t|\mathbf{x},y)$: how much more likely is $(\mathbf{x},y)$ to be drawn from target than from auxiliary density.



Bickel, Bogojeska, Lengauer, Scheffer

Estimation of Density Ratio

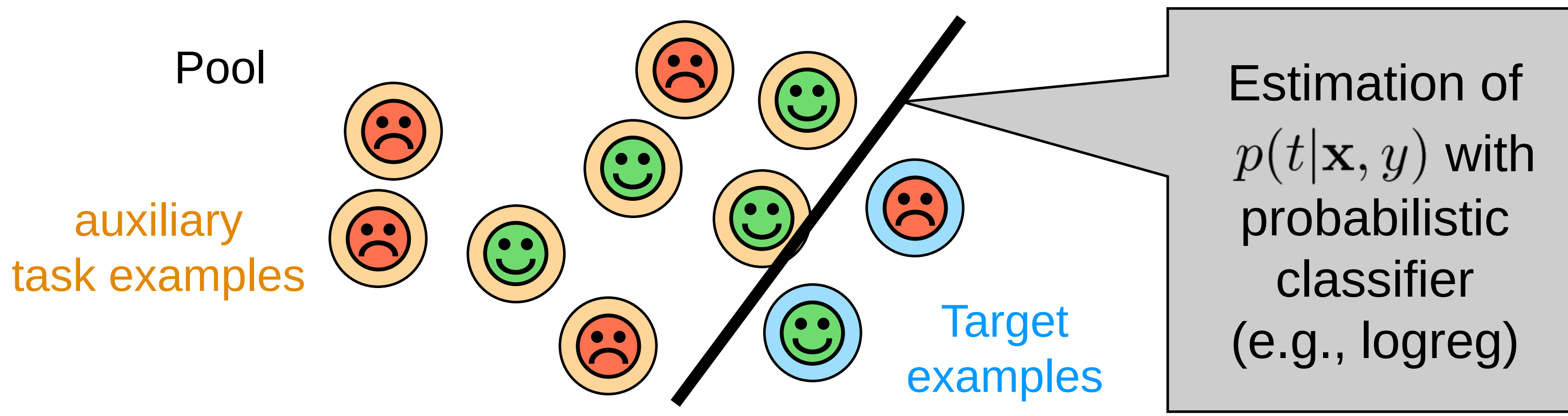
- Goal: Minimize loss under target distribution.

$$\mathbf{E}_{(\mathbf{x},y)\sim\text{Target}}[\ell(f(\mathbf{x}),y)] = \mathbf{E}_{(\mathbf{x},y)\sim\text{Pool}}\left[\frac{p(\mathbf{x},y|t)}{\sum_z p(z)p(\mathbf{x},y|z)}\ell(f(\mathbf{x}),y)\right]$$

- Theorem:

$$\frac{p(\mathbf{x},y|t)}{\sum_z p(z)p(\mathbf{x},y|z)} = \frac{p(t|\mathbf{x},y)}{p(t)}$$

- Intuition of $p(t|\mathbf{x},y)$: how much more likely is $(\mathbf{x},y)$ to be drawn from target than from auxiliary density.



Bickel, Bogojeska, Lengauer, Scheffer

Estimation of Density Ratio

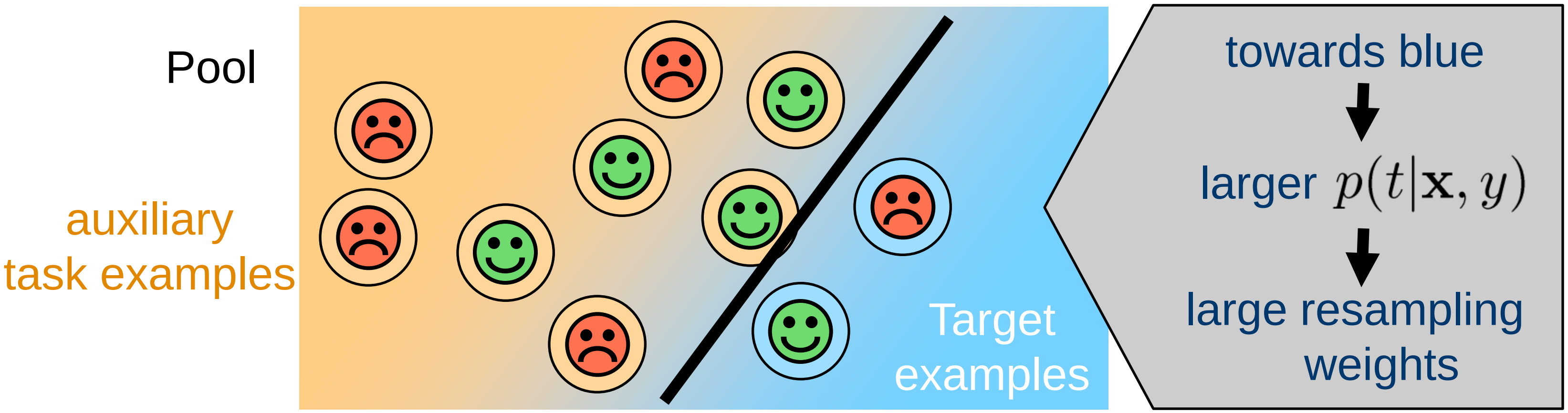
- Goal: Minimize loss under target distribution.

$$\mathbf{E}_{(\mathbf{x},y)\sim\text{Target}}[\ell(f(\mathbf{x}),y)] = \mathbf{E}_{(\mathbf{x},y)\sim\text{Pool}}\left[\frac{p(\mathbf{x},y|t)}{\sum_z p(z)p(\mathbf{x},y|z)}\ell(f(\mathbf{x}),y)\right]$$

- Theorem:

$$\frac{p(\mathbf{x},y|t)}{\sum_z p(z)p(\mathbf{x},y|z)} = \frac{p(t|\mathbf{x},y)}{p(t)}$$

- Intuition of $p(t|\mathbf{x},y)$: how much more likely is $(\mathbf{x},y)$ to be drawn from target than from auxiliary density.



Bickel, Bogojeska, Lengauer, Scheffer

Prior Knowledge on Task Similarity

- Prior knowledge in task similarity kernel $k(z, z')$.
- Encoding of prior knowledge in Gaussian prior

$$\mathbf{v} \sim N(0, \Sigma)$$

  on parameters $\mathbf{v}$ of a multi-class logistic regression model for the resampling weights.

- Main diagonal entries of Σ set to $\sigma_{\mathbf{v}}^2$ (standard regularizer),
- Diagonals of sub-matrices set to $k(z, z')\rho\sigma_{\mathbf{v}}^2$.

Distribution Matching – Algorithm

1. Weight Model:
   Train Logreg of target vs. auxiliary data with task similarity in Σ.

$$\text{Over } \mathbf{v}, \text{ maximize} \sum_{(\mathbf{x}_i, y_i, z_i) \in \text{Pool}} \log(p(z_i | \mathbf{x}_i, y_i, \mathbf{v})) - \mathbf{v}^\mathsf{T} \Sigma^{-1} \mathbf{v}$$

2. Target Model:
   Minimize regularized empirical loss on pool weighted by $\frac{p(t|\mathbf{x}_i, y_i, \mathbf{v})}{p(t)}$.

$$\text{For task } t, \text{ over } \mathbf{w}_t, \text{ minimize} \sum_{(\mathbf{x}_i, y_i) \in \text{Pool}} \frac{p(t|\mathbf{x}_i, y_i, \mathbf{v})}{p(t)} \ell(f(\mathbf{x}_i, \mathbf{w}_t), y_i) + \frac{\mathbf{w}_t^\mathsf{T} \mathbf{w}_t}{2\sigma_{\mathbf{w}}^2}$$

Result of step 1:
weight model

Overview

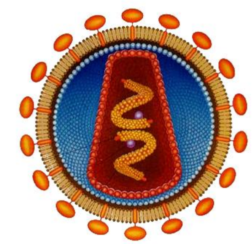
- Motivation.
  - HIV therapy screening.
  - Multiple tasks with differing distributions.

- Multi-task learning by distribution matching.
  - Problem Setting.
  - Density ratio matches pool to target distribution.
  - Discriminative estimation of matching weights.

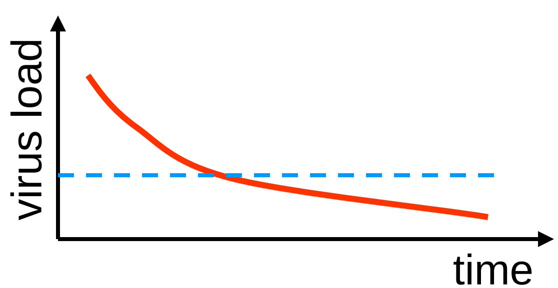
- Case study:
  - HIV therapy screening.

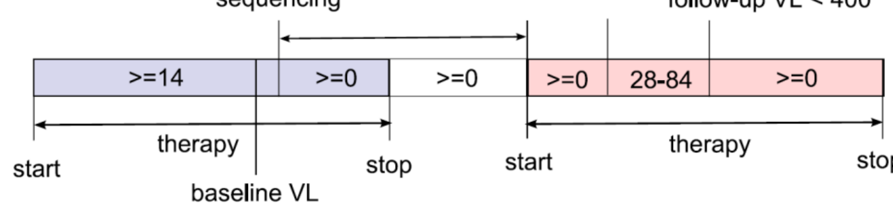
HIV Therapy Screening – Prediction Problem

- Information about each patient x, binary vector
  - of resistance-relevant virus mutations and
  - of previously given drugs.



- Drug combination selected out of 17 drugs.
  - Drug combinations correspond to tasks z.
- Target label y (success or failure of therapy).
  - 2 different labelings (virus load and multi-conditional).





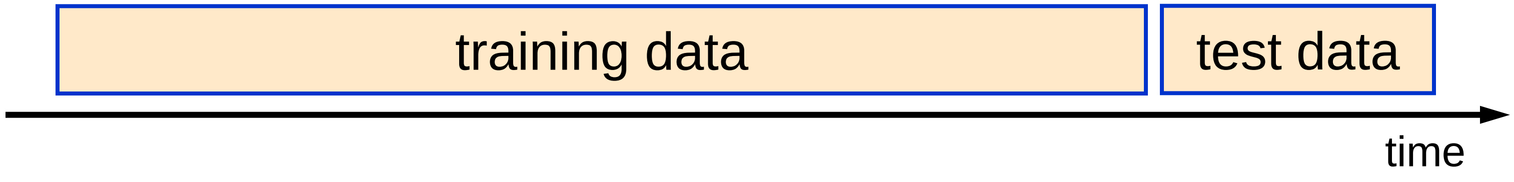
conditions

Bickel, Bogojeska, Lengauer, Scheffer

HIV Therapy Screening – Data

- Patients from hospitals in Italy, Germany, and Sweden.
  - 3260 labeled treatments.
  - 545 different drug combinations (tasks).
  - 50% of combinations with only one labeled treatment.

- Similarity of drug combinations: task kernel.
  - *Drug feature kernel*: product of drug indicator vectors.
  - *Mutation table kernel:* similarity of mutations that render drug ineffective.
- 80/20 training/test split, consistent with time stamps.



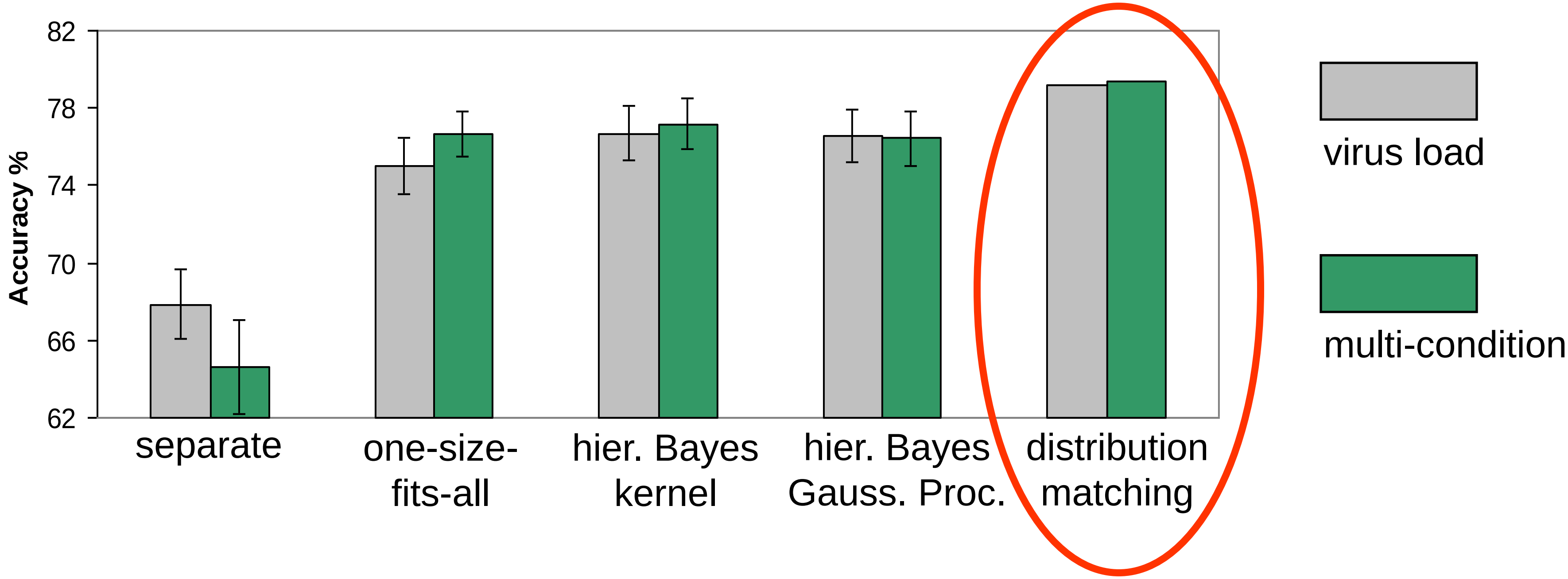
Bickel, Bogojeska, Lengauer, Scheffer

Reference Methods

- Independent models (separately trained).
- One-size-fits-all, product of task and feature kernel,
  - Bonilla, Agakov, and Williams (2007).
- Hierarchical Bayesian Kernel,
  - Evgeniou & Pontil (2004).
- Hierarchical Bayesian Gaussian Process
  - Yu, Tresp, and Schwaighofer (2005).

- Logistic regression is target model (except for Gaussian process model).
- RBF kernels.

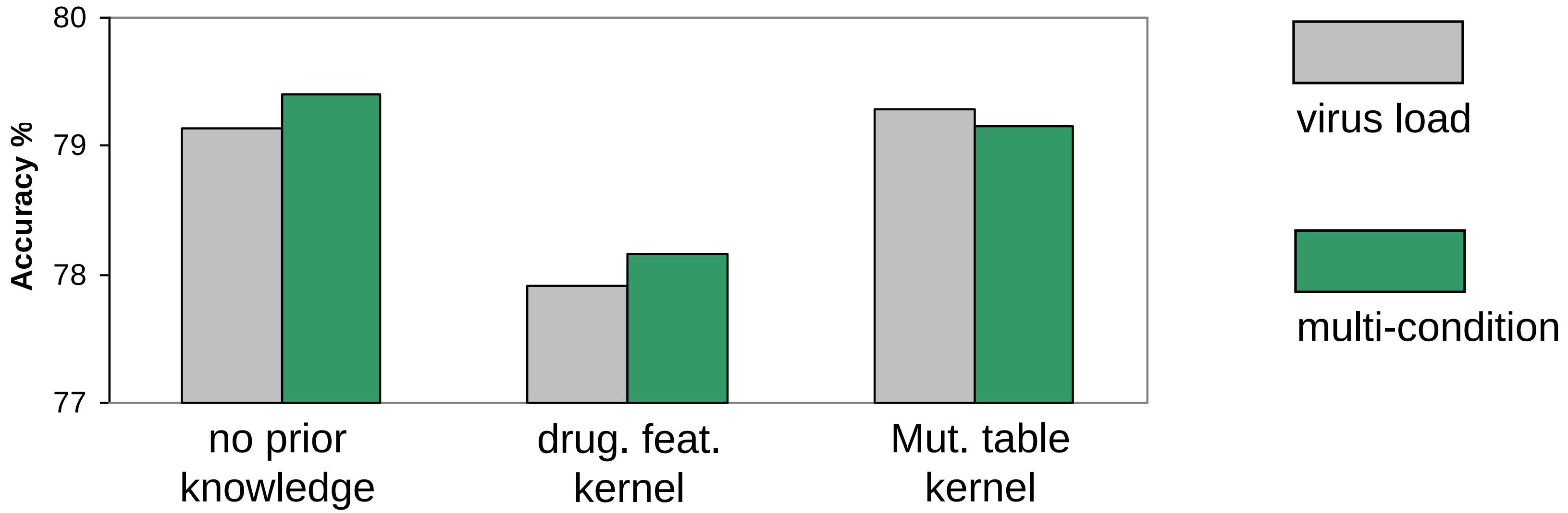
Results – Distribution Matching vs. Other



- Distribution matching always best (17 of 20 cases stat. significant) or as good as best reference method.
- Improvement over separately trained models 10-14%.

Bickel, Bogojeska, Lengauer, Scheffer

Results – Benefit of Prior Knowledge



- The common prior knowledge on similarity of drug combinations does not improve accuracy of distribution matching.

Conclusions

- Multi-task Learning:
  - Multiple problems with different distributions.
- Distribution matching:
  - Weighted pool distribution matches target distribution.
  - Discriminative estimation of weights with Logreg.
  - Training of target model with weighted loss terms.
- Case study: HIV therapy screening.
  - Distribution matching beats *iid* learning and hier. Bayes.
  - Benefit over separately trained models 10-14%.