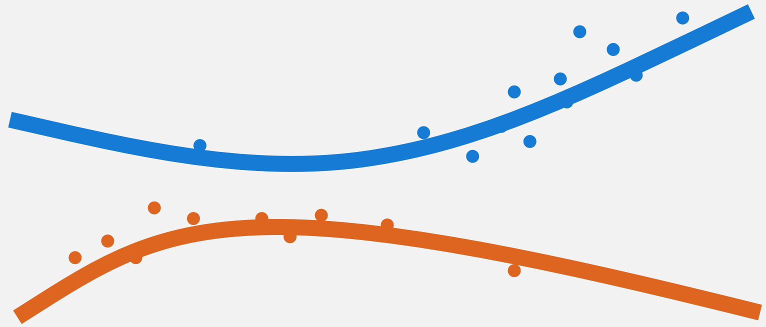


Potential outcomes, counterfactuals & conditional treatment effects, Pt. II

Causal Inference & Deep Learning

MIT IAP 2018

Fredrik D. Johansson

1. Potential outcomes framework

2. Supervised learning (risk minimization)

3. **Adjusting for distributional shift**

Recap: Estimating potential outcomes

- We consider learning hypotheses $f(x, t) \approx \mathbb{E}[\,Y \mid x, t\,]$
- Under ignorability, $\mathbb{E}[\,Y \mid x, t\,] = \mathbb{E}[Y(t) \mid x]$
- Want to do **risk minimization for** each outcome

$$\boxed{f(\cdot, 0) = \operatorname*{argmin}_{h_0} R_p^0(h_0)\,, \qquad R_p^0(h_0) := \mathbb{E}_p\left[\big(h_0(x) - Y(0)\big)^2\right]}$$

$$f(\cdot, 1) = \operatorname*{argmin}_{h_1} R_p^1(h_1)\,, \qquad R_p^1(h_1) := \mathbb{E}_p\left[\big(h_1(x) - Y(1)\big)^2\right]$$

Estimating potential outcomes

- We don't have samples of $\mathbb{E}_p\left[\left(h_0(x) - Y(0)\right)^2\right]$ from $p(X, T)$, only from $p(X \mid T = 0)$! *We only see* ***Y(0) for controls, t=0***

- No guarantee that

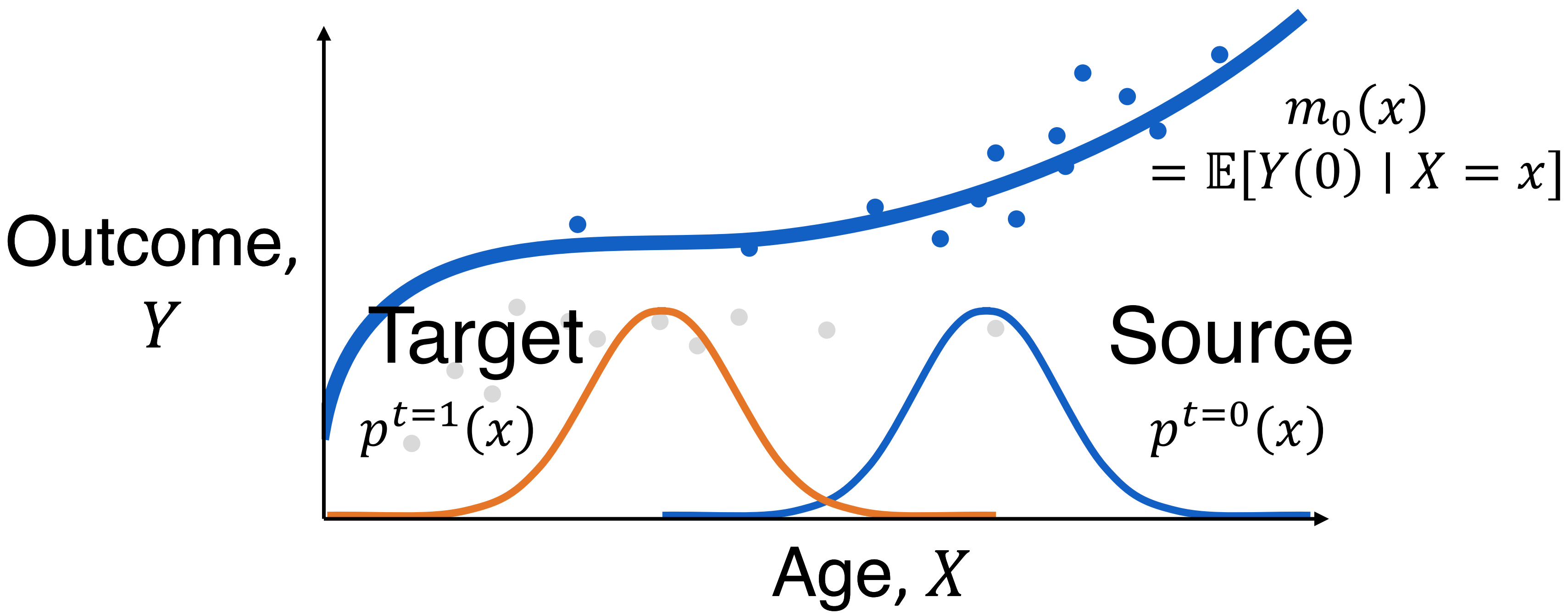
?

$$\mathbb{E}_p\left[\left(h_0(x) - Y(0)\right)^2\right] \approx \frac{1}{n} \sum_{i:t_i=0} (h_0(x_i) - y_i)^2$$

Only control group

Counterfactuals

- We can think of this as learning from one distribution (factual), and predicting in another (counterfactual)



Unsupervised domain adaptation

- This is exactly the setting of unsupervised domain adaptation

Labeled **Unlabeled**

- Observe $(x_1, y_1), \dots, (x_n, y_n) \sim \boldsymbol{p(x)}$ and $(x'_1), \dots, (x'_m) \sim \boldsymbol{q(x)}$

- Predict for $x' \sim \boldsymbol{q(x)}$

- Covariate-shift assumption: $q(Y \mid X) = p(Y \mid X)$

  Here, this follows from ignorability: $Y(0) \perp\!\!\!\perp T \mid X$

Bound idea

Mentioned yesterday!

1. Show that $R_p^\tau(f)$ is bounded by outcome risks $R_p^0(f)$, $R_p^1(f)$
2. Show that $R_p^t(f)$ decomposes to factual and counterfactual components $R_p^t(f) = R_F^t(f) + R_{CF}^t(f)$
3. Show that counterfactual loss is bounded by observable quantities and an **Integral Probability Metric** (IPM)
4. $\frac{1}{2} R_p^\tau(f) \leq R_F^0(f) + R_F^1(f) + \mathrm{IPM}_G\left(p^{t=0}(x), p^{t=1}(x)\right) - 2\sigma^2$

CATE and potential outcome MSE

- We can relate the CATE MSE to potential outcome losses

$$R_p^\tau(f) = \int_x \left(\hat{\tau}_f(x) - \tau(x)\right)^2 p(x)dx$$

$$= \int_x (f(x,1) - f(x,0) - \tau(x))^2 p(x)dx$$

$$\leq 2\int_x \left[(f(x,0) - m_0(x))^2 + \left(f(x,1) - m_1(x)\right)^2\right] p(x)dx$$

Unobserved!

- We let $\ell_f(x,t) = (f(x,t) - Y(t))^2$ be the loss at (x,t) (an RV)

CATE and potential outcome MSE

- We can now relate the CATE MSE to potential outcome losses

$$R_p^\tau(f) \leq 2\int_x \left[\big(f(x,0)-m_0(x)\big)^2 + \big(f(x,1)-m_1(x)\big)^2\right] p(x)dx$$

$$= 2\int_{x,y} \ell_f(x,0)p(x,y)dxdy + 2\int_{x,y} \ell_f(x,1)p(x,y)dxdy - \underbrace{4\sigma^2}_{\text{Variance in } Y}$$

$$=: \; 2\Big(\underbrace{R_p^0(f)}_{\text{Risk in predicting } Y(0)} + \underbrace{R_p^1(f)}_{\text{Risk in predicting } Y(1)}\Big) - 4\sigma^2$$

- We assume that $Y(t)$ are deterministic, going forward, for simplicity

Factual and counterfactual

- Each potential outcome loss decomposes by treatment group

$$R_p^0(f) = \int_x \ell_f(x,0)p(x)dx$$

$$= p(T=0)\int_x \ell_f(x,0)\underbrace{p(x \mid T=0)}_{\text{Control group}}dx \quad \text{— Factual}$$

$$+\, p(T=1)\int_x \ell_f(x,0)\underbrace{p(x \mid T=1)}_{\text{Treatment group}}dx \quad \text{— Counterfactual}$$

$$=: (1-u)\cdot R_F^0(f) + u \cdot R_{CF}^0(f)$$

- Where $u = p(T=1)$

Bounding counterfactual loss

- Let $p^{t=t'}(x) = p(X = x \mid T = t')$. Then,

$$R_{CF}^0 = \int_x \ell_f(x,0)p^{t=1}(x)dx$$

Unsupervised domain adaptation

Estimate $\mathbb{E}_q[\ell]$ based on labeled samples from p

$$\mathbb{E}_q[\ell] \leq \mathbb{E}_p[\ell] + \mathrm{IPM}(p,q)$$

gap!

$$\leq R_F^0 + \sup_{g \in G} \int_x g(x)|p^{t=1}(x) - p^{t=0}(x)|dx$$

Assumption that $\ell_f \in G$

$$\boxed{R_{CF}^0(f) \leq R_F^0(f) + \mathrm{IPM}_G\left(p^{t=0}(x), p^{t=1}(x)\right)}$$

Bringing it all together

- Add results for each potential outcome (and let Y be stochastic again)

$$\frac{1}{2}\left(R_p^{\tau} + 4\sigma^2\right) \leq R_p^0 + R_p^1$$

$$= (1-u) \cdot R_F^0 + u \cdot R_{CF}^0 + u \cdot R_F^1 + (1-u) \cdot R_{CF}^1$$

$$\leq (1-u+u)R_F^0 + (1-u+u)R_F^1$$

$$+(1-u+u)\mathrm{IPM}_G\left(p^{t=0}(x), p^{t=1}(x)\right)$$

$$\boxed{\frac{1}{2}R_p^{\tau} \leq R_F^0 + R_F^1 + \mathrm{IPM}_G\left(p^{t=0}(x), p^{t=1}(x)\right) - 2\sigma^2}$$

What do we take from this bound?

$$\frac{1}{2} R_p^{\tau}(f) \leq R_F^{0}(f) + R_F^{1}(f) + \mathrm{IPM}_G\left(p^{t=0}(x), p^{t=1}(x)\right) - 2\sigma^2$$

- We can bound an **unobservable** quantity using **observable** quantities (*under unverifiable assumptions*)
- The factual error is more **representative** of the counterfactual when treatment groups are similar (surprise! ☺)
- No immediate algorithmic framework! *How can we shrink IPM?*

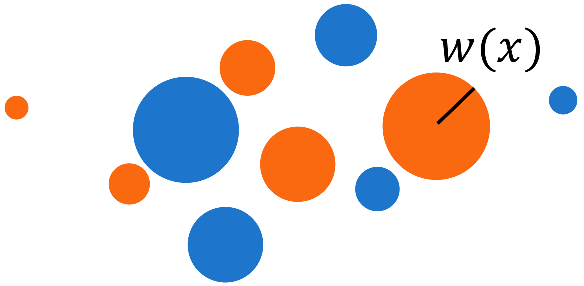
Idea 1. Sample re-weighting

- Re-weight treatment groups to adjust for bias

- **Importance sampling** principle

$$\mathbb{E}_{q(x)}[f(x)] = E_{p(x)}\left[\frac{q(x)}{p(x)} f(x)\right] \approx \frac{1}{n}\sum_{i=1}^{n} \frac{q(x_i)}{p(x_i)} f(x_i)$$

- Used to estimate e.g. average control outcome on treated population



What happens to our bound?

- For any distributions p, q, we have that $\mathrm{IPM}_G\left(p, \frac{p}{q} q\right) = 0$

- So with $w_t(x) = \frac{p(x)}{p(x|T=t)} = \frac{p(T=t)}{p(T=t|x)}$ Propensities!

  and $R^0_{p,w} = \mathbb{E}_p[w(\mathrm{x})\ell_f(x,0)]$, we have

$$\frac{1}{2} R^\tau_p(f) \leq R^0_{F,w_0}(f) + R^1_{F,w_1}(f) + \underbrace{\cancel{\mathrm{IPM}_G(w_0 p^{t=0}, w_1 p^{t=1})}}_{=0} - 2\sigma^2$$

- This is an example of inverse propensity weighting (IPW)

Problems with importance sampling

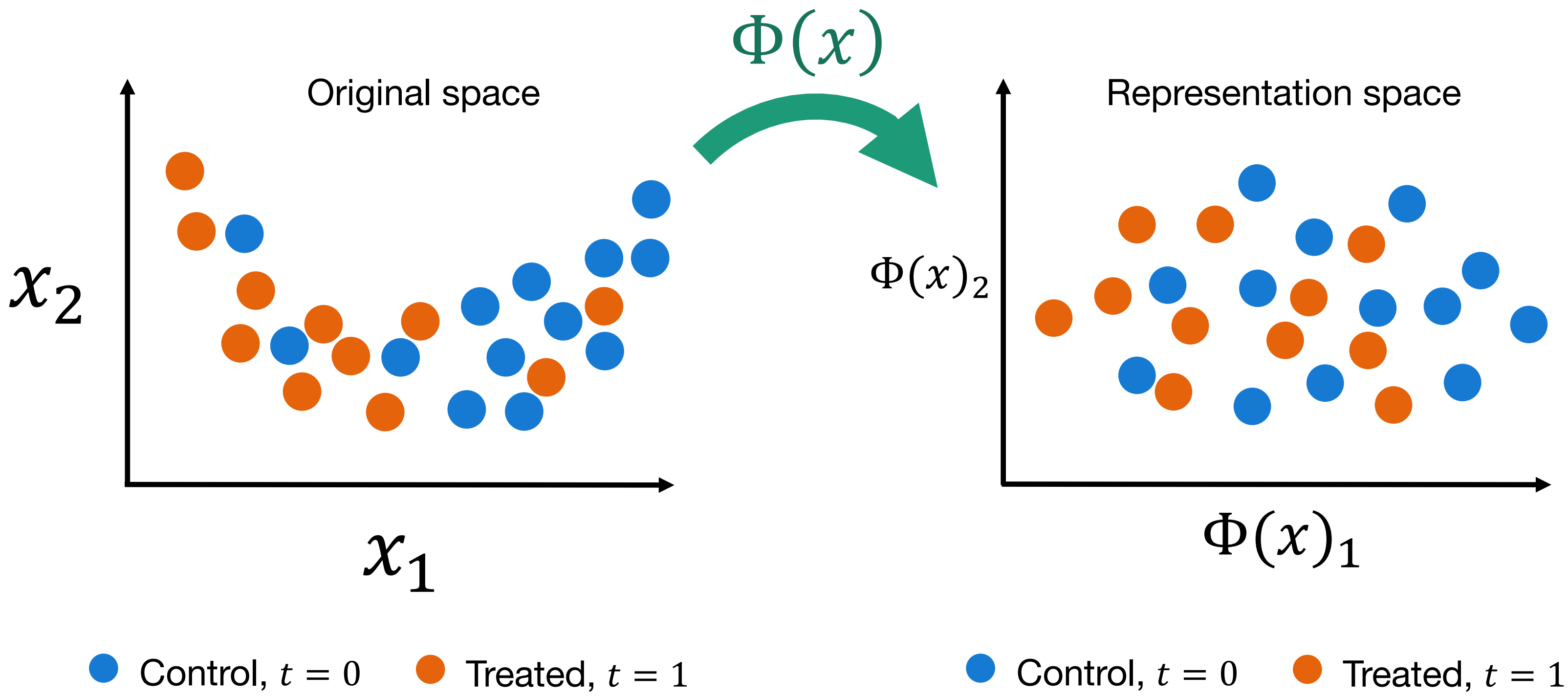
- If we knew the weighting functions w_0 and w_1

$$\frac{1}{2} R_p^{\tau}(f) \leq R_{F,w_0}^{0}(f) + R_{F,w_1}^{1}(f) - 2\sigma^2$$

- Great! However…
  - Importance sampling has severe variance problems, because of small effective sample size
  - We don't know the functions w_0 and w_1

Idea 2: Representation learning

- A shared representation helps identify meaningful interactions



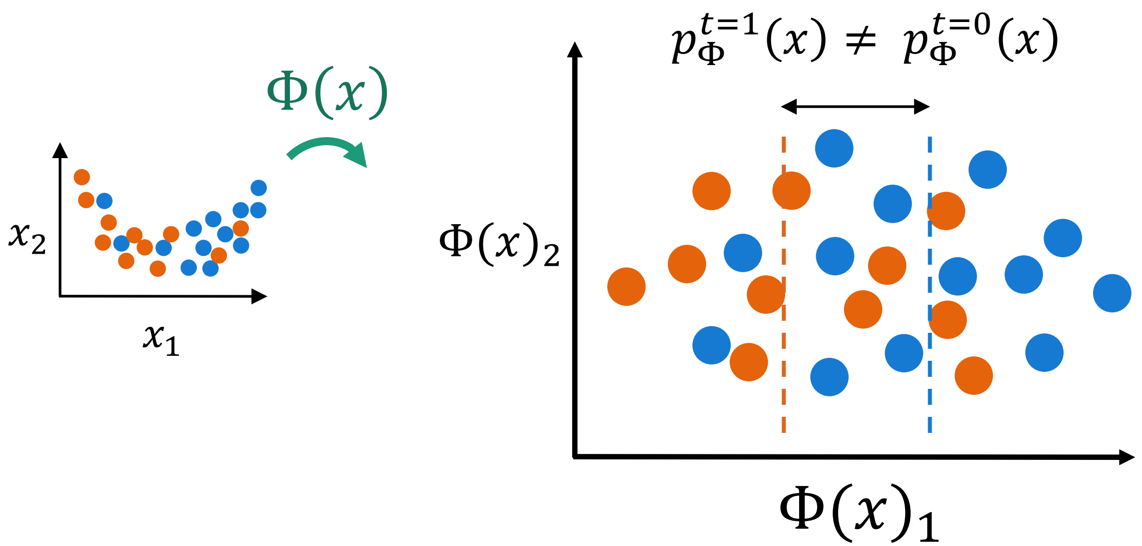
---

[1] J., Shalit, Sontag, *ICML* 2016

Imbalance in representation space

- In general, treatment groups still not representative of population

Treatment group imbalance



Control, $t = 0$ Treated, $t = 1$

---

[1]J., Shalit, Sontag, *ICML* 2016

CATE generalization bound

- **Theorem***: (Representation learning)

$$\underbrace{R_p^\tau}_{\text{Effect risk}} \leq 2 \sum_{t \in \{0,1\}} \Big(\underbrace{R_{p,w_t}^t(\Phi, h)}_{\text{Re-weighted factual loss}} + \underbrace{B_\Phi \, \mathrm{IPM}_G\big(p_\Phi^{1-t}(x), w_t \, p_\Phi^t(x)\big)}_{\text{Imbalance of re-weighted representations}} \Big)$$

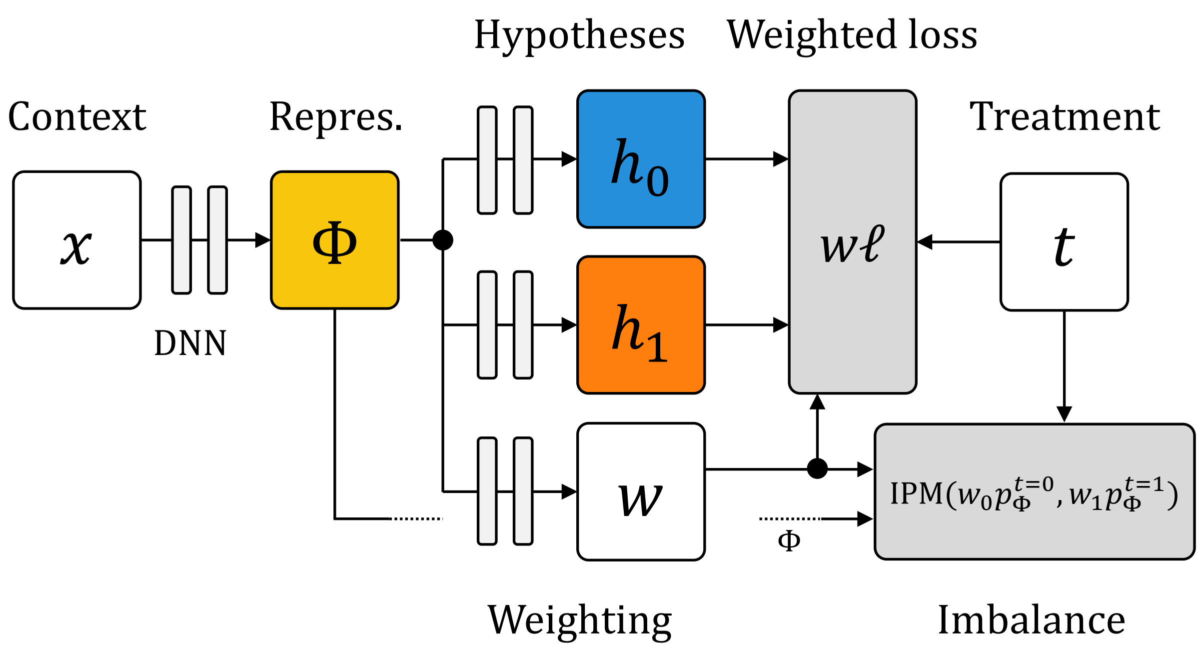
- Letting $\Phi(x) = x$, and $w_t(x)$ be inverse propensity weights, we recover the original result

- Work in submission on combining the two, and learning $w_{t(x)}$

*Extension to finite samples available

Trading off accuracy for balance

- Our full architecture learns a representation $\Phi(x)$, a re-weighting $w_t(x)$ and hypotheses $h_t(\Phi)$ to trade-off between the re-weighted loss $w\ell$ and imbalance between re-weighted representations



Evaluating CATE estimates

- **No ground truth**, similar to off-policy evaluation in reinforcement learning
- Requires either:
  - Knowledge of the true outcome (synthetic)
  - Knowledge of the logging policy (e.g. a randomized controlled trial)
- Our framework has proven effective in both settings

[1]Swaminathan & Joachims, *ICML* 2015

Empirical results: IHDP

- IHDP[1] is a widely used benchmark for causal effect estimation
- Original **randomized** study examined the effect of home-visits and high-quality child care on child cognitive test scores
- Feature set contains aspects of the child, mother, pregnancy etc
- The benchmark was made **observational** by removing all non-white mothers from the dataset.
- The outcome was synthesized based on original features and treatment

---

[1] Hill, *Journal of Computational and Graphical Statistics* 2011

Empirical results: IHDP

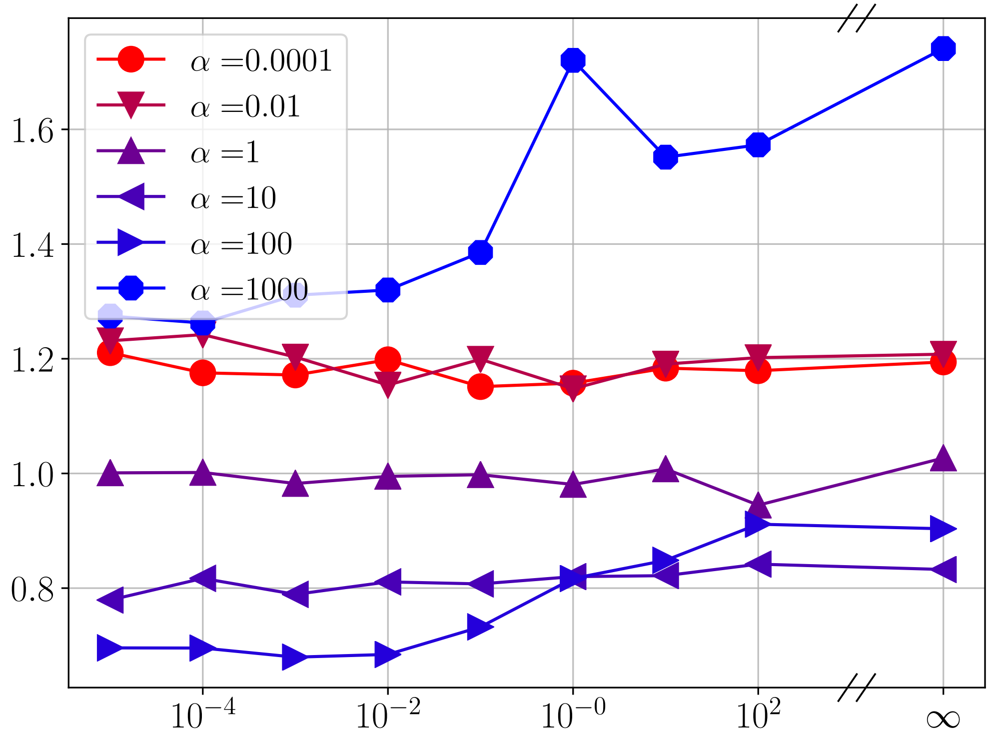
- Results on **held-out** units:

		Error in conditional effect	Error in average effect
		IHDP	
		$\sqrt{\epsilon_{\text{CATE}}}$	ϵ_{ATE}
	OLS/LR$_1$	$5.8 \pm .3$	$.94 \pm .06$
	OLS/LR$_2$	$2.5 \pm .1$	$.31 \pm .02$
	BLR	$5.8 \pm .3$	$.93 \pm .05$
	k-NN	$4.1 \pm .2$	$.79 \pm .05$
	TMLE	†	†
	BART	$2.3 \pm .1$	$.34 \pm .02$
	R.For.	$6.6 \pm .3$	$.96 \pm .06$
	C.For.	$3.8 \pm .2$	$.40 \pm .03$
Concatenating Φ and T	BNN	$2.1 \pm .1$	$.42 \pm .03$
Twin-head neural net ($\alpha = 0$)	TARNet	$.95 \pm .02$	$.28 \pm .01$
+ IPM regularization	CFR$_{\text{MMD}}$	$.78 \pm .02$	$.31 \pm .01$
	CFR$_{\text{WASS}}$	$.76 \pm .02$	$.27 \pm .01$
+ Re-weighting	RCFR	$.67 \pm .05$	—

Effect of imbalance penalty

Effect prediction error

(lower is better)



Re-weighting variance penalty

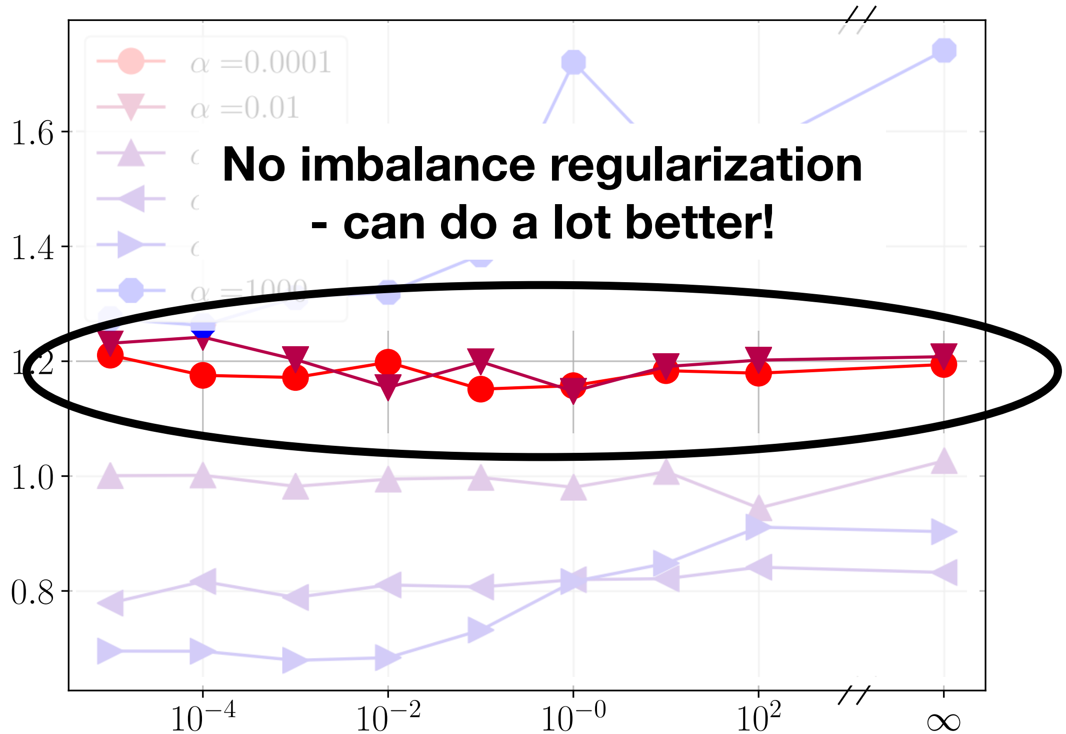
Color indicates imbalance penalty

Blue = High
Red = Low

---

[1]Swaminathan & Joachims, *ICML* 2015

Effect of imbalance penalty



Effect prediction error

(lower is better)

Re-weighting variance penalty

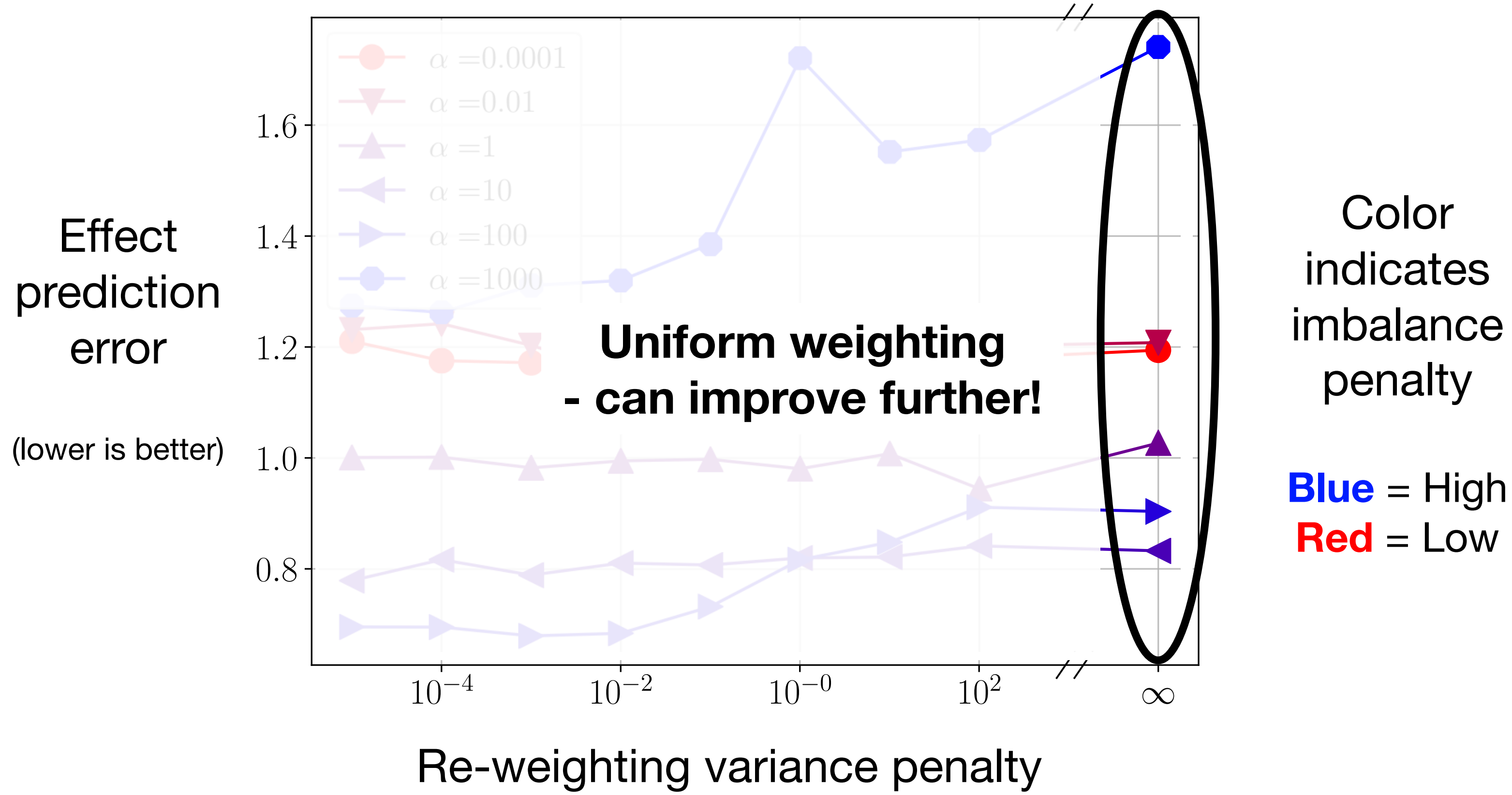
Color indicates imbalance penalty

Blue = High
Red = Low

---

[1]Swaminathan & Joachims, *ICML* 2015

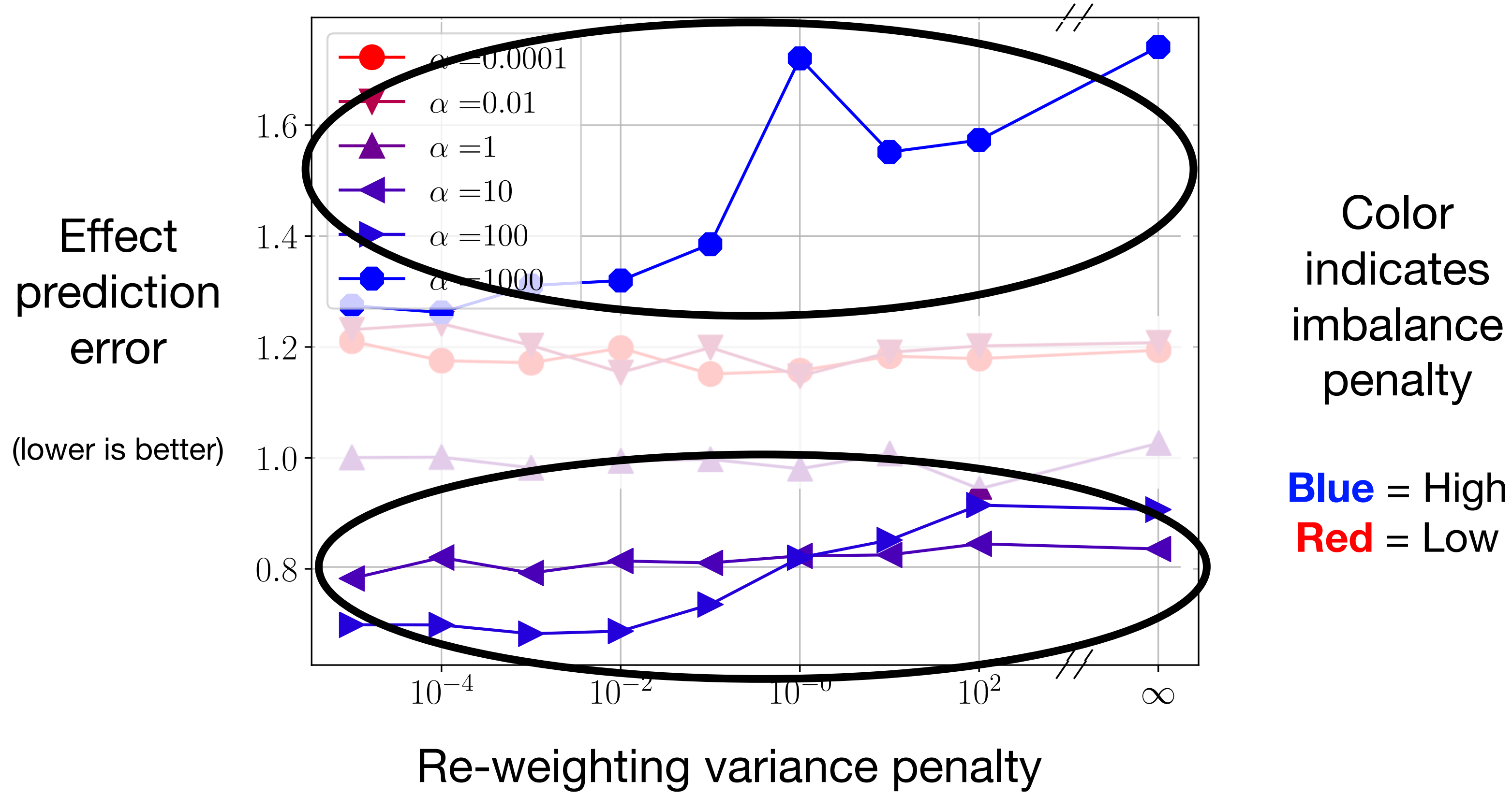
Effect of imbalance penalty



[1]Swaminathan & Joachims, *ICML* 2015

Effect of imbalance penalty

Re-weighting matters when we regularize imbalance!



---

[1]Swaminathan & Joachims, *ICML* 2015

Conclusions

- Causal inference involve interesting estimation problems
- Deep learning alone can help with solving them
- Empirical risk does not bound counterfactual risk
- Domain adaptation inspires “solutions” and theory
- Many open questions!

Counterfactuals

- Tabular records of patients

Age	**Treatment**	**Outcome** (Observed)	**Counterfactual** (Unobserved)	**Effect**
26	B	High	High	0
24	A	Low	High	1
…	…	…	…	…

Observe **Predict**