Stable Prediction across Unknown Environments

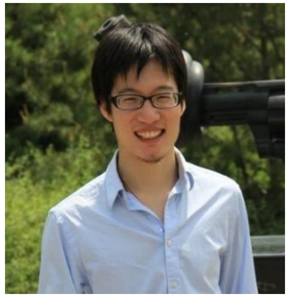
Kun Kuang
Tsinghua U

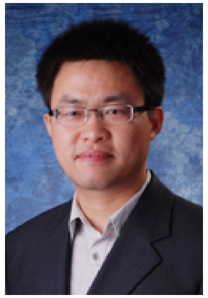
Peng Cui
Tsinghua U

Susan Athey
Stanford U

Ruoxuan Xiong
Stanford U

Bo Li
Tsinghua U

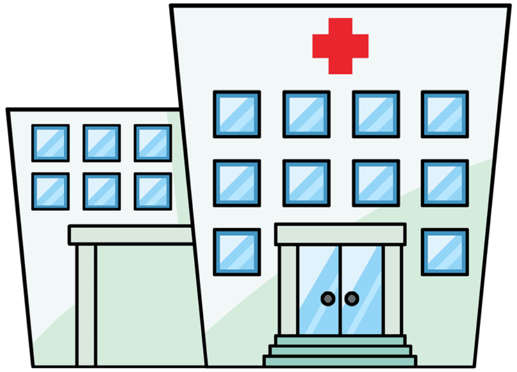
OUTLINE

Background

- Cancer survival rate prediction

Features:
- Body status
- Income
- Treatments
- Medications

Training Data



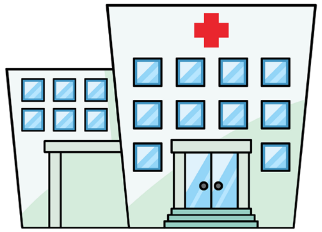
City Hospital

Higher income, higher survival rate.

Stable

Predictive Model

Unknown Testing Data



City Hospital

University Hospital

Survival rate is not so correlated with income.

- The performance of traditional predictive model is not stable

Why would a predictive model not be stable?

- Prediction / Classification
  - X: vector of features; $Y = \{0,1\}$
  - Environment: joint distribution of X and Y, denoted as $P(XY)$
- Suppose $X = \{S, V\}$, and $Y = f(S) + \varepsilon$
  - S: set of stable (causal) features, such as treatments, medications
  - V: set of noisy features, such as income, location
  - Assumption: $P(Y|S)$ is stable across environments, that is $P(Y|X) = P(Y|S)$
- **Why would a predictive model not be stable**?
  - **Dependence issue**, Y is not independent with V
  - **Environment shift issue**, $P(XY)_{training} \neq P(XY)_{testing}$

Why would a predictive model not be stable?

- **Dependence issue**
  - $X = \{S, V\}$, and $Y = f(S) + \varepsilon$
  - Diagram (b) & (c):
    - Y is not independent with V
  - Diagram (a): $Y \perp V$
    - Selection bias, leading to Y is not independent with V
  - **Some v $\subseteq V$ would be learned as important predictors**

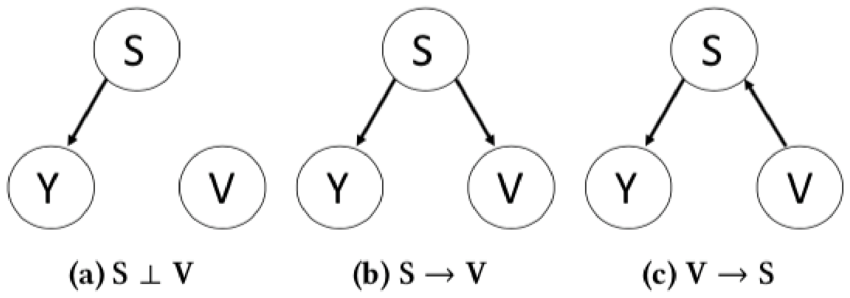


Figure 1: Three diagrams for stable features S, noisy features V, and response variable Y.

- **Environment shift issue**
  - $P(XY) = P(Y|X)P(X) = P(Y|S)P(X)$ (assume P(Y|S) is stable)
  - Selection bias → $P(X)_{training} \neq P(X)_{testing}$

    Y is not independent with V ⇒ $Corr(V_{training}, Y_{training}) \neq Corr(V_{testing}, Y_{testing})$

Problem – Stable Prediction

- **Given** **one** training environment $e \in \mathcal{E}$ with dataset $D^e = \{X^e, Y^e\}$
- **Task**: to learn a predictive model with **stable** performance across **unknown** environments $\mathcal{E}$.

- **Stability of the predictive model**:
  - Average_Error: $$Average\_Error = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} Error(D^e), \quad (1)$$
  - Stability_Error: $$Stability\_Error = \sqrt{\frac{1}{|\mathcal{E}|-1} \sum_{e \in \mathcal{E}} (Error(D^e) - Average\_Error)^2}, \quad (2)$$

OUTLINE

Related Work – address env. shift problem

- Covariate shift
  - Kernel mean matching [1], maximum entropy [2], robust bias-aware [3]
  - Importance weights: mimic the distribution of testing data to training data

$$\lim_{n\to\infty} \min_h \mathbb{E}_{f^{(n)}_{\text{training}}(x)\tilde{f}(y|x)} \left[\boxed{\frac{f_{\text{testing}}(\mathbf{X})}{f_{\text{training}}(\mathbf{X})}} (Y - h(\mathbf{X}))^2 \right]$$
$$= \min_h \mathbb{E}_{f_{\text{testing}}(x)\tilde{f}(y|x)} \left[(Y - h(\mathbf{X})^2) \right]$$

- These methods require prior knowledge of testing data
- These methods ignore the dependence issue

Related Work

- Invariant Component Learning
  - Invariant prediction [4], domain generalization [5]
  - Assume $P(Y|S)$ is stable across environments
  - Finding a subset/representation of features S', such that $P(Y|S')$ is invariant across all observed **multiple** environments

  - They could still have dependence issue on V', if $P(Y|V')$ is also invariant across observed environments

Challenges

- **Dependence challenge**
  - Y is not independent with V
  - **Some v $\subseteq V$ would be learned as important predictors**
- **Environment shift challenge**
  - The joint distribution $P(XY)$ is different across environments.
  - $\boldsymbol{Corr(V_{training}, Y_{training}) \neq Corr(V_{testing}, Y_{testing})}$
  - Can be addressed if $V \perp Y$ on training environment
- **Unknown testing environments challenge**

Key Challenge: How to make $V \perp Y$

OUTLINE

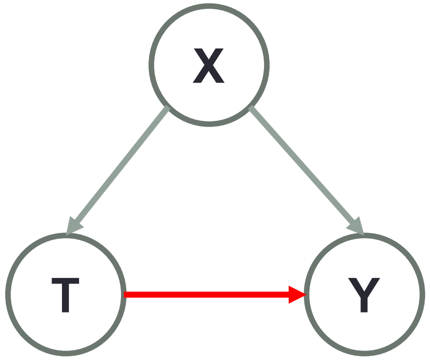
1. Background and Problem
2. Existing work and Challenges
3. Our GBR and DGBR Algorithms
4. Experiments

Our idea - link to causality

- Outcome generating mechanism
  - $Y = f(S) + \varepsilon,\ X = \{S, V\}$
- Difference between S and V
  - S has causal effect on Y,
  - but V has no causal effect on Y.

- **Our idea**: Recover causation between X and Y, such that $V \perp Y$, and only S is correlated with Y

Our idea - link to causality



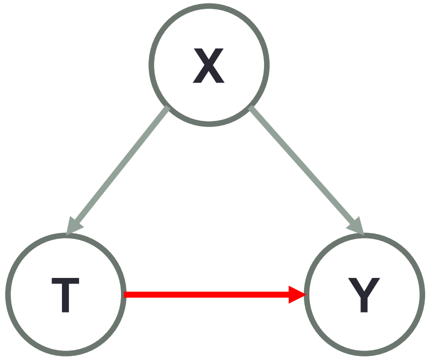
Causal Framework

- Causal inference with observational data
  - IPW [6], Entropy balancing [7], Approximate residual balancing[8], Differentiated Confounder Balancing [9]
  - **Sample reweighting** for variables balancing between $T = 1$ and $T = 0$, such that $T \perp X$.

$$W = \arg\min_{W} \left\| \frac{\sum_{i:T_i=1} W_i \cdot \mathbf{X}_i}{\sum_{i:T_i=1} W_i} - \frac{\sum_{i:T_i=0} W_i \cdot \mathbf{X}_i}{\sum_{i:T_i=0} W_i} \right\|_2^2$$

  - **After sample reweighting with W, the correlation between T and Y should be their causation.**
  - But they are limited to estimate the causal effect of one variable.

Our idea – Causality Regularizer



Causal Framework

- Approximate Global Balancing:
  - Motivation: Recovering causation between X and Y.
  - Sequentially learn causation between all X and Y via global sample weights *W* by minimizing:

$$\sum_{j=1}^{p} \left\| \frac{\mathbf{X}_{\cdot,-j}^{T} \cdot (W \odot \mathbf{X}_{\cdot,j})}{W^{T} \cdot \mathbf{X}_{\cdot,j}} - \frac{\mathbf{X}_{\cdot,-j}^{T} \cdot (W \odot (1-\mathbf{X}_{\cdot,j}))}{W^{T} \cdot (1-\mathbf{X}_{\cdot,j})} \right\|_2^2,$$

Loss function when learning the causation between X_j and Y

Sample reweighting with W → recovery causation → $V \perp Y$ → stable prediction

Our Algorithm 1 - GBR

- Global Balancing Regression (GBR) algorithm

$$\begin{aligned} \min \quad & \sum_{i=1}^{n} W_i \cdot \log(1 + \exp((1 - 2Y_i) \cdot (\mathbf{X}_i \boxed{\beta}))), \qquad (5) \\ s.t. \quad & \sum_{j=1}^{p} \left\| \frac{\mathbf{X}_{\cdot,-j}^T \cdot (W \odot \mathbf{X}_{\cdot,j})}{W^T \cdot \mathbf{X}_{\cdot,j}} - \frac{\mathbf{X}_{\cdot,-j}^T \cdot (W \odot (1 - \mathbf{X}_{\cdot,j}))}{W^T \cdot (1 - \mathbf{X}_{\cdot,j})} \right\|_2^2 \leq \lambda_1, \quad W \succeq 0, \\ & \|W\|_2^2 \leq \lambda_2, \quad \|\beta\|_2^2 \leq \lambda_3, \quad \|\beta\|_1 \leq \lambda_4, \quad \left(\sum_{k=1}^{n} W_k - 1\right)^2 \leq \lambda_5 \end{aligned}$$

Sample re-weighted logistic loss

Approximate Global Balancing

Causality Coefficients

- Other challenges: High-dimensional, and Non-linear prediction

Our Algorithm 2 - DGBR

- Deep Global Balancing Regression (DGBR) Algorithm

$$
\begin{aligned}
\min \quad & \sum_{i=1}^{n} W_i \cdot \log(1 + \exp((1 - 2Y_i) \cdot (\phi(\mathbf{X}_i)\beta))), \qquad (7) \\
s.t. \quad & \sum_{j=1}^{p} \left\| \frac{\phi(\mathbf{X}_{\cdot,-j})^T \cdot (W \odot \mathbf{X}_{\cdot,j})}{W^T \cdot \mathbf{X}_{\cdot,j}} - \frac{\phi(\mathbf{X}_{\cdot,-j})^T \cdot (W \odot (1 - \mathbf{X}_{\cdot,j}))}{W^T \cdot (1 - \mathbf{X}_{\cdot,j})} \right\|_2^2 \leq \lambda_1, \\
& \|(W \cdot \mathbf{1}) \odot (X - \hat{X})\|_F^2 \leq \lambda_2, \quad W \geq 0, \quad \|W\|_2^2 \leq \lambda_3, \\
& \|\beta\|_2^2 \leq \lambda_4, \quad \|\beta\|_1 \leq \lambda_5, \quad (\sum_{k=1}^{n} W_k - 1)^2 \leq \lambda_6 \\
& \sum_{k=1}^{K} (\|A^{(k)}\|_F^2 + \|\hat{A}^{(k)}\|_F^2) \leq \lambda_7,
\end{aligned}
$$

Deep Auto-Encoder

Global Balancing

Stable Prediction

Theoretical Analysis

$$\sum_{j=1}^{p} \left\| \frac{\mathbf{X}_{\cdot,-j}^{T} \cdot (W \odot \mathbf{X}_{\cdot,j})}{W^{T} \cdot \mathbf{X}_{\cdot,j}} - \frac{\mathbf{X}_{\cdot,-j}^{T} \cdot (W \odot (1 - \mathbf{X}_{\cdot,j}))}{W^{T} \cdot (1 - \mathbf{X}_{\cdot,j})} \right\|_{2}^{2}, \quad (4)$$

- The components of X could be mutually independent in the reweighted data.

PROPOSITION 1. *If* $0 < \hat{P}(\mathbf{X}_i = x) < 1$ *for all* x*, where* $\hat{P}(\mathbf{X}_i = x) = \frac{1}{n} \sum_i \mathbb{I}(\mathbf{X}_i = x)$, *there exists a solution* W^* *satisfies equation (4) equals 0 and variables in* $\mathbf{X}$ *are independent after balancing by* W^*.

- **Our GBR algorithm can make $V \perp Y$**

PROPOSITION 2. *If* $0 < \hat{P}(\mathbf{X}_i^e = x) < 1$ *for all* x *in environment* e, $Y^{e'}$ *and* $\mathbf{V}^{e'}$ *are independent when the joint probability mass function of* $(\mathbf{X}^{e'}, Y^{e'})$ *is given by reweighting the distribution from environment* e *using weights* W^**, so that* $p^{e'}(x, y) = p^e(y|x) \cdot (1/|\mathcal{X}|)$.

Theoretical Analysis

$$\sum_{j=1}^{p} \left\| \frac{\mathbf{X}_{\cdot,-j}^{T} \cdot (W \odot \mathbf{X}_{\cdot,j})}{W^{T} \cdot \mathbf{X}_{\cdot,j}} - \frac{\mathbf{X}_{\cdot,-j}^{T} \cdot (W \odot (1 - \mathbf{X}_{\cdot,j}))}{W^{T} \cdot (1 - \mathbf{X}_{\cdot,j})} \right\|_{2}^{2}, \quad (4)$$

- The components of X could be mutually independent in the reweighted data.

PROPOSITION 1. *If* $0 < \hat{P}(\mathbf{X}_i = x) < 1$ *for all* x*, where* $\hat{P}(\mathbf{X}_i = x) = \frac{1}{n}\sum_i \mathbb{I}(\mathbf{X}_i = x)$, *there exists a solution* W^* *satisfies equation (4) equals 0 and variables in* $\mathbf{X}$ *are independent after balancing by* W^*.

- **Our GBR algorithm can make $V \perp Y$**

PROPOSITION 2. *If* $0 < \hat{P}(\mathbf{X}_{\cdot}^{e} = x) < 1$ *for all* x *in environ-*

Propositions 1&2 suggest that our GBR algorithm can make a stable prediction across unknown environments

Theoretical Analysis

- Our DGBR algorithm can preserve all properties of the GBR algorithm while making the overlap property easier to satisfy and reducing the variance of balancing weights.

- Our DGBR algorithm can enable more accurate estimation of $P(Y|S)$.

- More details could be found in our paper.

OUTLINE

1. Background and Problem
2. Existing work and Challenges
3. Our GBR and DGBR Algorithms
4. **Experiments**

Experiments

- Baselines:
  - Logistic Regression (LR)
  - Deep Logistic Regression (DLR): LR + Deep Auto Encoder
- Evaluation Metric:
  - RMSE, Average_Error, Stability_Error

$$Average\_Error = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} Error(D^e), \quad (1)$$

$$Stability\_Error = \sqrt{\frac{1}{|\mathcal{E}|-1} \sum_{e \in \mathcal{E}} (Error(D^e) - Average\_Error)^2}, \quad (2)$$

Experiments on Synthetic Data

- Data generating
  - $X = \{S, V\}$ is binary.
  - $Y = h(f(S) + \epsilon)$ is also binary.

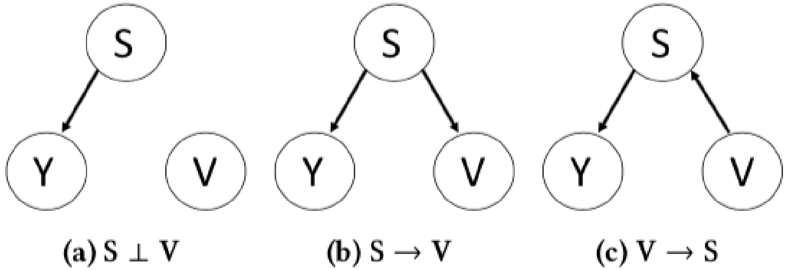


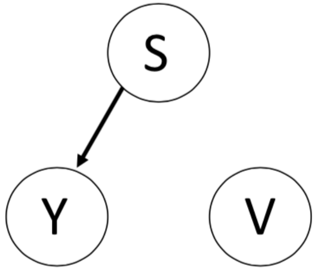
Figure 1: Three diagrams for stable features S, noisy features V, and response variable Y.

- Environments generating
  - Changing P_{XY} by sample selection with the **bias rate: r**
  - **Varying $P(Y|V)$: leading to $P(Y|X) \neq P(Y|S)$**
    - if $V = Y$, then $p(selected) = r$, otherwise $p(selected) = 1 - \text{r}$.
  - Note that: $r > 0.5$ implies $Corr(V, Y)$ is positive

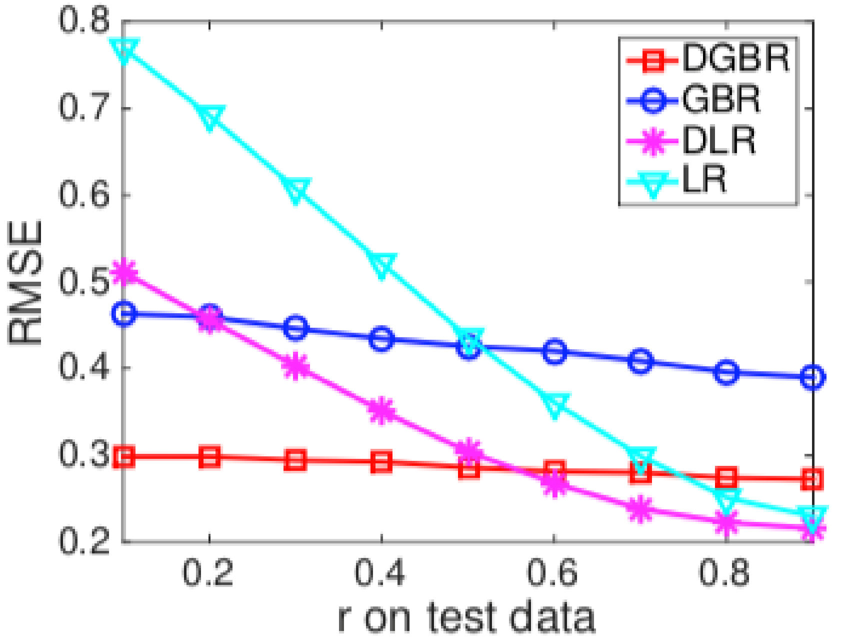
Experiments on Synthetic Data

- Setting $S \perp V$
  - **Trained** on one environment $\boldsymbol{r} = \mathbf{0.85}$, and **tested** on all environments $\boldsymbol{r} = \{\mathbf{0.1}, \dots, \mathbf{0.9}\}$
  - **Different r means different environment**

- Traditional LR and DLR failed
- GBR (dark blue) is more stable than LR
- DGBR (Red) is more stable than DLR
- DGBR is more stable and precise than GBR



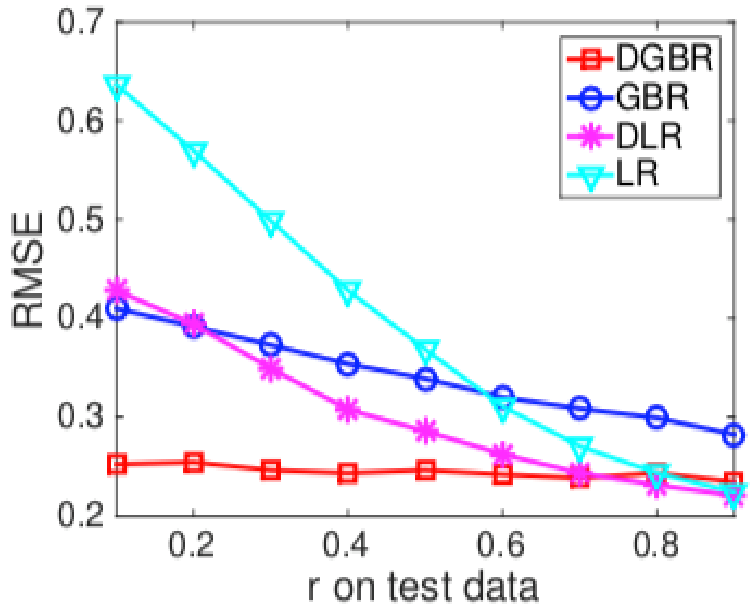
(a) S ⊥ V



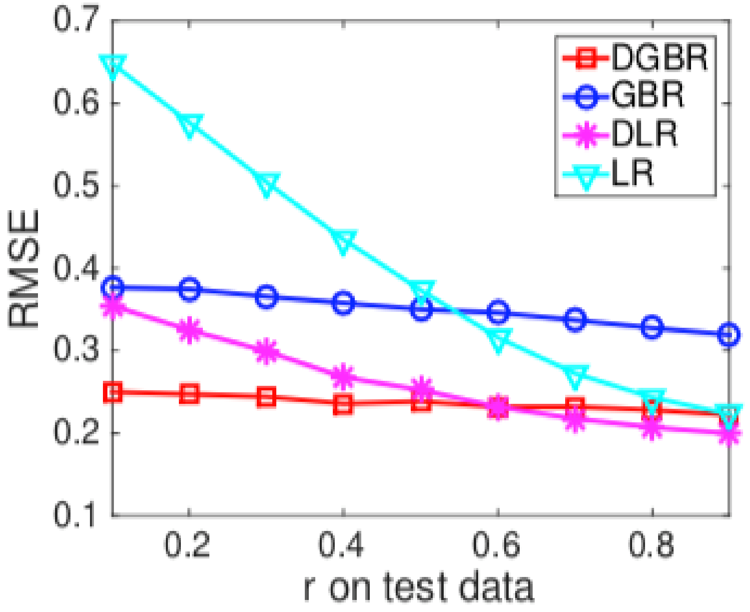
(f) Trained on $n = 2000$, $p = 20$, $r = 0.85$

Experiments on Synthetic Data

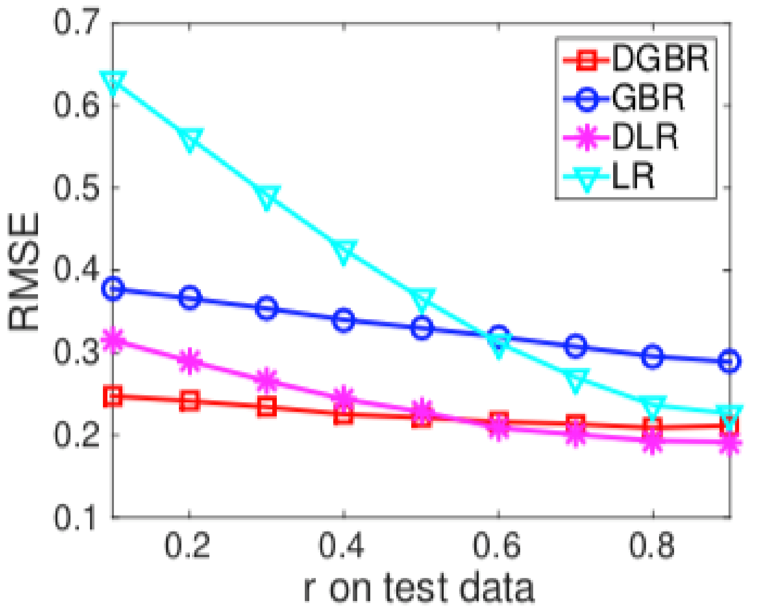
- More settings: varying n, p, and r



(b) Trained on $n = 1000$, $p = 20$, $r = 0.75$



(e) Trained on $n = 2000$, $p = 20$, $r = 0.75$

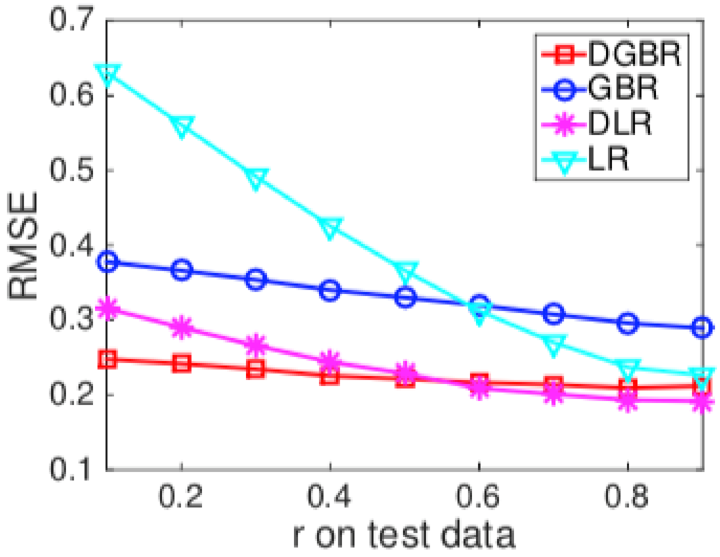


(h) Trained on $n = 4000$, $p = 20$, $r = 0.75$

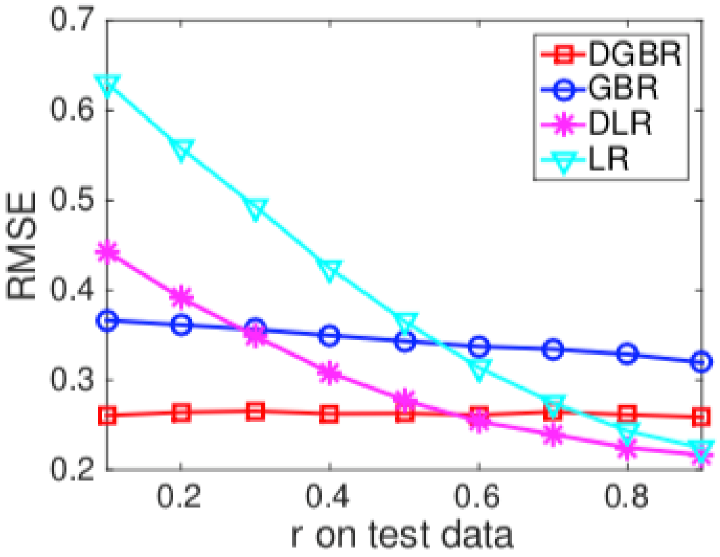
Vary sample size n

Experiments on Synthetic Data

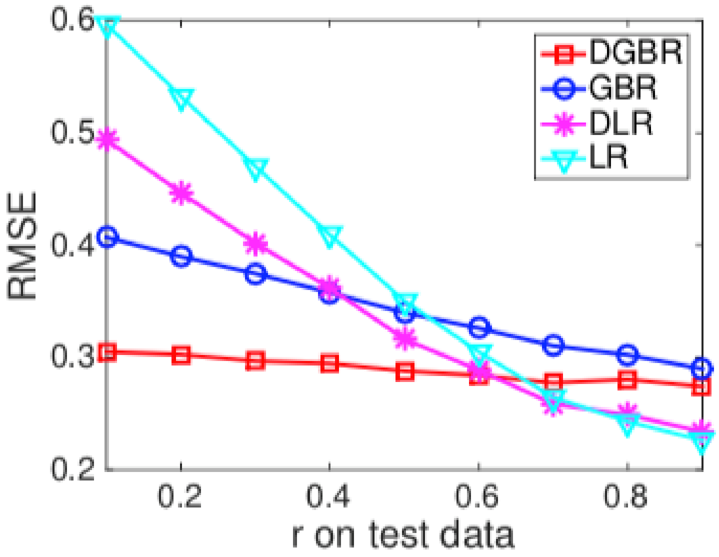
- More settings: varying n, p, and r



(a) **Trained on** $n = 4000$, $p = 20$, $r = 0.75$



(b) **Trained on** $n = 4000$, $p = 40$, $r = 0.75$

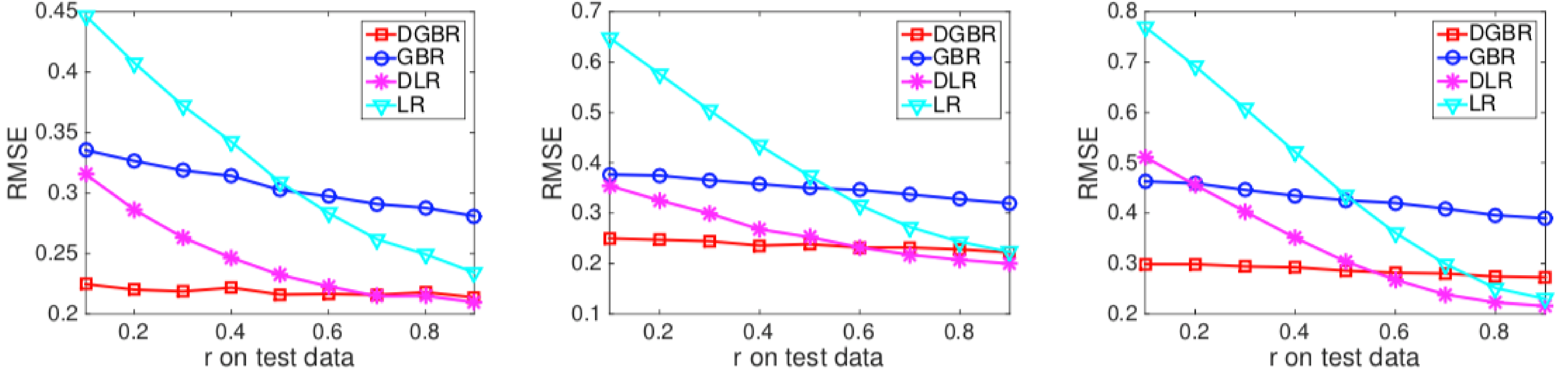


(c) **Trained on** $n = 4000$, $p = 80$, $r = 0.75$

Vary variables' dimension p

Experiments on Synthetic Data

- More settings: varying n, p, and r

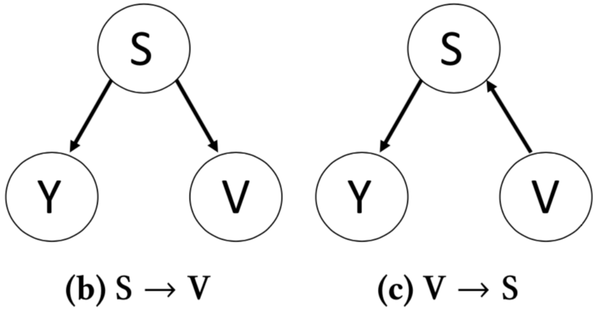


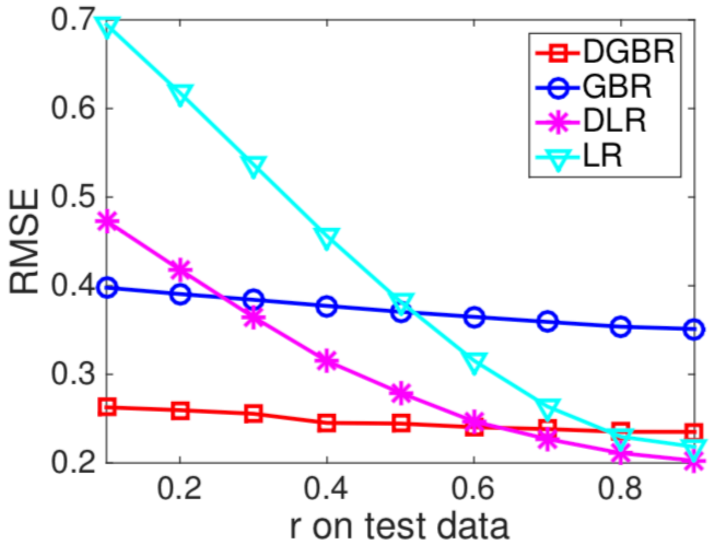
(d) Trained on $n = 2000$, $p = 20$, $r = 0.65$ (e) Trained on $n = 2000$, $p = 20$, $r = 0.75$ (f) Trained on $n = 2000$, $p = 20$, $r = 0.85$

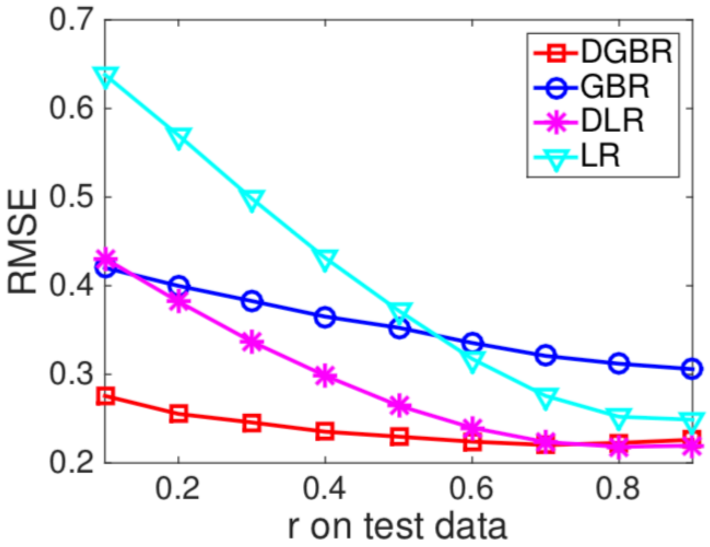
Vary bias rate r on training environment

Experiments on Synthetic Data

- More settings: setting S→V (S is the cause of V)
  setting V→S (V is the cause of S)







The RMSE of DGBR is consistently stable and small across environments under all settings.

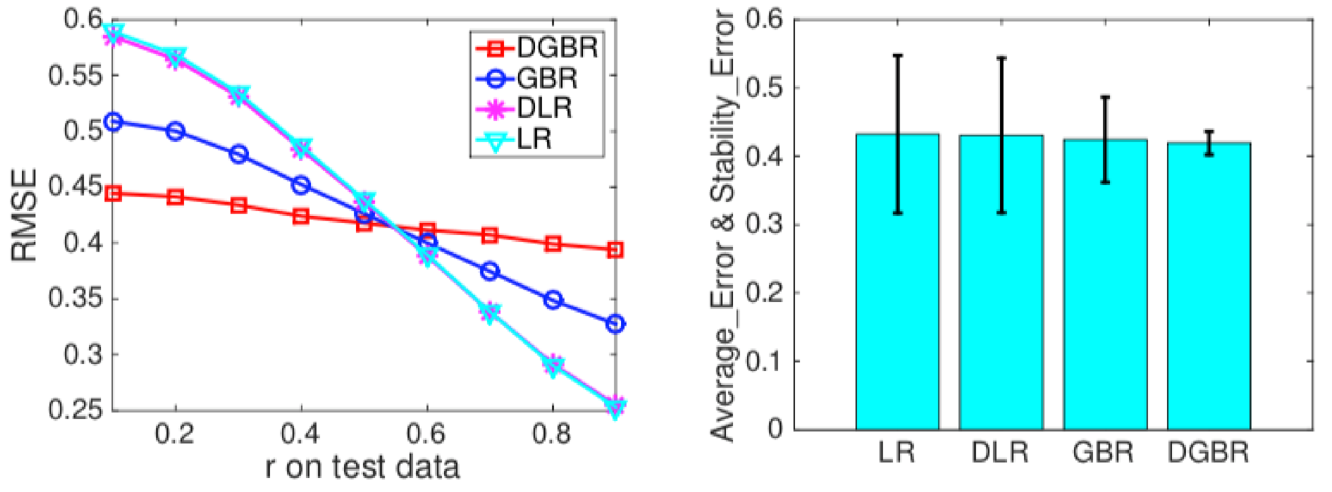
Experiments on Real World Data

2015

- Dataset Description:
  - Online advertising campaign (LONGCHAMP)
  - Users Feedback: 14,891 LIKE; 93,108 DISLIKE
  - 56 Features for each user
    - Age, gender, #friends, device, user setting on WeChat
- Experimental Setting:
  - Outcome Y: users feedback ⟵ Y = 1, if LIKE; Y = 0, if DISLIKE
  - Setting1: generating environment with bias rate r.
  - Setting2: generating environment with users' age.

Experiments on Real World Data – setting 1

- Environments generating:
  - Pre-selecting some noisy features V, then generating environments by varying P(Y|V) with bias rate r. (Models are trained with r=0.6)



(a) RMSE (b) Average_Error & Stability_Error

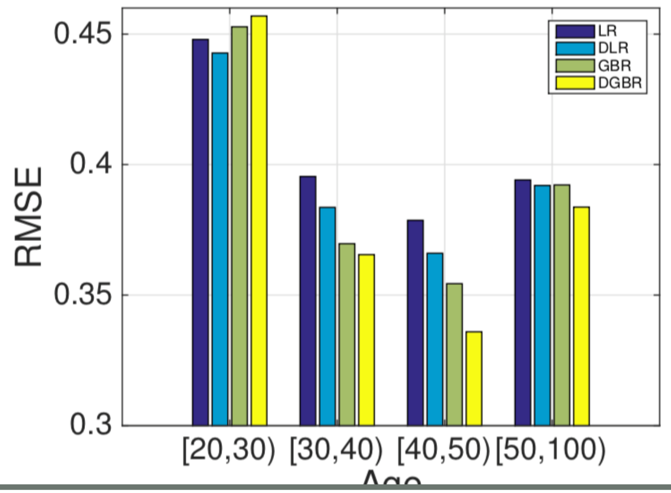
Fig. 13: Our proposed DGBR algorithm makes the most stable prediction on whether user will like or dislike an advertisement.

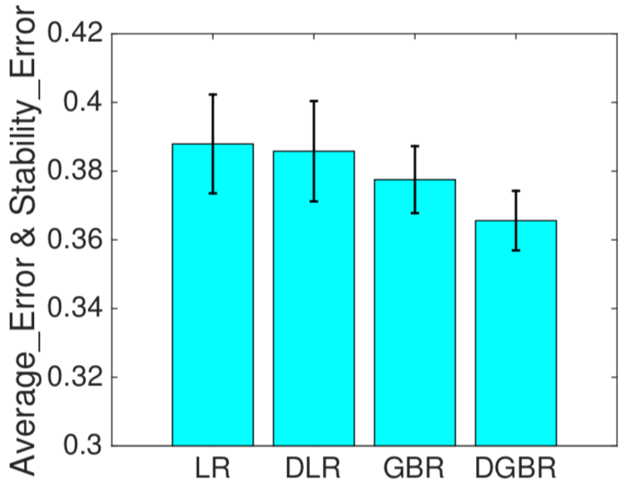
Average_Error: blue bar
Stability_Error: dark line

Our DGBR algorithm can make a more stable prediction across unknown environments.

Experiments on Real World Data – setting 2

- Environments generating:
  - Separate the whole dataset into 4 environments by users' age, including $Age \in [20,30)$, $Age \in [30,40)$, $Age \in [40,50)$, and $Age \in [50,100)$.





Average_Error: blue bar
Stability_Error: dark line

Our DGBR algorithm can make a more stable and precise prediction across unknown environments.

Conclusion

- Stable prediction across unknown environments.
  - Dependence issue, Y is not independent with V
  - Environment shift issue, $\boldsymbol{Corr}(\boldsymbol{V}_{\boldsymbol{training}}, \boldsymbol{Y}_{\boldsymbol{training}}) \neq \boldsymbol{Corr}(\boldsymbol{V}_{\boldsymbol{testing}}, \boldsymbol{Y}_{\boldsymbol{testing}})$
  - Unknown testing environments
- We proposed **Global Balancing Regression** and **Deep Global Balancing Regression** algorithms for stable prediction.
- We show, both **theoretically** and with **empirical experiments**, that our algorithms can make stable prediction across unknown environments

Reference

- [1] Huang J, Gretton A, Borgwardt K M, et al. Correcting sample selection bias by unlabeled data[C]//Advances in neural information processing systems. 2007: 601-608.
- [2] Dudík M, Phillips S J, Schapire R E. Correcting sample selection bias in maximum entropy density estimation[C]//Advances in neural information processing systems. 2006: 323-330.
- [3] Liu A, Ziebart B. Robust classification under sample selection bias[C]//Advances in neural information processing systems. 2014: 37-45.
- [4] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. 2016. Causal inference by using invariant prediction: identification and confidence intervals. Journal ofthe Royal Statistical Society: Series B 78, 5 (2016), 947–1012.
- [5] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. 2013. Domain generalization via invariant feature representation. In Proceedings of the 30thInternational Conference on Machine Learning (ICML-13). 10–18.
- [6] Kun Kuang, Peng Cui, Bo Li, Meng Jiang, Shiqiang Yang, and Fei Wang. 2017. Treatment Effect Estimation with Data-Driven Variable Decomposition.. In AAAI. 140–146.
- [7] Jens Hainmueller. 2012. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. Polit- ical Analysis 20, 1 (2012), 25–46.
- [8] Susan Athey, Guido Imbens, and Stefan Wager. Approximate Residual Balancing: De-Biased Inference of Average Treatment Effects in High Dimensions. *Journal of the Royal Statistical Society: Series B*, forthcoming.
- [9] Kun Kuang, Peng Cui, Bo Li, Meng Jiang, and Shiqiang Yang. 2017. Estimating Treatment Effect in the Wild via Differentiated Confounder Balancing. In Proceed- ings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 265–274.

Thanks!





Kun Kuang
kkun2010@gmail.com

KDD2018