STANFORD UNIVERSITY

Feature Noise Induces Loss Discrepancy Across Groups

Fereshte Khani

Percy Liang

Motivation

- Standard learning procedures work well in average

Gender Classifier	Overall Accuracy on all Subjects in Pilot Parlaiments Benchmark (2017)
Microsoft	93.7%
FACE++	90.0%
IBM	87.9%

Motivation

- Standard learning procedures work well in average
- Performance is different across groups

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

Motivation

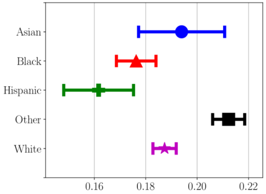
- Standard learning procedures work well in average
- Performance is different across groups
- Especially problematic for critical applications and protected groups

Search query	Work experience	Education experience	Profile views	Candidate	Xing ranking
Brand Strategist	146	57	12992	male	1
Brand Strategist	327	0	4715	female	2
Brand Strategist	502	74	6978	male	3
Brand Strategist	444	56	1504	female	4
Brand Strategist	139	25	63	male	5
Brand Strategist	110	65	3479	female	6
Brand Strategist	12	73	846	male	7
Brand Strategist	99	41	3019	male	8
Brand Strategist	42	51	1359	female	9
Brand Strategist	220	102	17186	female	10

Hiring

	WHITE	AFRICAN AMERICAN
Didn't Re-Offend	23.5%	44.9%
Did Re-Offend	47.7%	28.0%

Courts



Health Care

Why do such loss discrepancies exist?

Previous work

- Training data is biased

(Rothwell, 2014; Madras et al., 2019)

Previous work

- Training data is biased

  (Rothwell, 2014; Madras et al., 2019)

- Groups have different true functions

  (Dwork et al., 2018)

Virtual Reality Is Sexist: But It Does Not Have to Be

Kay Stanney[1*], Cali Fidopiastis[1] and Linda Foster[2]

[1]Design Interactive, Inc., Orlando, FL, United States
[2]Lockheed Martin Corporate, Washington, DC, United States

Previous work

- Training data is biased

  (Rothwell, 2014; Madras et al., 2019)

- Groups have different true functions

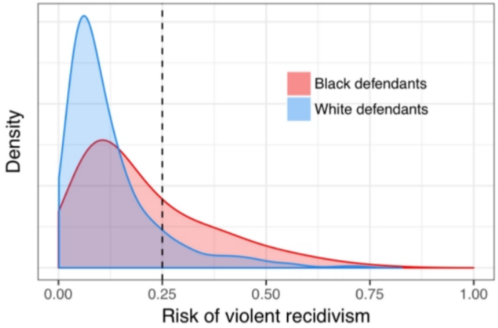
  (Dwork et al., 2018)

- Minority/generalization issues

  (Chen et al., 2018)

Previous work

- Training data is biased
  (Rothwell, 2014; Madras et al., 2019)
- Groups have different true functions
  (Dwork et al., 2018)
- Minority/generalization issues
  (Chen et al., 2018)
- From soft classifiers to hard decisions
  (Canetti et al., 2019; Corbett-Davies and Goel, 2018)
- Groups have different amount of noise
  (Corbett-Davies and Goel, 2018; Corbett-Davies et al., 2017)



Previous work

- Training data is biased
  (Rothwell, 2014; Madras et al., 2019)
- Groups have different true functions
  (Dwork et al., 2018)
- Minority/generalization issues
  (Chen et al., 2018)
- From soft classifiers to hard decisions
  (Canetti et al., 2019; Corbett-Davies and Goel, 2018)
- Groups have different amount of noise
  (Corbett-Davies and Goel, 2018; Corbett-Davies et al., 2017)

This work

- No biased training data
- Same true function for both groups
- Infinite data
- Linear regression setup
- Same amount of noise

Even under the most favorable condition $\left\{\begin{array}{l}\text{No biased training data}\\ \text{Same true function}\\ \text{Infinite data}\\ \text{Linear regression setup}\\ \text{Same amount of noise}\end{array}\right\}$ there is still loss discrepancy.

Even under the most favorable condition

- No biased training data
- Same true function
- Infinite data
- Linear regression setup
- Same amount of noise

there is still loss discrepancy.

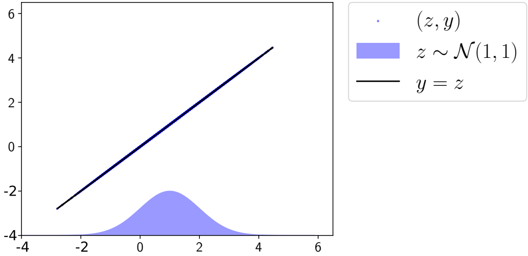
Main Takeaway

Same amount of **feature noise** on all individuals affects groups differently.

Outline

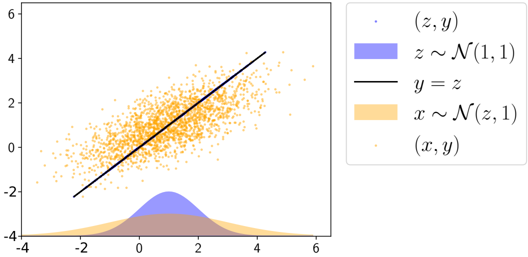
- Background on feature noise in linear regression
- Setup
- Feature noise induces loss discrepancy
- Experiments

Background: Feature noise in Linear Regression



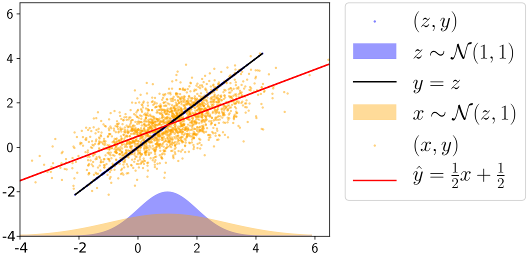
- **Setup:**
  $z \sim \mathcal{P}_z, \quad y = \beta^\top z + \alpha,$

Background: Feature noise in Linear Regression



- **Setup:**
  $z \sim \mathcal{P}_z, \quad y = \beta^\top z + \alpha,$
  $\mathbb{E}[u] = 0$ and u is independent of y and z

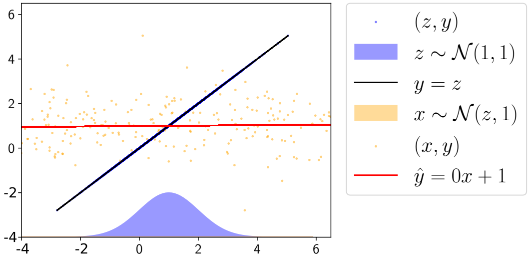
Background: Feature noise in Linear Regression



- **Setup:**
  $z \sim \mathcal{P}_z, \quad y = \beta^\top z + \alpha, \quad x = z + u$
  $\mathbb{E}[u] = 0$ and u is independent of y and z

- **Method:**
  $\hat{y} = \hat{\beta}^\top x + \hat{\alpha}, \quad$ Least squares estimator

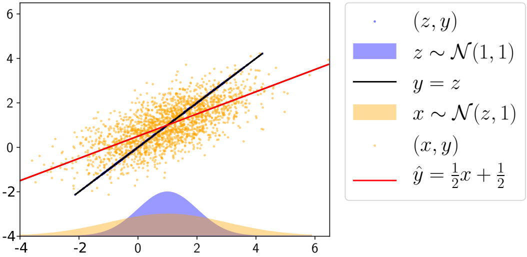
Background: Feature noise in Linear Regression



- **Setup:**
  $z \sim \mathcal{P}_z, \quad y = \beta^\top z + \alpha, \quad x = z + u$
  $\mathbb{E}[u] = 0$ and u is independent of y and z

- **Method:**
  $\hat{y} = \hat{\beta}^\top x + \hat{\alpha}$, Least squares estimator

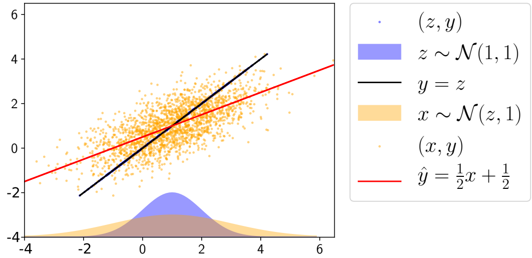
Background: Feature noise in Linear Regression



- **Setup:**
  $z \sim \mathcal{P}_z, \quad y = \beta^\top z + \alpha, \quad x = z + u$
  $\mathbb{E}[u] = 0$ and u is independent of y and z

- **Method:**
  $\hat{y} = \hat{\beta}^\top x + \hat{\alpha}, \quad$ Least squares estimator

Background: Feature noise in Linear Regression



- **Setup:**
  $z \sim \mathcal{P}_z, \quad y = \beta^\top z + \alpha, \quad x = z + u$
  $\mathbb{E}[u] = 0$ and u is independent of y and z

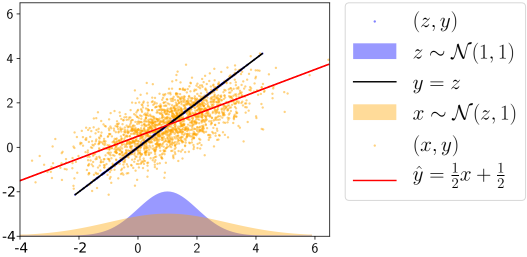
- **Method:**
  $\hat{y} = \hat{\beta}^\top x + \hat{\alpha}, \quad$ Least squares estimator

- **Analysis:**

Let Λ denotes noise to signal ratio

$$\Lambda \stackrel{\text{def}}{=} (\Sigma_z + \Sigma_u)^{-1}\Sigma_u$$

Background: Feature noise in Linear Regression



- **Setup:**
  $z \sim \mathcal{P}_z, \quad y = \beta^\top z + \alpha, \quad x = z + u$
  $\mathbb{E}[u] = 0$ and u is independent of y and z
- **Method:**
  $\hat{y} = \hat{\beta}^\top x + \hat{\alpha}$, Least squares estimator
- **Analysis:**

  Let Λ denotes noise to signal ratio

  $$\Lambda \stackrel{\text{def}}{=} (\Sigma_z + \Sigma_u)^{-1} \Sigma_u$$

  $$\hat{\beta} = \beta - \Lambda\beta$$

  $$\hat{\alpha} = \alpha + (\Lambda\beta)^\top \mathbb{E}[z]$$

Outline

✓ Background on feature noise in linear regression

- Setup
- Feature noise induces loss discrepancy
- Experiments

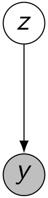
Setup

true (latent) features $z \in \mathbb{R}^n$

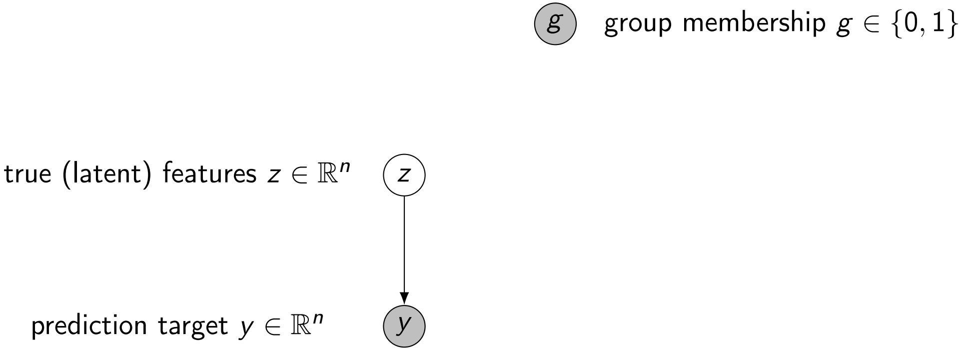
Setup

true (latent) features $z \in \mathbb{R}^n$

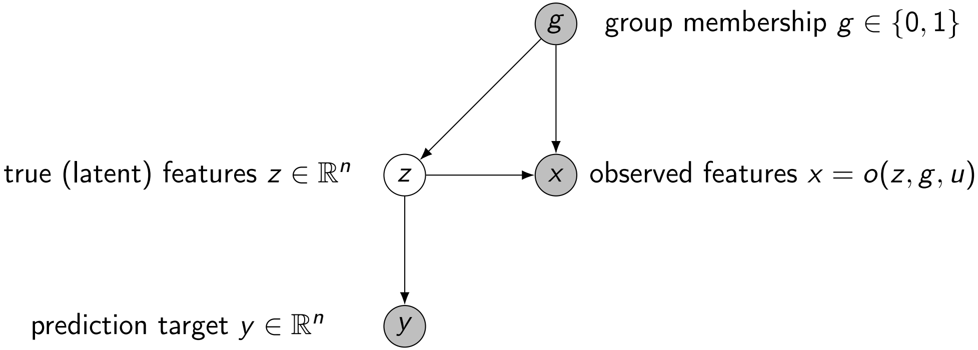
prediction target $y \in \mathbb{R}^n$



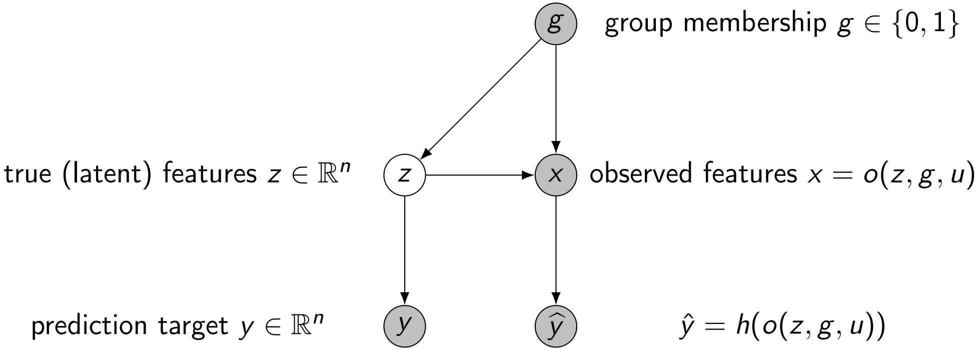
Setup


g
group membership $g \in \{0, 1\}$
true (latent) features $z \in \mathbb{R}^n$
z
prediction target $y \in \mathbb{R}^n$
y

Setup

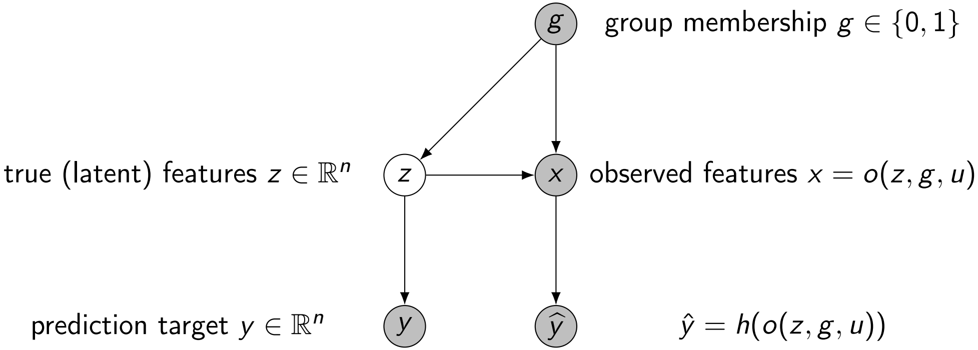

group membership $g \in \{0, 1\}$
true (latent) features $z \in \mathbb{R}^n$
z
x
observed features $x = o(z, g, u)$
prediction target $y \in \mathbb{R}^n$
y
g

Setup



group membership $g \in \{0, 1\}$
true (latent) features $z \in \mathbb{R}^n$
observed features $x = o(z, g, u)$
prediction target $y \in \mathbb{R}^n$
$\hat{y} = h(o(z, g, u))$
g
z
x
y
$\widehat{y}$

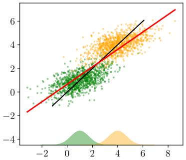
Setup



loss $\ell(\hat{y}, y)$: impact of the predictor for an individual

Outline

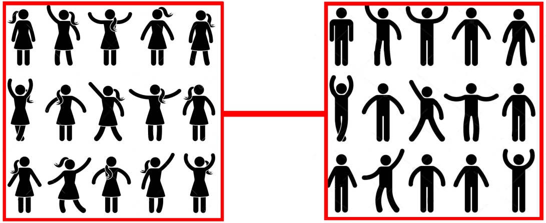
- ✓ Background on feature noise in linear regression
- ✓ Setup
- Feature noise induces loss discrepancy
- Experiments



Outline: noise induces loss discrepancy

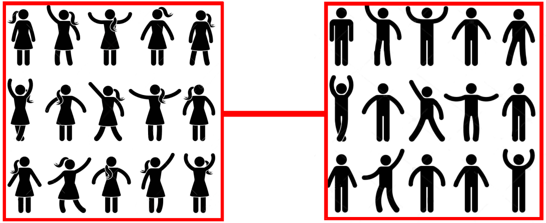
loss discrepancy / observation function	?	?
?	?	?
?	?	?

Statistical Loss Discrepancy[1]



[1] (Hardt et al., 2016; Agarwal et al., 2018; Woodworth et al., 2017; Pleiss et al., 2017; Khani et al., 2019)

Statistical Loss Discrepancy[1]



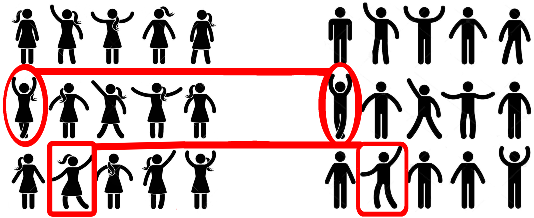
Definition (Statistical Loss Discrepancy (SLD))

For a predictor h, observation function o, and loss function ℓ, statistical loss discrepancy is the difference between the expected loss between two groups:

$$\mathrm{SLD}(h, o, \ell) = |\mathbb{E}[\ell \mid g = 1] - \mathbb{E}[\ell \mid g = 0]|$$

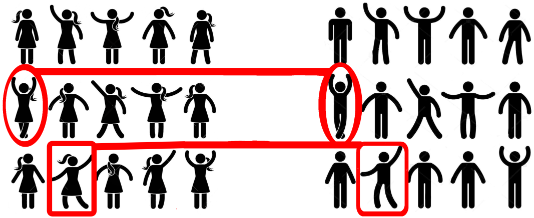
[1] (Hardt et al., 2016; Agarwal et al., 2018; Woodworth et al., 2017; Pleiss et al., 2017; Khani et al., 2019)

Counterfactual Loss Discrepancy [2]



[2] (Kusner et al., 2017; Chiappa, 2019; Loftus et al., 2018; Nabi and Shpitser, 2018; Kilbertus et al., 2017)

Counterfactual Loss Discrepancy [2]



Definition (Counterfactual Loss Discrepancy (CLD))

For a predictor h, observation function o, and loss function ℓ, counterfactual loss discrepancy is the expected difference between the loss of an individual and its counterfactual counterpart:

$$\mathrm{CLD}(h, o, \ell) = \mathbb{E}\left[|L_0 - L_1|\right],$$

where $L_{g'} = \mathbb{E}[\ell(h(o(z, g', u)), y)|z]$.

[2] (Kusner et al., 2017; Chiappa, 2019; Loftus et al., 2018; Nabi and Shpitser, 2018; Kilbertus et al., 2017)

Loss functions

- **Residual:** measures the amount of underestimation.

$$\ell_{\text{res}}(y, \hat{y}) \stackrel{\text{def}}{=} y - \hat{y}$$

Loss functions

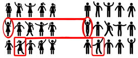
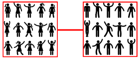
- **Residual:** measures the amount of underestimation.

$$\ell_{\mathsf{res}}(y, \hat{y}) \stackrel{\text{def}}{=} y - \hat{y}$$

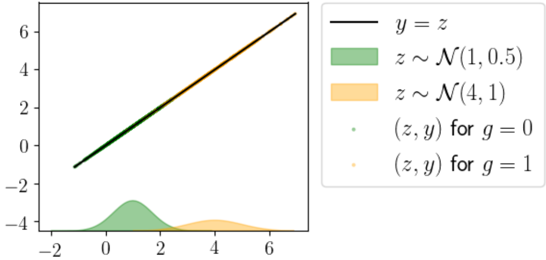
- **Squared error:** measures the overall performance.

$$\ell_{\mathsf{sq}}(y, \hat{y}) \stackrel{\text{def}}{=} (y - \hat{y})^2$$

Summary

loss discrepancy / observation function	**CLD** 	**SLD** 
?	?	?
?	?	?

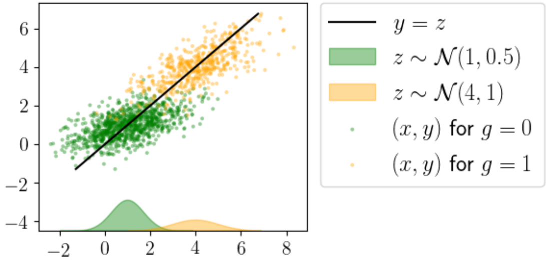
Independent noise without group information



- **Setup:**

$$z \sim \mathcal{P}_z, \quad y = \beta^\top z + \alpha$$

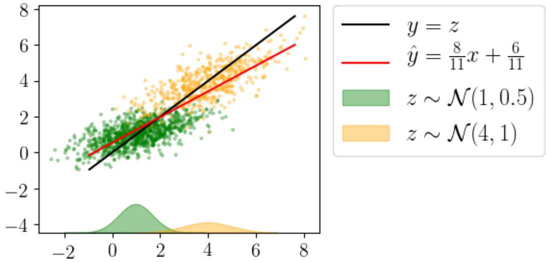
Independent noise without group information



- **Setup:**

$$z \sim \mathcal{P}_z, \quad y = \beta^\top z + \alpha$$
$$x = o_{-\mathrm{g}}(z, g, u) = z + u$$

Independent noise without group information



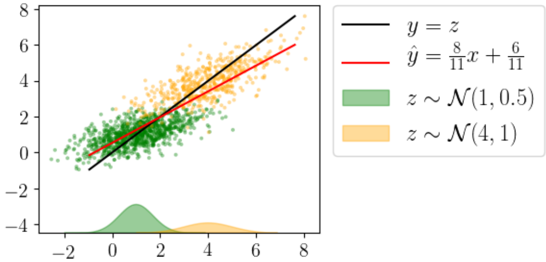
- **Setup:**

$$z \sim \mathcal{P}_z, \quad y = \beta^\top z + \alpha$$
$$x = o_{-\mathrm{g}}(z, g, u) = z + u$$

- **Method:**

$$\hat{y} = \hat{\beta} x + \hat{\alpha}, \quad \text{Least squares estimator}$$

Independent noise without group information



- **Setup:**

$$z \sim \mathcal{P}_z, \quad y = \beta^\top z + \alpha$$
$$x = o_{-\mathrm{g}}(z, g, u) = z + u$$

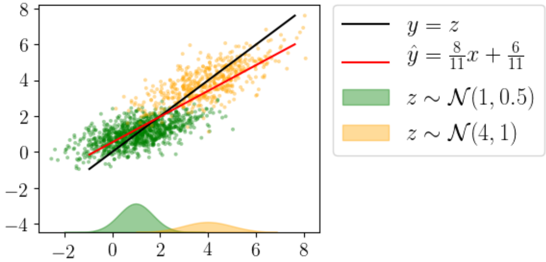
- **Method:**

$$\hat{y} = \hat{\beta}x + \hat{\alpha}, \quad \text{Least squares estimator}$$

- **Analysis:**

$$\mathrm{CLD}(o_{-\mathrm{g}}, \ell_{\mathrm{res}}) = 0$$

Independent noise without group information



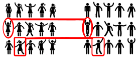
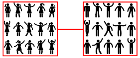
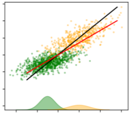
- **Important factors in statistical loss discrepancy (SLD)**
  1. noise ratio

$$\Lambda = (\Sigma_z + \Sigma_u)^{-1}\Sigma_u$$

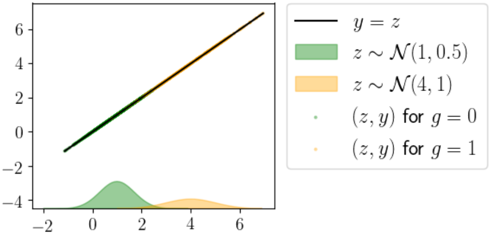
  2. difference in means

$$\Delta\mu = \mathbb{E}[z \mid g = 1] - \mathbb{E}[z \mid g = 0]$$

Summary

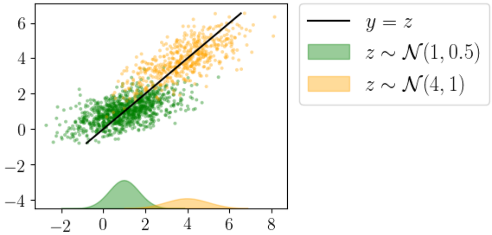
loss discrepancy / observation function	CLD 	SLD 
 $o_{-\mathrm{g}} = z + u$	0	$\lvert(\Lambda\beta)^\top \Delta\mu_z\rvert$
?	?	?

Independent noise with group information



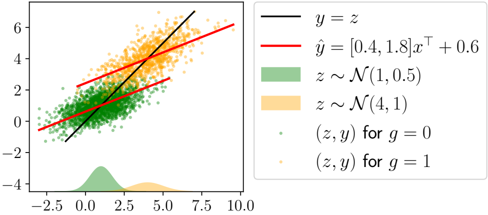
- **Setup:**
  $z \sim \mathcal{P}_z, \quad y = \beta^\top z + \alpha$

Independent noise with group information



- **Setup:**
  $z \sim \mathcal{P}_z, \quad y = \beta^\top z + \alpha$
  $x = o_{+\mathrm{g}}(z, g, u) = [z + u, g]$

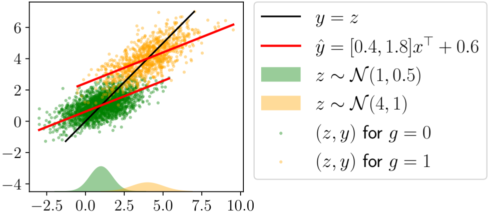
Independent noise with group information



- **Setup:**
  $z \sim \mathcal{P}_z, \quad y = \beta^\top z + \alpha$
  $x = o_{+\mathrm{g}}(z, g, u) = [z + u, g]$

- **Method:**
  $\hat{y} = \hat{\beta} x + \hat{\alpha}, \quad$ Least squares estimator

Independent noise with group information



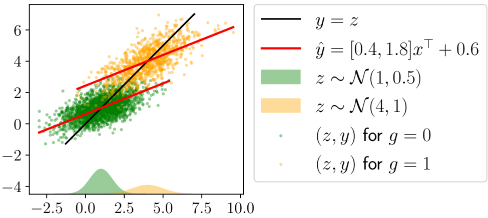
- **Setup:**
  $z \sim \mathcal{P}_z, \quad y = \beta^\top z + \alpha$
  $x = o_{+\mathrm{g}}(z, g, u) = [z + u, g]$

- **Method:**
  $\hat{y} = \hat{\beta}x + \hat{\alpha}, \quad$ Least squares estimator

- **Analysis:**

$$\mathrm{SLD}(o_{+\mathrm{g}}, \ell_{\mathrm{res}}) = 0$$

Independent noise with group information



Important factors in counterfactual loss discrepancy (CLD)

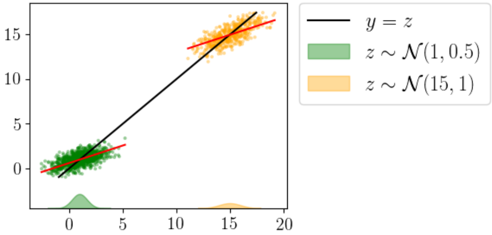
1. noise ratio

$$\Lambda' = (\Sigma_{z|g} + \Sigma_u)^{-1}\Sigma_u$$

2. difference in means

$$\Delta\mu = \mathbb{E}[z \mid g = 1] - \mathbb{E}[z \mid g = 0]$$

Independent noise with group information



Important factors in counterfactual loss discrepancy (CLD)

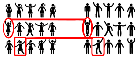
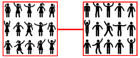
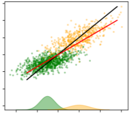
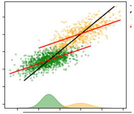
1. noise ratio

$$\Lambda' = (\Sigma_{z|g} + \Sigma_u)^{-1}\Sigma_u$$

2. difference in means

$$\Delta\mu = \mathbb{E}[z \mid g = 1] - \mathbb{E}[z \mid g = 0]$$

Summary

observation function \ loss discrepancy	CLD 	SLD 
 $o_{-\mathrm{g}} = z + u$	0	$\lvert(\Lambda\beta)^\top \Delta\mu_z\rvert$
 $o_{+\mathrm{g}} = [z + u, g]$	$\lvert(\Lambda'\beta)^\top \Delta\mu_z\rvert$	0

Outline

- ✓ Background on feature noise in linear regression
- ✓ Setup
- ✓ Feature noise induces loss discrepancy
- Experiments

Datasets

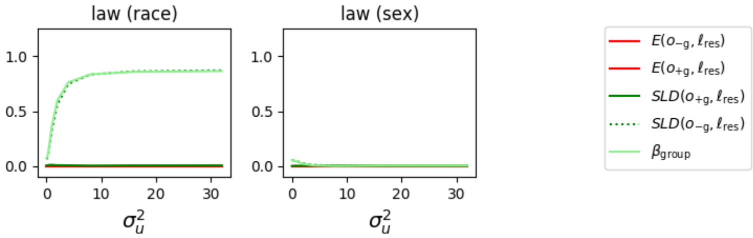
name	#records	#features	target	features example	group	$\mathbb{P}[g = 1]$	$\Delta\mu_y$	$\Delta\sigma_y^2$	$\|\Delta\mu_x\|_2$	$\|\Delta\Sigma_x\|_F$
C&C	1994	91	crime rate	#homeless, average income, ...	race	0.50	1.10	0.96	5.62	12.75
law	20798	25	final GPA	undergraduate GPA, LSAT, ...	race	0.86	0.87	0.01	2.24	2.79
					sex	0.56	0.005	0.04	0.42	0.51
students	649	33	final grade	study time, #absences, ...	sex	0.59	0.26	0.12	1.40	2.26

Datasets

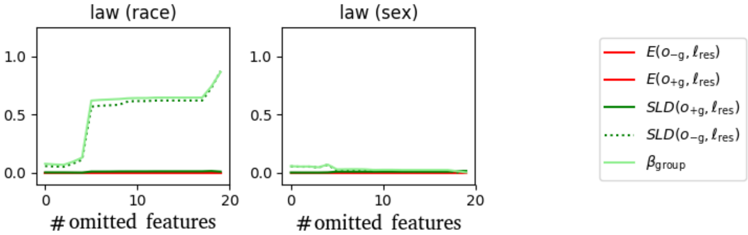
name	#records	#features	target	features example	group	$\mathbb{P}[g=1]$	$\Delta\mu_y$	$\Delta\sigma_y^2$	$\|\Delta\mu_x\|_2$	$\|\Delta\Sigma_x\|_F$
C&C	1994	91	crime rate	#homeless, average income, ...	race	0.50	1.10	0.96	5.62	12.75
law	20798	25	final GPA	undergraduate GPA, LSAT, ...	race	0.86	0.87	0.01	2.24	2.79
					sex	0.56	0.005	0.04	0.42	0.51
students	649	33	final grade	study time, #absences, ...	sex	0.59	0.26	0.12	1.40	2.26

Experiments (ℓ_{res})

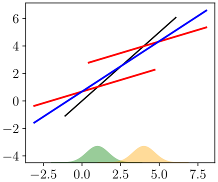
Equal noise

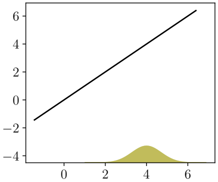


Omitting features



In the paper but not in this talk

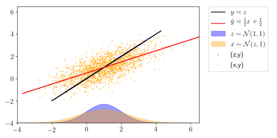




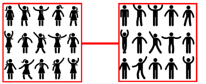
different distributions $\implies$ high loss discrepancy   Same distributions $\implies$ no loss discrepancy

We studied theoretically and experimentally the time it takes for a classifier to adapt to this shift.

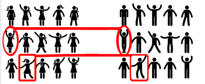
Noise causes attenuation bias



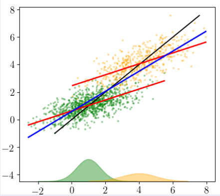
SLD



CLD

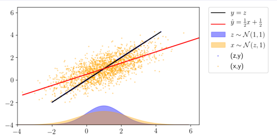


Noise induces loss discrepancy

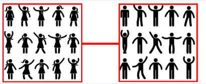


	CLD 	SLD 
ℓ_{res}	$o_{-\text{g}} : 0$ $o_{+\text{g}} : \lvert(\Lambda'\beta)^\top \Delta\mu_z\rvert$	$o_{-\text{g}} : \lvert(\Lambda\beta)^\top \Delta\mu_z\rvert$ $o_{+\text{g}} : 0$

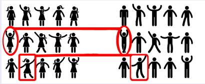
Noise causes attenuation bias



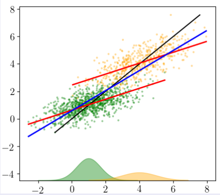
SLD



CLD



Noise induces loss discrepancy



	CLD	SLD
ℓ_{res}	$o_{-\text{g}} : 0$ $o_{+\text{g}} : \lvert(\Lambda'\beta)^\top \Delta\mu_z\rvert$	$o_{-\text{g}} : \lvert(\Lambda\beta)^\top \Delta\mu_z\rvert$ $o_{+\text{g}} : 0$

Thank You!

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. (2018). A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453*.

Canetti, R., Cohen, A., Dikkala, N., Ramnarayan, G., Scheffler, S., and Smith, A. (2019). From soft classifiers to hard decisions: How fair can we be? In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 309–318.

Chen, I., Johansson, F. D., and Sontag, D. (2018). Why is my classifier discriminatory? In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3539–3550.

Chiappa, S. (2019). Path-specific counterfactual fairness. In *Association for the Advancement of Artificial Intelligence (AAAI)*, volume 33, pages 7801–7808.

Corbett-Davies, S. and Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 797–806.

Dwork, C., Immorlica, N., Kalai, A. T., and Leiserson, M. (2018). Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency*, pages 119–133.

Hardt, M., Price, E., and Srebo, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3315–3323.

Khani, F., Raghunathan, A., and Liang, P. (2019). Maximum weighted loss discrepancy. *arXiv preprint arXiv:1906.03518*.

Kilbertus, N., Carulla, M. R., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 656–666.

Kusner, M. J., Loftus, J. R., Russell, C., and Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4069–4079.

Loftus, J. R., Russell, C., Kusner, M. J., and Silva, R. (2018). Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859*.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. (2019). Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 349–358.

Nabi, R. and Shpitser, I. (2018). Fair inference on outcomes. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. (2017). On fairness and calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5684–5693.

Rothwell, J. (2014). *How the war on drugs damages Black social mobility*. Brookings Institution.

Woodworth, B., Gunasekar, S., Ohannessian, M. I., and Srebro, N. (2017). Learning non-discriminatory predictors. In *Conference on Learning Theory (COLT)*, pages 1920–1953.