Explainable AI & Leverage Score Sampling

R. Teal Witter

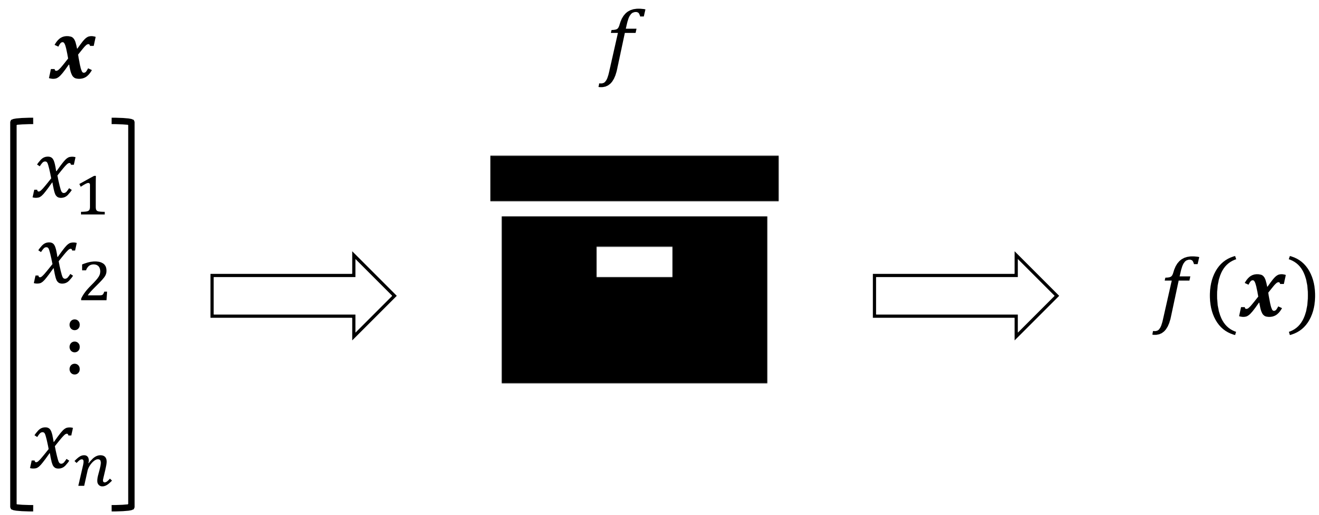
New York University

Shapley Values & Leverage Score Sampling

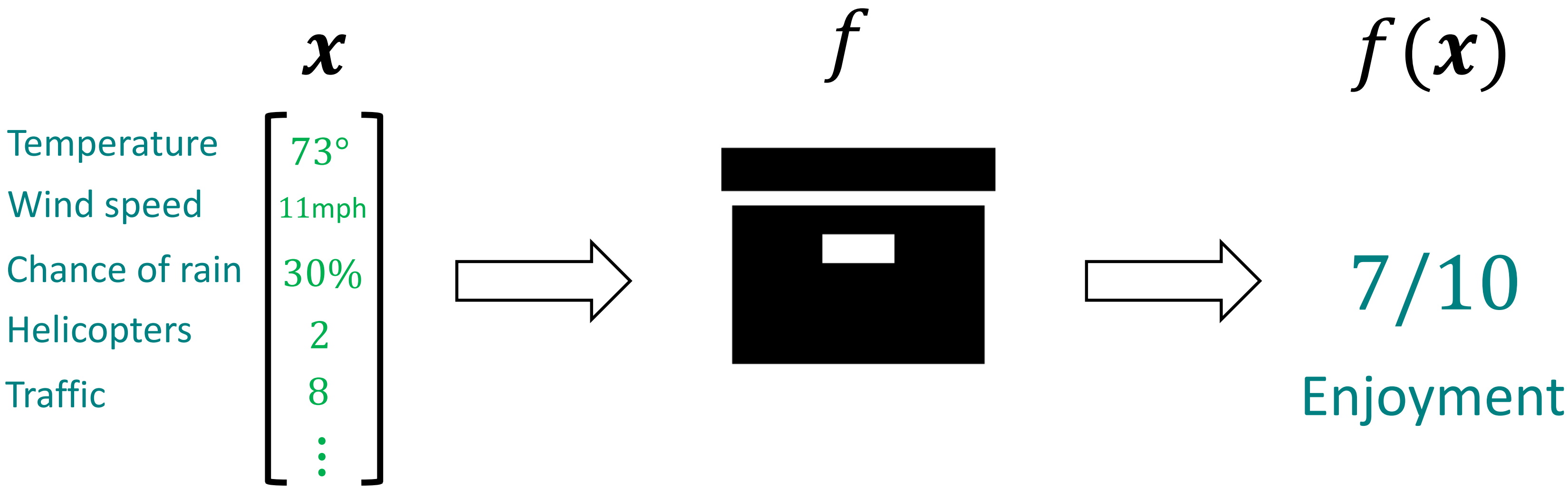
Joint work with

Christopher Musco
New York University

AI Predictions



Example:



$\boldsymbol{x}$
f
$f(\boldsymbol{x})$
Temperature
Wind speed
Chance of rain
Helicopters
Traffic
73°
11mph
30%
2
8
⋮
7/10
Enjoyment

Explaining Predictions

Attribute the prediction to features relative to a baseline

“Since the traffic is 8 instead of 3, the ride is 1.7 less enjoyable.”

Attribution value!

Explaining Predictions

Attribute the prediction to features relative to a baseline

"Since the traffic is 8 instead of 3, the ride is 1.7 less enjoyable."

	Explicand	Baseline
Temperature	73°	89°
Wind speed	11mph	1mph
Chance of rain	30%	0%
Helicopters	2	5
Traffic	8	3
	⋮	⋮

					$f(x)$
89°	11mph	30%	5	3	5/10
89°	11mph	30%	5	8	4/10
73°	1mph	0%	5	3	6/10
73°	1mph	0%	5	8	8/10

Desirable Properties

Dummy: *If a feature never changes the prediction, then its attribution value is 0*

Symmetry: *If two features always induce the same change, then their attribution values are the same*

Linearity: *For two predictive functions, the attribution value of a feature in the combined function is the sum of the attribution values for each function*

Efficiency: *The attribution values sum to the difference between the predictions on the explicand and baseline*

$\Leftrightarrow$ Shapley values!

Lloyd Shapley received the Nobel Prize in Economics for formulating Shapley values.

Shapley Values for Feature Attribution

Let $S \in [n]$ and define $v(S) = f(\boldsymbol{x}^S)$ where

S						$f(\boldsymbol{x}^S)$
$\{2,3\}$	$89°$	11mph	30%	5	3	$5/10$
$\{2,3,5\}$	$89°$	11mph	30%	5	8	$4/10$

The SHAP paper (one of the first to use Shapley values in explainable AI) has 25k+ citations.

Shapley Values for Feature Attribution

Let $S \in [n]$ and define $v(S) = f(\boldsymbol{x}^S)$ where

S						$f(\boldsymbol{x}^S)$
$\{2,3\}$	89°	11mph	30%	5	3	5/10
$\{2,3,5\}$	89°	11mph	30%	5	8	4/10

For a set function $v: 2^{[n]} \to \mathbb{R}$, the ith Shapley value is

$$\phi_i = \frac{1}{n} \sum_{S \in [n] \setminus \{i\}} \frac{v(S \cup \{i\}) - v(S)}{\binom{n-1}{|S|}}$$

Estimating Shapley values

Monte Carlo Sampling: Sample $S, S \cup \{i\}$ to use $v(S \cup \{i\}) - v(S)$
… but samples only used for one Shapley value

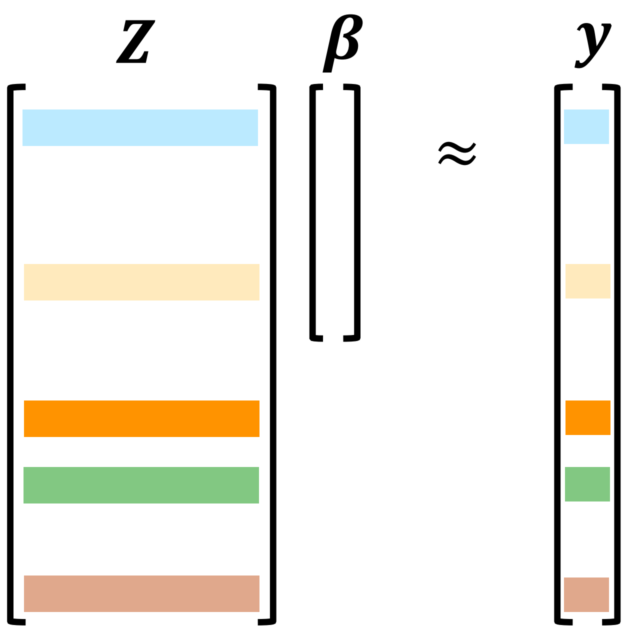
Maximum Reuse Sampling: Sample S to either add/subtract $v(S)$ for all i
… but magnitude of $v(S)$ is much larger than magnitude of $v(S \cup \{i\}) - v(S)$

Permutation Sampling: Sample $S_1 \subset S_2 \subset \cdots \subset S_n$ to use $v(S_{\ell+1}) - v(S_\ell)$
... but only 2x reuse

Regression Formulation

Constraint to satisfy *efficiency* property.

$$\boldsymbol{\phi} = \arg \min_{\boldsymbol{\beta}:\langle \beta, 1\rangle = v([n]) - v(\emptyset)} ||\boldsymbol{Z\beta} - \boldsymbol{y}||_2$$



Connection known since 80's [CGKR 1988]

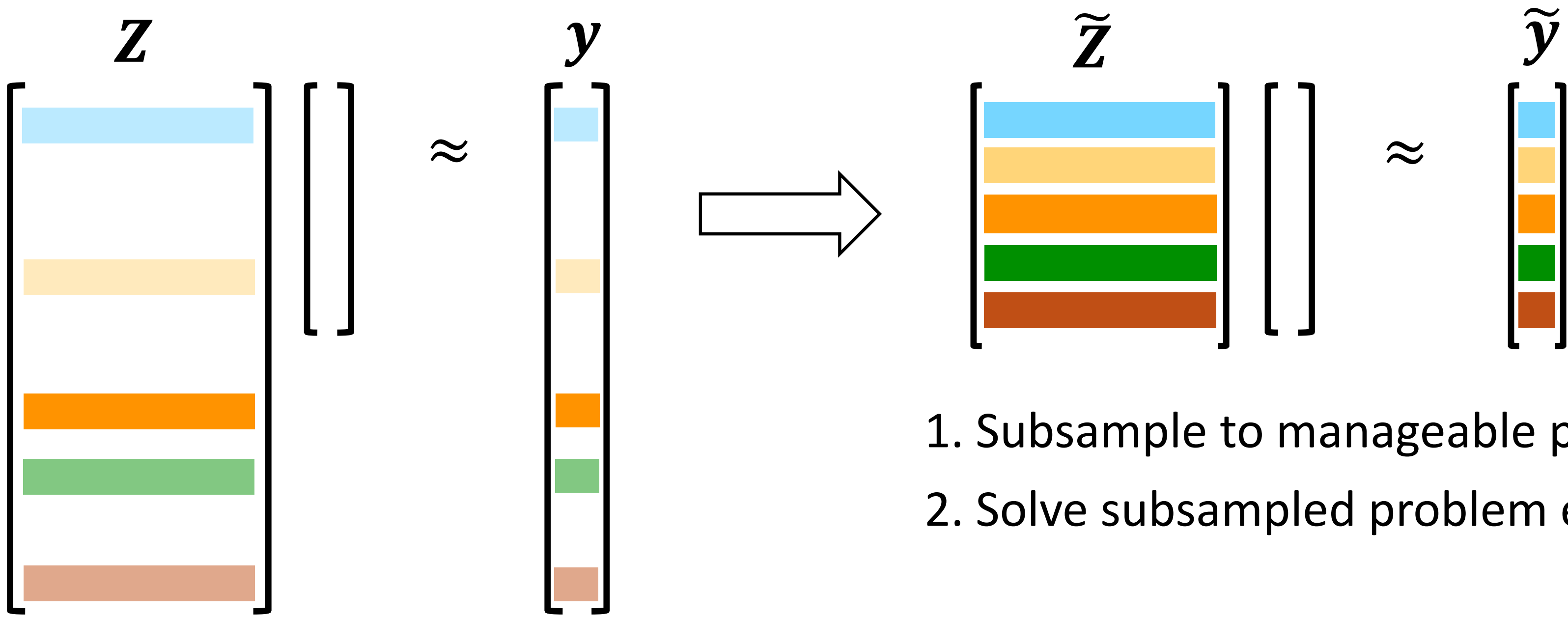
Each row/entry corresponds to binary vector

$$\boldsymbol{z} \in \{0,1\}^n : 0 < \ ||\boldsymbol{z}||_1 < n$$

Weighted by $\quad w(||\boldsymbol{z}||_1) = \dfrac{1}{\binom{n}{||\boldsymbol{z}||_1}(n - ||\boldsymbol{z}||_1)||\boldsymbol{z}||_1}$

Kernel SHAP

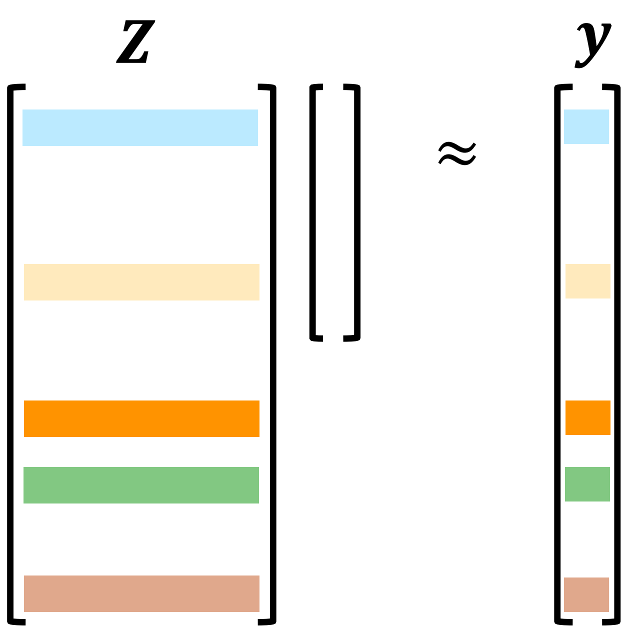
$$\tilde{\boldsymbol{\phi}} = \arg \min_{\boldsymbol{\beta}:\langle \beta, 1\rangle = v([n]) - v(\emptyset)} ||\tilde{\boldsymbol{Z}}\boldsymbol{\beta} - \tilde{\boldsymbol{y}}||_2$$

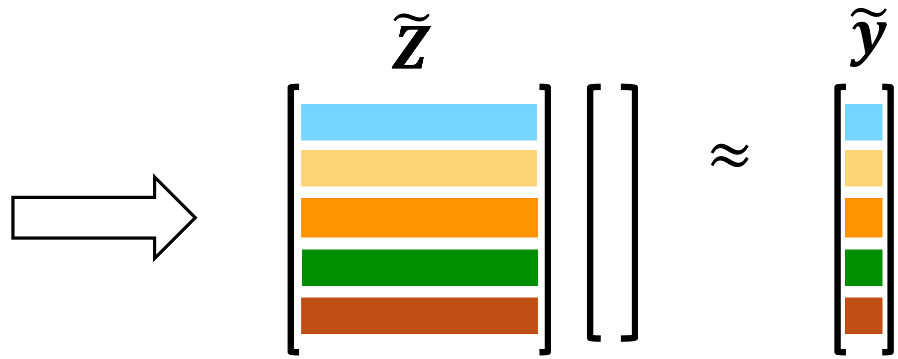


1. Subsample to manageable problem
2. Solve subsampled problem exactly

Kernel SHAP

$$\tilde{\boldsymbol{\phi}} = \arg \min_{\boldsymbol{\beta}:\langle \beta, 1 \rangle = v([n]) - v(\emptyset)} ||\tilde{\boldsymbol{Z}}\boldsymbol{\beta} - \tilde{\boldsymbol{y}}||_2$$





How should we subsample?

Kernel SHAP selects row $\boldsymbol{z}$ with probability proportional to $w(||\boldsymbol{z}||)$

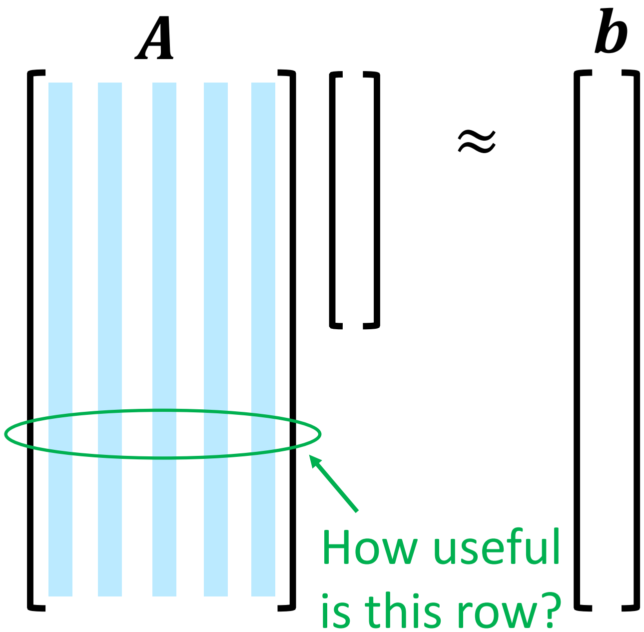
Constrained to Unconstrained Regression

$$\begin{aligned}\boldsymbol{\phi} &= \arg \min_{\boldsymbol{\beta}:\langle \beta, 1 \rangle = v([n]) - v(\emptyset)} ||\boldsymbol{Z\beta} - \boldsymbol{y}||_2 \\ &= \arg \min_{\boldsymbol{\beta}} ||\boldsymbol{A\beta} - \boldsymbol{b}||_2 + \mathbf{1}\,\frac{v([n]) - v(\emptyset)}{n}\end{aligned}$$

- By constraint, we know the component in the $\mathbf{1}$ direction
- Only optimize to residual target in space orthogonal to $\mathbf{1}$

Formulate as unconstrained problem so we can apply our favorite tools!

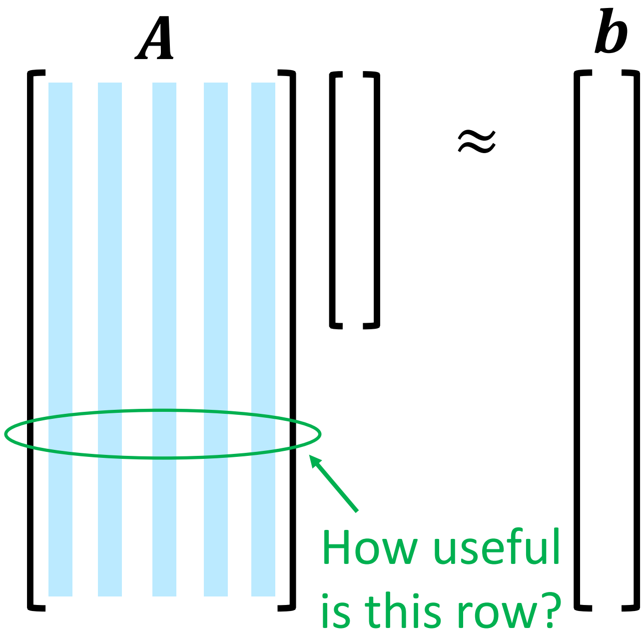
Leverage Scores



Row $\boldsymbol{z}$ has “leverage”:

$$\ell_{\boldsymbol{z}} = \boldsymbol{A}_{\boldsymbol{z}}(\boldsymbol{A}_{\boldsymbol{z}}^{\top}\boldsymbol{A}_{\boldsymbol{z}})^{+}\boldsymbol{A}_{\boldsymbol{z}}^{\top}$$

Leverage Scores and Shapley values



Row $\boldsymbol{z}$ has "leverage":

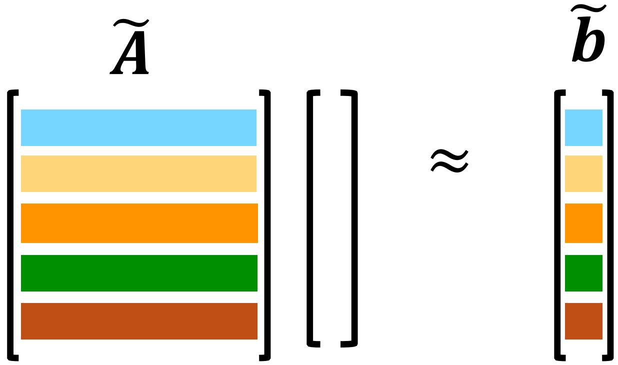
$$\ell_{\boldsymbol{z}} = \boldsymbol{A}_{\boldsymbol{z}}(\boldsymbol{A}_{\boldsymbol{z}}^{\top}\boldsymbol{A}_{\boldsymbol{z}})^{+}\boldsymbol{A}_{\boldsymbol{z}}^{\top}$$

$$\ell_{z} = \binom{n}{||\boldsymbol{z}||}^{-1}$$

Very similar to weighting in Shapley value definition!

Leverage SHAP

$$\tilde{\boldsymbol{\phi}} = \arg\min_{\boldsymbol{\beta}} ||\tilde{\boldsymbol{A}}\boldsymbol{\beta} - \tilde{\boldsymbol{b}}||_2 + \mathbf{1}\,\frac{v([n]) - v(\emptyset)}{n}$$

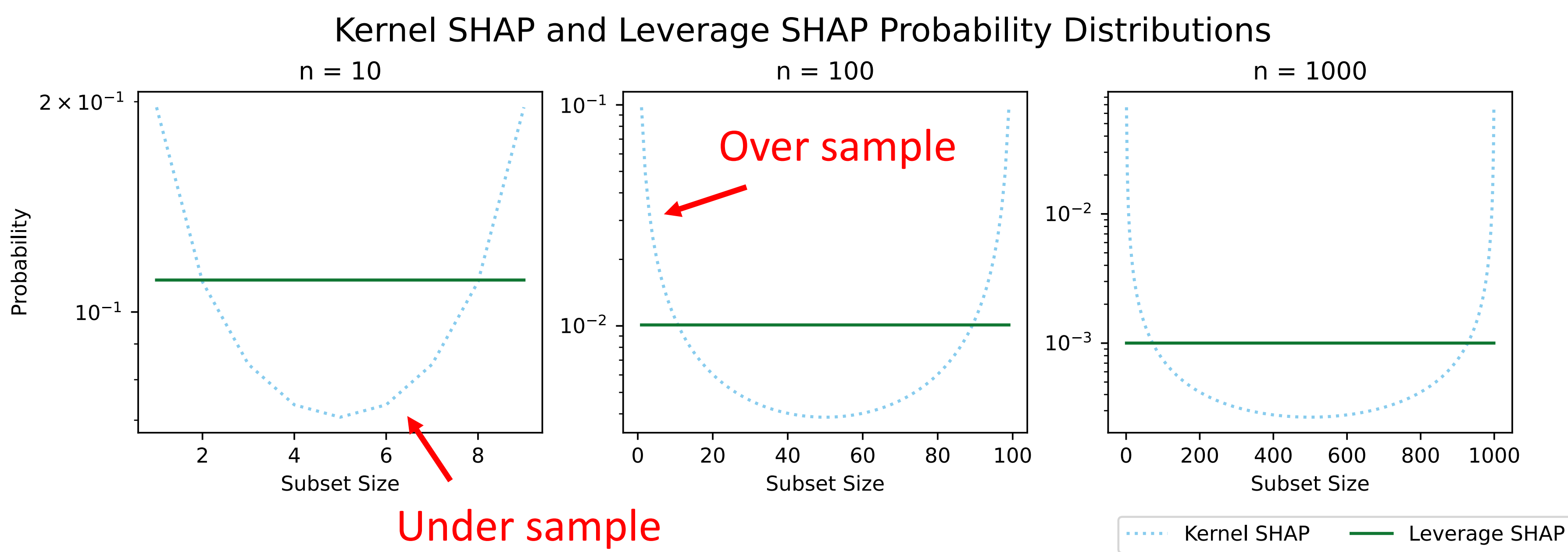


Leverage SHAP selects row $\boldsymbol{z}$ with probability proportional to leverage score!

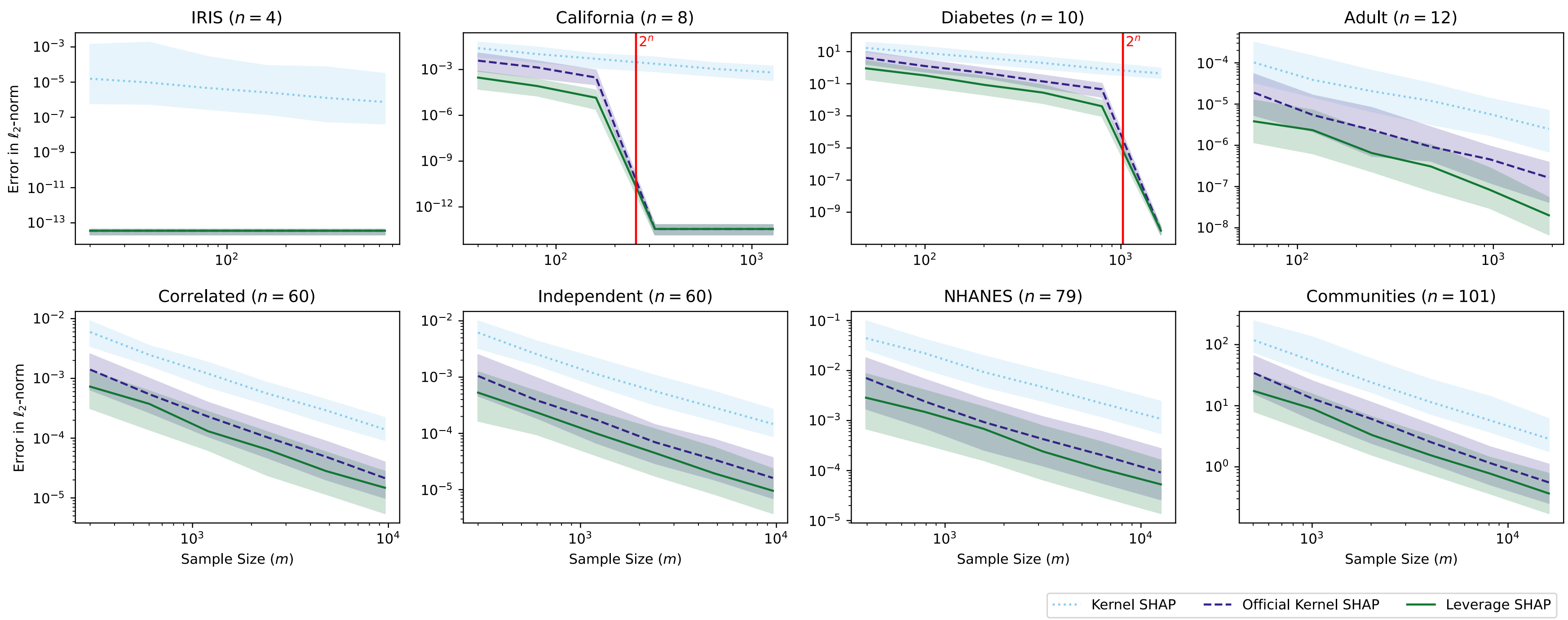
+ Paired Sampling

+ Bernoulli Sampling

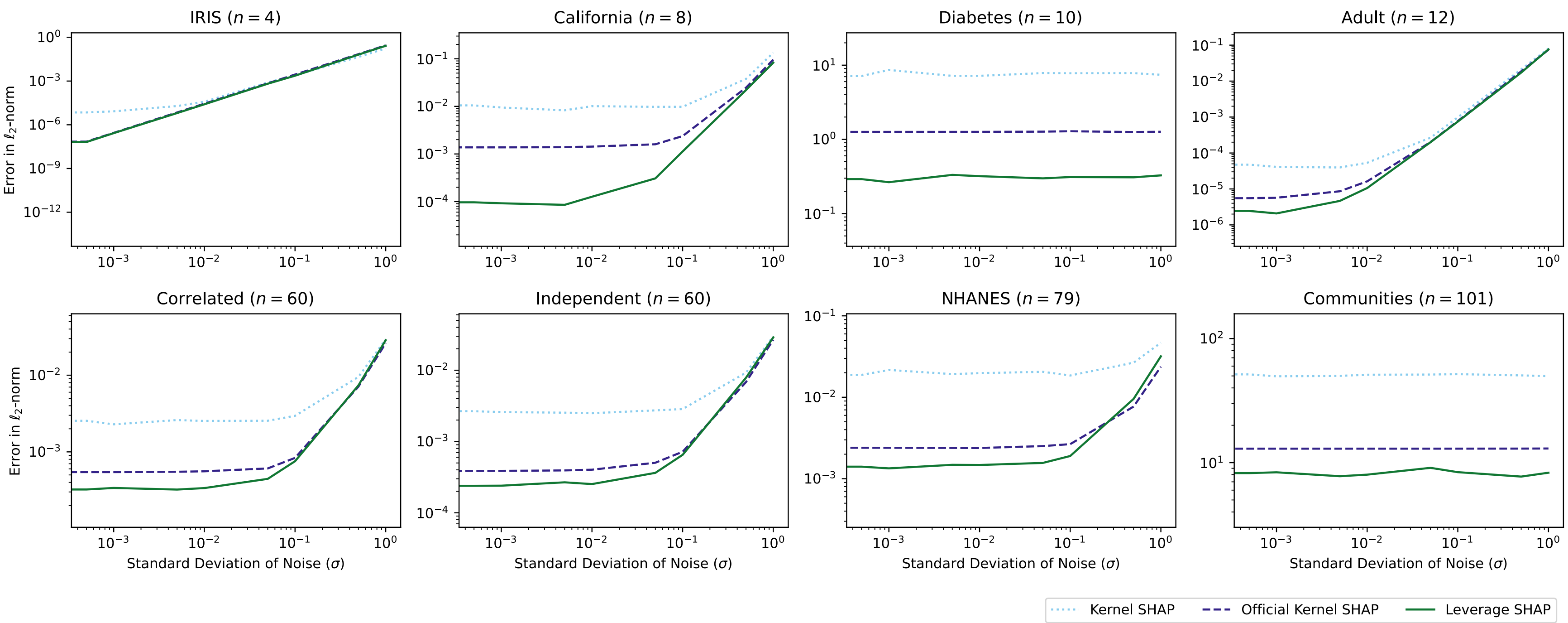
Leverage SHAP vs Kernel SHAP Probabilities



Accuracy by Sample Size



Accuracy by Noise



Robustness is useful, e.g., $v(S) = \mathbb{E}_{\boldsymbol{x}^S}[f(\boldsymbol{x}^S)]$

Theoretical Guarantee

As long as $m = \tilde{O}\left(\frac{n}{\epsilon}\right)$, the Leverage SHAP solution $\widetilde{\boldsymbol{\phi}}$ satisfies

$$||\boldsymbol{A}\widetilde{\boldsymbol{\phi}} - \boldsymbol{b}||_2^2 \leq (1+\epsilon)||\boldsymbol{A}\boldsymbol{\phi} - \boldsymbol{b}||_2^2$$

with high probability

Guarantee similar to standard leverage analysis but proof requires

- Modifications for paired sampling
- Modifications for sampling without replacement

Interpretable Corollary

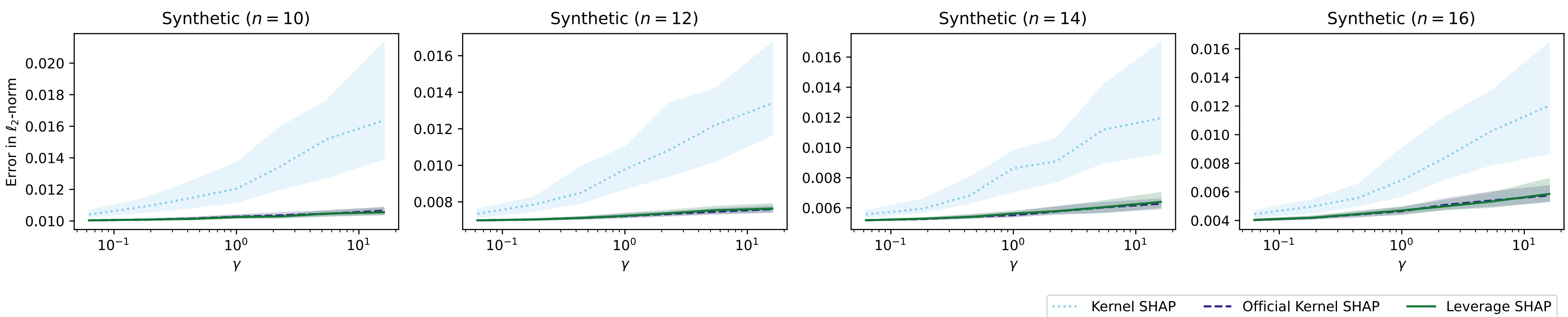
As long as $m = \tilde{O}\left(\frac{n}{\epsilon}\right)$, the Leverage SHAP solution $\widetilde{\boldsymbol{\phi}}$ satisfies

$$||\widetilde{\boldsymbol{\phi}} - \boldsymbol{\phi}||_2^2 \leq \epsilon\, \gamma ||\boldsymbol{\phi}||_2^2$$

with high probability where $\gamma = \frac{||\boldsymbol{A}\, \boldsymbol{\phi} - \boldsymbol{b}||_2^2}{||\boldsymbol{A}\, \boldsymbol{\phi}||_2^2} \in [0, \infty)$

Intuition: We can find $\widetilde{\boldsymbol{\phi}}$ close to the optimal in objective value but, when optimal solution is bad, $\widetilde{\boldsymbol{\phi}}$ will be far from $\boldsymbol{\phi}$

γ in Practice*



Takeaway: γ is a parameter of regression (not artifact of analysis)

*Computing γ requires exponential time so experiments are small

Banzhaf Values & Leverage Score Sampling

Joint work with

Yurong Liu

NYU

Flip Korn

Google

Tarfah Alrashed

Google

Dimitris Paparas

Google

Juliana Freire

NYU

Desirable Properties

Dummy: *If a feature never changes the prediction, then its attribution value is 0*

Symmetry: *If two features always induce the same change, then their attribution values are the same*

Linearity: *For two predictive functions, the attribution value of a feature in the combined function is the sum of the attribution values for each function*

2-Efficiency: *If two features are combined, their combined attribution values is the sum of their separate attribution values*

$\Longleftrightarrow$ Banzhaf values!

Banzhaf Values

For a set function $v: 2^{[n]} \to \mathbb{R}$, the ith Banzhaf value is

$$\phi_i = \frac{1}{2^{n-1}} \sum_{S \in [n] \setminus \{i\}} v(S \cup \{i\}) - v(S)$$

Banzhaf values are

- Simpler
- Empirically easier to approximate

Estimating Banzhaf values

Monte Carlo (MC): Sample $S, S \cup \{i\}$ to use $v(S \cup \{i\}) - v(S)$

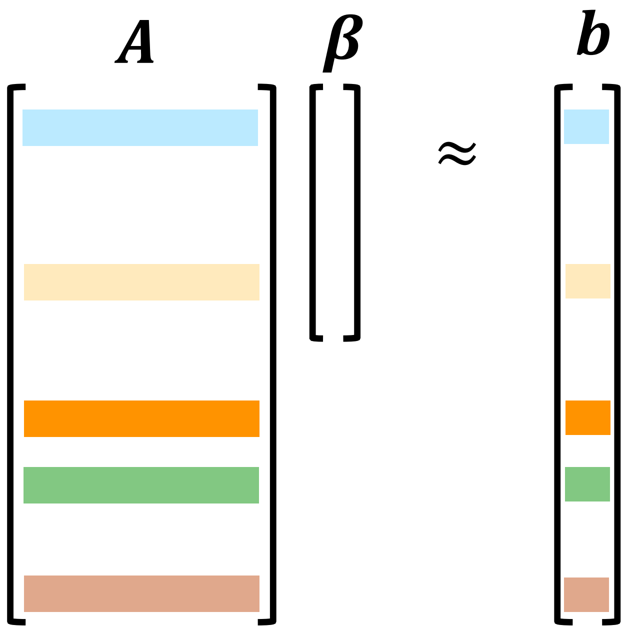
… but samples only used for one Shapley value

Maximum Sampling Reuse (MSR): Sample S to either add/subtract $v(S)$ for all i

… but magnitude of $v(S)$ is much larger than magnitude of $v(S \cup \{i\}) - v(S)$

Regression Formulation

$$\boldsymbol{\phi} = \arg\min_{\boldsymbol{\beta}} ||\boldsymbol{A\beta} - \boldsymbol{b}||_2$$

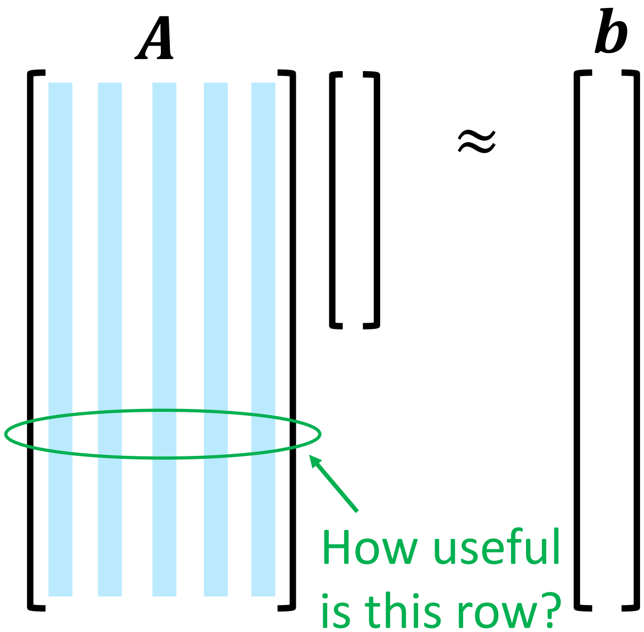


Special case known since 90's [HH 1992]

Each row/entry corresponds to binary vector

$$\boldsymbol{z} \in \left\{-\frac{1}{2}, \frac{1}{2}\right\}^n$$

Leverage Scores and Banzhaf values



Row $\boldsymbol{z}$ has “leverage”:

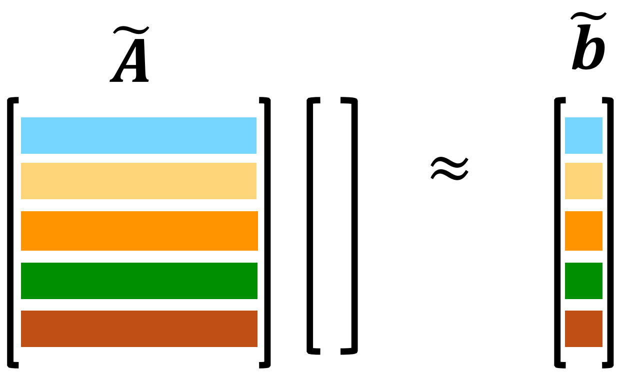
$$\ell_z = \boldsymbol{A}_{\boldsymbol{z}}(\boldsymbol{A}_{\boldsymbol{z}}^\top \boldsymbol{A}_{\boldsymbol{z}})^+ \boldsymbol{A}_{\boldsymbol{z}}^\top$$

$$\ell_z = \frac{n}{2^n}$$

Very similar to weighting in Banzhaf value definition!

Kernel Banzhaf

$$\tilde{\boldsymbol{\phi}} = \arg\min_{\boldsymbol{\beta}} ||\tilde{\boldsymbol{A}}\boldsymbol{\beta} - \tilde{\boldsymbol{b}}||_2$$

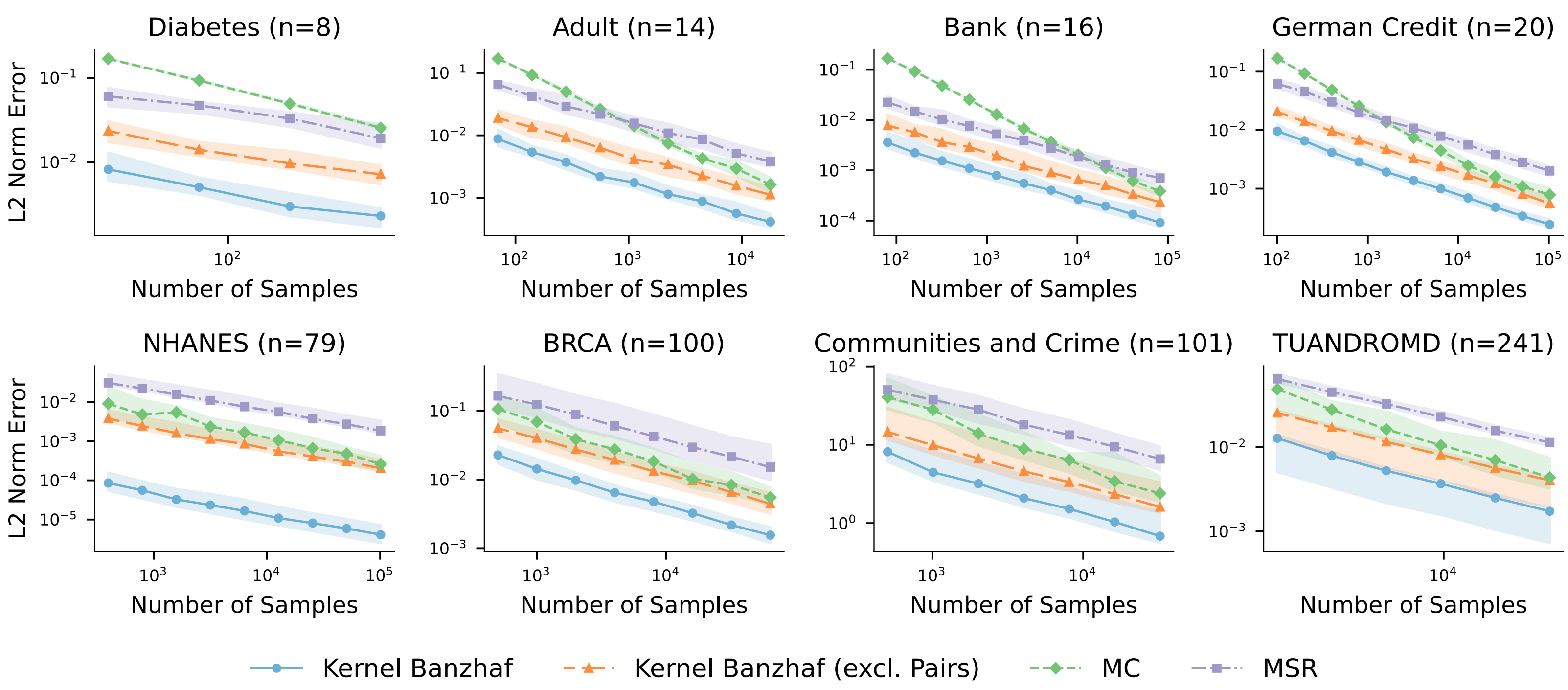


Kernel SHAP selects row $\boldsymbol{z}$ with probability proportional to leverage score!

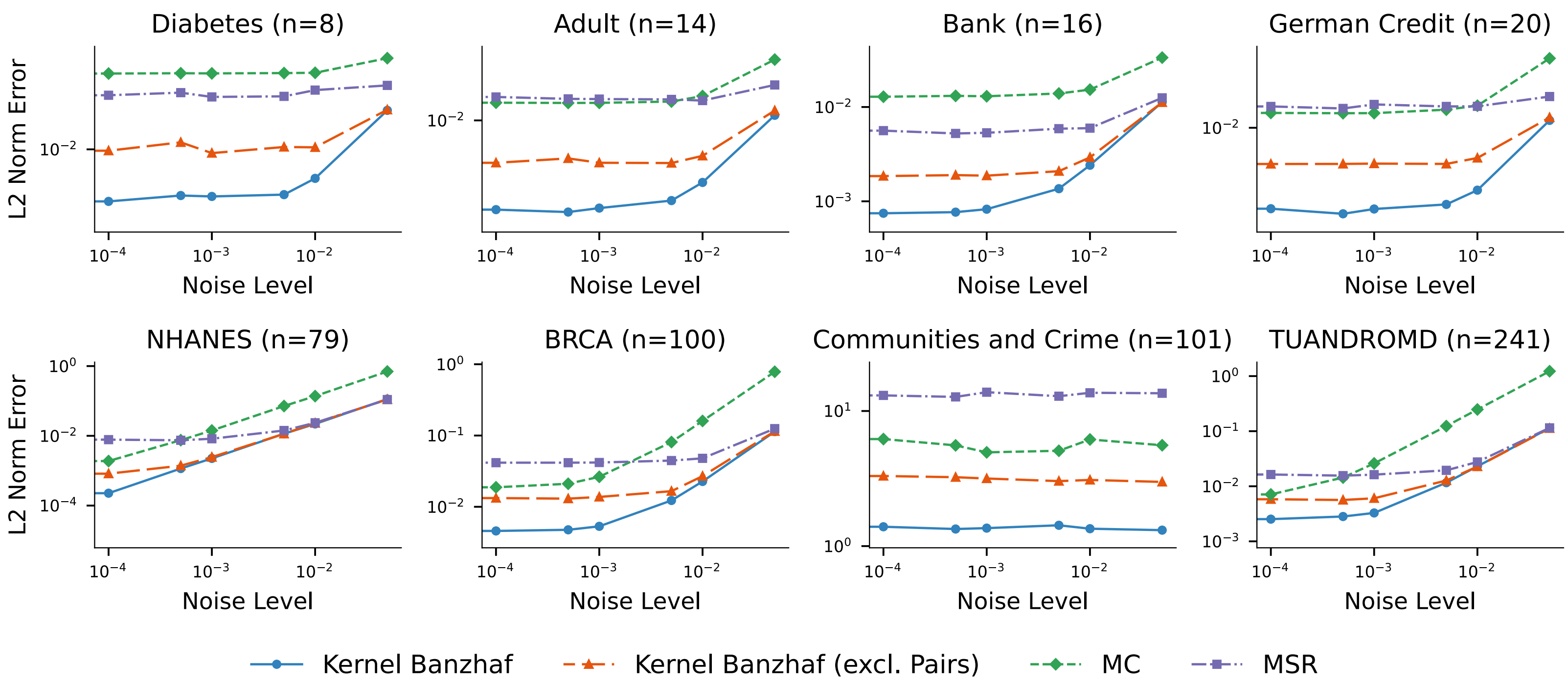
+ Paired Sampling

(But not Bernoulli Sampling)

Accuracy by Number of Samples



Accuracy by Noise



Theoretical Guarantees

As long as $m = \tilde{O}\left(\frac{n}{\epsilon}\right)$, the Kernel Banzhaf solution $\widetilde{\boldsymbol{\phi}}$ satisfies

$$||\boldsymbol{A}\widetilde{\boldsymbol{\phi}} - \boldsymbol{b}||_2^2 \leq (1+\epsilon)||\boldsymbol{A}\boldsymbol{\phi} - \boldsymbol{b}||_2^2$$

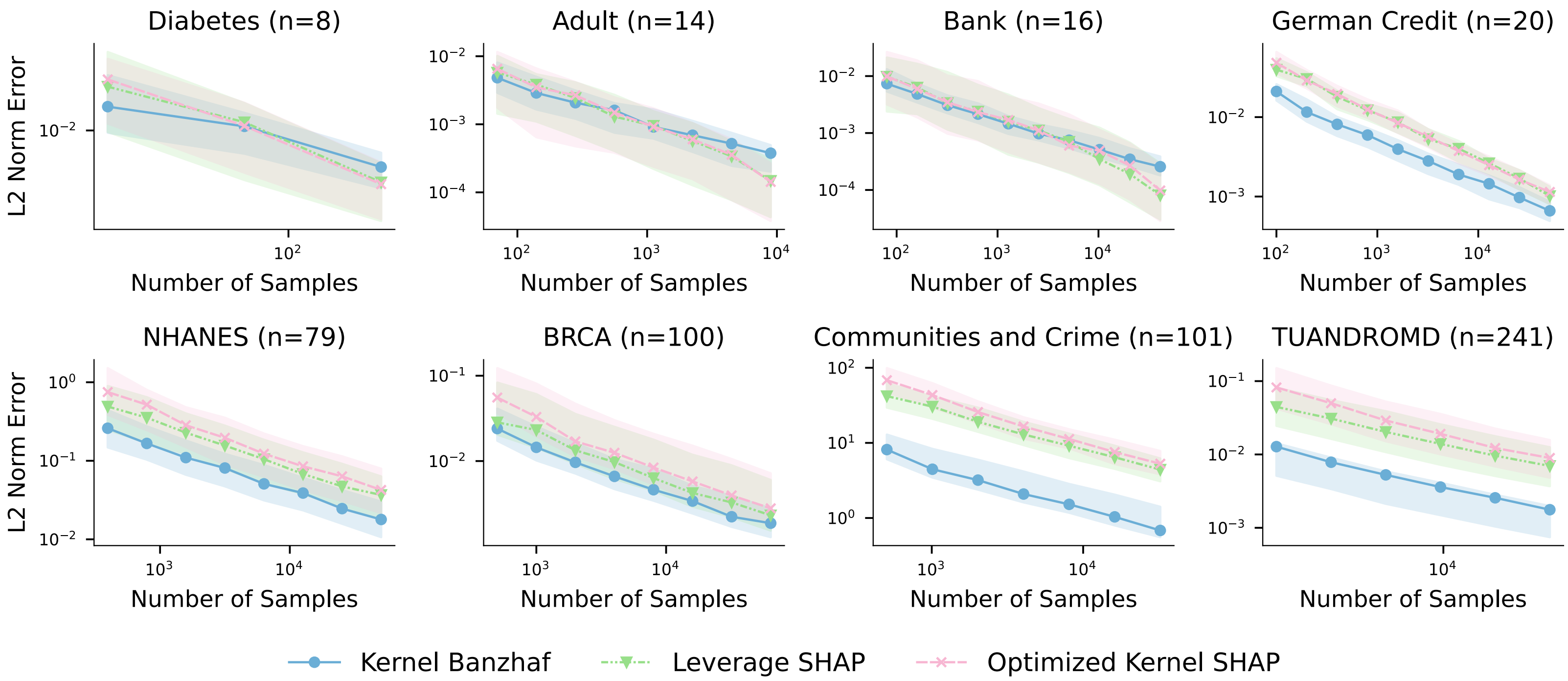
and, for $\gamma = \frac{||\boldsymbol{A}\ \boldsymbol{\phi} - \boldsymbol{b}||_2^2}{||\boldsymbol{A}\ \boldsymbol{\phi}||_2^2} \in [0, \infty)$, $\Updownarrow$ Equivalent for Banzhaf values

$$||\widetilde{\boldsymbol{\phi}} - \boldsymbol{\phi}||_2^2 \leq \epsilon\, \gamma ||\boldsymbol{\phi}||_2^2$$

with high probability

Guarantee similar to standard leverage analysis but proof requires modifications for paired sampling

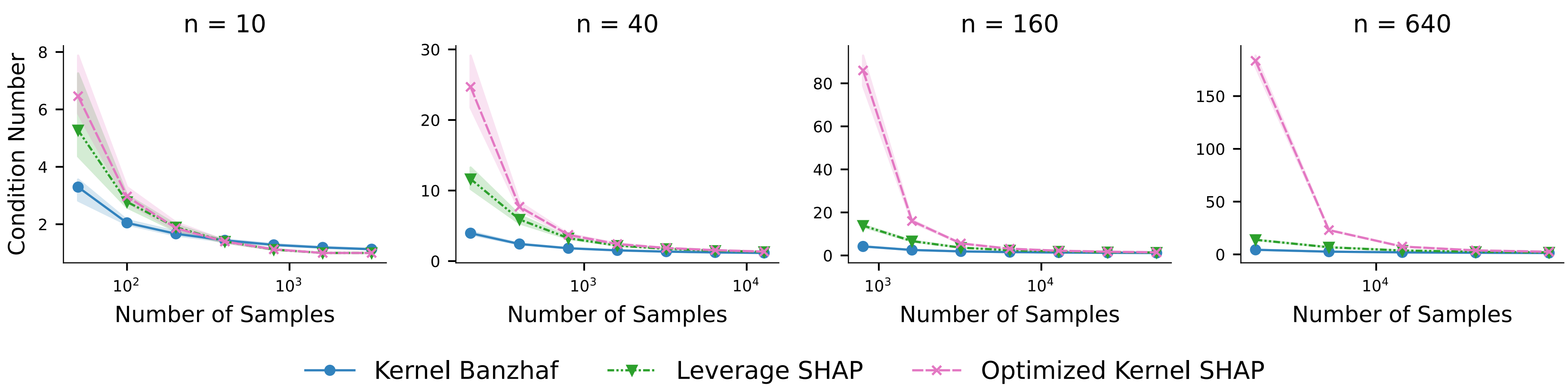
Shapley vs Banzhaf Estimators



Why do Leverage SHAP and Kernel Banzhaf perform differently?

Condition Number of $\tilde{A}$

Subsampled Banzhaf problem is more well-conditioned.



Since the full problem is well-conditioned, a subsampled problem with large condition number means weights won't generalize.

Thank you!

Any questions or comments?

Let me know! I’m submitting both papers tonight!

Any ideas for future work?

At 7am ET tomorrow, I’ll start looking for new projects ☺

Any feedback on presentation?

I plan to present this as my job talk