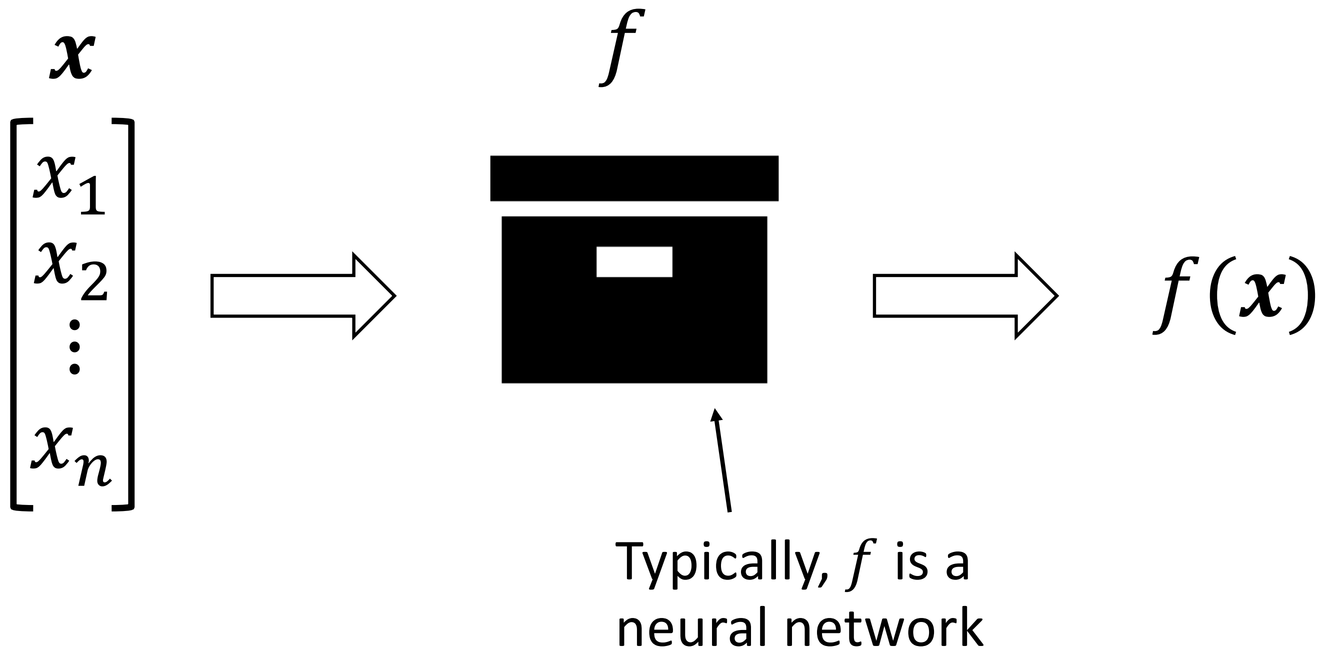
Provably Accurate Shapley Value Estimation via Leverage Score Sampling

Christopher Musco & R. Teal Witter

ICLR 2025 (Spotlight)

Explaining AI Outputs



Shapley Values

"Because of x_i, the model output $f(\boldsymbol{x})$ increased by ϕ_i."

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

Prediction **with** feature i

Prediction **without** feature i

Technically, $v(S) = \mathbb{E}[\mathrm{f}(\mathbf{x}^S)]$ where $x_i^S = \begin{cases} x_i, & i \in S \\ \text{sampled}, & i \notin S \end{cases}$

Computing Shapley Values

"Because of x_i, the model output $f(\boldsymbol{x})$ increased by ϕ_i."

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

Challenge: Exponentially many terms for each Shapley value!

Today: *Regression-based estimators*

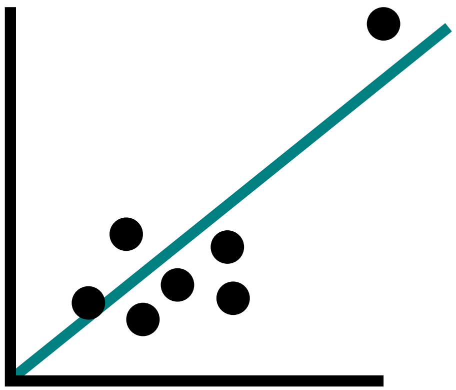
Regression Formulation

Lemma [CGKR '88]:

$$\begin{bmatrix}\phi_1\\ \phi_2\\ \vdots\\ \phi_n\end{bmatrix} = \text{argmin}_{\boldsymbol{\beta}} \left|\left|\boldsymbol{A}\boldsymbol{\beta} - \boldsymbol{b}\right|\right|_2^2 = \left\| \underset{(2^n-2)\times n}{\overset{\boldsymbol{A}}{\begin{bmatrix} \quad \\ \quad \end{bmatrix}}} \underset{n}{\overset{\boldsymbol{\beta}}{\begin{bmatrix} \quad \end{bmatrix}}} - \underset{2^n - 2}{\overset{\boldsymbol{b}}{\begin{bmatrix} \quad \\ \quad \end{bmatrix}}} \right\|_{\mathbf{2}}^{\mathbf{2}}$$

Using the Regression Formulation

Goal: Sample only a few points and (approximately) recover the line



Kernel SHAP: samples each S w.p. $\frac{1}{\binom{n}{|S|}|S|(n-|S|)}$

Beyond Kernel SHAP

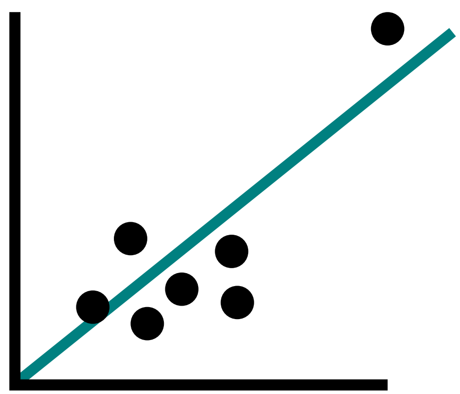
Question: How *should* we sample points?

Ideally, we want:

- Good performance
- Theoretical guarantees

Leverage Scores

Challenge of Sampling: Which points preserve the line?



Leverage Scores

Challenge of Sampling: Which points preserve the line?

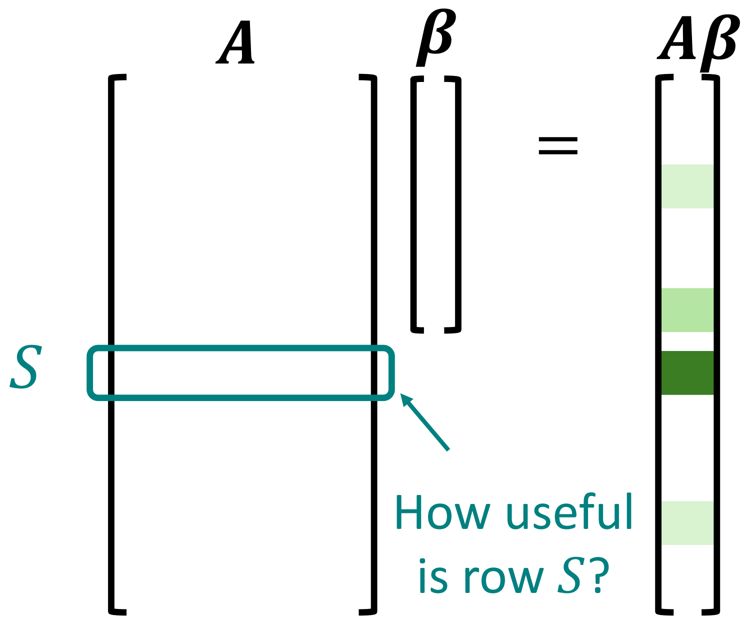
+ Without the high-leverage point, we find a *very* different line

Leverage Scores

Challenge of Sampling: Which points preserve the line?

+ With the high-leverage point, we find a close line

Leverage Scores



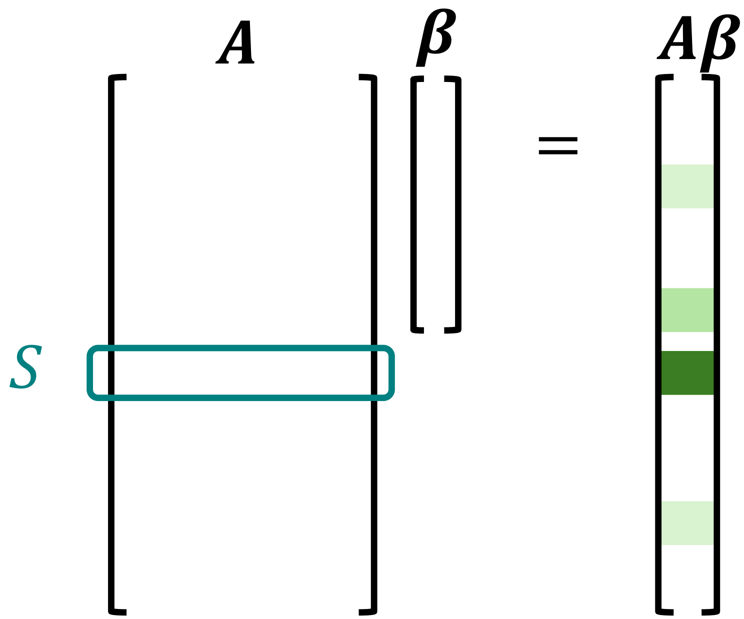
How useful is row S?

If there is an $\boldsymbol{A\beta}$ like this, then row S is "by itself".

Row S has "leverage":

$$\ell_S = \max_{\boldsymbol{\beta}} \frac{(\boldsymbol{A\beta})^2{}_S}{||\boldsymbol{A\beta}\,||_2^2}$$

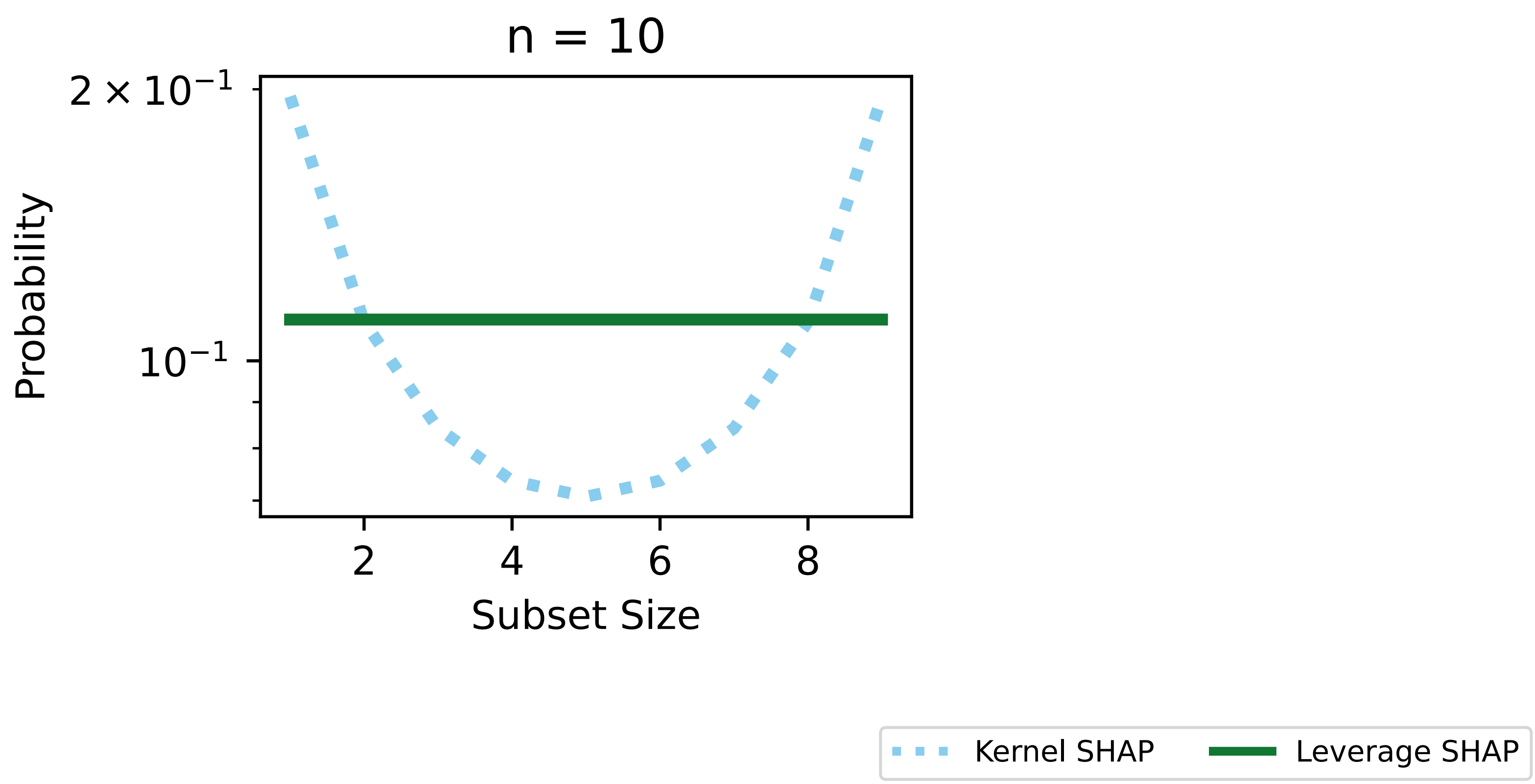
Leverage Scores



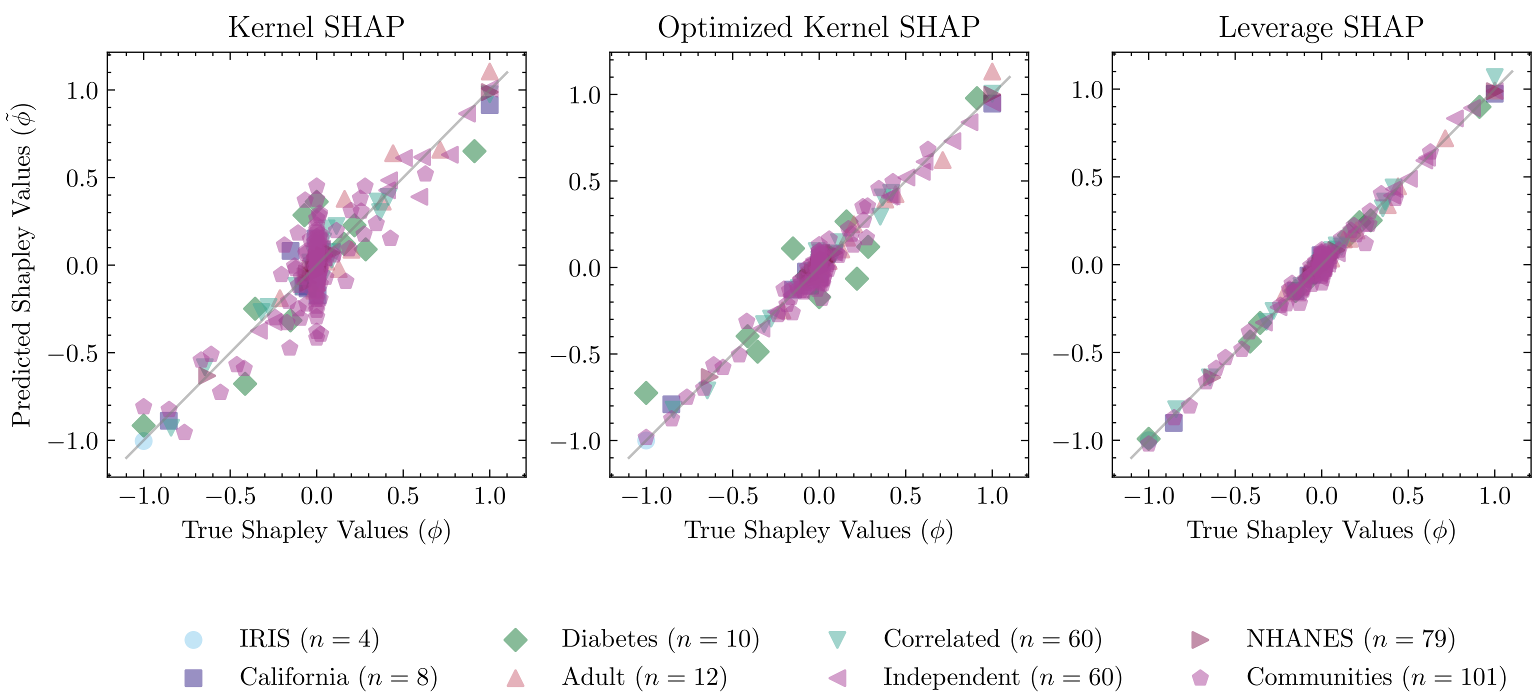
Row S has “leverage”:

$$\ell_S = \max_{\beta} \frac{(A\beta)^2{}_S}{||A\beta||_2^2} = \frac{1}{\binom{n}{|S|}}$$

Leverage SHAP vs Kernel SHAP Sampling



Leverage SHAP vs Kernel SHAP Qualitatively



Leverage SHAP Guarantee

Lemma: Let $\gamma = \frac{||\boldsymbol{A}\,\boldsymbol{\phi} - \boldsymbol{b}||_2^2}{||\boldsymbol{A}\,\boldsymbol{\phi}||_2^2}$ and $\epsilon > 0$. With $O\left(n \log n + \frac{n}{\epsilon}\right)$ samples and with probability 99/100, the Leverage SHAP solution $\widetilde{\boldsymbol{\phi}}$ satisfies

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

Leverage SHAP Performance

ℓ_2-error: $||\boldsymbol{\phi} - \tilde{\boldsymbol{\phi}}||_2^2 / ||\boldsymbol{\phi}||_2^2$

	IRIS	California	Diabetes	Adult	Correlated	Independent	NHANES	Communities
Kernel SHAP								
Mean	0.026	0.0266	0.0553	0.0673	0.0465	0.0264	0.0604	0.12
1st Quartile	1.61e-05	0.00829	0.0116	0.0182	0.0244	0.0134	0.0202	0.0563
2nd Quartile	0.000898	0.0236	0.0229	0.0345	0.0404	0.0217	0.0388	0.089
3rd Quartile	0.0328	0.0424	0.0524	0.0936	0.056	0.0303	0.0843	0.149
Optimized Kernel SHAP								
Mean	4.84e-09	0.00342	0.0093	0.00989	0.0117	0.00474	0.00758	0.0233
1st Quartile	1.66e-13	0.000802	0.00161	0.00187	0.00499	0.00194	0.00156	0.00962
2nd Quartile	2.17e-13	0.00238	0.00356	0.00489	0.00916	0.00391	0.00425	0.0173
3rd Quartile	2.69e-10	0.00489	0.00868	0.0122	0.015	0.00695	0.00871	0.0325
Leverage SHAP								
Mean	4.84e-09	0.000311	0.0023	0.00477	0.00716	0.00288	0.00532	0.0156
1st Quartile	1.66e-13	4.47e-05	0.000215	0.000477	0.00289	0.000843	0.000995	0.0062
2nd Quartile	2.17e-13	0.000133	0.000969	0.00124	0.00528	0.00257	0.00288	0.0104
3rd Quartile	2.69e-10	0.000366	0.00241	0.00354	0.00891	0.00417	0.00554	0.0225

In the paper…

- Performance by sample size
- Performance by noise
- Exploration of γ in theoretical guarantee
- Ablation experiments
- More ☺

Thank you!!

rtealwitter@gmail.com