Finding Heavily-Weighted Features with the Weight-Median Sketch
Kai Sheng Tai, Vatsal Sharan, Peter Bailis, Gregory Valiant
Stanford University

**Background**

**Sketches**
- Sketches are useful for approximate query processing under memory constraints
- Examples: top-$k$ frequent items, quantiles, # of distinct items

**Memory-Constrained Online Learning**
- ML on mobile devices, wearables, appliances: update models online on locally-observed data
- Examples: language modeling on phones, human activity classifiers on wearables

**Problems**

**High-Dimensional Feature Spaces**
- Memory usage can increase over time as new features are observed
- Example: Spam classifier on text stream continually observes new n-grams
- Over time, models can grow to exceed memory constraints

**Loss of Interpretability**
- Feature hashing fixes the feature dimension, but at cost of interpretability
- Difficult to debug and perform analyses on learned weights

**The Weight-Median Sketch (WM-Sketch)**

**Sketched Linear Classifier**
- Learns a compressed classifier within a fixed space budget
- Input feature vectors $x_t$ compressed via random projection

**Efficient Weight Recovery**
- Supports efficient estimation of high-magnitude weights
- Identify most influential features in model

**Algorithm**

**Random Projection**
- Sparse Count-Sketch projection $R$
- Efficiently implemented via hashing

**Updates**
- Project example $x_t \mapsto Rx_t$
- Update sketch state using gradient descent

**Estimating weight of feature $i$**
- Take median of buckets that feature $i$ hashes to (+ random sign flips)
- Maintain heap of top-$K$ weights

**Theoretical Guarantees**

**Theorem (informal)** Given feature dimension $d$, failure probability $\delta$, $\epsilon > 0$, set sketch size $k = O\left(\epsilon^{-1} \log^4 (d/\delta)\right)$ and sketch depth $s = O\left(\epsilon^{-2} \log^2 (d/\delta)\right)$. Then:

$$\|w_\star - \hat{w}_\text{est}\|_\infty \leq \epsilon \|w_\star\|_1,$$

where $w_\star$ is the optimal weight vector and $\hat{w}_\text{est}$ are the recovered weights.

**Remark.** Sketch size has only polylogarithmic dependence on $d$!

**Evaluation**

- Lower error in weight recovery than baselines
- Comparable classification accuracy with best baseline methods
- Consistent improvement in classification accuracy vs. feature hashing

**Applications**

- Enables streaming analyses that can be formulated as classification tasks
- Online feature selection
- Streaming data explanation
- Detecting large relative differences between data streams
- Streaming pointwise mutual information for finding highly-correlated pairs

**Paper:** arXiv:1711.02305