Streaming Weak Submodularity: Interpreting Neural Networks on the Fly

Ethan R. Elenberg, Alexandros G. Dimakis — The University of Texas at Austin
Moran Feldman — The Open University of Israel
Amin Karbasi — Yale University

BACKGROUND

- **Streaming Algorithm:**
  - One pass over N input elements
  - Maintain at most $o(N)$ elements in memory
  - Worst case/random stream order
  - Randomized/deterministic algorithm
  - Approximation ratio $\mathbb{E}[f(S)] \geq R \cdot f(OPT)$

$$\max_{|S| \leq k} f(S)$$

- **Assumptions:**
  - Nonnegative $f(A) \geq 0, \forall A$
  - Monotone $f(B \mid A) \geq 0, \quad \forall A, B$
  - γ_k-weakly submodular $\gamma_r \triangleq \min_{\substack{L, S \subseteq \mathcal{N}: \\ |L|, |S \setminus L| \leq r}} \frac{\sum_{j \in S \setminus L} f(j \mid L)}{f(S \mid L)}$

PROOF TECHNIQUES

- **Example Function:** $f_k(S) = \min\{2 \cdot |S \cap U| + 1, \;\; 2 \cdot |S \cap V|\}$

$U = \{u_1, \ldots, u_k\}$ $\quad V = \{v_1, \ldots, v_k\}$ $\quad D = \{w_1, \ldots, w_d\}$

$$(\gamma_{2k} = 0.5 \;\;, \;\; f_{\max} = 2k)$$

  - Worst case order begins with only elements from $U \cup D$
  - Sublinear streaming algorithms must drop many u before any v arrive
  - Approximation ratio is arbitrarily small for large k
- **Approximation Ratios:**
  - Let $\mathcal{E}$ be the event $f(S) < \tau$ (balanced if $\Pr[\mathcal{E}] = 2 - \sqrt{2 - e^{-\gamma/2}}$)

$$\mathbb{E}[f(S)] \geq (1 - \Pr[\mathcal{E}]) \cdot \tau$$

$$\mathbb{E}[f(S)] \geq \frac{1}{2} \cdot \left(\gamma \cdot [\Pr[\mathcal{E}] - e^{-\gamma/2}] \cdot f(OPT) - 2\tau\right)$$

  - Show one instance is guaranteed to be a good approximation

FUTURE WORK

- Tighten approximation bounds
- Analyze additional classes of algorithms: randomized, γ input
- Combinatorial interpretability for fairness, adversarial examples, ...

SUMMARY

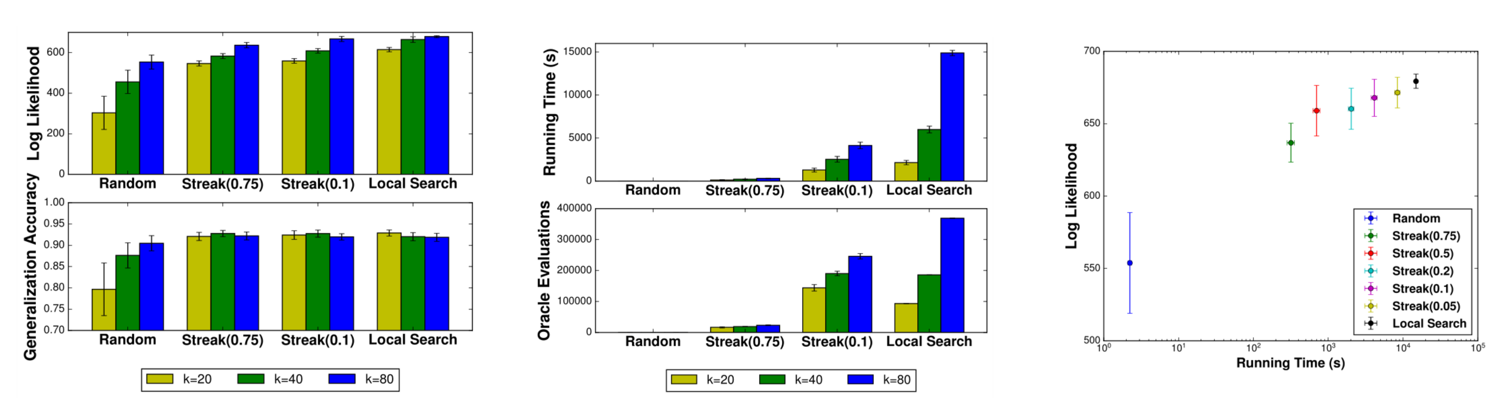
Many discrete optimization applications have a very large ground set or an expensive function evaluation oracle. We design and analyze streaming algorithms for the general class of *weakly submodular* set functions:

- **Worst case stream order:** No randomized streaming algorithm using sublinear memory can maximize a 0.5-weakly submodular function with constant approximation ratio
- **Random stream order:** Greedy, deterministic streaming algorithm for weak submodular maximization with constant approximation ratio
- **Experimental Evaluation:** Nonlinear sparse regression and interpretability of black-box neural networks

EXPERIMENTAL RESULTS

Sparse logistic regression: Compute pairwise products of features as needed

Phishing Dataset, N≈4.7k, 40 iterations

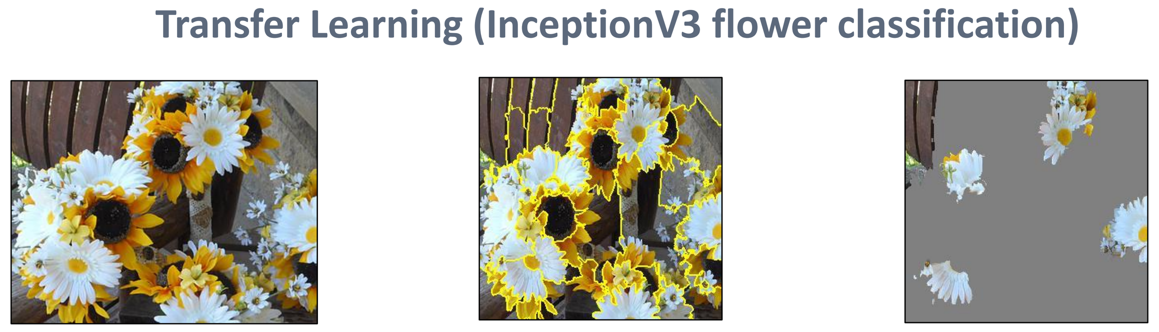


Performance Cost Tradeoff

Interpretability: Select image segments which maximize label's likelihood

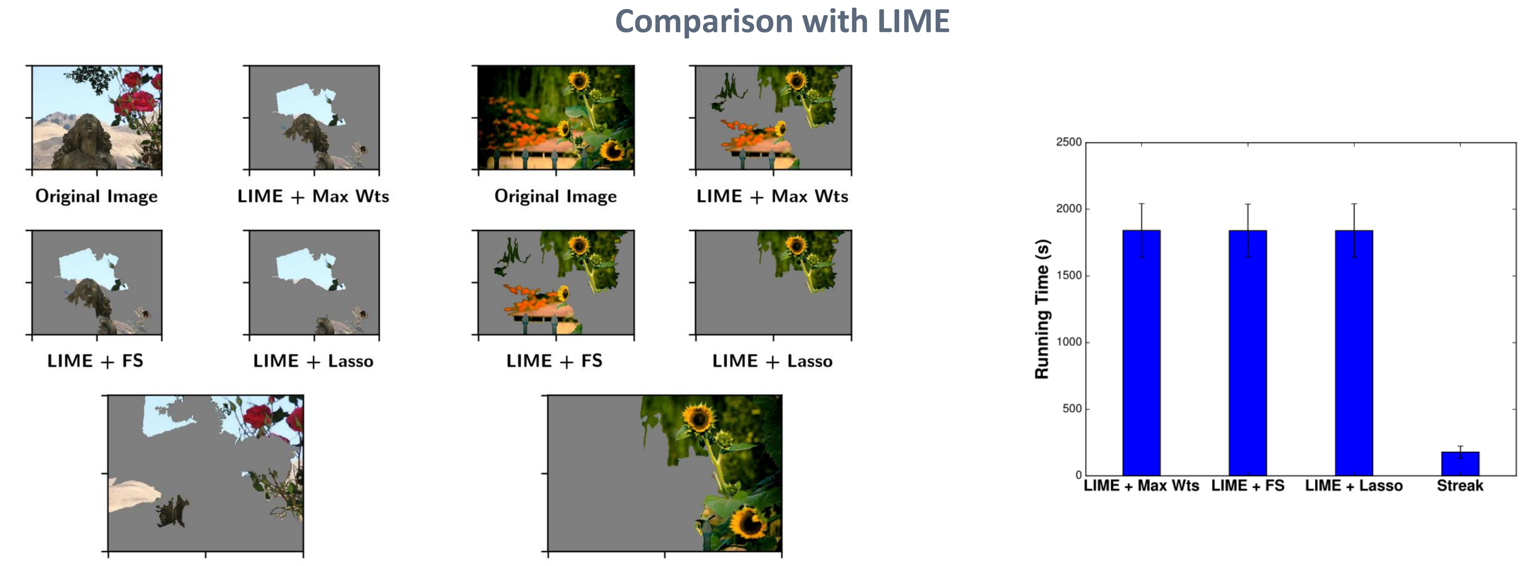
$$\max_{|S| \leq k} \text{softmax\_score}(\text{Image}_S)$$

Transfer Learning (InceptionV3 flower classification)



Original Image Segmented Image Interpretation for "daisy" (top label)

Comparison with LIME



STREAMING GREEDY ALGORITHMS

Discrete Derivative of a test element w.r.t. current solution:

$$f(i \mid A) \triangleq f(A \cup i) - f(A)$$

ThresholdGreedy

- **Initialize** $S = \varnothing$
- **Add** incoming element u if discrete derivative exceeds threshold

$$|S| < k \quad \text{and} \quad f(u \mid S) \geq \tau / k$$

STREAK

- **Compute** running maximum singleton $f(u_m) = m$
- **Run** and update $\mathcal{O}(\varepsilon^{-1} \log k)$ instances of **ThresholdGreedy**, with exponentially spaced thresholds

$$\tau \in \{(1-\varepsilon)^i \mid i \in \mathbb{Z} \text{ and } (1-\varepsilon)m/(9k^2) \leq (1-\varepsilon)^i \leq mk\}$$

- **Return** the output of best instance or the best singleton

$$\max\{S_{I^*}, u_m\}$$

MAIN RESULTS

Worst Case Impossibility

- For every constant $c \in (0, 1]$, there exists a 0.5-weakly submodular set function $f(S)$ such that any randomized algorithm which uses $o(N)$ memory to solve $\max_{|S| \leq k} f(S)$ has an approximation ratio less than c .

Average Case Guarantees

Algorithm	ThresholdGreedy	Streak
Approximation Ratio	$\tau \cdot (\sqrt{2 - e^{-\gamma/2}} - 1)$	$(1-\varepsilon)\gamma \cdot \frac{3 - e^{-\gamma/2} - 2\sqrt{2 - e^{-\gamma/2}}}{2}$
Memory	$\mathcal{O}(k)$	$\mathcal{O}(\varepsilon^{-1} k \log k)$
Running Time	$\mathcal{O}(Nf)$	$\mathcal{O}(Nf\varepsilon^{-1} \log k)$

REFERENCES

[1] Ashwinkumar Badanidiyuru, Baharan Mirzasoleiman, Amin Karbasi, and Andreas Krause. "Streaming Submodular Meets Maximization: Massive Data Summarization on the Fly," in KDD, 2014.

[2] Abhimanyu Das and David Kempe. "Submodular meets Spectral: Greedy Algorithms for Subset Selection," in ICML, 2011.

[3] Ethan R. Elenberg, Rajiv Khanna, Alexandros G. Dimakis, and Sahand Negahban. "Restricted Strong Convexity Implies Weak Submodularity," in NIPS workshop on Learning in High Dimensions with Structure, 2016. https://arxiv.org/abs/1612.00804

[4] Marco Rulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You? Explaining the Predictions of Any Classifier," in KDD, 2016.