Adaptive Submodularity: A New Approach to Active Learning and Stochastic Optimization

Daniel Golovin
Joint with Andreas Krause & Debajyoti Ray
Influence in Social Networks
[Kempe, Kleinberg, & Tardos, KDD `03]

Who should get promotional offers?

\[ V = \{\text{Alice, Bob, Charlie, Daria, Eric, Fiona}\} \]

\[ F(A) = \text{Expected } \# \text{ of people influenced when targeting } A \]
Adaptively select promotion targets, see which of their friends are influenced

• Goal: Find optimal policy, rather than a set
Adaptive Submodularity

[G & Krause, COLT 2010]

Objective function $f(A, x_V)$
- $A$: Set of actions you’ve done.
- $x_V$: realization of world state (from known prior $P$)

Conditional Expected Marginal Benefit:
\[ \Delta(e \mid x_A) := \mathbb{E}[f(A \cup \{e\}, x_V) - f(A, x_V) \mid x_A] \]

Adaptive Submodularity:
\[ \Delta(e \mid x_A) \geq \Delta(e \mid x_B) \text{ for all } x_A \preceq x_B \]

“$x_B$ contains all observations in $x_A$”

Adaptive Monotonicity:
\[ \Delta(e \mid x_A) \geq 0 \text{ for all } e, x_A \]
What’s it good for?

Fundamental new tool for designing and analyzing greedy near-optimal algorithms for

- active learning,
- experimental design,
- and many other adaptive optimization problems

It allows us to “lift” many results for submodular function optimization to the adaptive realm.
Example: Submodular Maximization under a Cardinality Constraint

Problem: Find \( A^* = \arg \max \{ f(A) : |A| \leq k \} \)

Initialize \( A = \emptyset \)
For \( i = 1, 2, \ldots, k \)
\[ e_i = \arg \max_e \{ f(A \cup \{ e \}) - f(A) \} \]
\( A = A \cup \{ e_i \} \)
Select \( e_i \)

Theorem [Nemhauser et al ‘78]
Given a monotone submodular function \( f \), \( f(\emptyset) = 0 \),
the greedy algorithm selects a set \( A_{\text{greedy}} \) such that
\[ f(A_{\text{greedy}}) \geq (1 - 1/e) \max_{|A| \leq k} f(A) \]
The Adaptive-Greedy Algorithm

Problem: Find $\pi^* = \arg \max \{F(\pi) : |\pi| \leq k\}$

Initialize $A = \emptyset$
For $i = 1, 2, \ldots, k$
    $e_i = \arg \max_e \Delta(e | x_A)$
    $A = A \cup \{e_i\}$
    Select $e_i$ and observe outcome $x_{e_i}$ for it.

Theorem [G & Krause, COLT ‘10]
Given an adaptive monotone submodular function $f$ with $f(\emptyset) = 0$, the adaptive greedy algorithm returns $\pi^{\text{greedy}}$ such that

$$F(\pi^{\text{greedy}}) \geq (1 - 1/e) \max_{\pi : |\pi| \leq k} F(\pi)$$
**Theorem:** Objective is adaptive monotone submodular.

Hence, adapt-greedy is a $(1 - 1/e) \approx 63\%$ approximation to the optimal policy.

We also get $(1 - 1/e)$ for Stochastic Submodular Maximization, generalizing [Asadpour et al, ‘08]
Stochastic Min Cost Cover

- Example: “Get 25% market penetration using min # of free phone giveaways”
- Adaptively get a threshold amount of value.
- Minimize expected number of actions.
- If objective is adapt-submod and monotone, we get a logarithmic approximation.

$$\ln(n) + 1$$ for Stochastic Set Cover

[Goemans & Vondrak, LATIN ‘06], [Liu et al., SIGMOD ‘08]

c.f., Interactive Submodular Set Cover [Guillory & Bilmes, ICML ‘10]
Optimal Decision Trees

- Prior over diseases $P(Y)$
- Likelihood of outcomes $P(X_v | Y)$
- Suppose that $P(X_v | Y)$ is deterministic (noise free)
- Each test eliminates hypotheses $y$
- How should we test to eliminate all incorrect hypotheses?

$$\Delta(t | x_A) = \mathbb{E} \left[ \begin{array}{c} \text{mass ruled out} \\ \text{by } t \text{ if we} \\ \text{know } x_A \end{array} \right]$$

“Generalized binary search”
Equivalent to max. info-gain
Objective = probability mass of hypotheses you have ruled out.

It’s Adaptive Submodular.

\[
b_0 := \mathbb{P}(\text{blue})
\]

\[
g_0 := \mathbb{P}(\text{green})
\]

\[
\Delta_{\text{init}}(x) = \frac{2g_0b_0}{g_0 + b_0}
\]

\[
b_1 := \mathbb{P}(\text{blue})
\]

\[
g_1 := \mathbb{P}(\text{green})
\]

\[
\Delta_{\text{final}}(x) = \frac{2g_1b_1}{g_1 + b_1}
\]

\[
b_0 \geq b_1, \quad g_0 \geq g_1
\]

Not hard to show \(\Delta_{\text{final}}(x) \leq \Delta_{\text{init}}(x)\)
Optimal Decision Trees

“Diagnose the patient as cheaply as possible (w.r.t. expected cost)"

Adaptive-Greedy is an \((\ln(1/p_{\text{min}}) + 1)\) approximation, where \(p_{\text{min}} := \min_h \{P(h)\}\)

Result requires that tests are exact (no noise)!

Garey & Graham, 1974; Loveland, 1985; Arkin et al., 1993; Kosaraju et al., 1999; Dasgupta, 2004; Guillory & Bilmes, 2009; Nowak, 2009; Gupta et al., 2010
In practice, observations are typically noisy

Results for noise-free case do not generalize

**Key problem**: Tests no longer rule out hypotheses (only make them less likely)

Intuitively, want to gather *enough information* to make the right decision!
Noisy active learning

Suppose we run all tests, i.e., see $X_V$

Best we can do is to maximize expected utility

$$a^* = \arg \max_a \sum_y P(y | x_V) U(a, y)$$

Key question:
How should we cheaply test to guarantee we choose $a^*$?

Existing approaches:
- Generalized binary search?
- Maximize information gain?
- Maximize value of information?

Not adaptive submodular in the noisy setting!

**Theorem:** All previous approaches require exponentially more tests than the optimum
A new criterion for nonmyopic VOI

- **Strategy**: Replace noisy problem with noiseless problem
- **Key idea**: Make test outcomes part of the hypothesis

\[ X_V = [X_1, X_2, X_3] \]

Test \( X_1 \)

Same high-level strategy used independently by Bellala et al., NIPS 2010
A new criterion for nonmyopic VOI

Only need to distinguish between noisy hypotheses that lead to different decisions!

Tests \([X_1, X_2, X_3]\)

Suppose we find \(X_1 = 1\)

Weight of edge = product of incident hypotheses' probabilities
Theoretical guarantees

**Theorem**: The edge-cutting objective is adaptive submodular and adaptive monotone.

Let $\rho_{\min} = \min\{P(x_V)\}$. Then

$$\text{Cost}(\pi_{\text{Greedy}}) \leq (2 \ln \left( \frac{1}{\rho_{\min}} \right) + 1) \text{Cost}(\pi^*)$$

First approximation guarantee for Bayesian active learning with noisy observations!
Example: The Iowa Gambling Task

Various competing theories on how people make decisions under uncertainty

- Maximize expected utility [von Neumann & Morgenstern]
- Prospect theory [Kahnemann & Tversky]
- Portfolio optimization [Hanoch & Levy]

How should we design tests to distinguish theories?

Which would you prefer?
Iowa Gambling as B.E.D.

Every possible test $X_s = (g_{s,1}, g_{s,2})$ is a pair of gambles

Theories parameterized by $\theta$

Each theory predicts utility for each gamble $U(g, y, \theta)$

$$P(X_s = 1 \mid y, \theta) = \frac{1}{1 + \exp(U(g_{s,1}, y, \theta) - U(g_{s,2}, y, \theta))}$$
Simulation Results

Adaptive submodular criterion outperforms existing approaches
Preliminary Experimental Results [in collaboration with Colin Camerer]

- Run designs on 11 naïve subjects
- Indication of heterogeneity in population
- Algorithm useful in real-time settings
Conclusions

- Fundamental new tool for the design & analysis of greedy, near-optimal adaptive algorithms
- Recovers and generalizes many known results in a unified manner. (We can also handle costs, and approximate or noisy maximization of $\Delta(e \mid x_v)$)
- Other benefits, e.g., speedup via lazy evaluations of $\Delta$, data dependent upper bounds, … (see the papers)