Budgeted stream-based active learning via adaptive submodular maximization

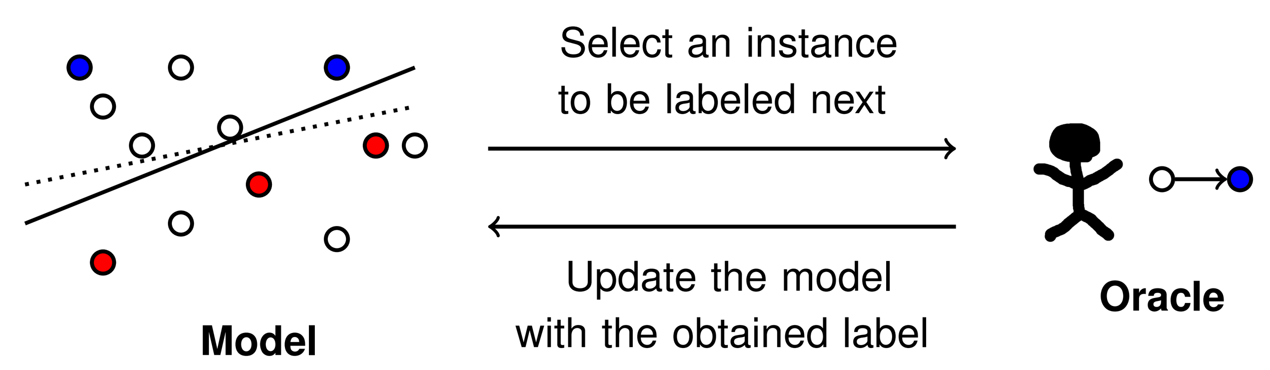
Kaito Fujii and Hisashi Kashima (Kyoto University)

Overview

- Stream-based active learning is reduced to **online adaptive stochastic optimization** problem.
- **Policy-adaptive submodularity**, a new property of stochastic functions, is proposed.
- **Constant-factor competitive algorithms** are proposed.

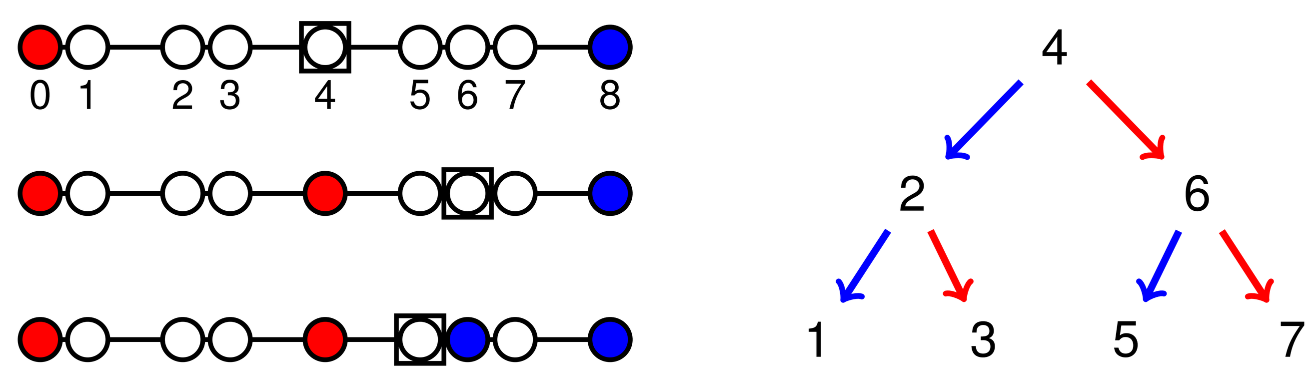
Adaptive Stochastic Optimization

The learner can query the labels of unlabeled instances.



- V a set of unlabeled instances, $\mathcal{Y}$ a set of possible labels.
- **A prior distribution** $p(\phi)$ on **oracle** $\phi\colon V \to \mathcal{Y}$ is given.

Look for the optimal policy, not the optimal subset.



A decision tree encoding an algorithm, called **policy**

Maximize the expected value of the objective function.

- The objective function $f\colon 2^V \times \mathcal{Y}^V \to \mathbb{R}_{\geq 0}$ returns quality of a set of selected instances with ϕ (the labels of all instances).

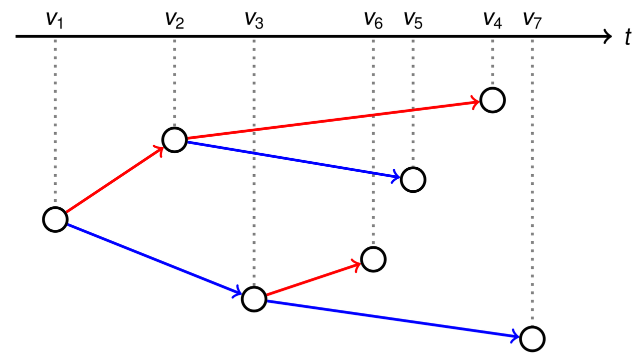
Maximize$_\pi$ $\boxed{\mathbb{E}_\phi[f(E(\pi,\phi),\phi)]}$ subject to $\boxed{|E(\pi,\phi)| \leq k \, (\forall \phi)}$

expected value of f — size constraint

where $E(\pi,\phi)$ is the instances selected by policy π under ϕ.

Online Adaptive Stochastic Optimization

In stream-based active learning, unlabeled instances arrive sequentially.



Select k instances to be labeled out of n unlabeled instances arriving in random order.

- **stream setting**: a limited amount of memory is available.
- **secretary setting**: irrevocable decision at each arrival.

Policy-Adaptive Submodularity

Diminishing return property of the gain of any policy.

Instances $A = \{v_1, \cdots, v_l\}$, Observations $\psi = \{(v_1, y_1), \cdots, (v_l, y_l)\}$

- **The expected marginal gain of an instance** v

$$\Delta(v|\psi) = \mathbb{E}_{\phi \sim p(\phi|\psi)}[f(A \cup \{v\}, \phi) - f(A,\phi)]$$

- **The expected marginal gain of a policy** π

$$\Delta(\pi|\psi) = \mathbb{E}_{\phi \sim p(\phi|\psi)}[f(A \cup E(\pi,\phi), \phi) - f(A,\phi)]$$

Submodularity

$f(A \cup \{v\}) - f(A) \geq f(B \cup \{v\}) - f(B)$ $(\forall A \subseteq B,\ v \in V \setminus B)$ $\longleftarrow$ / $\longrightarrow$ $f(A \cup P) - f(A) \geq f(B \cup P) - f(B)$ $(\forall A \subseteq B,\ P \subseteq V \setminus B)$

Extension to the adaptive setting

Adaptive Submodularity
$\Delta(v|\psi_A) \geq \Delta(v|\psi_B)$
$(\forall \psi_A \subseteq \psi_B,\ v \in V \setminus B)$

$\longleftarrow$ / $\not\longrightarrow$

Policy-Adaptive Submodularity
$\Delta(\pi|\psi_A) \geq \Delta(\pi|\psi_B)$
$(\forall \psi_A \subseteq \psi_B,\ \bigcup_\phi E(\pi,\phi) \subseteq V \setminus B)$

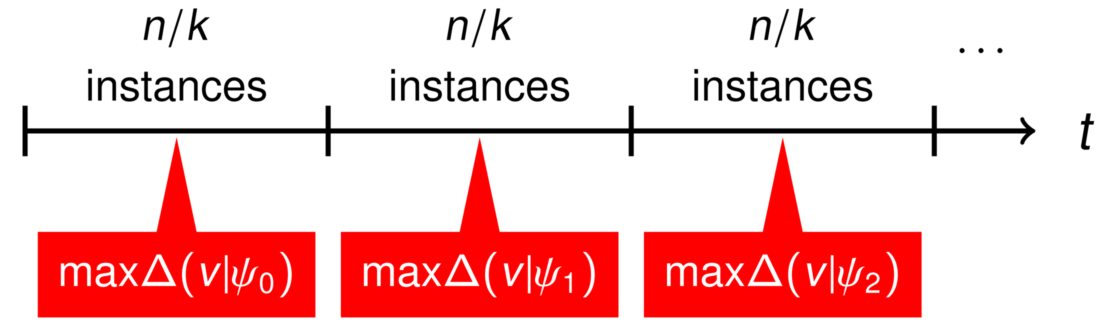
Many existing adaptive submodular functions satisfy it.

- Generalized binary search [Golovin-Krause'10]
- EC^2 [Golovin+'10]
- ALuMA [Gonen+'10]
- maximum Gibbs error criterion [Cuong+'13]
- In cases where labels of instances are independent

Algorithms

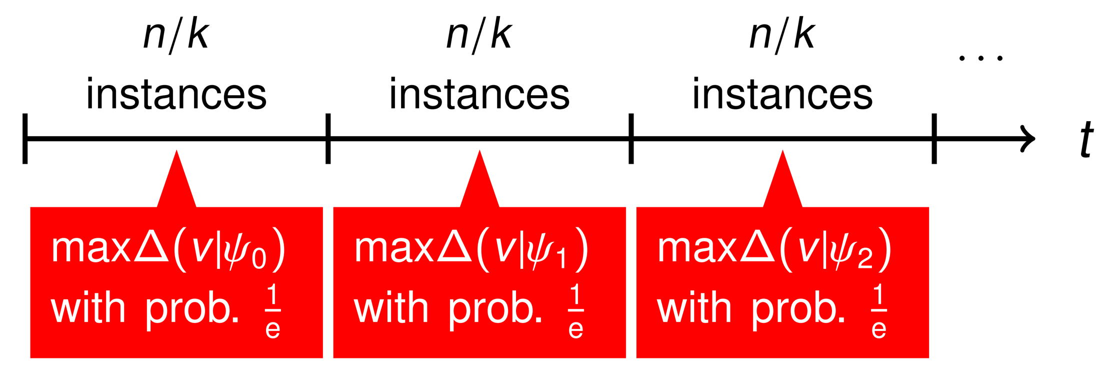
stream setting $(2-\sqrt{3})(1-1/e) \approx 0.16$-competitive

Select the instance of the largest expected marginal gain from each segment.



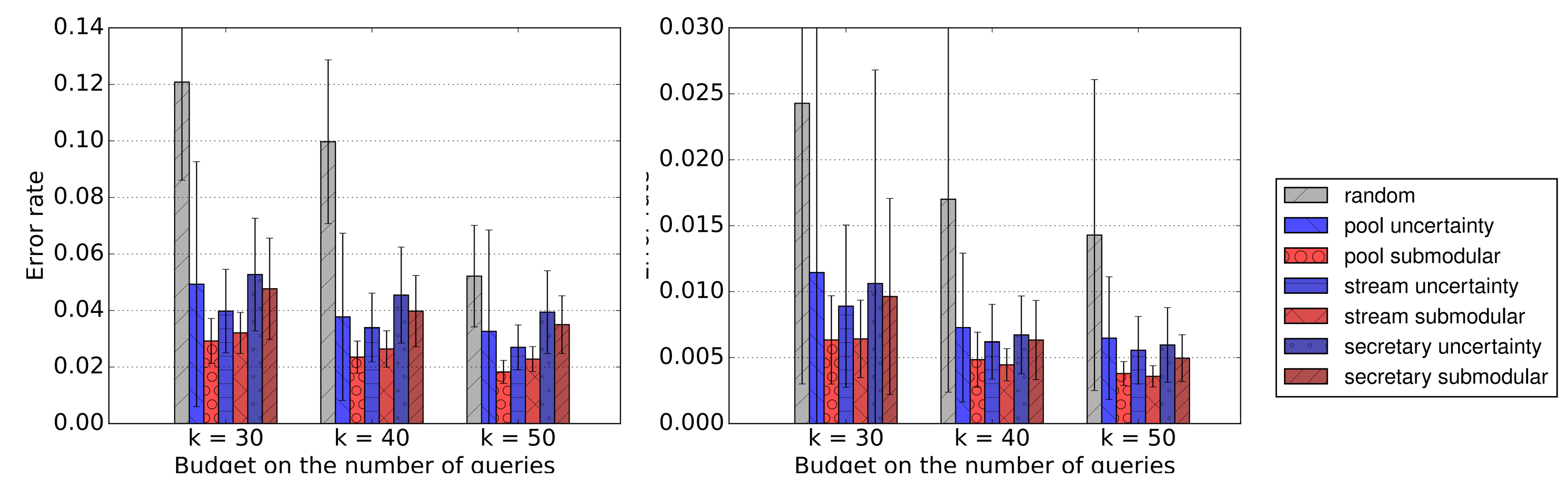
secretary setting $\frac{1-1/e}{2e\sqrt{1+2/e}} \approx 0.08$-competitive

Apply the classical secretary algorithm to each segment.



Experiments

- **Datasets**: WDBC (596 instances, 32 dimensions), MNIST (14780 instances, 10 dimensions (reduced by PCA))
- **Algorithms**: Compare our proposed methods (online versions of the noise-tolerant ALuMA algorithm [Gonen+'11, Chen-Krause'13]) with existing heuristics (random, uncertainty sampling)
- **Settings**: pool-based, stream and secretary settings.
- **Test classifier**: linear SVM for all methods.



(a) WDBC dataset (b) MNIST dataset