The Knowledge Gradient for Sequential Decision Making with Stochastic Binary Feedbacks

Yingfei Wang, Chu Wang and Warren B. Powell

Princeton University
Sequential Decision Problems

- $M$ discrete alternatives
- Unknown truth $\mu_x$
- Each time $n$, the learner chooses an alternative $x^n$, receives reward $W_{x^n}$.

  offline objective $\max \mathbb{E}^\pi [\mu_{x^N}]$

  online objective $\max \mathbb{E}^\pi \sum_{n=0}^{N-1} [\mu_{x^n}]$

Overview

- Numerous Communities
  - Multi-armed bandits
  - Ranking and selection
  - Stochastic search
  - Control theory
  - ......

- Various Applications
  - Recommendations: ads, news
  - Packet routing
  - Revenue management
  - Laboratory experiments guidance:
  - ......
Applications with binary outputs

- Revenue management: whether or not a customer books a room.
- Health analytics: success (patient does not need to return for more treatment) or failure (patient does need followup care).
- Production of single or double-walled nanotubes: controllable parameters: catalyst, laser power, Hydrogen, pressure, temperature, Ar/CO2, ethylene etc.
Applications with binary outputs

- Revenue management: whether or not a customer books a room.
- Health analytics: success (patient does not need to return for more treatment) or failure (patient does need followup care).
- Production of single or double-walled nanotubes: controllable parameters: catalyst, laser power, Hydrogen, pressure, temperature, Ar/CO2, ethylene etc.

Single Wall

Double Wall
Outline

1. Sequential Decision Problems with Binary Outputs
2. The Knowledge Gradient Policy
3. Experimental Results
1. **Sequential Decision Problems with Binary Outputs**
   Model
   Bayesian linear classification and Laplace approximation

2. **The Knowledge Gradient Policy**

3. **Experimental Results**
Model

- A finite set of alternatives $x \in \mathcal{X} = \{x_1, \ldots, x_M\}$.
- Binary outcome $y \in \{-1, +1\}$ with unknown probability $p(y = +1|x)$.
- Goal: given a limited budget $N$, choose the measurement policy $(x^0, \ldots, x^{N-1})$ and the implementation decision that maximizes $p(y = +1|x)$.
- Generalized linear model for modeling probability

$$p(y = +1|x, w) = \sigma(w^T x),$$

where $\sigma(a) = \frac{1}{1+\exp(-a)}$ or $\sigma(a) = \Phi(a) = \int_{-\infty}^{a} \mathcal{N}(s|0, 1^2)ds$. 
Logistic and probit regression

- Training set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$
- Likelihood $p(\mathcal{D}|\mathbf{w}) = \prod_{i=1}^n \sigma(y_i \cdot \mathbf{w}^T \mathbf{x}_i)$.
- $\hat{\mathbf{w}} = \arg \min_{\mathbf{w}} \sum_{i=1}^n - \log(\sigma(y_i \cdot \mathbf{w}^T \mathbf{x}_i))$.

Bayesian logistic and probit regression

- $p(\mathbf{w}|\mathcal{D}) = \frac{1}{Z} p(\mathcal{D}|\mathbf{w}) p(\mathbf{w}) \propto p(\mathbf{w}) \prod_{i=1}^n \sigma(y_i \cdot \mathbf{w}^T \mathbf{x}_i)$.
- Extend to leverage for sequential model updates:
  
  $p(\mathbf{w}|\mathcal{D}^0) \xrightarrow{\mathbf{x}^0,y^1} p(\mathbf{w}|\mathcal{D}^1) \xrightarrow{\mathbf{x}^1,y^2} p(\mathbf{w}|\mathcal{D}^2) \cdots$

- Exact Bayesian inference for linear classifier is intractable.
- Monte Carlo sampling or analytic approximations to the posterior: Laplace approximation.
Laplace approximation

- $\Psi(w) = \log p(D|w) + \log p(w)$.
- Second-order Taylor expansion to $\Psi$ around its MAP (maximum a posteriori) solution $\hat{w} = \arg \max_w \Psi(w)$:

  $$\Psi(w) \approx \Psi(\hat{w}) - \frac{1}{2}(w - \hat{w})^T H(w - \hat{w}), \quad H = -\nabla^2 \Psi(w)|_{w=\hat{w}}.$$

- Laplace approximation to the posterior $p(w|D) \approx \mathcal{N}(w|\hat{w}, H^{-1})$. 
Online Bayesian linear classification based on Laplace approximation

- Extend to leverage for sequential model updates:
  Laplace approximated posterior serves as prior for the next available data.
- \( p(w_j) = \mathcal{N}(w_j|m_j^0, (q_j^0)^{-1}) \)
- \((m_j^n, q_j^n) \xrightarrow{\{x^n, y^{n+1}\}} (m_j^{n+1}, q_j^{n+1})\)
- \( \hat{t} := \frac{\partial^2 \log \sigma(y_i w_i^T x)}{\partial f^2} \bigg|_{f=\hat{w}^T x} \)

\[
m^{n+1} = \arg \max_w -\frac{1}{2} \sum_{i=1}^{d} q_i^n (w_i - m_i^n)^2 + \log(\sigma(y w^T x))
\]

\[
q_j^{n+1} = q_j^n - \hat{t} x_j^2
\]
Online Bayesian linear classification based on Laplace approximation

\[
\arg \max_w -\frac{1}{2} \sum_{i=1}^{d} q_i(w_i - m_i)^2 + \log(\sigma(yw^T x)).
\]

- 1-dimensional bisection method:
  Set \(\partial \Psi / \partial w_i = 0\). Define \(p\) as \(p := \frac{\sigma'(yw^T x)}{\sigma(yw^T x)}\). Then we have \(w_i = m_i + yp \frac{x_i}{q_i}\).

\[
p = \frac{\sigma'(p \sum_{i=1}^{d} x_i^2 / q_i + ym^T x)}{\sigma(p \sum_{i=1}^{d} x_i^2 / q_i + ym^T x)}.
\]

The equation has a unique solution in interval \([0, \sigma'(ym^T x) / \sigma(ym^T x)]\).
 Sequential Decision Problems with Binary Outputs

 The Knowledge Gradient Policy
   Knowledge Gradient Policy for Lookup Table Model
   Knowledge Gradient Policy for Linear Bayesian Classification

 Experimental Results
Characteristics of our problems

- Expensive experiments.
- Small samples.
- Requiring that we learn from our decisions as quickly as possible.
Knowledge gradient policy for lookup table model [3]

- $M$ discrete alternatives, unknow truth $\mu_x$, $W_x = \mu_x + \epsilon$
- $\mu_x \mid \mathcal{F}^n \sim \mathcal{N}(\theta^n_x, \sigma^n_x)$
- Knowledge state $S^n = (\theta^n, \sigma^n)$, $V(s) = \max_x \theta_x$

\[ \nu^KG_x(S^n) = \mathbb{E}[V(S^{n+1}(x)) - V(S^n)] = \mathbb{E}\left[\max_{x'} \theta_x^{n+1}(x') - \max_{x'} \theta_x^n \mid S^n\right]. \]

- The Knowledge Gradient (KG) policy $X^KG(S^n) = \arg \max_x \nu^KG_x(S^n)$.

Change which produces a change in the decision.

Change in estimate of value of alternative 5 due to measurement.
Knowledge gradient policy for linear Bayesian classification belief model

- \( y_x | \mathbf{w} \sim \text{Bernoulli}(\sigma(\mathbf{w}^T \mathbf{x})) \)
- \( w_j | \mathcal{F}^n \sim \mathcal{N}(m_j^n, (q_j^n)^{-1}) \)
- Knowledge state \( S^n = (\mathbf{m}^n, \mathbf{q}^n) \)
- \( V(s) = \max_x p(y_x = +1 | x, s) \)

\[
\nu^KG_x(S^n) = \mathbb{E}[V(S^{n+1}(x, y)) - V(S^n)|S^n]
= \mathbb{E}[\max_{x'} p(y_{x'} = +1 | x', S^{n+1}(x, y)) - \max_{x'} p(y_{x'} = +1 | x', S^n)|S^n]
\]

- The **Knowledge Gradient (KG) policy** \( X^{KG}(S^n) = \arg \max_x \nu^KG_x(S^n) \).
- The knowledge gradient policy can work with any choice of link function \( \sigma(\cdot) \) and approximation procedures by adjusting the transition function \( S^{n+1}(x, \cdot) \) accordingly.
- Online learning [7]: \( X^{OLKG}(S^n) = \arg \max_x p(y = +1 | x, S^n) + (N - n)\nu^KG_x(S^n) \).
1 Sequential Decision Problems with Binary Outputs

2 The Knowledge Gradient Policy

3 Experimental Results
   Behavior of the KG policy
   Comparison with other Policies
Sampling behavior of the KG policy
Absolute class distribution error

Figure: Absolute distribution error.
Competing policies

- random sampling (Random)
- a myopic method that selects the most uncertain instance each step (MostUncertain)
- discriminative batch-mode active learning (Disc) [4] with batch size set to 1
- expected improvement (EI) [8] with an initial fit of 5 examples
- Thompson sampling (TS) [2]
- UCB on the latent function $\mathbf{w}^T \mathbf{x}$ (UCB) [6]

Metric

Opportunity Cost (OC)

$$OC := \max_{\mathbf{x} \in \mathcal{X}} p(y = +1 | \mathbf{x}, \mathbf{w}^*) - p(y = +1 | \mathbf{x}^{N+1}, \mathbf{w}^*).$$
Comparison with other Policies

(a) sonar

(b) glass identification

(c) blood transfusion

(d) survival

(e) breast cancer (wpbc)

(f) planning relax

Figure: Opportunity cost on UCI.
Thank you! Questions?
Xavier Boyen and Daphne Koller.
Tractable inference for complex stochastic processes.

Olivier Chapelle and Lihong Li.
An empirical evaluation of thompson sampling.

A knowledge-gradient policy for sequential information collection.

Yuhong Guo and Dale Schuurmans.
Discriminative batch mode active learning.
In Advances in neural information processing systems, pages 593–600, 2008.

Steffen L Lauritzen.
Propagation of probabilities, means, and variances in mixed graphical association models.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire.
A contextual-bandit approach to personalized news article recommendation.
Ilya O Ryzhov, Warren B Powell, and Peter I Frazier.  
The knowledge gradient algorithm for a general class of online learning problems.  

Matthew Tesch, Jeff Schneider, and Howie Choset.  
Expensive function optimization with stochastic binary outcomes.  