The Knowledge Gradient for Sequential Decision Making with Stochastic Binary Feedbacks

Yingfei Wang, Chu Wang and Warren B. Powell

Princeton University

Sequential Decision Problems

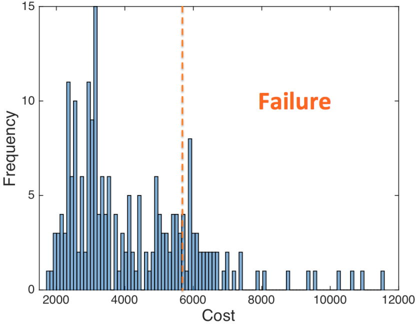
- M discrete alternatives
- Unknow truth μ_x
- Each time n, the learner chooses an alternative x^n, receives reward $W^n_{x^n}$.
- offline objective $\max \mathbb{E}^\pi[\mu_{x^N}]$
  online objective $\max \mathbb{E}^\pi \sum_{n=0}^{N-1}[\mu_{x^n}]$

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:
  - ......

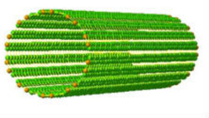
Applications with binary outputs

- 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.

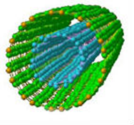


Applications with binary outputs

- 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

Model

- A finite set of alternatives $\boldsymbol{x} \in \mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$.
- Binary outcome $y \in \{-1, +1\}$ with unknown probability $p(y = +1|\boldsymbol{x})$.
- Goal: given a limited budget N, choose the measurement policy $(\boldsymbol{x}^0, \ldots, \boldsymbol{x}^{N-1})$ and the implementation decision that maximizes $p(y = +1|\boldsymbol{x})$.
- Generalized linear model for modeling probability

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

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

Bayesian logistic and probit regression

- $p(\boldsymbol{w}|\mathcal{D}) = \frac{1}{Z} p(\mathcal{D}|\boldsymbol{w}) p(\boldsymbol{w}) \propto p(\boldsymbol{w}) \prod_{i=1}^n \sigma(y_i \cdot \boldsymbol{w}^T \boldsymbol{x}_i)$.
- Extend to leverage for sequential model updates:
  $p(\boldsymbol{w}|\mathcal{D}^0) \xrightarrow{\boldsymbol{x}^0, y^1} p(\boldsymbol{w}|\mathcal{D}^1) \xrightarrow{\boldsymbol{x}^1, y^2} p(\boldsymbol{w}|\mathcal{D}^2) \cdots$
- Exact Bayesian inference for linear classifier is intractable.
- Monte Carlo sampling or analytic approximations to the posterior: Laplace approximation.

Laplace approximation

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

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

- Laplace approximation to the posterior $p(\boldsymbol{w}|\mathcal{D}) \approx \mathcal{N}(\boldsymbol{w}|\hat{\boldsymbol{w}}, \boldsymbol{H}^{-1})$.

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.
- $p(w_j) = \mathcal{N}\left(w_j | m_j^0, (q_j^0)^{-1}\right)$
- $(m_j^n, q_j^n) \xrightarrow{\{x^n, y^{n+1}\}} (m_j^{n+1}, q_j^{n+1})$
- $\hat{t} := \frac{\partial^2 \log \sigma(y_i \boldsymbol{w}_i^T \boldsymbol{x})}{\partial f^2}\Big|_{f=\hat{\boldsymbol{w}}^T \boldsymbol{x}}$

$$
\begin{aligned}
\boldsymbol{m}^{n+1} &= \arg\max_{\boldsymbol{w}} -\frac{1}{2} \sum_{i=1}^{d} q_i^n (w_i - m_i^n)^2 + \log(\sigma(y \boldsymbol{w}^T \boldsymbol{x})) \\
q_j^{n+1} &= q_j^n - \hat{t} x_j^2
\end{aligned}
$$

Online Bayesian linear classification based on Laplace approximation

$$\arg\max_{\boldsymbol{w}} -\frac{1}{2}\sum_{i=1}^{d} q_i(w_i - m_i)^2 + \log(\sigma(y\boldsymbol{w}^T\boldsymbol{x})).$$

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

$$p = \frac{\sigma'(p\sum_{i=1}^{d} x_i^2/q_i + y\boldsymbol{m}^T x)}{\sigma(p\sum_{i=1}^{d} x_i^2/q_i + y\boldsymbol{m}^T x)}.$$

The equation has a unique solution in interval $[0, \sigma'(y\boldsymbol{m}^T x)/\sigma(y\boldsymbol{m}^T x)]$.

Characteristics of our problems

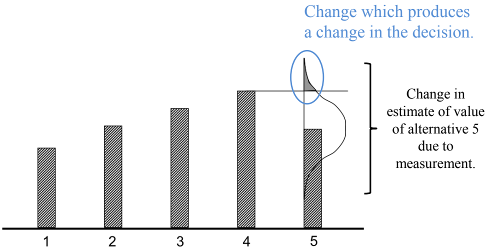
- Expensive experiments.
- Small samples.
- Requiring that we learn from our decisions as quickly as possible.

Knowledge gradient policy for lookup table model [3]

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

$$\nu_x^{KG}(S^n) = \mathbb{E}[V(S^{n+1}(x)) - V(S^n)] = \mathbb{E}[\max_{x'} \theta_{x'}^{n+1}(x) - \max_{x'} \theta_{x'}^n | S^n].$$

- The Knowledge Gradient (KG) policy $X^{KG}(S^n) = \arg\max_x \nu_x^{KG}(S^n)$.



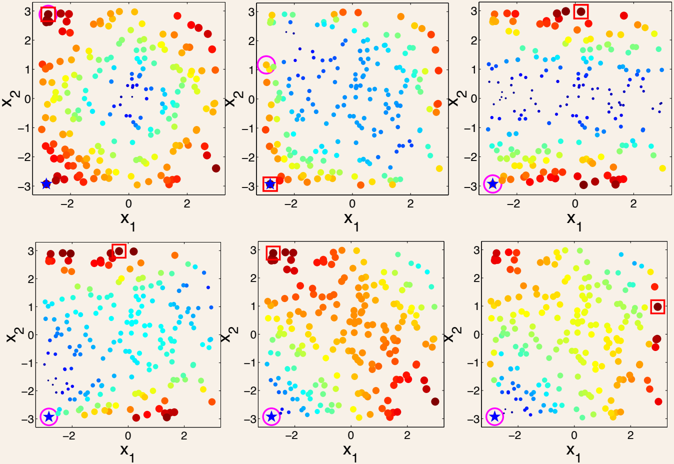
Knowledge gradient policy for linear Bayesian classification belief model

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

$$\begin{aligned}
\nu_{\boldsymbol{x}}^{KG}(S^n) &= \mathbb{E}\left[V(S^{n+1}(\boldsymbol{x}, y)) - V(S^n) | S^n\right] \\
&= \mathbb{E}\left[\max_{\boldsymbol{x}'} p\big(y_{\boldsymbol{x}'} = +1 | \boldsymbol{x}', S^{n+1}(\boldsymbol{x}, y)\big) - \max_{x'} p(y_{\boldsymbol{x}'} = +1 | \boldsymbol{x}', S^n) | S^n\right]
\end{aligned}$$

- The Knowledge Gradient (KG) policy $X^{KG}(S^n) = \arg\max_{\boldsymbol{x}} \nu_{\boldsymbol{x}}^{KG}(S^n)$.
- 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]: $X^{OLKG}(S^n) = \arg\max_x p(y = +1 | x, S^n) + (N - n)\nu_x^{KG}(S^n)$.

Sampling behavior of the KG policy



Absolute class distribution error

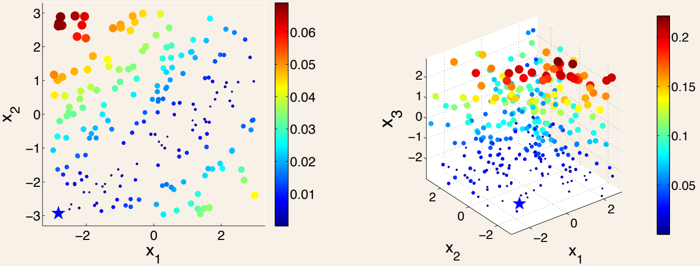


Figure: Absolute distribution error.

Competing policies

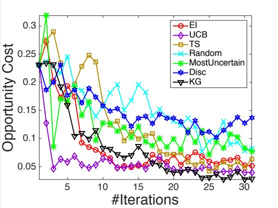
- random sampling (Random)
- a myopic method that selects the most uncertain instance each step (MostUncertain)
- discriminative batch-mode active learning (Disc) [4] with batch size set to 1
- expected improvement (EI) [8] with an initial fit of 5 examples
- Thompson sampling (TS) [2]
- UCB on the latent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

Metric

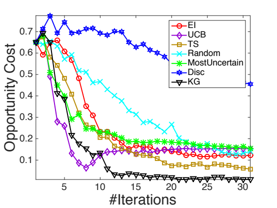
Opportunity Cost (OC)

$$\text{OC} := \max_{\boldsymbol{x}\in\mathcal{X}} p(y=+1|\boldsymbol{x},\boldsymbol{w}^*) - p(y=+1|\boldsymbol{x}^{N+1},\boldsymbol{w}^*).$$

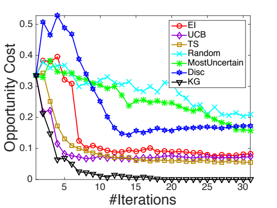
Comparison with other Policies



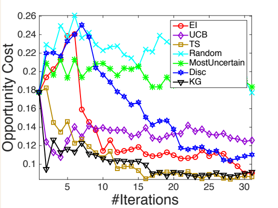
(a) sonar



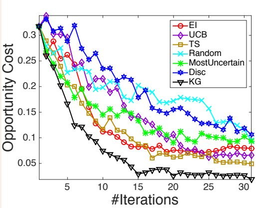
(b) glass identification



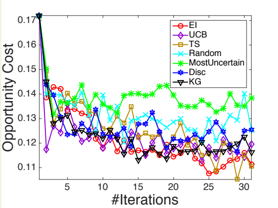
(c) blood transfusion



(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.
In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 33–42. Morgan Kaufmann Publishers Inc., 1998.

Olivier Chapelle and Lihong Li.
An empirical evaluation of thompson sampling.
In *Advances in neural information processing systems*, pages 2249–2257, 2011.

Peter I Frazier, Warren B Powell, and Savas Dayanik.
A knowledge-gradient policy for sequential information collection.
SIAM Journal on Control and Optimization, 47(5):2410–2439, 2008.

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.
Journal of the American Statistical Association, 87(420):1098–1108, 1992.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire.
A contextual-bandit approach to personalized news article recommendation.
In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

Ilya O Ryzhov, Warren B Powell, and Peter I Frazier.
The knowledge gradient algorithm for a general class of online learning problems.
Operations Research, 60(1):180–195, 2012.

Matthew Tesch, Jeff Schneider, and Howie Choset.
Expensive function optimization with stochastic binary outcomes.
In *Proceedings of The 30th International Conference on Machine Learning*, pages 1283–1291, 2013.