On the robustness of a one-period look-ahead policy in multi-armed bandit problems

Ilya O. Ryzhov\textsuperscript{1}  Peter I. Frazier\textsuperscript{2}  Warren B. Powell\textsuperscript{1}

\textsuperscript{1}Operations Research and Financial Engineering  
Princeton University  
Princeton, NJ 08540, USA

\textsuperscript{2}Operations Research and Information Engineering  
Cornell University  
Ithaca, NY 14853, USA

International Workshop on Computational Stochastics  
June 1, 2010
Outline

1. Introduction

2. Mathematical model

3. The knowledge gradient (KG) approach
   - Review of KG policy
   - Adjusting the KG policy

4. Computational experiments

5. Conclusions
Outline

1. Introduction

2. Mathematical model

3. The knowledge gradient (KG) approach
   - Review of KG policy
   - Adjusting the KG policy

4. Computational experiments

5. Conclusions
Motivation: clinical drug trials

- We are testing experimental diabetes treatments on human patients
- We want to find the best treatment, but we also care about the effect on the patients
- How can we allocate groups of patients to treatments?
Motivation: energy portfolios

- We are refitting residential buildings with new energy-saving technologies.
- Promising technologies include (McKinsey & Company 2007):
  - Residential lighting
  - Energy-efficient water heaters
  - Improved ventilation systems
  - And many others...
- How can we find the most energy-efficient technologies while improving each individual building?
The multi-armed bandit problem

- There are $M$ different arms (diabetes treatments or energy portfolios)
- The effectiveness of each arm is unknown, but we have a Bayesian belief about it
- We can measure an arm (by prescribing a treatment to a group of patients) and observe a result that changes our beliefs
- We need to strike a balance between exploration (trying a treatment that might be effective) and exploitation (relying on a treatment that already seems to work well)
Outline

1 Introduction

2 Mathematical model

3 The knowledge gradient (KG) approach
   - Review of KG policy
   - Adjusting the KG policy

4 Computational experiments

5 Conclusions
The multi-armed bandit problem

At first, we believe that

\[ \mu_x \sim \mathcal{N} \left( \mu_x^0, (\sigma_x^0)^2 \right). \]

We measure arm \( x^1 \) and observe a reward

\[ \hat{\mu}_x^1 \sim \mathcal{N} \left( \mu_x^1, \lambda_{x^1}^2 \right). \]

As a result, our beliefs change:

\[
\mu_x^1 = \frac{(\sigma_x^0)^{-2} \mu_x^0 + \lambda_x^{-2} \hat{\mu}_x^1}{(\sigma_x^0)^{-2} + \lambda_x^{-2}} \\
(\sigma_x^1)^2 = \left[ (\sigma_x^0)^{-2} + \lambda_x^{-2} \right]^{-1}
\]

for \( x = x^1 \). If \( x \neq x^1 \), then \( \mu_x^1 = \mu_x^0 \) and \( \sigma_x^1 = \sigma_x^0 \).
The multi-armed bandit problem

- After $n$ measurements, our beliefs about the alternatives are encoded in the knowledge state:
  
  $$s^n = (\mu^n, \sigma^n)$$

- A decision rule $X^n$ is a function that maps the knowledge state $s^n$ to an arm $X^n(s^n) \in \{1, \ldots, M\}$.

- A learning policy $\pi$ is a sequence of decision rules $X^{\pi,1}, X^{\pi,2}, \ldots$

**Objective function**

Choose a measurement policy $\pi$ to achieve

$$\sup_{\pi} \mathbb{E}^{\pi} \sum_{n=0}^{N-1} \mu_{X^{\pi,n}(s^n)}$$

for some finite measurement budget $N$. 
The index policy approach

The bandit literature studies index policies with decision rules of the form

$$X^{\pi,n}(s^n) = \arg \max_x I_x^{\pi}(\mu_x^n, \sigma_x^n)$$

where the index $I_x^{\pi}$ depends on our beliefs about $x$, but not $y \neq x$.

Examples of index policies

- Interval estimation (Kaelbling 1993):

  $$X^{IE,n}(s^n) = \arg \max_x \mu_x^n + z \cdot \sigma_x^n$$

- Gittins indices (Gittins & Jones 1974):

  $$X^{Gitt,n}(s^n) = \arg \max_x G(\mu_x^n, \sigma_x^n, \lambda_x, \gamma)$$

- Upper confidence bound policies (Lai 1987, Auer et al. 2002)
Outline

1. Introduction
2. Mathematical model
3. The knowledge gradient (KG) approach
   - Review of KG policy
   - Adjusting the KG policy
4. Computational experiments
5. Conclusions
The knowledge gradient policy

- Developed by Gupta & Miescke (1996) and Frazier et al. (2008) for offline learning
- Extended to multi-armed bandit problems by Ryzhov & Powell (2009) and Ryzhov et al. (2010)
- One-period look-ahead policy: how much will the next measurement improve our expected objective value?

**KG decision rule**

\[ X^{KG,n}(s^n) = \arg \max_x \mu_n^x + (N - n - 1) \mathbb{E}^n \left( \max_{x'} \mu^{n+1}_{x'} - \max_{x'} \mu^x_{x'} \mid x^n = x \right) \]

Because the uncertainty bonus contains \( \max_{x'} \mu^n_{x'} \), KG is not an index policy.
Lemma

Define the expected improvement made by measuring x exactly m times in a row, starting at time n:

\[ \nu_{x, n}^{n,m} = \mathbb{E}^n \left( \max_{x'} \mu_{x', n}^{n+1} - \max_{x'} \mu_{x', n}^n \mid x^n = \ldots = x^{n+m-1} = x \right) \]

Then,

\[ \nu_{x, n}^{n,m} = \tilde{\sigma}_x^n (m) f \left( \frac{\left| \mu_x^n - \max_{x' \neq x} \mu_{x', n}^n \right|}{\tilde{\sigma}_x^n (m)} \right) \]

where \( f (z) = z \Phi (z) + \phi (z) \), \( \Phi \) and \( \phi \) are the standard Gaussian cdf and pdf, and

\[ \tilde{\sigma}_x^n (m) = \frac{(\sigma_x^n)^2 m}{\left( \lambda_x^2 / (\sigma_x^n)^2 \right) + m}. \]
Computing the knowledge gradient

Proof.

Given \( x^n = \ldots = x^{n+m-1} = x \), we have \( \mu_{x'}^{n+m} = \mu_x^n \) for \( x' \neq x \), due to the independence of the arms. It can be shown that the conditional distribution of \( \mu_{x}^{n+m} \) is \( \mathcal{N} \left( \mu_x^n, (\tilde{\sigma}_x^n(m))^2 \right) \).

As a result,

\[
\nu_{x}^{n,m} = \mathbb{E} \max \left( \max_{x' \neq x} \mu_x^n, \mu_x^n + \tilde{\sigma}_x^n(m) \cdot Z \right) - \max_{x'} \mu_x^n,
\]

for \( Z \sim \mathcal{N} (0,1) \). The closed-form solution of this expectation is the formula for \( \nu_{x}^{n,m} \) given in the Lemma.
Computing the knowledge gradient

The KG decision rule can now be written in closed form as

\[
X^{KG,n}(s^n) = \arg \max_x \mu_x^n + (N - n - 1) \nu_{x,1}^n.
\]

The quantity \( \nu_{x,1}^n \) represents the extra benefit per time period obtained by measuring arm \( x \) once.
Non-concavity of the value of information

- When $\nu_{x}^{n,m}$ is non-concave in $m$, the KG factor $\nu_{x}^{n,1}$ undervalues the benefits of measuring $x$ (Frazier & Powell 2010).

- **Example:** Arm 0 has known value 0, whereas arm 1 has a prior $\mathcal{N}(-1,25)$. When $N = 10^5$ and $\lambda^2 = 10^4$, KG always chooses arm 0.

![Graph](image.png)
Non-concavity of the value of information

- Since KG always chooses arm 0, it achieves an expected value of 0.
- However, if we measure arm 1 in the first 10 iterations, then exploit (choose whichever arm seems to be the best) until the end of the time horizon, we would expect to collect a reward of
  \[10 \cdot \mu_1^0 + (N - 10) \nu_{1,10}^0 = 3690.\]
- We use this idea to adjust the KG policy for non-concave problems.
Adjusting the KG policy

- Define $\mu^n_* = \max_{x'} \mu^n_{x'}$. If we exploit starting at time $n$, we expect our total reward to be $(N - n) \mu^n_*$. 

- If we first measure $x$ an additional $m$ times, and then exploit, we expect the reward to be

$$m \mu^n_x + (N - n - m) (\mu^n_* + \nu^{n,m}_x).$$

- Taking the difference and dividing by $m$ yields

$$\mu^n_x - \mu^n_* + \frac{1}{m} (N - n - m) \nu^{n,m}_x,$$

the average improvement in the reward obtained from each measurement of $x$. 
Adjusting the KG policy

- We now optimize $m$:

$$m^*(x) = \arg \max_{m=1,\ldots,N-n} \mu^n_x - \mu^*_n + \frac{1}{m} (N - n - m) v^n_{x,m}.$$

- The adjusted KG policy, denoted by KG(*), chooses an arm under the assumption that it will be measured $m^*$ times:

$$X^{KG(*),n}(s^n) = \arg \max_{x} \mu^n_x - \mu^*_n + \frac{1}{m^*(x)} (N - n - m^*(x)) v^n_{x,m^*(x)}.$$
Effects of adjustment

- In most cases, KG(*) offers no improvement over KG
- KG(*) offers a significant improvement when either $N$ or $\lambda^2$ is large enough
Effects of adjustment

- In most cases, KG(*) offers no improvement over KG
- KG(*) offers a significant improvement when either $N$ or $\lambda^2$ is large enough
Summary of adjusted policy

- The KG(*) policy approximates a multi-step look-ahead by computing the value of $m^*(x)$ measurements for each arm $x$.
- The example suggests that KG(*) and KG perform similarly in many problem settings, which shows a kind of robustness of the one-step look-ahead.
- However, there is a particular class of problems (high $\lambda^2$ or high $N$) where KG(*) offers significant improvement.
Outline

1 Introduction

2 Mathematical model

3 The knowledge gradient (KG) approach
   - Review of KG policy
   - Adjusting the KG policy

4 Computational experiments

5 Conclusions
Computational experiments

- We compared KG(*) to KG and five index policies on 100 problems with $M = 100$, $\lambda_x^2 = 100$ and $N = 50$
- Performance measure: $\sum_{n=0}^{N-1} \mu_{X^{KG(*)},n}(s^n) - \mu_{X^{\pi},n}(s^n)$
Effect of $\lambda^2$ on the comparison

- KG(*) offers little improvement over KG (but outperforms the index policies) on the base problems.
- We varied the magnitude of $\lambda^2$ and observed the impact on performance:

![Graph showing the effect of $\lambda^2$ on suboptimality of policy](image)
Effect of $N$ on the comparison

- We also considered the impact of the number of measurements $N$ on performance.
- For $N > 10^3$, KG(*) starts to offer a significant improvement.
Outline

1. Introduction
2. Mathematical model
3. The knowledge gradient (KG) approach
   - Review of KG policy
   - Adjusting the KG policy
4. Computational experiments
5. Conclusions
Conclusions

- The KG policy is a one-period look-ahead approach for multi-armed bandit problems
- The KG(*) policy adjusts KG to account for non-concavity in the value of information, by approximating a multi-step look-ahead
- In many problems, KG(*) offers only a small improvement over KG, meaning that the KG policy is generally robust
- We have identified the particular class of problems where KG(*) offers significant improvement (high $N$)


