Knowledge Gradient

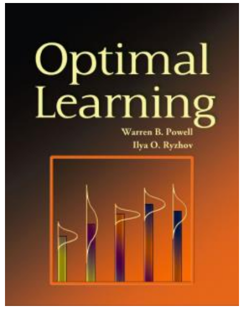
A Partial Recipe for Interpretable Reinforcement Learning

Donghun Lee
Dept. of Math, Korea University

17 November 2021

AIML@K
KOREA UNIVERSITY

Knowledge Gradient?



- Knowledge Gradient (KG): (Powell and Ryzhov 2013)

$$v_x^{KG,n} := \mathbb{E}\left[V^{n+1}\left(S^{n+1}(x)\right) - V^n(s^n) \middle| s^n\right]$$

- A bit of details
  - Time (iteration) counter $n \in \{0,1,2,\cdots\}$
  - Decision $x \in \mathcal{X}$ (finite decision set $\mathcal{X}$)
  - State s^n, at time n (state space: $\mathcal{S}$, such that $\forall n: s^n \in \mathcal{S}$)
  - State "transition function" $S^{n+1}(x): \mathcal{X} \to \mathcal{S}$

Problem Setting

- Sequential Decision Making
  - Relatively new problem (in math)

*A SEQUENTIAL DECISION PROBLEM WITH A FINITE MEMORY**

By Herbert Robbins

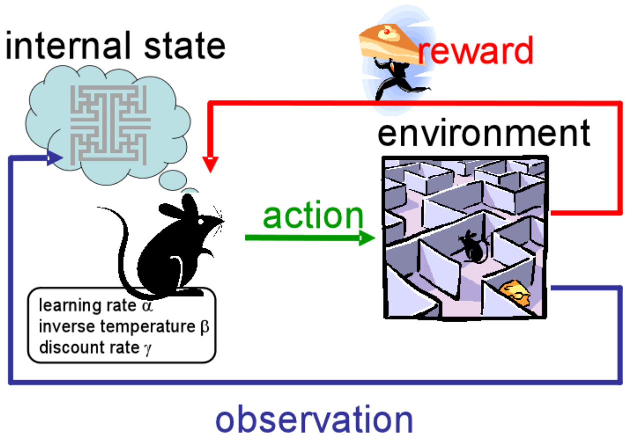
COLUMBIA UNIVERSITY

Communicated by Paul A. Smith, October 1, 1956

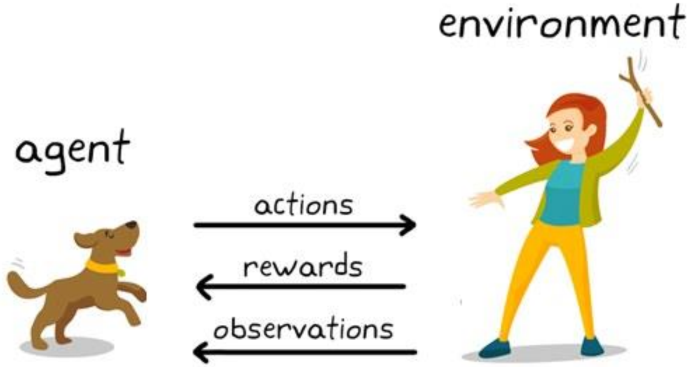
1. *Summary.*—We consider the problem of successively choosing one of two ways of action, each of which may lead to success or failure, in such a way as to maximize the long-run proportion of successes obtained, the choice each time being based on the results of a fixed number of the previous trials.

AIML@K KOREA UNIVERSITY

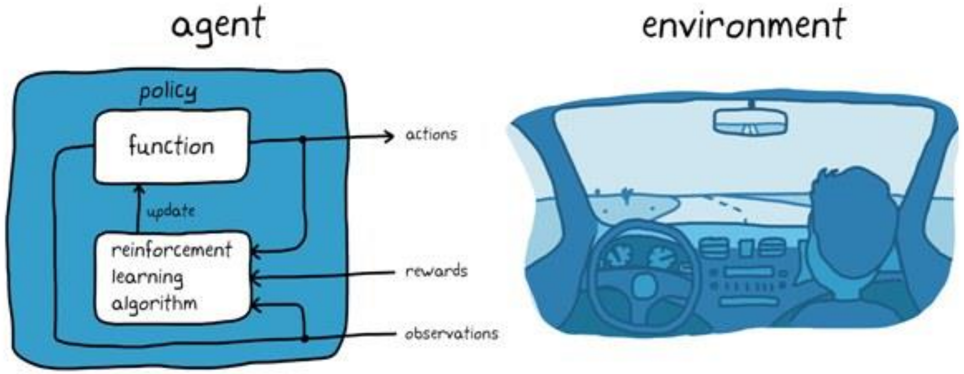
Problem Setting



- Sequential Decision Making
  - Sounds similar to …
    - Reinforcement Learning
    - Control Theory



- How to make "best" decision?
  - Now, and also later
  - Based on finite interactions
  - With the aim of optimizing some fn.



AIML@K KOREA UNIVERSITY

Real World Example

- Decision: x^n "what to eat for lunch, on n-th day"
  - Example: finite decision set $X = \{0,1,2,3,4\} = [5]$

$x^n = 0$

$x^n = 1$

$x^n = 2$

$x^n = 3$

$x^n = 4$





Real World Example

- $R(x)$: "reward" for choosing x for lunch (a random var.)
  - Choose $x^n = x$, and then observe a realization $\hat{R}^{n+1}$
  - $\forall n$: $R^{n+1} = R(x^n) \sim ?$ (unknown dist.)

- Daily welfare from lunch choices $x^0, x^1, \cdots, x^{N-1}$

$$\sum_{n=0}^{N-1} R(x^n)$$





Real World Example

- $R(x)$: "reward" for choosing x for lunch (a random var.)
  - Choose $x^n = x$, and then observe a realization $\hat{R}^{n+1}$
  - $\forall n$: $R^{n+1} = R(x^n) \sim ?$ (unknown dist.)

- Scenario: What's for lunch over N days?
  - What to decide? $x^0, x^1, \cdots, x^{N-1}$
  - Objective? maximize $\mathbb{E}R(x^N)$
  - What to learn?





Real World Example

- Scenario: What's for lunch over N days?
  - What to decide? $x^0, x^1, \cdots, x^{N-1}$
  - Objective? choose x^N to maximize $\mathbb{E}R(x^N)$
  - What to learn?
    - Consider magic 8 ball function $\pi: \mathcal{S} \mapsto \mathcal{X}$
      - that makes decisions based on something. i.e. $x^n = \pi(s^n)$
      - Call it "policy function"

Real World Example

- Scenario: What's for lunch over N days?
  - What to decide? $x^0, x^1, \cdots, x^{N-1}$
  - Objective: choose $x^0, x^1, \cdots, x^{N-1}$ and x^N to maximize $\mathbb{E}[R(x^N)|s^N]$
  - What to learn?
    - Consider magic 8 ball function $\pi: \mathcal{S} \mapsto \mathcal{X}$
      - that makes decisions based on something. i.e. $x^n = \pi(s^n)$
      - Call it "policy function"





Real World Example

- Scenario: What's for lunch over N days?
  - What to decide? $x^0, x^1, \cdots, x^{N-1}$
  - Objective: choose $x^0, x^1, \cdots, x^{N-1}$ and x^N to maximize $\mathbb{E}[R(x^N)|s^N]$
  - What to learn?
    - Consider magic 8 ball function $\pi: \mathcal{S} \mapsto \mathcal{X}$
      - that makes decisions based on something. i.e. $x^n = \pi(s^n)$
      - Call it "policy function"
  - Objective of <u>learning</u>: <u>how to choose</u> π to maximize $\mathbb{E}\left[R\left(\pi(s^N)\right)\middle|s^N\right]$

Real World Example





- $R(x)$: "reward" for choosing x for lunch (a random var.)
  - Choose $x^n = x$, and then observe a realization $\hat{R}^{n+1}$
  - $\forall n$: $R^{n+1} = R(x^n) \sim ?$ (unknown dist.)

- Greedily choice based on "KG" maximizes

$$\mathbb{E}\left[R\left(\pi(s^N)\right)\middle|s^N\right]$$

  - Instead of maximizing rewards from N observations, maximize what you "learn" from N observations

AIML@K KOREA UNIVERSITY

Real World Example

- $R(x)$: "reward" for choosing x for lunch (a random var.)
  - Choose $x^n = x$, and then observe a realization $\hat{R}^{n+1}$
  - $\forall n$: $R^{n+1} = R(x^n) \sim ?$ (unknown dist.)

- Knowledge gradient "policy"

$$\pi^{KG}(s^n) \coloneqq \underset{x \in \mathcal{X}}{\operatorname{argmax}}\, v_x^{KG,n}$$
$$= \underset{x \in \mathcal{X}}{\operatorname{argmax}}\, \mathbb{E}\left[V^{n+1}\left(S^{n+1}(x)\right) - V^n(s^n) \middle| s^n\right]$$

  - Makes a myopically optimal choice (given current knowledge s^n)
  - Asymptotically converges to the optimal choice (assuming some model on $R(x^n)$)

About “Knowledge” in KG

- Knowledge Gradient (KG):

$$v_x^{KG,n} := \mathbb{E}\left[V^{n+1}\left(S^{n+1}(x)\right) - V^n(s^n)\middle|s^n\right]$$

  - “Expected increment of $\boldsymbol{V}$ of observing a reward from decision x”

- What is V^n?

$$V^n(s^n) = \max_{x\in\mathcal{X}} \mathbb{E}[\tilde{R}^n(x)|s^n]$$

  - “Knowledge” on (which decision would give) the largest expected R
    - $\tilde{R}^n$ is our belief on R at time n
      - i.e. after observing $\hat{R}^n \sim R(x^{n-1})$ incurred by x^{n-1} (n-th choice)

About “Knowledge” in KG

- Knowledge Gradient (KG):

$$v_x^{KG,n} \coloneqq \mathbb{E}\left[V^{n+1}\left(S^{n+1}(x)\right) - V^n(s^n)\middle|s^n\right]$$

  - “Expected increment of $\boldsymbol{V}$ of observing a reward from decision x”

- What is V^n?

$$V^n(s^n) = \max_{x\in\mathcal{X}} \mathbb{E}[\tilde{R}^n(x)|s^n]$$

  - “Knowledge” on (which decision would give) the largest expected R
    - If you know $\forall x$: $\mathbb{E}[R(x)]$, you have all the knowledge to make the “best” decision

About "Knowledge" in KG

- Knowledge Gradient (KG):

$$v_x^{KG,n} := \mathbb{E}\left[V^{n+1}\left(S^{n+1}(x)\right) - V^n(s^n)\middle|s^n\right]$$

  - "Expected increment of $\boldsymbol{V}$ of observing a reward from decision x"

- What is V^n?

$$V^n(s^n) = \max_{x \in \mathcal{X}} \mathbb{E}[\tilde{R}^n(x)|s^n]$$

  - "Knowledge" on (which decision would give) the largest expected R
    - If you know $\forall x$: $\mathbb{E}[R(x)]$, you have all the knowledge to make the "best" decision
    - What was R? $\forall n$: $R^{n+1} = R(x^n) \sim ?$ (unknown dist.), so here comes modeling

Frazier, Peter, and Warren Powell. 2007. "The Knowledge Gradient Policy for Offline Learning with Independent Normal Rewards." *2007 IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning*. https://doi.org/10.1109/adprl.2007.368181.

Offline **KG**

- The first result with "KG" name (Frazier and Powell, 2007)
- Modeling assumption
  - $\forall n$: $R^{n+1} = R(x^n) \sim \mathcal{N}(\mu_x^n, \beta_x^n)$ (Gaussian, independent given x)

Frazier, Peter, and Warren Powell. 2007. “The Knowledge Gradient Policy for Offline Learning with Independent Normal Rewards.” *2007 IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning*. https://doi.org/10.1109/adprl.2007.368181.

Offline **KG**

- The first result with “KG” name (Frazier and Powell, 2007)
- Modeling assumption
  - $\forall n$: $R^{n+1} = R(x^n) \sim \mathcal{N}(\mu_x^n, \beta_x^n)$ (Gaussian, independent given x)

- Key theoretical contributions
  - Analytic computation of $v_x^{KG,n} := \mathbb{E}\left[V^{n+1}\left(S^{n+1}(x)\right) - V^n(s^n) \middle| s^n\right]$
    $= \tilde{\sigma}(\beta_x^n)[\zeta_x^n \Phi(\zeta_x^n) + \phi(\zeta_x^n)]$

$$\zeta_x^n = -\frac{\left|\bar{\mu}_t^x - \max_{x' \neq x} \mu_t^{x'}\right|}{\tilde{\sigma}_t^x} \qquad \tilde{\sigma}_t^x = \frac{\bar{\sigma}_t^x}{\sqrt{1 + \left(\frac{\sigma^\epsilon}{\bar{\sigma}_t^x}\right)^2}}$$

AIML@K KOREA UNIVERSITY

Frazier, Peter, and Warren Powell. 2007. “The Knowledge Gradient Policy for Offline Learning with Independent Normal Rewards.” *2007 IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning*. https://doi.org/10.1109/adprl.2007.368181.

Offline **KG**

- The first result with “KG” name (Frazier and Powell, 2007)
- Modeling assumption
  - $\forall n$: $R^{n+1} = R(x^n) \sim \mathcal{N}(\mu_x^n, \beta_x^n)$ (Gaussian, independent given x)
- Key theoretical contributions
  - Analytic computation of $v_x^{KG,n} := \mathbb{E}\left[V^{n+1}\left(S^{n+1}(x)\right) - V^n(s^n) \middle| s^n\right]$
    $= \tilde{\sigma}(\beta_x^n)[\zeta_x^n \Phi(\zeta_x^n) + \phi(\zeta_x^n)]$
  - Myopically optimal choice, assuming correct model
  - Asymptotic convergence to optimal choice

Frazier, Peter, Warren Powell, and Savas Dayanik. 2009. "The Knowledge-Gradient Policy for Correlated Normal Beliefs." *INFORMS Journal on Computing* 21 (4): 599–613.

Offline KG, Correlated Belief

- Extended modeling assumption (Frazier et al. 2008)
- Modeling assumption
    - $\forall n$: $R^{n+1} = R(x^n) \sim \mathcal{N}(\boldsymbol{\mu}^{\boldsymbol{n}}, \Sigma^n)$ (|$\mathcal{X}$|-variate Gaussian)
- Key theoretical contributions
    - $O(|\mathcal{X}|)$ algorithm to compute $v_x^{KG,n} := \mathbb{E}\left[V^{n+1}\left(S^{n+1}(x)\right) - V^n(s^n) \middle| s^n\right]$
    - Myopically optimal choice, assuming correct model
    - Asymptotic convergence to optimal choice

Scott, Warren, Peter Frazier, and Warren Powell. 2010. "The Correlated Knowledge Gradient for Maximizing Expensive Continuous Functions with Noisy Observations Using Gaussian Process Regression." *Operations Res. Financial Eng. , Princeton, NJ*. https://www.academia.edu/download/4837036/scottpowell-akg_2010_05_11.pdf.

Offline KG, Corr. Belief, Continuous x

- Another extension, on $\mathcal{X}$ (Scott et al. 2010)
- Modeling assumption
  - $\forall n\colon [\boldsymbol{R}^{1:n}, R^{n+1}]^\top \sim \mathcal{N}(\boldsymbol{\mu}^{n+1}, \Sigma^{n+1})$ $((n+1)$-variate Gaussian)
- Key idea
  - Use Gaussian Process to learn (potentially infinite-dimensional) $\boldsymbol{\mu}, \Sigma$
- Key contribution
  - Approximation of KG w/ GP (instead of $\max_{x \in \mathcal{X}}$, computed $\max_{x \in \{x^0, x^1, \cdots x^{n-1}\}}$)

Mes, Martijn R. K. , Powell, Warren B, and Frazier, Peter I. 2011. “Hierarchical Knowledge Gradient for Sequential Sampling.” 2011. https://www.jmlr.org/papers/volume12/mes11a/mes11a.pdf.

Offline KG, Hierarchical Belief

- Yet another extension on modeling (Mes et al. 2011)
- Modeling assumption
  - $\forall n$: $R^{n+1} = R(x) \sim \mathcal{N}\left(\sum_{g\in\mathcal{G}}^{2} w_x^{g,n} \mu_x^{g,n}, \Sigma^n\right)$
    - Explicit multi-level aggregation on mean-parameter of the model

- Key theoretical contributions
  - Algorithm to compute $v_x^{KG,n} := \mathbb{E}\left[V^{n+1}\left(S^{n+1}(x)\right) - V^n(s^n)\middle|s^n\right]$
  - Asymptotic convergence to optimal choice

AIML@K KOREA UNIVERSITY

From Offline KG to Online KG

- Up so far: KG variations, all maximizing

$$0 \cdot \sum_{n=0}^{N-1} \mathbb{E}\left[\hat{R}^{n+1}\left(\pi(s^n)\right)\right] + \mathbb{E}\left[R\left(\pi(s^N)\right)\middle| s^N\right]$$

  - Recall: Instead of maximizing rewards from N observations, maximize what you "learn" from N observations

- How about our daily lunch welfare for those N days?

$$0 \cdot \sum_{n=0}^{N-1} \mathbb{E}\left[\hat{R}^{n+1}\left(\pi(s^n)\right)\right] + \mathbb{E}\left[\hat{R}^{N}\left(\pi(s^{N-1})\right)\right]$$

Online Learning with KG

- No KG-based algorithm that robustly maximizes

$$\sum_{n=0}^{N-1} \mathbb{E}\left[\hat{R}^{n+1}\left(\pi(s^n)\right)\right]$$

  - Standard attack: build a KG-basd algorithm with sublinear regret
    - Implies asymptotically zero time-amortized regret
  - Why still KG?

$$v_x^{KG,n} := \mathbb{E}\left[V^{n+1}\left(S^{n+1}(x)\right) - V^n(s^n)\middle|s^n\right]$$

    - “Expected increment of V of observing a reward from decision x”

Online Learning with KG

- No KG-based algorithm that robustly maximizes

$$\sum_{n=0}^{N-1} \mathbb{E}\left[\hat{R}^{n+1}\left(\pi(s^n)\right)\right]$$

  - Standard attack: build a KG-basd algorithm with sublinear regret
    - Implies asymptotically zero time-amortized regret
      - Implies no prior knowledge of N

Online Learning with KG

- Online KG-based algorithm that robustly maximizes

$$\sum_{n=0}^{N-1} \mathbb{E}\left[\hat{R}^{n+1}\left(\pi(s^n)\right)\right]$$

  - Can provide

$$v_x^{KG,n} := \mathbb{E}\left[V^{n+1}\left(S^{n+1}(x)\right) - V^n(s^n)\middle| s^n\right]$$

    - "Expected increment of $\boldsymbol{V}$ of observing a reward from decision x"
      - At any time n, without prior knowledge of N
    - … can be used to answer …
    - "What experiment setup $x \in \mathcal{X}$ to test today?"

AIML@K KOREA UNIVERSITY





Online Learning with KG

- $R(x)$: "reward" for choosing x for lunch (a random var.)
  - Choose $x^n = x$, and then observe a realization $\hat{R}^{n+1}$
  - $\forall n$: $R^{n+1} = R(x^n) \sim ?$ (unknown dist.)
- Daily welfare from lunch choices $x^0, x^1, \cdots, x^{N-1}$

$$\sum_{n=0}^{N-1} \mathbb{E}\left[\hat{R}^{n+1}\left(\pi(s^n)\right)\right]$$

  - Also, may be useful to answer: "what $x \in \mathcal{X}$ to have for lunch today?"

Featured Select Works by

- The advisor … and some of his students


Warren Powell


Peter Frazier


Ilya Ryzhov


Donghun Lee


Yingfei Wang

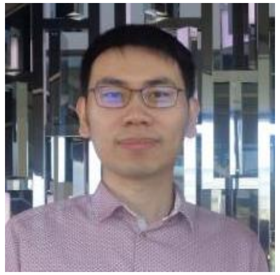
time t

In order of joining Powell Lab


Martijn Mes


Warren Scott


Daniel Jiang


Weidong Han

AIML@K
KOREA UNIVERSITY