Knowledge Gradient for Selection with Covariates: Consistency and Computation

Haihui Shen

Sino-US Global Logistics Institute
Shanghai Jiao Tong University

Joint work with Liang Ding (Texas A&M), Jeff Hong (Fudan), and Xiaowei Zhang (HKU)

@ 2020 INFORMS Annual Meeting (Virtual)

November 7-13, 2020

上海交通大學
Shanghai Jiao Tong University
董浩云航运与物流研究院
CY TUNG Institute of Maritime and Logistics
中美物流研究院
Sino-US Global Logistics Institute

Contents

Selection of the Best

- Select the best from a finite set of alternatives, whose performances are unknown and can only be learned by sampling.
- The samples may come from computer simulation or real experiments.
- E.g., select the best medicine (treatment), advertisement (recommendation), production line, inventory management, etc.

Selection of the Best

- Select the best from a finite set of alternatives, whose performances are unknown and can only be learned by sampling.
- The samples may come from computer simulation or real experiments.
- E.g., select the best medicine (treatment), advertisement (recommendation), production line, inventory management, etc.

- Sampling may be expensive (in time and/or money), thereby budget-constrained.
- **Goal**: a sampling strategy to learn the performances and identify the best as efficiently as possible.

Selection with Covariates

- In many cases, "the best" is not universal but depends on the covariates (contextual information).
- In the example of personalized medicine, the covariates may be *gender*, *age*, *weight*, *medical history*, *drug reaction*, etc.
- In the example of customized advertisement, the covariates may be *gender*, *age*, *location*, *education*, *browsing history*, etc.

Selection with Covariates

- In many cases, “the best” is not universal but depends on the covariates (contextual information).
- In the example of personalized medicine, the covariates may be *gender*, *age*, *weight*, *medical history*, *drug reaction*, etc.
- In the example of customized advertisement, the covariates may be *gender*, *age*, *location*, *education*, *browsing history*, etc.

- **Goal**: a sampling strategy to learn the performance surfaces (functions) as efficiently as possible.
  - With the learned performance surfaces, we can identify the best alternative once the covariates are given (or observed).

Knowledge Gradient

- Knowledge Gradient (KG), introduced in Frazier et al. (2008), is a sequential sampling strategy under Bayesian perspective.

Knowledge Gradient

- Knowledge Gradient (KG), introduced in Frazier et al. (2008), is a sequential sampling strategy under Bayesian perspective.
- For selection of the best (*without* covariates):
  - KG-based sampling strategies are widely used;
  - the performance is often competitive with or outperforms other sampling strategies (Ryzhov 2016).
- For selection of the best *with* covariates:
  - KG-based sampling strategies are emerging (Pearce and Branke 2017);
  - the theory is not complete yet, e.g., no theoretical analysis of the asymptotic behavior of such strategies.

Knowledge Gradient

- Knowledge Gradient (KG), introduced in Frazier et al. (2008), is a sequential sampling strategy under Bayesian perspective.

- For selection of the best (*without* covariates):
  - KG-based sampling strategies are widely used;
  - the performance is often competitive with or outperforms other sampling strategies (Ryzhov 2016).

- For selection of the best *with* covariates:
  - KG-based sampling strategies are emerging (Pearce and Branke 2017);
  - the theory is not complete yet, e.g., no theoretical analysis of the asymptotic behavior of such strategies.

What We Did?

- In this research, we
  - propose a sampling strategy based on the integrated KG, which is suitable for more general situation;
  - provide a theoretical analysis of the asymptotic behavior of the sampling strategy;
  - propose a stochastic gradient ascent (SGA) algorithm to solve the sampling strategy.

What We Did?

- In this research, we
  - propose a sampling strategy based on the integrated KG, which is suitable for more general situation;
  - provide a theoretical analysis of the asymptotic behavior of the sampling strategy;
  - propose a stochastic gradient ascent (SGA) algorithm to solve the sampling strategy.

	Pearce and Branke (2017)	Our Work
Sampling Noise	homoscedastic	can be heteroscedastic
Sampling Cost	constant	can be different

What We Did?

- In this research, we
  - propose a sampling strategy based on the integrated KG, which is suitable for more general situation;
  - provide a theoretical analysis of the asymptotic behavior of the sampling strategy;
  - propose a stochastic gradient ascent (SGA) algorithm to solve the sampling strategy.

	Pearce and Branke (2017)	Our Work
Sampling Noise	homoscedastic	can be heteroscedastic
Sampling Cost	constant	can be different
Asymptotic Analysis	numerical	theoretical

What We Did?

- In this research, we
  - propose a sampling strategy based on the integrated KG, which is suitable for more general situation;
  - provide a theoretical analysis of the asymptotic behavior of the sampling strategy;
  - propose a stochastic gradient ascent (SGA) algorithm to solve the sampling strategy.

	Pearce and Branke (2017)	Our Work
Sampling Noise	homoscedastic	can be heteroscedastic
Sampling Cost	constant	can be different
Asymptotic Analysis	numerical	theoretical
To Solve	sample average approximation	SGA

Setting

- M competing alternatives with *unknown* performance surface $\theta_i(\boldsymbol{x})$, $i = 1, \ldots, M$.
- The covariates $\boldsymbol{x} = (x_1, \ldots, x_d)^\intercal \in \mathcal{X} \subset \mathbb{R}^d$ has density $\gamma(\boldsymbol{x})$.
- We want to learn *offline*: $\operatorname{argmax}_{1 \le i \le M} \theta_i(\boldsymbol{x})$, for $\boldsymbol{x} \in \mathcal{X}$.

Setting

- M competing alternatives with *unknown* performance surface $\theta_i(\boldsymbol{x})$, $i = 1, \dots, M$.
- The covariates $\boldsymbol{x} = (x_1, \dots, x_d)^\intercal \in \mathcal{X} \subset \mathbb{R}^d$ has density $\gamma(\boldsymbol{x})$.
- We want to learn *offline*: $\operatorname{argmax}_{1 \le i \le M} \theta_i(\boldsymbol{x})$, for $\boldsymbol{x} \in \mathcal{X}$.

- For simplification purpose, in this presentation we just consider the constant sampling cost ($\equiv 1$), which is not necessary.
- The budget is N samples.
- Sample on alternative i at location $\boldsymbol{x}$ has *independent* normal distribution with *unknown* mean $\theta_i(\boldsymbol{x})$ and *known* variance $\lambda_i(\boldsymbol{x})$.

Setting

- M competing alternatives with *unknown* performance surface $\theta_i(\boldsymbol{x})$, $i = 1, \ldots, M$.
- The covariates $\boldsymbol{x} = (x_1, \ldots, x_d)^\intercal \in \mathcal{X} \subset \mathbb{R}^d$ has density $\gamma(\boldsymbol{x})$.
- We want to learn *offline*: $\operatorname{argmax}_{1 \le i \le M} \theta_i(\boldsymbol{x})$, for $\boldsymbol{x} \in \mathcal{X}$.
- For simplification purpose, in this presentation we just consider the constant sampling cost ($\equiv 1$), which is not necessary.
- The budget is N samples.
- Sample on alternative i at location $\boldsymbol{x}$ has *independent* normal distribution with *unknown* mean $\theta_i(\boldsymbol{x})$ and *known* variance $\lambda_i(\boldsymbol{x})$.
- We need a good strategy to guide the sampling decision (on *which alternative* and at *what location*) until the N samples are taken.

Bayesian Perspective

- Assign prior for $\{\theta_1(\boldsymbol{x}), \ldots, \theta_M(\boldsymbol{x})\}$, under which $\theta_i(\boldsymbol{x})$'s are independent Gaussian processes with:
  - mean function $\mu_i^0(\boldsymbol{x}) := \mathbb{E}[\theta_i(\boldsymbol{x})|\mathcal{F}^0]$;
  - covariance function $k_i^0(\boldsymbol{x}, \boldsymbol{x}') := \mathrm{Cov}[\theta_i(\boldsymbol{x}), \theta_i(\boldsymbol{x}')|\mathcal{F}^0]$.

Bayesian Perspective

- Assign prior for $\{\theta_1(\boldsymbol{x}), \ldots, \theta_M(\boldsymbol{x})\}$, under which $\theta_i(\boldsymbol{x})$'s are independent Gaussian processes with:
    - mean function $\mu_i^0(\boldsymbol{x}) := \mathbb{E}[\theta_i(\boldsymbol{x})|\mathcal{F}^0]$;
    - covariance function $k_i^0(\boldsymbol{x}, \boldsymbol{x}') := \mathrm{Cov}[\theta_i(\boldsymbol{x}), \theta_i(\boldsymbol{x}')|\mathcal{F}^0]$.

- After n samples, $\{\theta_1(\boldsymbol{x}), \ldots, \theta_M(\boldsymbol{x})\}$ are still independent Gaussian processes under the posterior with:
    - mean function $\mu_i^n(\boldsymbol{x}) := \mathbb{E}[\theta_i(\boldsymbol{x})|\mathcal{F}^n]$;
    - covariance function $k_i^n(\boldsymbol{x}, \boldsymbol{x}') := \mathrm{Cov}[\theta_i(\boldsymbol{x}), \theta_i(\boldsymbol{x}')|\mathcal{F}^n]$.

Bayesian Perspective

- Assign prior for $\{\theta_1(\boldsymbol{x}), \ldots, \theta_M(\boldsymbol{x})\}$, under which $\theta_i(\boldsymbol{x})$'s are independent Gaussian processes with:
  - mean function $\mu_i^0(\boldsymbol{x}) := \mathbb{E}[\theta_i(\boldsymbol{x}) | \mathcal{F}^0]$;
  - covariance function $k_i^0(\boldsymbol{x}, \boldsymbol{x}') := \mathrm{Cov}[\theta_i(\boldsymbol{x}), \theta_i(\boldsymbol{x}') | \mathcal{F}^0]$.

- After n samples, $\{\theta_1(\boldsymbol{x}), \ldots, \theta_M(\boldsymbol{x})\}$ are still independent Gaussian processes under the posterior with:
  - mean function $\mu_i^n(\boldsymbol{x}) := \mathbb{E}[\theta_i(\boldsymbol{x}) | \mathcal{F}^n]$;
  - covariance function $k_i^n(\boldsymbol{x}, \boldsymbol{x}') := \mathrm{Cov}[\theta_i(\boldsymbol{x}), \theta_i(\boldsymbol{x}') | \mathcal{F}^n]$.
- $\mu_i^n(\boldsymbol{x})$ is used as our estimator (or predictor) of $\theta_i(\boldsymbol{x})$, and $k_i^n(\boldsymbol{x}, \boldsymbol{x})$ characterizes the uncertainty at $\theta_i(\boldsymbol{x})$.

Bayesian Perspective

- Assign prior for $\{\theta_1(\boldsymbol{x}), \ldots, \theta_M(\boldsymbol{x})\}$, under which $\theta_i(\boldsymbol{x})$'s are independent Gaussian processes with:
  - mean function $\mu_i^0(\boldsymbol{x}) := \mathbb{E}[\theta_i(\boldsymbol{x})|\mathcal{F}^0]$;
  - covariance function $k_i^0(\boldsymbol{x}, \boldsymbol{x}') := \mathrm{Cov}[\theta_i(\boldsymbol{x}), \theta_i(\boldsymbol{x}')|\mathcal{F}^0]$.
- After n samples, $\{\theta_1(\boldsymbol{x}), \ldots, \theta_M(\boldsymbol{x})\}$ are still independent Gaussian processes under the posterior with:
  - mean function $\mu_i^n(\boldsymbol{x}) := \mathbb{E}[\theta_i(\boldsymbol{x})|\mathcal{F}^n]$;
  - covariance function $k_i^n(\boldsymbol{x}, \boldsymbol{x}') := \mathrm{Cov}[\theta_i(\boldsymbol{x}), \theta_i(\boldsymbol{x}')|\mathcal{F}^n]$.
- $\mu_i^n(\boldsymbol{x})$ is used as our estimator (or predictor) of $\theta_i(\boldsymbol{x})$, and $k_i^n(\boldsymbol{x}, \boldsymbol{x})$ characterizes the uncertainty at $\theta_i(\boldsymbol{x})$.
- Updating Equation: if the n-th sample y is taken on i at $\boldsymbol{v}$, then

$$\mu_i^n(\boldsymbol{x}) = \mu_i^{n-1}(\boldsymbol{x}) + k_i^{n-1}(\boldsymbol{x}, \boldsymbol{v})[k_i^{n-1}(\boldsymbol{v}, \boldsymbol{v}) + \lambda_i(\boldsymbol{v})]^{-1}[y - \mu_i^{n-1}(\boldsymbol{v})],$$

$$k_i^n(\boldsymbol{x}, \boldsymbol{x}') = k_i^{n-1}(\boldsymbol{x}, \boldsymbol{x}') - k_i^{n-1}(\boldsymbol{x}, \boldsymbol{v})[k_i^{n-1}(\boldsymbol{v}, \boldsymbol{v}) + \lambda_i(\boldsymbol{v})]^{-1}k_i^{n-1}(\boldsymbol{v}, \boldsymbol{x}').$$

Objective of Sampling Strategy

- After N samples, we will estimate $\operatorname{argmax}_{1\le i\le M}\theta_i(\boldsymbol{x})$ via $\operatorname{argmax}_{1\le i\le M}\mu_i^N(\boldsymbol{x})$.

Objective of Sampling Strategy

- After N samples, we will estimate $\mathrm{argmax}_{1\le i\le M}\, \theta_i(\boldsymbol{x})$ via $\mathrm{argmax}_{1\le i\le M}\, \mu_i^N(\boldsymbol{x})$.
- View $\max_i \mu_i^N(\boldsymbol{x})$ as a terminal reward under Bayesian perspective:
  - its expected value depends on a sampling strategy π;
  - we want to maximize this expected reward.

Objective of Sampling Strategy

- After N samples, we will estimate $\mathrm{argmax}_{1\leq i\leq M}\,\theta_i(\boldsymbol{x})$ via $\mathrm{argmax}_{1\leq i\leq M}\,\mu_i^N(\boldsymbol{x})$.

- View $\max_i \mu_i^N(\boldsymbol{x})$ as a terminal reward under Bayesian perspective:
    - its expected value depends on a sampling strategy π;
    - we want to maximize this expected reward.

- The objective becomes

$$\max_{\pi} \int_{\mathcal{X}} \mathbb{E}^{\pi}\left[\max_{1\leq i\leq M} \mu_i^N(\boldsymbol{x})\right]\gamma(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

Integrated Knowledge Gradient

- Let $(a^n, \boldsymbol{v}^n)$ denote the n-th sampling decision, i.e., on alternative a^n at location $\boldsymbol{v}^n$, and $S^n := (\mu_1^n, \ldots, \mu_M^n, k_1^n, \ldots, k_M^n)$ the random *state* after the n-th sample.

Integrated Knowledge Gradient

- Let $(a^n, \boldsymbol{v}^n)$ denote the n-th sampling decision, i.e., on alternative a^n at location $\boldsymbol{v}^n$, and $S^n := (\mu_1^n, \ldots, \mu_M^n, k_1^n, \ldots, k_M^n)$ the random *state* after the n-th sample.

- If $N = 1$, the optimal strategy is

$$\operatorname*{argmax}_{1\le i\le M, \boldsymbol{x}\in\mathcal{X}} \int_{\mathcal{X}} \mathbb{E}\left[\max_{1\le a\le M} \mu_a^1(\boldsymbol{v}) \,\middle|\, S^0, a^1 = i, \boldsymbol{v}^1 = \boldsymbol{x}\right] \gamma(\boldsymbol{v}) \mathrm{d}\boldsymbol{v}.$$

Integrated Knowledge Gradient

- Let $(a^n, \boldsymbol{v}^n)$ denote the n-th sampling decision, i.e., on alternative a^n at location $\boldsymbol{v}^n$, and $S^n := (\mu_1^n, \ldots, \mu_M^n, k_1^n, \ldots, k_M^n)$ the random *state* after the n-th sample.

- If $N = 1$, the optimal strategy is

$$\operatorname*{argmax}_{1\le i\le M, \boldsymbol{x}\in\mathcal{X}} \int_{\mathcal{X}} \mathbb{E}\left[\max_{1\le a\le M} \mu_a^1(\boldsymbol{v}) \,\middle|\, S^0, a^1 = i, \boldsymbol{v}^1 = \boldsymbol{x}\right]\gamma(\boldsymbol{v})\mathrm{d}\boldsymbol{v}.$$

- Myopic Strategy: Treat each time as if there were only one sample left, and allocate the n-th sample according to

$$\operatorname*{argmax}_{1\le i\le M, \boldsymbol{x}\in\mathcal{X}} \int_{\mathcal{X}} \mathbb{E}\left[\max_{1\le a\le M} \mu_a^n(\boldsymbol{v}) \,\middle|\, S^{n-1}, a^n = i, \boldsymbol{v}^n = \boldsymbol{x}\right]\gamma(\boldsymbol{v})\mathrm{d}\boldsymbol{v}.$$

Integrated Knowledge Gradient

- Recall the myopic strategy:

$$\operatorname*{argmax}_{1\le i\le M, \boldsymbol{x}\in\mathcal{X}} \int_{\mathcal{X}} \mathbb{E}\left[\max_{1\le a\le M} \mu_a^n(\boldsymbol{v}) \,\middle|\, S^{n-1}, a^n = i, \boldsymbol{v}^n = \boldsymbol{x}\right]\gamma(\boldsymbol{v})\mathsf{d}\boldsymbol{v}.$$

Integrated Knowledge Gradient

- Recall the myopic strategy:

$$\underset{1\le i\le M, \boldsymbol{x}\in\mathcal{X}}{\operatorname{argmax}} \int_{\mathcal{X}} \mathbb{E}\left[\max_{1\le a\le M} \mu_a^n(\boldsymbol{v}) \,\Big|\, S^{n-1}, a^n = i, \boldsymbol{v}^n = \boldsymbol{x}\right]\gamma(\boldsymbol{v})\mathrm{d}\boldsymbol{v}.$$

- It is equivalent to maximizing

$$\int_{\mathcal{X}} \mathbb{E}\left[\max_{1\le a\le M} \mu_a^n(\boldsymbol{v}) - \underbrace{\max_{1\le a\le M} \mu_a^{n-1}(\boldsymbol{v})}_{\text{irrelevant to } (i,\, \boldsymbol{x})} \,\Big|\, S^{n-1}, a^n = i, \boldsymbol{v}^n = \boldsymbol{x}\right]\gamma(\boldsymbol{v})\mathrm{d}\boldsymbol{v}.$$

Integrated Knowledge Gradient

- Recall the myopic strategy:

$$\underset{1\le i\le M, \boldsymbol{x}\in\mathcal{X}}{\operatorname{argmax}} \int_{\mathcal{X}} \mathbb{E}\left[\max_{1\le a\le M} \mu_a^n(\boldsymbol{v}) \,\middle|\, S^{n-1}, a^n = i, \boldsymbol{v}^n = \boldsymbol{x}\right] \gamma(\boldsymbol{v}) \mathrm{d}\boldsymbol{v}.$$

- It is equivalent to maximizing

$$\int_{\mathcal{X}} \mathbb{E}\left[\max_{1\le a\le M} \mu_a^n(\boldsymbol{v}) - \underbrace{\max_{1\le a\le M} \mu_a^{n-1}(\boldsymbol{v})}_{\text{irrelevant to } (i,\, \boldsymbol{x})} \,\middle|\, S^{n-1}, a^n = i, \boldsymbol{v}^n = \boldsymbol{x}\right] \gamma(\boldsymbol{v}) \mathrm{d}\boldsymbol{v}.$$

- It is the expected value of information gained by sampling $(i, \boldsymbol{x})$, integrated over the domain of covariates.

Integrated Knowledge Gradient

- Recall the myopic strategy:

$$\underset{1\le i\le M, \boldsymbol{x}\in\mathcal{X}}{\operatorname{argmax}} \int_{\mathcal{X}} \mathbb{E}\left[\max_{1\le a\le M} \mu_a^n(\boldsymbol{v}) \,\middle|\, S^{n-1}, a^n = i, \boldsymbol{v}^n = \boldsymbol{x} \right] \gamma(\boldsymbol{v}) \mathrm{d}\boldsymbol{v}.$$

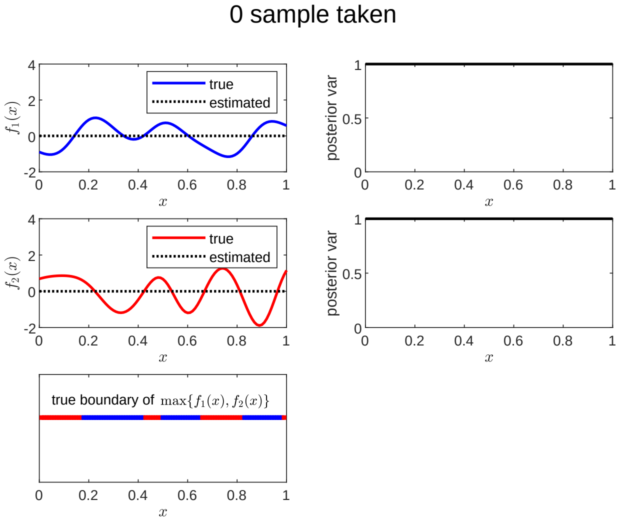
- It is equivalent to maximizing

$$\int_{\mathcal{X}} \mathbb{E}\left[\max_{1\le a\le M} \mu_a^n(\boldsymbol{v}) - \underbrace{\max_{1\le a\le M} \mu_a^{n-1}(\boldsymbol{v})}_{\text{irrelevant to } (i,\, \boldsymbol{x})} \,\middle|\, S^{n-1}, a^n = i, \boldsymbol{v}^n = \boldsymbol{x} \right] \gamma(\boldsymbol{v}) \mathrm{d}\boldsymbol{v}.$$

- It is the expected value of information gained by sampling $(i, \boldsymbol{x})$, integrated over the domain of covariates.

- We always search for $(i, \boldsymbol{x})$ that maximizes such integrated expected information gain, thus refer it as Integrated Knowledge Gradient (IKG) sampling strategy.

Numerical Illustration



Theoretical Result

Theorem 1

Under some mild assumptions, the IKG sampling strategy is consistent, that is, as $N \to \infty$, for all $\boldsymbol{x} \in \mathcal{X}$,

(i) $k_i^N(\boldsymbol{x}, \boldsymbol{x}) \to 0$ *a.s. for* $i = 1, \ldots, M$;

(ii) $\mu_i^N(\boldsymbol{x}) \to \theta_i(\boldsymbol{x})$ *a.s. for* $i = 1, \ldots, M$;

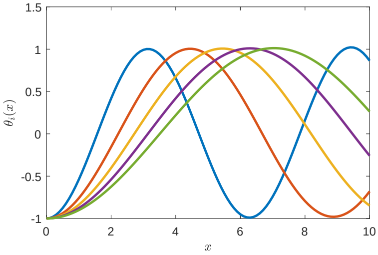
(iii) $\operatorname{argmax}_i \mu_i^N(\boldsymbol{x}) \to \operatorname{argmax}_i \theta_i(\boldsymbol{x})$ *a.s.*

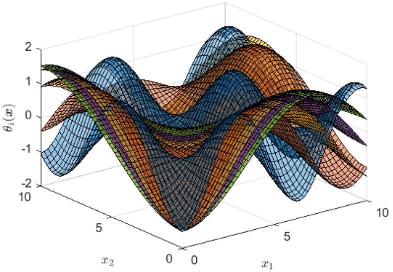
Synthetic Problem

- We consider $M = 5$ alternatives with mean performance surfaces

$$\theta_i(\boldsymbol{x}) = \sum_{j=1}^{d} \frac{x_j^2}{4000} - 1.5^{d-1} \prod_{j=1}^{d} \cos\left(\frac{x_j}{\sqrt{ij}}\right), \quad \boldsymbol{x} \in \mathcal{X} = [0, 10]^d, i = 1, \ldots, 5.$$

- Visualization of the 5 surfaces for $d = 1, 2$



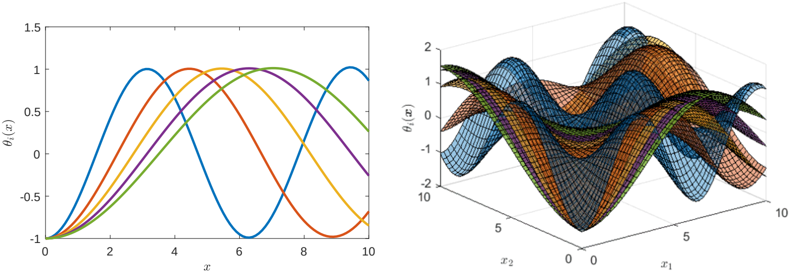


Synthetic Problem

- We consider $M = 5$ alternatives with mean performance surfaces

$$\theta_i(\boldsymbol{x}) = \sum_{j=1}^{d} \frac{x_j^2}{4000} - 1.5^{d-1} \prod_{j=1}^{d} \cos\left(\frac{x_j}{\sqrt{ij}}\right), \quad \boldsymbol{x} \in \mathcal{X} = [0, 10]^d, i = 1, \ldots, 5.$$

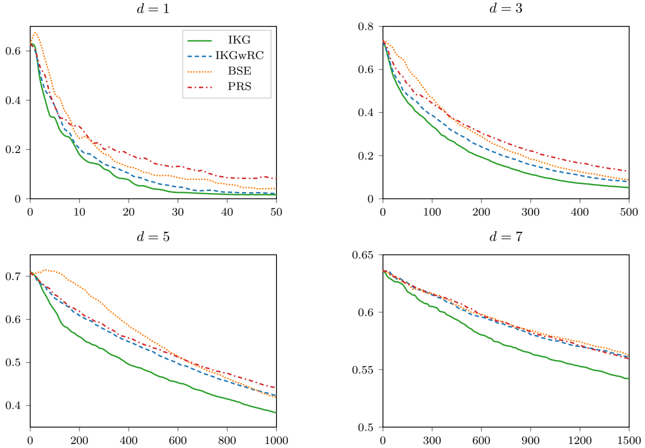
- Visualization of the 5 surfaces for $d = 1, 2$



- Sampling variance $\lambda_i(\boldsymbol{x}) \equiv 0.01$; Uniformly distributed covariates $\boldsymbol{x}$.

Results

- Take prior $\mu_i^0(\boldsymbol{x}) \equiv 0$ and $k_i^0(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\frac{1}{d}\|\boldsymbol{x} - \boldsymbol{x}'\|^2)$.



Estimated Opportunity Cost (vertical) as a function of the Sampling Budget (horizontal)

Concluding Remarks

- We propose an IKG sampling strategy, which is suitable for more general situation.
- We provide a theoretical analysis of the asymptotic behavior of the sampling strategy.
- We propose a SGA algorithm to solve the sampling strategy.

References

- Frazier, P. I., W. Powell, and S. Dayanik (2008). A knowledge gradient policy for sequential information collection. *SIAM J. Control Optim.* 47(5), 2410-2439.
- L'Ecuyer, P. (1995). Note: On the interchange of derivative and expectation for likelihood ratio derivative estimators. *Manag. Sci.* 41(4), 738-747.
- Pearce, M. and J. Branke (2017). Efficient expected improvement estimation for continuous multiple ranking and selection. In *Proc. 2017 Winter Simulation Conf.*, 2161-2172.
- Ryzhov, I. O. (2016). On the convergence rates of expected improvement methods. *Oper. Res.* 64(6), 1515-1528.

Thank you for your attention!

The full paper is available at https://arxiv.org/abs/1906.05098

Haihui Shen
shenhaihui@sjtu.edu.cn

November, 2020