Bayesian Sequential Learning for Contextual Selection

Xiaowei Zhang (City University of Hong Kong)

Mostly OM 2019

Joint work with Liang Ding (HKUST), L. Jeff Hong (Fudan), and Haihui Shen (CityU)

Selection of the Best

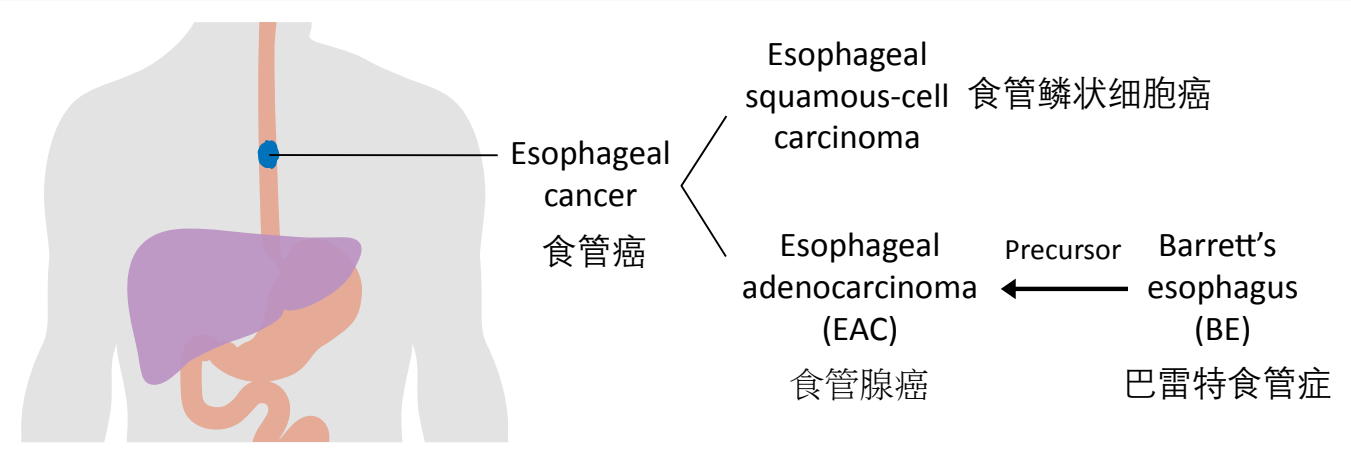
- Select the best from a finite set of alternatives, whose performances are unknown and can only be learned by sampling
- E.g., treatment selection, display advertising, inventory management

- Sampling is expensive, thereby budget-constrained
  - can afford some time to make a good sampling decision
- **Goal:** a sampling strategy to learn the performances and identify the best as efficiently as possible

Contextual Information

- Covariates: age, gender, browsing history, location
- "Best" is not universal but depends on the context
  - general v.s. specific
- Personalized decision-making emerges (big data and advanced IT)
  - precision medicine
  - customized advertisement
  - robo-advisor
  - smart building

Personalized Cancer Prevention

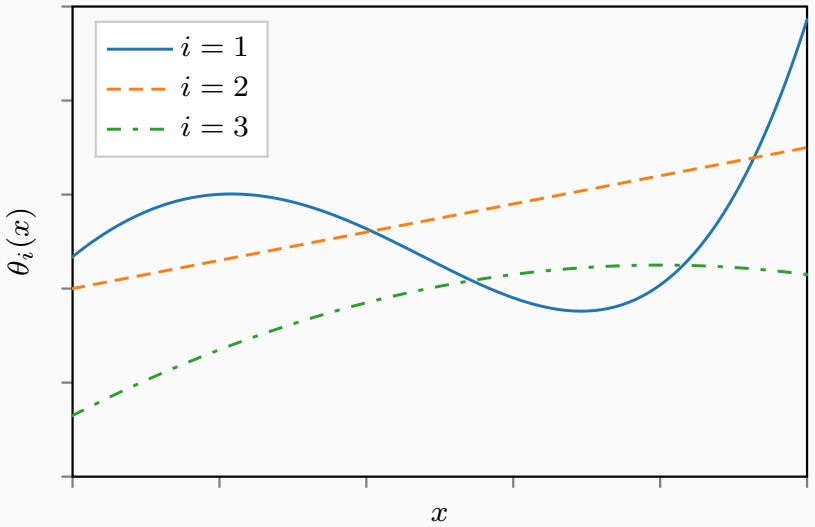


Esophageal cancer is a leading cancer among males (4th in China and 7th in U.S.)

- BE is a precursor to EAC and its management has drawn much attention
- 3 treatment regimens:
  (1) no drug
  (2) aspirin chemoprevention
  (3) statin chemoprevention
- 4 covariates:
  (1) age
  (2) annual progression rate of BE to EAC
  (3) effect of aspirin
  (4) effect of statin

Contextual Selection

- Consider M alternatives with performances $\theta_1(\boldsymbol{x}), \ldots, \theta_M(\boldsymbol{x})$, where $\boldsymbol{x} = (x_1, \ldots, x_d)^\intercal$ are the covariates

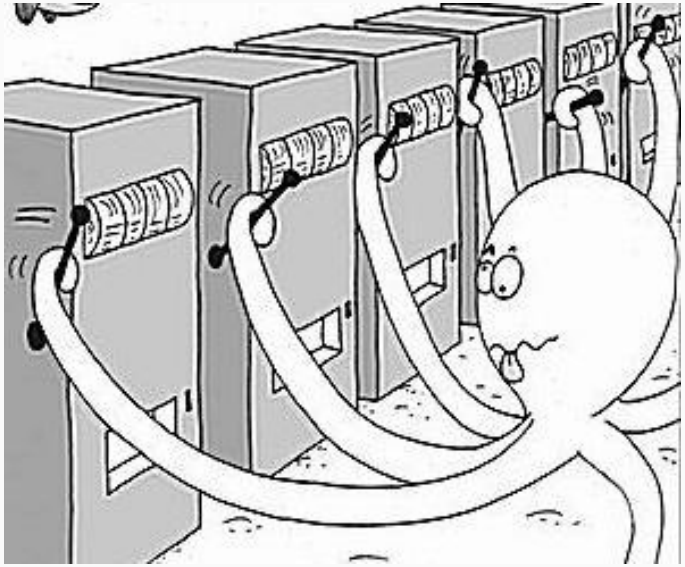


- **Goal:** develop a sequential sampling strategy to learn the decision rule

$$i^*(\boldsymbol{x}) := \underset{1 \leq i \leq M}{\arg\max}\, \theta_i(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathcal{X}$$

  - The decision rule is estimated *offline* and is then applied *online* to subsequent arriving individuals
- A sampling strategy specifies where to take a sample, namely $(i, \boldsymbol{x})$, given available information at time n

Related Problem: Multi-armed Bandit



- Classical framework for sequential decision-making (Robbins, 1952)
  - select an alternative
  - take a (random) sample
  - update estimates
- “Arm”: treatment regimen, display of ads, inventory policy

Related Literature

- Contextual bandit
  - Goldenshluger and Zeevi (2013); Perchet and Rigollet (2013)
  - minimize "regret" relative to an oracle, rather than identify the best
  - covariates arrive randomly, instead of being actively selected
  - sampling decision chooses i only, instead of $(i, \boldsymbol{x})$
- Individualized treatment rules (ITR)
  - Qian and Murphy (2011); Zhao et al. (2012)
  - learn ITR from the given data of (covariate, treatment, response)
  - does not involve design of sampling strategy
- Ranking and selection
  - Kim and Nelson (2001); Frazier et al. (2008)
  - no covariates: $\max_i \theta_i$
- Bayesian optimization
  - Shahriari et al. (2016)
  - optimize one single unknown function: $\max_{\boldsymbol{x}} \theta(\boldsymbol{x})$
- Active learning
  - Settles (2012)
  - binary responses and discrete-valued covariates

Typical Structure of Bayesian Sequential Learning

(i) Calculate the posterior distribution of $(\theta_1, \ldots, \theta_M)$ based on the samples collected so far

(ii) Use the posterior to decide the next sampling location via certain criterion
   - often formulated as an optimization problem

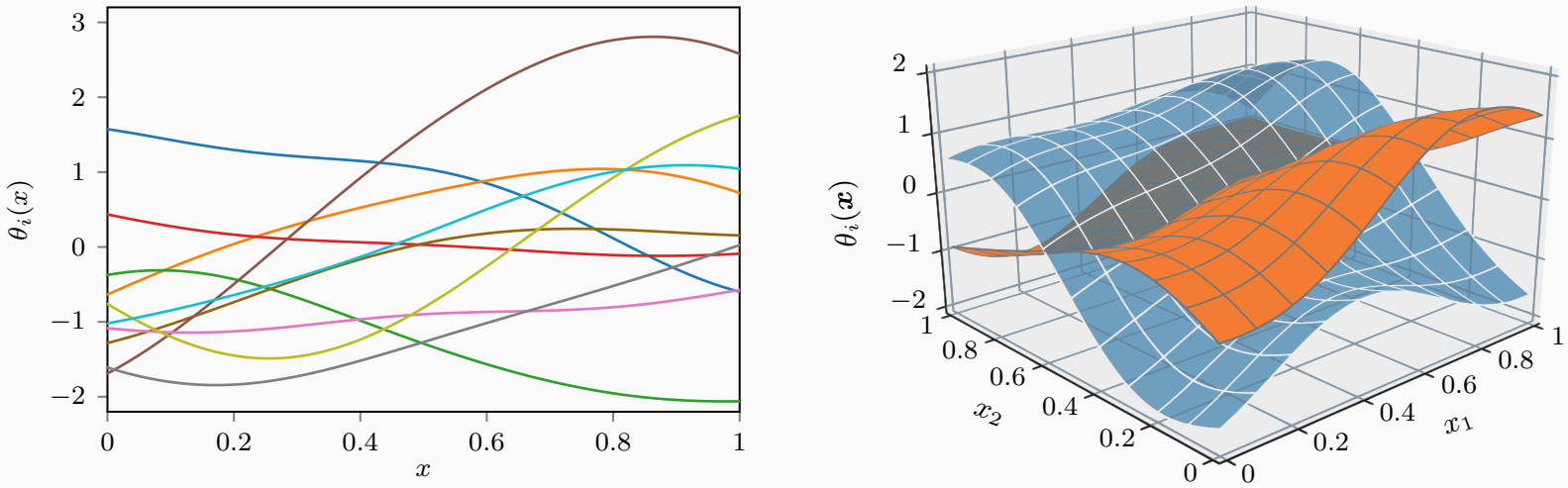
(iii) Take a (noisy) sample at the chosen location

(iv) Iterate until the sampling budget is exhausted

- Sampling strategies differ in
  - model for $(\theta_1, \ldots, \theta_M)$: linear, nonparametric
  - model for the sampling noise
  - deterministic or randomized sampling decision
  - criterion for choosing the sampling decision

Nonparametric Bayesian Formulation

- Treat $\{\theta_1(\boldsymbol{x}), \ldots, \theta_M(\boldsymbol{x})\}$ as random functions and assume a prior under which they are independent Gaussian processes (GPs)
- GP is specified by mean function $\mu(\boldsymbol{x})$ and covariance function $k(\boldsymbol{x}, \boldsymbol{x}')$



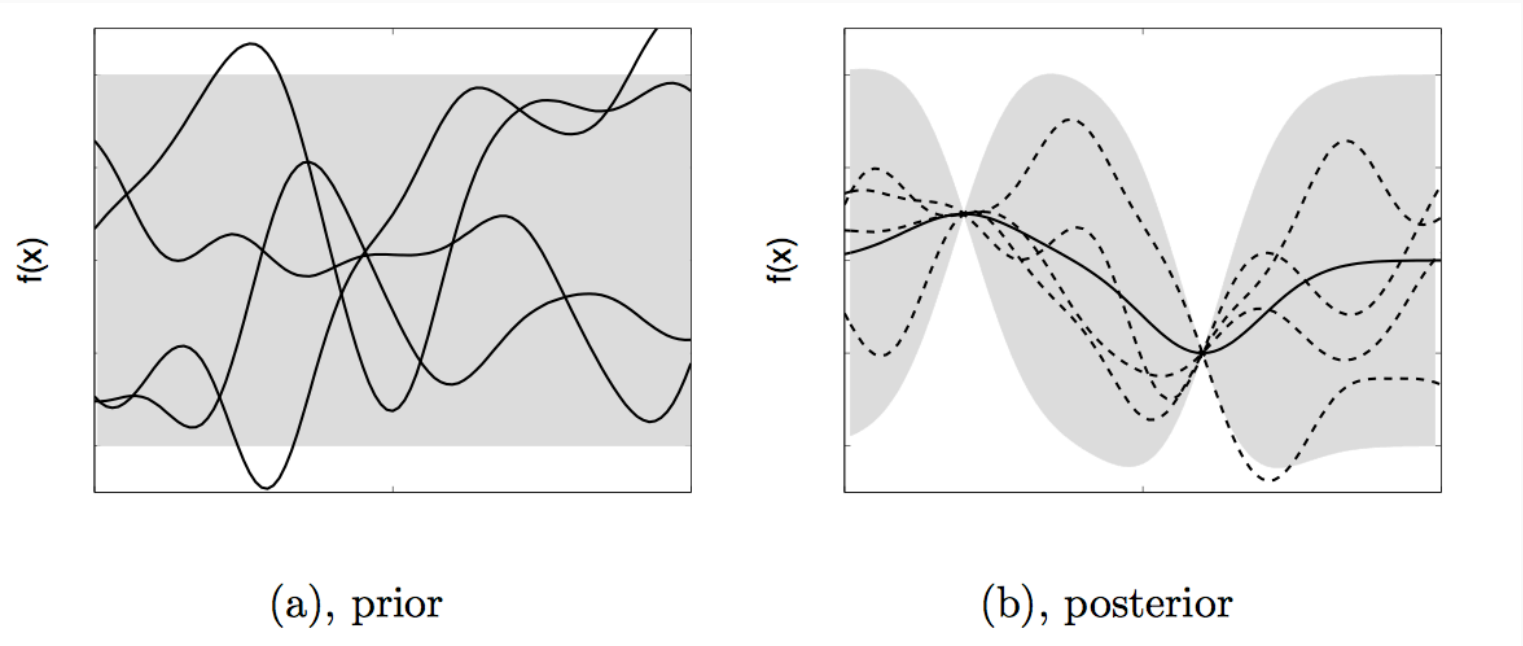
GP Realizations in 1-D and 2-D

Gaussian Process Regression

- Sampling decision at time n is $(a^n, \boldsymbol{v}^n)$: take a sample of $\theta_{a^n}(\boldsymbol{v}^n)$
  - alternative $a^n \in \{1, \ldots, M\}$ at location $\boldsymbol{v}^n \in \mathcal{X}$
- The sample y^{n+1} is an independent, unbiased, and normally distributed

$$y^{n+1} \mid \theta_{a^n}(\boldsymbol{v}^n) \sim \theta_{a^n}(\boldsymbol{v}^n) + \mathcal{N}(0, \lambda_{a^n}).$$

- Under the posterior, $\{\theta_1(\boldsymbol{x}), \ldots, \theta_M(\boldsymbol{x})\}$ are independent GPs



(a), prior (b), posterior

Objective of Sampling Strategy

- After sampling budget N is exhausted, we would estimate $i^*(\boldsymbol{x})$ based on the posterior means of $\theta_i(\boldsymbol{x})$

$$i^*(\boldsymbol{x}) \approx \arg\max_{1\le i\le M} \mathbb{E}[\theta_i(\boldsymbol{x})|\mathcal{F}^N] = \arg\max_{1\le i\le M} \mu_i^N(\boldsymbol{x})$$

- View $\max_i \mu_i^N(\boldsymbol{x})$ as a "reward"
  - its expected value depends on $\pi = \{(a^n, \boldsymbol{v}^n) : n = 0, \dots, N-1\}$
  - maximize the reward $\iff$ minimize the "opportunity cost" $\max_i \theta_i(\boldsymbol{x}) - \max_i \mu_i^N(\boldsymbol{x})$
- The objective becomes

$$\sup_{\pi\in\Pi} \mathbb{E}^\pi \left[\int_{\mathcal{X}} \max_{1\le i\le M} \mu_i^N(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right]$$

  - only terminal reward, no intermediate rewards are collected
  - samples are expensive to acquire and N is usually not large
  - Focuses on identifying the best alternative, rather than minimizing the accumulated regret

Myopic Sampling Strategy

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

$$\underset{1\le i\le M, x\in\mathcal{X}}{\arg\max}\ \mathbb{E}\left[\int_{\mathcal{X}} \max_{1\le a\le M} \mu_a^1(v)\mathrm{d}v \,\middle|\, S^0 = s, a^0 = i, v^0 = x\right]$$

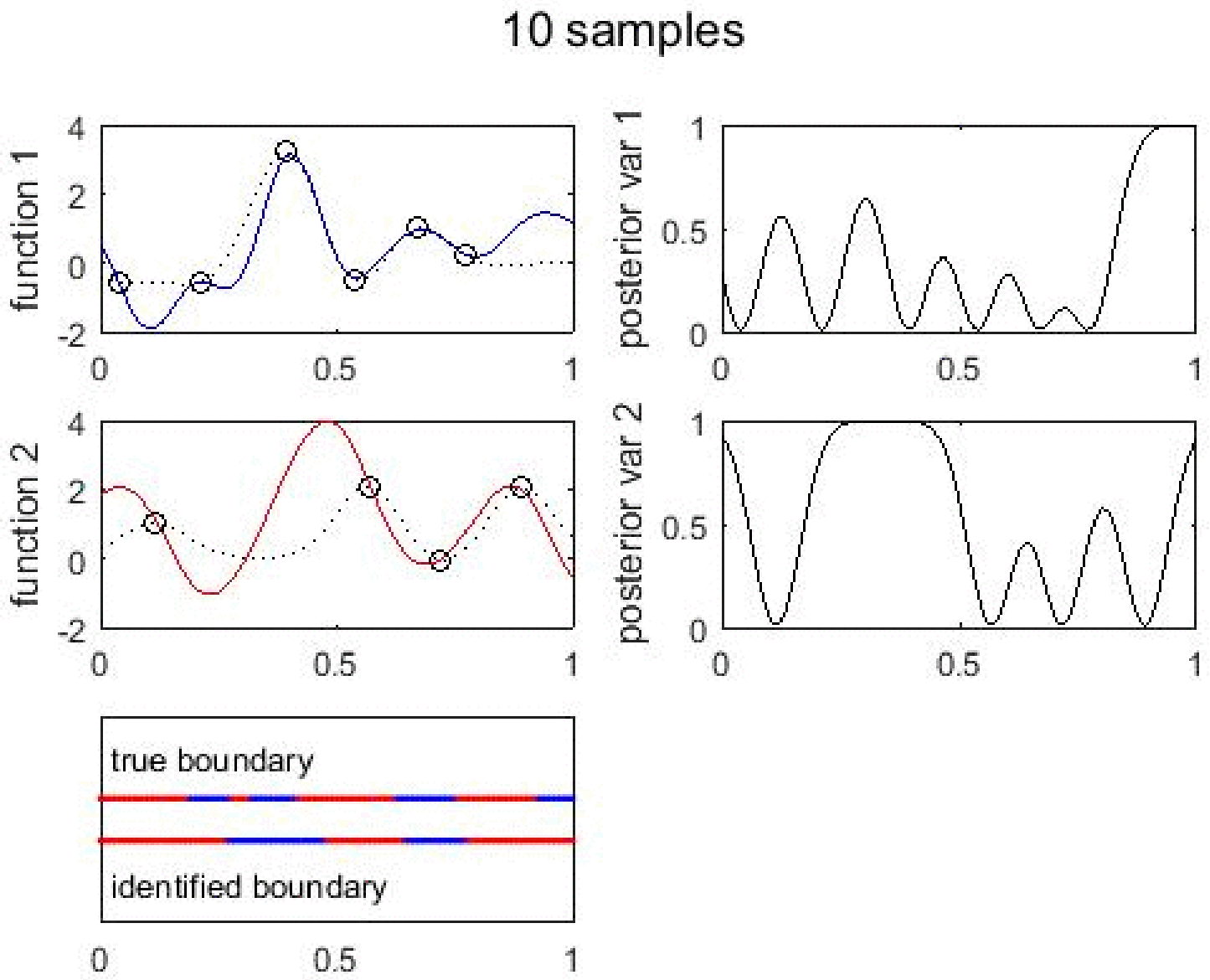
- Myopic: treat each time period as if there were only one sample left

$$\underset{1\le i\le M, x\in\mathcal{X}}{\arg\max}\ \mathbb{E}\left[\int_{\mathcal{X}} \max_{1\le a\le M} \mu_a^{n+1}(v)\mathrm{d}v \,\middle|\, S^n = s, a^n = i, v^n = x\right]$$

- This is equivalent to maximizing

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

  - increment in the expected value of information gained by sampling (i, x)


10 samples
function 1
function 2
posterior var 1
posterior var 2
true boundary
identified boundary
4
2
0
-2
0
0.5
1

Asymptotic Analysis

Theorem (Z., Shen, Hong, and Ding, 2019)

The myopic sampling strategy is consistent*:*

(i) $\mathrm{Var}^N[\theta_i(\boldsymbol{x})] \to 0$ *a.s.*

(ii) $\mu_i^N(\boldsymbol{x}) \to \theta_i(\boldsymbol{x})$ *a.s.*

(iii) $\arg\max_i \mu_i^N(\boldsymbol{x}) \to \arg\max_i \theta_i(\boldsymbol{x})$ *a.s.*

under the following assumptions

(i) $k_i^0(\boldsymbol{x}, \boldsymbol{x}') = \tau_i^2 \rho_i(|\boldsymbol{x} - \boldsymbol{x}'|)$ *and* ρ_i *is positive, continuous, decreasing*

(ii) μ_i^0 *and* $\lambda_i(\cdot)$ *are continuous*

(iii) $\mathcal{X}$ *is compact with nonempty interior*

Proof Sketch

- Step 1: Fix i. Assume an alternative i is sampled infinitely often. Show $\mathrm{Var}^N[\theta_i(\boldsymbol{x})] \to \infty$ for all $\boldsymbol{x}$ under the sampling strategy
  - a key, new technical result is to prove k_i^n converges *uniformly* as $n \to \infty$ using reproducing kernel Hilbert space theory

- Step 2: Show no alternatives are sampled only finite times as $N \to \infty$ under the sampling strategy

Generalization

- One may have prior knowledge with regard to the covariates
  - certain values may be more important or appear more frequently than others
- Suppose the prior knowledge is expressed by a probability density function $\gamma(\cdot)$ on $\mathcal{X}$
  - objective becomes

$$\sup_{\pi \in \Pi} \mathbb{E}^{\pi} \left[\int_{\mathcal{X}} \max_{1 \le i \le M} \mu_i^N(x) \gamma(x) \mathrm{d}x \right]$$

  - the myopic strategy becomes

$$\arg\max_{1 \le i \le M, x \in \mathcal{X}} \mathbb{E} \left[\int_{\mathcal{X}} \max_{1 \le a \le M} \mu_a^{n+1}(v) \gamma(v) \mathrm{d}v \,\middle|\, S^n = s, a^n = i, v^n = x \right]$$

- The asymptotic analysis still holds

Computation

- Key component is $h_i^n(\boldsymbol{v}, \boldsymbol{x}) := \mathbb{E}\left[\max_a \mu_a^{n+1}(\boldsymbol{v}) | S^n, (a^n, \boldsymbol{v}^n) = (i, \boldsymbol{x})\right]$
  - the myopic strategy is to solve

$$\max_{1 \le i \le M, \boldsymbol{x} \in \mathcal{X}} \int_{\mathcal{X}} h_i^n(\boldsymbol{v}, \boldsymbol{x}) \gamma(\boldsymbol{v}) \mathrm{d}\boldsymbol{v}$$

- Given S^n, the predictive distribution of $\mu_a^{n+1}(\boldsymbol{v})$ before sampling $(i, \boldsymbol{x})$ is

$$\mu_a^{n+1}(\boldsymbol{v}) = \begin{cases} \mu_a^n(\boldsymbol{v}) + \tilde{\sigma}_a^n(\boldsymbol{v}, \boldsymbol{x}) Z^{n+1}, & \text{if } a = i \\ \mu_a^n(\boldsymbol{v}), & \text{if } a \neq i \end{cases}$$

  where $\tilde{\sigma}_a^n(\boldsymbol{v}, \boldsymbol{x}) = \frac{k_a^n(\boldsymbol{v}, \boldsymbol{x})}{k_a^n(\boldsymbol{x}, \boldsymbol{x}) + \lambda_a}$

- $h_i^n(\boldsymbol{v}, \boldsymbol{x})$ is in the form of $\mathbb{E}[(\alpha + \beta Z) \vee \gamma]$ and can be expressed in terms of the standard normal distribution functions

$$h_i^n(\boldsymbol{v}, \boldsymbol{x}) = |\tilde{\sigma}_i^n(\boldsymbol{v}, \boldsymbol{x})| \phi\left(\left|\frac{\Delta_i^n(\boldsymbol{v})}{\tilde{\sigma}_i^n(\boldsymbol{v}, \boldsymbol{x})}\right|\right) - |\Delta_i^n(\boldsymbol{v})| \Phi\left(-\left|\frac{\Delta_i^n(\boldsymbol{v})}{\tilde{\sigma}_i^n(\boldsymbol{v}, \boldsymbol{x})}\right|\right)$$

Stochastic Gradient Ascent

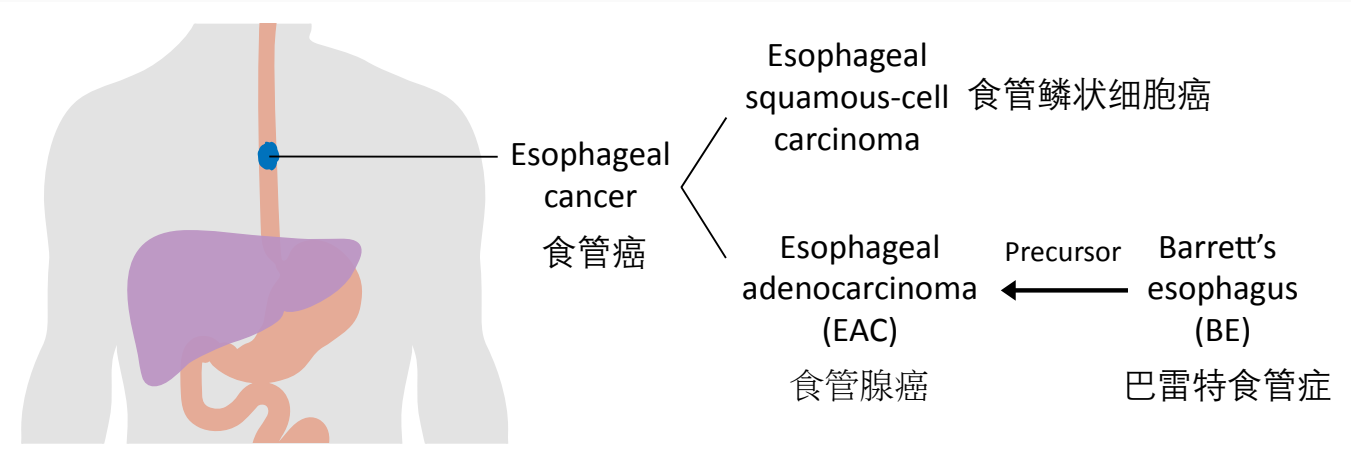
- For each i, we need to solve

$$\max_{x \in \mathcal{X}} \int_{\mathcal{X}} h_i^n(v, x)\gamma(v)\mathrm{d}v = \max_x \mathbb{E}[h_i^n(\xi, x)],$$

  where ξ is a $\mathcal{X}$-valued random variable with density $\gamma(\cdot)$

- Use SGA to solve the optimization problem
  - sample average approximation is too slow
- $\frac{\partial}{\partial x} h_i^n(\xi, x)$ is an unbiased estimator of $\frac{\partial}{\partial x}\mathbb{E}[h_i^n(\xi, x)]$ under certain regularity conditions, and it can be derived analytically for various prior covariance functions

Background



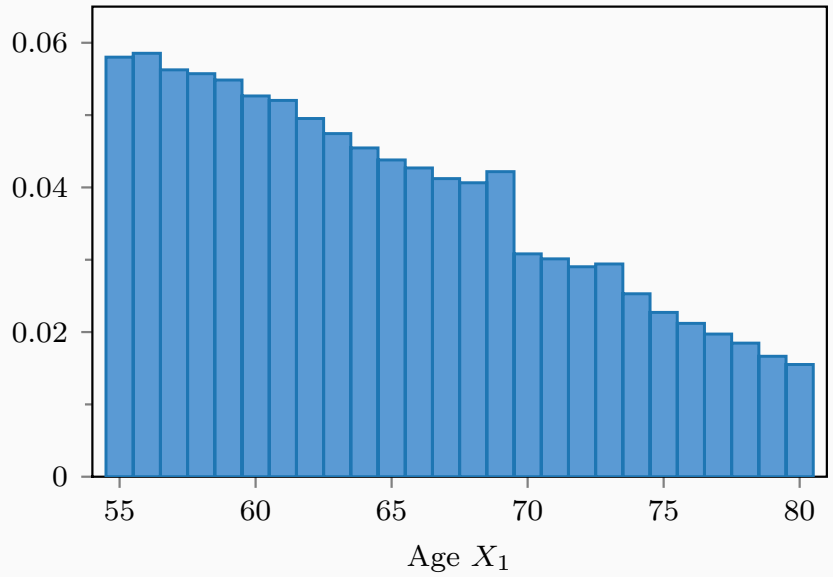
- Select the best among 3 treatment regimens for BE
  - (1) no drug
  - (2) aspirin chemoprevention
  - (3) statin chemoprevention

  as a function of 4 individual characteristics
  - X_1 age
  - X_2 annual progression rate of BE to EAC
  - X_3 effect of aspirin (i.e., progression reduction effect)
  - X_4 effect of statin

Distribution of Covariates

- Assume $X_1, \cdots, X_4$ are independent

Covariates	Distributions	Support	Mean
X_1	Discrete (Figure below)	$\{55, \cdots, 80\}$	64.56
X_2	Unif $(0, 0.1)$	$[0, 0.1]$	0.05
X_3	Triangular $(0, 0.59, 1)$	$[0, 1]$	0.53
X_4	Triangular $(0, 0.62, 1)$	$[0, 1]$	0.54

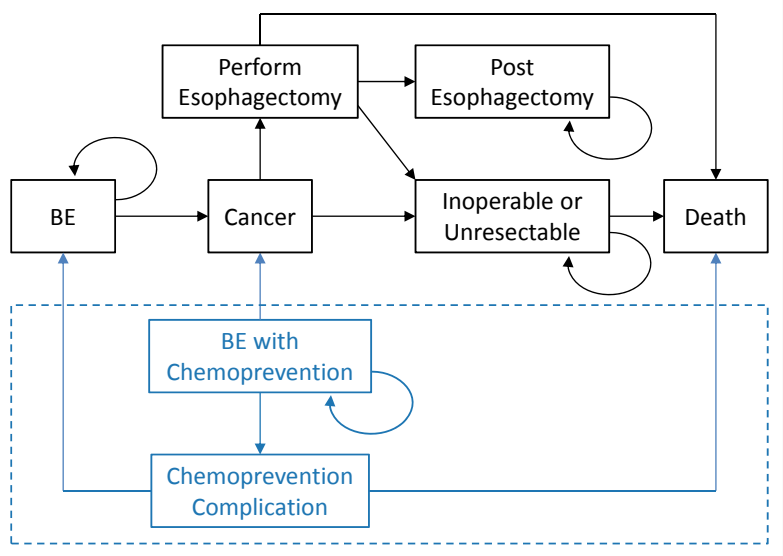


Probability mass function of X_1 (truncated).

Source: U.S. 2013 population data, U.S. Census Bureau.

Markov Model

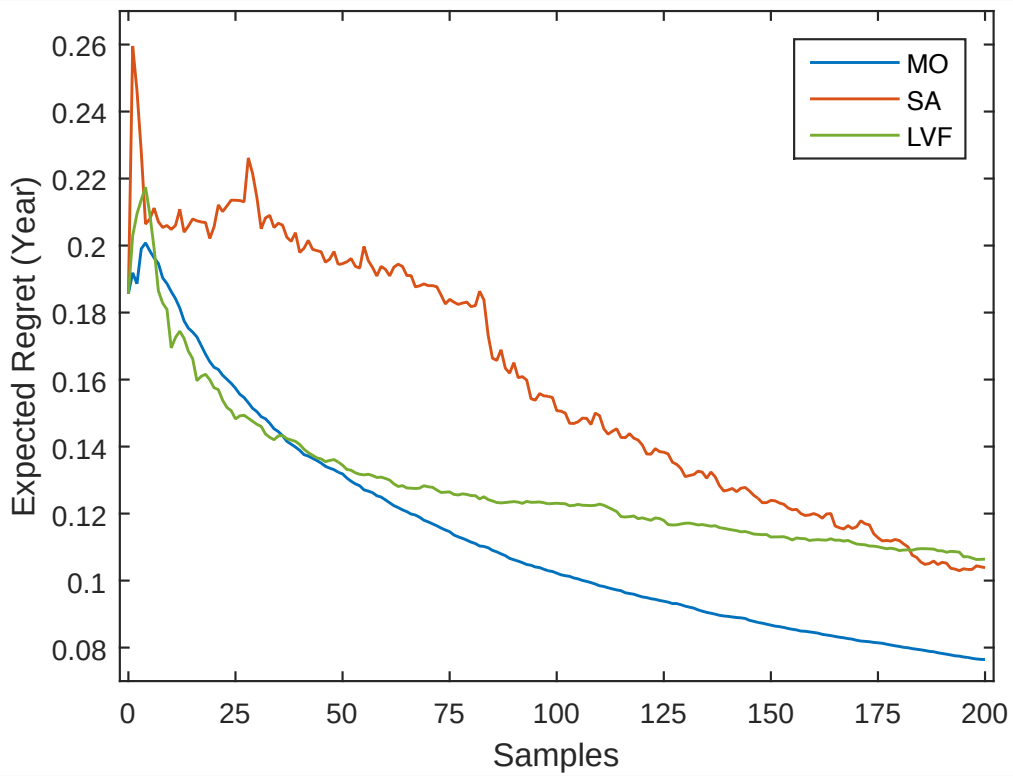
- A Markov chain model was developed by Hur et al. (2004) and Choi et al. (2014) to study the effectiveness of aspirin and statin chemoprevention against EAC



- A male with BE goes through various health state until death
- The person in each state can die from age-related all-cause mortality
- The time length between state transition is one month
- Detailed structure inside dotted box depends on drug
- Parameters are well calibrated

- Output $Y_i(\boldsymbol{X})$: Quality-adjusted life years (QALYs) after the starting age under treatment regimen i conditioning on $\boldsymbol{X}$.

Results



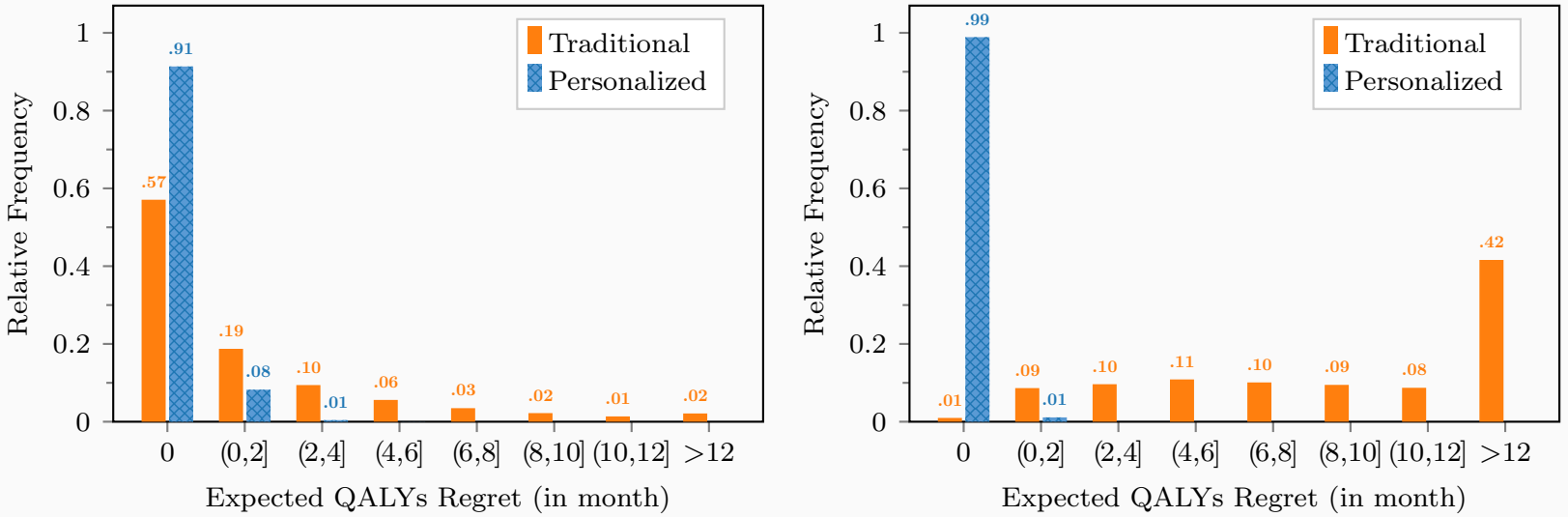
- *Regret*: difference in expected QALYs between the true optimal treatment and the selected treatment
- SA (successive allocation): choose i in a round-robin fashion and choose $\boldsymbol{x}$ uniformly
- LVF (large-variance-first): choose $(i, \boldsymbol{x})$ that maximizes $\mathrm{Var}^n(\theta_i(\boldsymbol{x}))$

Added Value of Contextual Information: Larger Expected QALYs

- Traditional: regret is $\theta_{i^*(X)}(X) - \theta_{i^\dagger}(X) = \max_i \theta_i(X) - \theta_{i^\dagger}(X)$, where

$$i^\dagger := \arg\max_{1 \le i \le M} \{\mathbb{E}[\mu_i(X)]\}$$

- Personalized: regret is $\max_i \theta_i(X) - \theta_{\widehat{i^*}(X)}(X)$, where $\widehat{i^*}(X)$ is computed via the myopic sampling strategy



Left: The *entire population*. Right: A *specific group* with $X = (X_1, X_2, 0.9, 0.2)^\intercal$.

Concluding Remarks

- Emergence of personalization/customization in business analytics motivates us to consider *contextual selection*
  - performance of each alternative is unknown
  - samples are noisy and expensive to acquire

- Develop a nonparametric Bayesian sampling strategy to learn the decision rule as a function of the covariates
  - the "intermediate regret" is discarded
  - assume we can choose the value of covariates
  - the decision rule is estimated *offline* and is then applied *online* to subsequent arriving individuals
  - more suitable for scenarios where sampling budget N is small or moderate

- Showcase the developed approach via personalized cancer prevention

Thanks!

References

S. E. Choi, K. E. Perzan, A. C. Tramontano, C. Y. Kong, and C. Hur. Statins and aspirin for chemoprevention in Barrett's esophagus: Results of a cost-effectiveness analysis. *Canc. Prev. Res.*, 7(3):341–350, 2014.

P. I. Frazier, W. B. Powell, and S. Dayanik. A knowledge-gradient policy for sequential information collection. *SIAM J. Control Optim.*, 47(5):2410–2439, 2008.

A. Goldenshluger and A. Zeevi. A linear response bandit problem. *Stoch. Syst.*, 3(1):230–261, 2013.

C. Hur, N. S. Nishioka, and G. S. Gazelle. Cost-effectiveness of aspirin chemoprevention for Barrett's esophagus. *J. Natl. Canc. Inst.*, 96(4):316–325, 2004.

S.-H. Kim and B. L. Nelson. A fully sequential procedure for indifference-zone selection in simulation. *ACM Trans. Model. Comput. Simul.*, 11(3):251–273, 2001.

V. Perchet and P. Rigollet. The multi-armed bandit problem with covariates. *Ann. Stat.*, 41(2):693–721, 2013.

M. Qian and S. A. Murphy. Performance guarantees for individualized treatment rules. *Ann. Stat.*, 39(2):1180–1210, 2011.

H. Robbins. Some aspects of the sequential design of experiments. *Bull. Am. Math. Soc.*, 58(5):527–535, 1952.

B. Settles. *Active Learning*. Morgan & Claypool, 2012.

B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas. Taking the human out of the loop: A review of Bayesian optimization. *Proc. IEEE*, 104 (1):148 – 175, 2016.

Y. Zhao, D. Zeng, A. J. Rush, and M. R. Kosorok. Estimating individualized treatment rules using outcome weighted learning. *J. Amer. Statist. Assoc.*, 107(499):1106–1118, 2012.