Bayesian Optimization via Simulation with Pairwise Sampling & Correlated Beliefs

Jing Xie, Peter I. Frazier, Stephen E. Chick

School of Operations Research & Information Engineering, Cornell University, USA
Technology & Operations Management Area, INSEAD, FRANCE

Motivating Problems

- What is the best staffing level in a hospital?
- What is the best admission control policy in a call center?
- What is the best ambulance positioning design in a city?

Introduction

More generally, we have a finite number k of alternative systems whose performance **cannot** be evaluated directly.

- We use a stochastic simulator to evaluate the mean performance θ_x for each $x = 1, \ldots, k$ with **noise**. However, simulation is **expensive**, and our sampling budget is **limited**.
- Our **goal** is to find: $x^* = \text{argmax}_x \, \theta_x$.

How to Allocate Simulation Effort Efficiently?

- We use a **Bayesian value-of-information (VOI)** approach (Howard 1966, Chick & Gans 2009, Chick & Frazier 2009) to design sampling strategies. They are called
  - **myopic one-step lookahead** (Jones et al. 1998, Lizotte et al. 2007, Frazier et al. 2009, Chick et al. 2010),
  - or **knowledge-gradient (KG)** policies (Frazier 2009).

Basic Idea:

- Given the function evaluations obtained, use Bayesian statistics (e.g., **Gaussian Process (GP) regression**) to get:
  - **estimates** of $\vec{\theta} = (\theta_1, \ldots, \theta_k)$.
  - **uncertainties** in these estimates.
- Then, use these estimates and uncertainties to
  - **assess** the VOI of possible future evaluations.
  - **decide** where to evaluate next.

Our Novelty:

- Allow for **correlation** in prior beliefs on $\vec{\theta}$ *and across samples*.
- Allow for sampling from a **pair** in each given stage.
- Handle **large-scale** problems with many alternatives.

Modeling

Our belief about the performance of the alternatives $\vec{\theta}$ is a Gaussian process prior over a discrete set, and thus is a **correlated** multivariate normal prior, i.e., $\vec{\theta} \sim \mathcal{N}(\mu_0, \Sigma_0)$.

We allow **correlated** simulation output through **common random numbers (CRN)**, and observe $Y \sim \mathcal{N}(\vec{\theta}, \Lambda)$ if sampling all alternatives together using CRN.

At each time n we:

- calculate the Bayesian posterior distribution on $\vec{\theta}$ given what we currently know, which is $\mathcal{N}(\mu_n, \Sigma_n)$.
- choose a subset of alternatives X_{n+1} to sample at time $n+1$.
- if $n = N$ (stopping time), report $\text{argmax}\{\mu_N\}$ as the "implementation decision", else increment n and iterate.

Value of Information

- At time n, the VOI of choosing a subset of alternatives X (cf. Ginsbourger *et al.*) to evaluate at time $n+1$ is

$$V_n(X) = \mathbb{E}_n\left[\max_x \mu_{n+1,x} \mid X_{n+1} = X\right] - \max_x \mu_{n,x}.$$

  - $\mu_{n,x}$ is the expected value of θ_x at time n.
  - $\max_x \mu_{n,x}$ is the best we can do at time n.
  - $\max_x \mu_{n+1,x}$ is the best we will be able to do after seeing X_{n+1}
- $V_n(X)$ is a **value of information (VOI)**.
- $V_n(X)$ is **similar to the expected improvement (EI)**.
- $V_n(X)$ **differs from EI** (as usually defined) because it considers the change in the posterior mean at points that are not sampled.
- Frazier et al. 2011 gives analytical computation for
  - $X = x$ (singleton)
  - $X = (x^{(1)}, x^{(2)})$ (observe the **difference** between a pair)

The KG Factor is the VOI Per Unit Cost

Define the **KG factor** of a subset X of alternatives to sample by

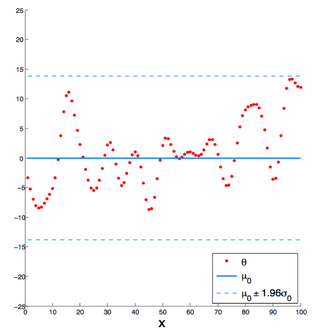
$$\nu_n^{KG}(X) := V_n(X)/c(X).$$

- $c(X)$ is the computational cost (time) for sampling X.
- $\nu_n^{KG}(X)$ is the average rate of earning VOI by sampling X.

Consider a 1-dim example with known μ_0, Σ_0 and Λ:

$k = 100$,
$\mu_0 = \vec{0}$,
$\Sigma_0(i,j) = 100 \exp\left[-\frac{(i-j)^2}{50}\right]$,
$\Lambda(i,i) = 50$,
$\Lambda(i,j) = 25$, for $i, j = 1, \ldots, 100$.

Truth & Prior Belief



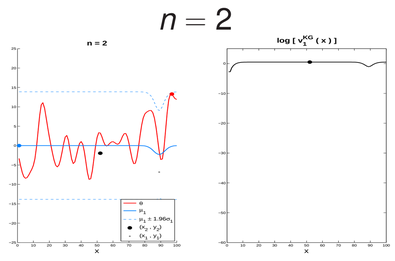
The Existing KG Algorithm

$$X_n^{KG} = \underset{x \in \{1,\ldots,k\}}{\text{argmax}} \, \nu_{n-1}^{KG}(x)$$

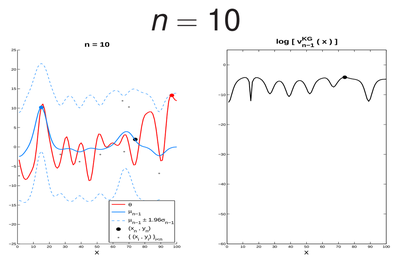
(Frazier et al. 2008, 2009; Chick & Frazier 2009)

Posterior at time $n-1$ (left) $\Rightarrow$ $\log[\nu_{n-1}(x)]$ (right) $\Rightarrow$ X_n (left).

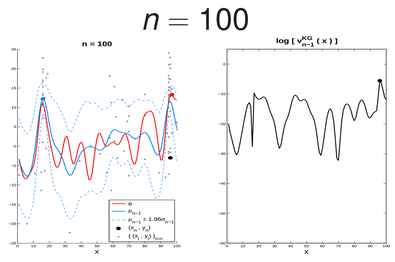
$n = 2$



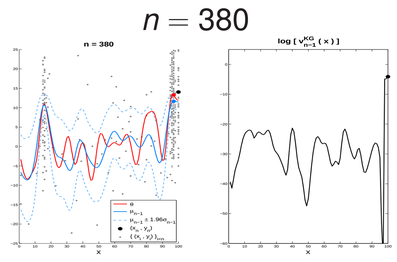
$n = 10$



$n = 100$



$n = 380$



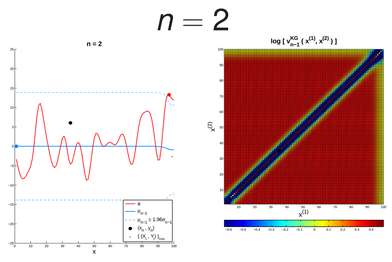
Our New KG^2 Algorithm

$$X_n^{KG} = \underset{X \in \mathcal{X}}{\text{argmax}} \, \nu_{n-1}^{KG}(X)$$

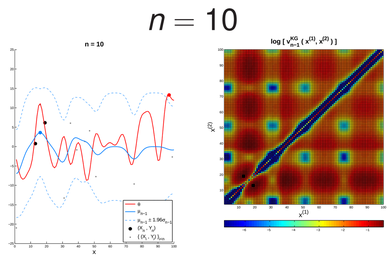
where $\mathcal{X} = \{1, \ldots, k\} \cup \{(x^1, x^2) \in \{1, \ldots, k\}^2 : x^1 < x^2\}$

Posterior at time $n-1$ (left) $\Rightarrow$ $\log[\nu_{n-1}(X)]$ (right) $\Rightarrow$ X_n (left).

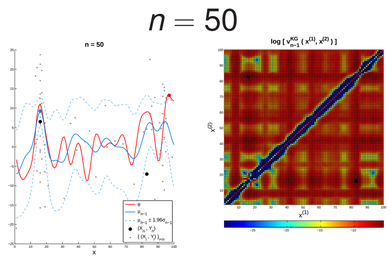
$n = 2$



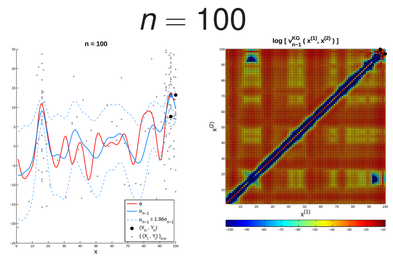
$n = 10$



$n = 50$



$n = 100$



Exploitation vs. Exploration

- KG factor ↑ as posterior mean ↑.
- KG factor ↑ as posterior variance ↑.
- KG factor balances these two tendencies in our sampling algorithm.

Why turn one optimization problem into many?

$\max_x \theta_x$ (1 opt. problem)

$\Downarrow$

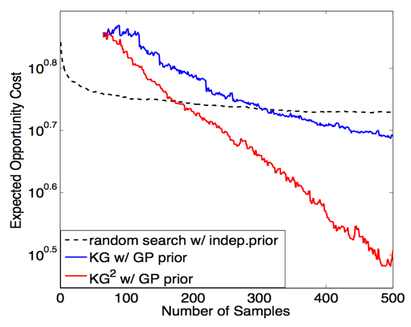
$\max_X \nu_n^{KG}(X)$ for $n \geq 1$ (many)

- Evaluating θ_x is **expensive** (minutes, hours, days), and derivative information is (often) unavailable!
- Evaluating $\nu_n^{KG}(X)$ is **quick** (microseconds), and derivative information is available (for **multi-start gradient ascent**)!

More Remarks

- **Asymptotic convergence**: $\mu_n \to \vec{\theta}$ almost surely as $n \to \infty$
- Can assess **"non-myopic"** VOI and KG factors by considering consecutive repeated samples.
- We use **maximum likelihood estimation** of the hyper-parameters in a GP prior and Λ if they are unknown.

Numerical Results: Ackley Problem with 10^5 Alternatives and "small" noise



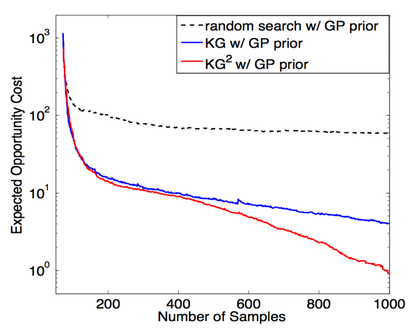
$$\theta_x = 20 \cdot \exp\left(-0.2\sqrt{\frac{1}{d}\sum_{i=1}^{d} z_i^2(x)}\right) + \exp\left(\frac{1}{d}\sum_{i=1}^{d}\cos(2\pi z_i(x))\right) - 20 - e,$$

where $\{z_i(x)\}_{i=1}^5$ are the grid coordinates of each alternative x.

$d = 5$ dimensions. $z_i \in \{-1.5, -1.\ldots, 3\}$ for $i = 1.\ldots, 5$.

$\Lambda(x, x') = 10 \cdot \exp\left[-\sum_{i=1}^5 [(z_i(x) - z_i(x'))/50]^2\right]$ for all x and x' (assumed unknown).

Numerical Results: Rosenbrock Problem with 10^6 Alternatives and "big" noise



$$\theta_x = -\sum_{i=1}^{d-1}\left[100\left[z_i(x)^2 - z_{i+1}(x)\right]^2 + [z_i(x) - 1]^2\right],$$

where $\{z_i(x)\}_{i=1}^5$ are the grid coordinates of each alternative x.

$d = 6$ dimensions. $z_i \in \{-0.8, -0.5.\ldots, 1.9\}$ for $i = 1, \ldots, 6$.

$\Lambda(x, x) = 125$ for all x and $\Lambda(x, x') = 50$ for all $x \neq x'$ (assumed unknown).

Conclusions

Our KG^2 algorithm takes advantage of two types of correlations, deals with large-scale problems, and significantly reduces the number of samples required to find a good alternative!

References

Chick, S.E., J. Branke, C. Schmidt. 2010. Sequential sampling to myopically maximize the expected value of information. *INFORMS Journal on Computing* **22**(1) 71–80.
Chick, S.E., P.I. Frazier. 2009. The conjunction of the knowledge gradient and economic approach to simulation selection. *Proceedings of the 2009 Winter Simulation Conference* 528–539.
Chick, S.E., N. Gans. 2009. Economic analysis of simulation selection problems. *Management Science* **55**(3) 421–437.
Frazier, P.I. 2009. Knowledge-gradient methods for statistical learning. Ph.D. thesis, Princeton University.
Frazier, P., W. B. Powell, S. Dayanik. 2008. A Knowledge Gradient Policy for Sequential Information Collection. *SIAM Journal on Control and Optimization* **47**(5) 2410–2439.
Frazier, P., W. B. Powell, and S. Dayanik. 2009. The Knowledge Gradient Policy for Correlated Normal Beliefs. *INFORMS Journal on Computing* **21** (4) 599–613.
Frazier, P.I., J. Xie, S.E. Chick. 2011. Value of information methods for pairwise sampling with correlations. *Proceedings of the 2011 Winter Simulation Conference* 3974–3986.
Ginsbourger, D., R. Le Riche, L. Carraro, 2007, A Multi-points Criterion for Deterministic Parallel Global Optimization based on Kriging, *Intl. Conf. on Nonconvex Programming, NCP07*, Rouen, France.
Jones, D.R., M. Schonlau, W.J. Welch. 1998. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization* **13**(4) 455–492.
Lizotte, D., T. Wang, M. Bowling, D. Schuurmans. 2007. Automatic gait optimization with gaussian process regression. *Proceedings of International Joint Conferences on Artificial Intelligence*. 944–949.

jx66@cornell.edu, pf98@cornell.edu, stephen.chick@insead.edu