Fast Bayesian Optimization via Conjugate Sampling

Yifei Ma* and Roman Garnett and Jeff Schneider***

*Carnegie Mellon University **Washington University in St. Louis

yifeim@cs.cmu.edu

Motivation

Traditional BO designs queries using the full Bayesian distribution.
Thompson sampling can be as efficient when only using a sample point [1].
However, besides conceptual simplicity, are there computational benefits by using sampling?

Background of Thompson Sampling

To maximize $f(\mathbf{x}) = f(\mathbf{x}; \boldsymbol{\theta})$, s.t. $\mathbf{x} \in \mathcal{X}$, where $\boldsymbol{\theta}$ is unknown,

assume prior for $p(\boldsymbol{\theta})$ and iterate:

sample $\tilde{\boldsymbol{\theta}} \propto p(\boldsymbol{\theta} \mid \mathbf{x}_\tau, y_\tau, \forall \tau \leq t)$

query $\mathbf{x}_{t+1} = \arg\max_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}; \tilde{\boldsymbol{\theta}})$

obtain $y_{t+1} = f(\mathbf{x}_{t+1}) + \varepsilon_{t+1},\ \varepsilon_{t+1} \sim \mathcal{N}(0, \sigma_n^2)$

Criterion: choose queries according to the probability that they are optimal.

Problem Formulation

Quickly sample from a standard form of Bayesian posterior distribution

$$\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \mathbf{A}^{-1}), \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{n \times n}$ is a positive-definite (PD) matrix.
Avoid performing the Cholesky decomposition of $\mathbf{A}$.

Table 1: Examples of Bayesian Posterior Distribution

	Bayesian Linear Regression (BLR)	Gaussian processes (GP) w/ fixed pool of choices†
constraint	$\forall \mathbf{x} \in \mathbb{R}^n, \text{s.t.} \|\mathbf{x}\|_2 \leq 1$	$\mathcal{X} = \{\mathbf{x}_1^*, \ldots, \mathbf{x}_n^*\}$
parameter	$\forall \boldsymbol{\theta} \in \mathbb{R}^n$	$\boldsymbol{\theta} = \big(f(\mathbf{x}_1^*), \ldots, f(\mathbf{x}_n^*)\big)^\top$
model	$y_\tau = \mathbf{x}_\tau^\top \boldsymbol{\theta} + \varepsilon_\tau$	$y_\tau = \boldsymbol{\theta}_{s_\tau} + \varepsilon_\tau$ for $\mathbf{x}_\tau = \mathbf{x}^*_{s_\tau}$
post. $\mathbf{A}$	$\boldsymbol{\Sigma}_p^{-1} + \frac{1}{\sigma_n^2} \sum_{\tau=1}^t \mathbf{x}_\tau \mathbf{x}_\tau^\top$	$\mathbf{K}_{**}^{-1} + \frac{1}{\sigma_n^2} \sum_{\tau=1}^t \mathbf{s}_\tau \mathbf{s}_\tau^\top$

*Assume $\boldsymbol{\Sigma}_p^{-1}$ and $\mathbf{K}_{**}^{-1}$ can be easily computed and stored.
†s_τ is the position of $\mathbf{x}_\tau$ in the list of feasible queries; $\mathbf{s}_\tau \in \mathbb{R}^n$ is the indicator vector of s_τ.

Proposed Method

Algorithm 1: Conjugate Sampling

Require: PD matrix $\mathbf{A}$, integer $k_{\max} \leq n$ ($k_{\max} = n$ for exact sampling)
Ensure: One sample point $\tilde{\boldsymbol{\eta}} \sim \mathcal{N}(\mathbf{0}, \mathbf{A}^{-1})$ or its approximation if $k < n$

let $\mathbf{x}_0 = \boldsymbol{\eta}_0 = \mathbf{p}_0 = \mathbf{0}$, $\alpha_0 = 0$, choose random $\mathbf{r}_0 = \mathbf{c} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

for $k = 1, \ldots, k_{\max}$ **do**

compute residual (i.e., negative gradient) $\mathbf{r}_k = \mathbf{r}_{k-1} - \alpha_{k-1} \mathbf{A} \mathbf{p}_{k-1}$

find conjugate direction $\mathbf{p}_k = \mathbf{r}_k + \beta_k \mathbf{p}_{k-1}$, where $\beta_k = \frac{\mathbf{r}_k^\top (\mathbf{r}_k - \mathbf{r}_{k-1})}{\mathbf{r}_{k-1}^\top \mathbf{r}_{k-1}}$

perform line search $\mathbf{x}_k = \mathbf{x}_{k-1} + \alpha_k \mathbf{p}_k$, where $\alpha_k = \frac{\mathbf{r}_k^\top \mathbf{r}_k}{\|\mathbf{p}_k\|_{\mathbf{A}}^2}$

cumulate random variables $\boldsymbol{\eta}_k = \boldsymbol{\eta}_{k-1} + \xi_k \mathbf{p}_k$, where $\xi_k \overset{\text{iid}}{\sim} \mathcal{N}(0, \|\mathbf{p}_k\|_{\mathbf{A}}^{-2})$

if $\mathbf{c} \approx \mathbf{A}\mathbf{x}_k$ **then**

break

output $\tilde{\boldsymbol{\eta}} = \sqrt{n/k}\, \boldsymbol{\eta}_k$, (also generates $\mathbf{x}_k \approx \mathbf{A}^{-1}\mathbf{c}$ for the chosen $\mathbf{c}$)

Properties

Theorem 1 ([2]). *The conjugate directions, denoted in matrix form by* $\mathbf{P} = (\mathbf{p}_1, \ldots, \mathbf{p}_k)$ *are A-orthogonal, such that* $\mathbf{P}^\top \mathbf{A} \mathbf{P} = \mathbf{D}$, *where* $\mathbf{D} = \operatorname{diag}(\|\mathbf{p}_1\|_{\mathbf{A}}^2, \ldots, \|\mathbf{p}_k\|_{\mathbf{A}}^2)$ *is positive-definite.*

Exact sampling if $\mathbf{A}$ has distinct eigenvalues and Alg 1 runs to $k = n$, because

$\tilde{\boldsymbol{\eta}} = \mathbf{P}\boldsymbol{\xi}$, where $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \mathbf{D}^{-1})$

The covariance is $\mathbf{P}\mathbf{D}^{-1}\mathbf{P}^\top = \mathbf{A}^{-1}$

Approximate sampling if stopped early

Scale the result to generate the same amount of variance

Computational Benefits

Applies when $\mathbf{A}$ is sparse or structured, assuming

$t_{\mathbf{A}} (\ll n^2)$ is the time complexity of matrix-vector multiplications;
$m_{\mathbf{A}} (\ll n^2)$ is the space complexity to store $\mathbf{A}$;
$\kappa_{\mathbf{A}} (\asymp \sqrt{n})$ is the condition number of $\mathbf{A}$.

Table 2: Comparison of Complexity of Posterior Sampling

Method	Time	Space
Thompson sampling (naive)	$O(n^3)$, fixed	$O(n^2)$, dense
Thompson sampling (online)	$O(n^2)$, fixed	$O(n^2)$, dense
Conjugate sampling	$O(\sqrt{\kappa_{\mathbf{A}}} t_{\mathbf{A}})$	$O(n + m_{\mathbf{A}})$
E.g., Kronecker-GP in D dimensions [3]	$O(Dn^{\frac{3}{2}+\frac{1}{D}} + n^{\frac{3}{2}})$	$O(n + Dn^{\frac{2}{D}})$

Simulations

BLR, maximizes $\mathbf{x}^\top \boldsymbol{\theta}$ s.t. $\|\mathbf{x}\|_2 \leq 1$.
Prior given by $\boldsymbol{\Theta} = \mathbf{I}$, $\sigma_n = 1$, $n = 100$.

✓ Conjugate sampling with $k_{\max} = 1$ comparable with Thompson sampling.
✓ Cumulative regret is $O(\sqrt{T \log T})$.
✓ For $k_{\max} = 1$, conjugate sampling scales any random exploration vector to balance exploration and exploitation.

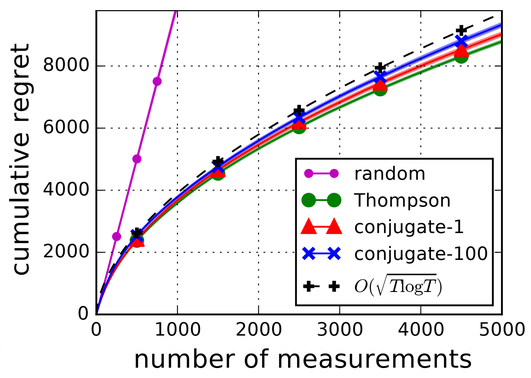

Figure 1: BLR. Cumulative regret

GP, maximize $f(\mathbf{x})$ s.t. $\mathbf{x} \in \{\mathbf{x}_1^*, \ldots, \mathbf{x}_n^*\}$. Assume square exponential kernel $\kappa(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\{-\frac{1}{2\ell^2}\|\mathbf{x} - \mathbf{x}'\|_2^2\}$ with $\ell = 0.3$, $\sigma_f = 1$, $\sigma_n = 1$. Feasible queries are 5^3 Cartesian grid points in $[0, 1]^3$. Use preconditioner derived from prior (Kronecker product of 1d kernel matrices).

✓ (Preconditioned) conjugate sampling with $k_{\max} = 1$ comparable with Thompson sampling; both converged as cumulative regret is $O(\sqrt{T \log T})$.
✓ Improved time and space complexity shown in Table 2.
× Increasing $k_{\max}$ limit does not decrease regret; suffer numerical instability?

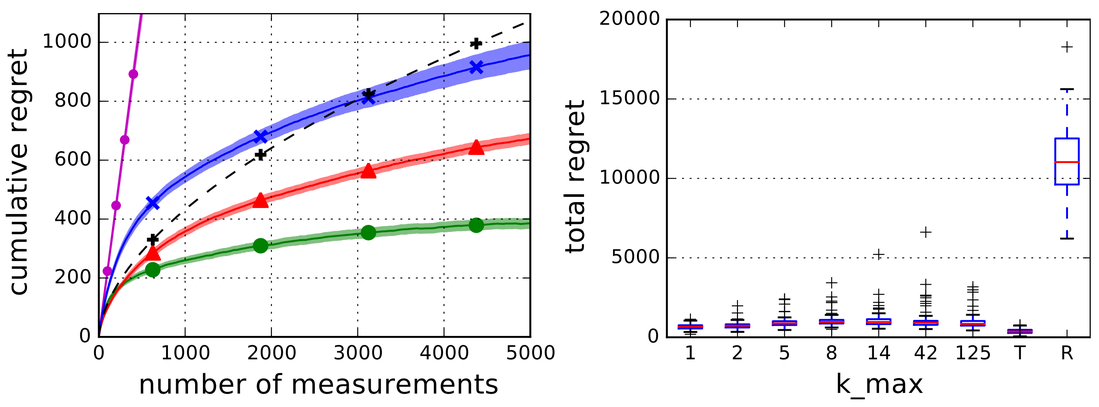

Figure 2: GP. Left: cumulative regret. Right: total regret at $T = 5000$. T for Thompson sampling and R for Random sampling

Future Work

- Theoretical properties of conjugate sampling, especially when $k_{\max}$ is small?
- Better preconditioner to avoid numerical instabilities?
- Other applications where matrix-vector multiplications are fast?

References

[1] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, pages 2249–2257, 2011.

[2] Yousef Saad. *Iterative methods for sparse linear systems*. Siam, 2003.

[3] Seth Flaxman, Andrew Gordon Wilson, Daniel B Neill, Hannes Nickisch, and Alexander J Smola. Fast kronecker inference in gaussian processes with non-gaussian likelihoods. In *International Conference on Machine Learning*, volume 2015, 2015.