Scalable Thompson Sampling using Sparse Gaussian Process Models

Sattar Vakili[*], **Henry Moss**[+], *Artem Artemev*[+], *Vincent Dutordoir*[+], *Victor Picheny*[+]

[*] *MediaTek Research, UK*

[+] *Secondmind Labs, UK*

Neurips 2021

Overview

- ◇ Thompson Sampling (TS) is a classical statistical learning method (1933, Thompson)
- ◇ Sample from the current belief
- ◇ Efficient sample complexity
- ◇ Sampling from approximate distribution for computational reasons
- ◇ That may invalidate performance guarantees [Phan et al., 2019]
- ◇ Our contribution: complete analytical and empirical study of a scalable TS using sparse approximation of GP models

Problem Formulation

- ◇ Consider an objective function $f : \mathcal{X} \to \mathbb{R}$, $\mathcal{X} \subset \mathbb{R}^d$
- ◇ A sequential learning policy selects a batch $\{x_{t,b}\}_{b\in[B]}$ of observations at each time $t = 1, 2, \ldots$
- ◇ Recieves noisy evaluation of f: $y_{t,b} = f(x_{t,b}) + \epsilon_{t,b}$
- ◇ Objective: minimize *regret*

$$R(T, B; f) = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{b=1}^{B} f(x^*) - f(x_{t,b})\right]$$

Regularity Assumptions

◇ Assumption 1: The function f is in the reproducing kernel Hilbert space (RKHS) of a positive definite kernle k

$$||f||_{H_k} \leq \mathcal{B}$$

◇ Assumption 2: $\epsilon_{t,b}$ are independent $R-$sub-Gaussian random variables

$$\mathbb{E}[e^{h\epsilon_{t,b}}] \leq \exp(\frac{h^2R^2}{2}), \ \ \forall h \in \mathbb{R}, \forall t, b \in \mathbb{N}.$$

◇ We provide our regret bounds under these two assumptions.

Surrogate Gaussian Process Model

◇ Provided data $\mathcal{H}_t = \{\mathbf{X}_t, \mathbf{y}_t\}$

◇ A surrogate GP model provides us with a posterior mean and covariance

$$\mu_t(x) = k_{\mathbf{X}_t,x}^\top (K_{\mathbf{X}_t,\mathbf{X}_t} + \tau \mathbf{I})^{-1} \mathbf{y}_t$$

$$k_t(x, x') = k(x, x') - k_{\mathbf{X}_t,x}^\top (K_{\mathbf{X}_t,\mathbf{X}_t} + \tau \mathbf{I})^{-1} k_{\mathbf{X}_t,x'}$$

◇ we may use this posterior distribution to sample from

Computational Complexity and Approximations

- ◇ TS using GP models has two computational bottlenecks
- ◇ $\mathcal{O}(tB)^3$ computational complexity of the posterior distribution (matrix inverse)
- ◇ $O(N^3)$ computational complexity of a joint sample on N points (Cholesky decomposition)
- ◇ These two can be resolved, respectively, by sparse variational GP (SVGP) [Titsias, 2009] and decoupled sampling [Wilson et al., 2020]
- ◇ Both methods introduce approximation errors which need careful treatment to guarantee performance

SVGP

◇ Inducing points $\mathbf{Z}_t = \{z_1, ..., z_{m_t}\}$

◇ Inducing variables $\mathbf{u}_t = \hat{f}(\mathbf{Z}_t)$

◇ A prior Gaussian density $q_t(\mathbf{u}_t) = \mathcal{N}(\mathbf{m}_t, \mathbf{S}_t)$

$$\mu_t^{(s)}(x) = k_{\mathbf{Z}_t,x}^\top K_{\mathbf{Z}_t,\mathbf{Z}_t}^{-1} \mathbf{m}_t$$

$$k_t^{(s)}(x, x') = k(x, x') + k_{\mathbf{Z}_t,x}^\top K_{\mathbf{Z}_t,\mathbf{Z}_t}^{-1} (\mathbf{S}_t - K_{\mathbf{Z}_t,\mathbf{Z}_t}) K_{\mathbf{Z}_t,\mathbf{Z}_t}^{-1} k_{\mathbf{Z}_t,x'}$$

◇ Computational complexity:

$$\mathcal{O}((tB)^3) \quad \rightarrow \quad \mathcal{O}(tbm_t^2)$$

SVGP with inducing features

◇ Inducing variables can be also be given with respect to integral transforms of $\hat{f}$: $u_{t,i} = \int_{\mathcal{X}} \hat{f}(x)\psi_i(x)dx$

◇ We choose the inducing features as the Mercer eigenfunctions of k

◇ Approximate posterior

$$\mu_t^{(s)}(x) = \boldsymbol{\phi}_{m_t}^\top(x)\mathbf{m}_t$$

$$k_t^{(s)}(x, x') = k(x, x') + \boldsymbol{\phi}_{m_t}^\top(x)(\mathbf{S}_t - \Lambda_{m_t})\boldsymbol{\phi}_{m_t}(x')$$

- $\boldsymbol{\phi}_m(x) \triangleq [\phi_1(x), ..., \phi_m(x)]^\intercal$

Decoupled Sampling with Inducing Points

◇ Sample from prior plus the effect of data [Wilson et al., 2020]

◇ Sample from prior: using truncated feature representations

$$\hat{f}(x) = \sum_{j=1}^{\infty} \sqrt{\lambda_j} w_j \phi_j(x)$$

$$\hat{f}(x) = \sum_{j=1}^{M} \sqrt{\lambda_j} w_j \phi_j(x)$$

◇ The effect of data: using SVGP

Scalable Thompson Sampling

◇ In addition to the decoupled sampling of Wilson et al. [2020], we scale the posterior variance with α_t, to ensure sufficient exploration

$$\tilde{f}_t(x) = \sum_{j=1}^{M} \alpha_t \sqrt{\lambda_j} w_j \phi_j(x) + \sum_{j=1}^{m_t} v_{t,j} k(x, z_j)$$

◇ $v_{t,j} = [K_{\mathbf{Z}_t,\mathbf{Z}_t}^{-1}(\alpha_t(\mathbf{u}_t - \mathbf{m}_t) + \mathbf{m}_t - \alpha_t \mathbf{\Phi}_{m_t,M} \Lambda_M^{\frac{1}{2}} \mathbf{w}_M)]_j$

◇ $\mathbf{\Phi}_{m_t,M} = [\boldsymbol{\phi}_M(z_1), ..., \boldsymbol{\phi}_M(z_{m_t})]^\intercal$

◇ $\mathbf{w}_M = [w_1, ..., w_M]^\intercal,\ w_i \sim \mathcal{N}(0, 1)$

◇ Computational complexity:

$$\mathcal{O}(N^3) \quad \rightarrow \quad \mathcal{O}((m_t + M)BN)$$

Scalable Thompson Sampling (inducing features)

◇ In addition to the decoupled sampling of Wilson et al. [2020], we scale the posterior variance with α_t, to ensure sufficient exploration

$$\tilde{f}_t(x) = \sum_{j=1}^{M} \alpha_t \sqrt{\lambda_j} w_j \phi_j(x) + \sum_{j=1}^{m_t} v_{t,j} \lambda_j \phi_j(x)$$

◇ $v_{t,j} = [\Lambda_{m_t}^{-1}(\alpha_t(\mathbf{u}_t - \mathbf{m}_t) + \mathbf{m}_t - \alpha_t \Lambda_{m_t}^{\frac{1}{2}} \mathbf{w}_{m_t})]_j$

◇ Λ_{m_t} is the diogonal matrix of eigenvalues

Regret Performance of Vanilla GP-TS

◇ For vanilla GP-TS Chowdhury and Gopalan [2017]:

$$R(T; F) = \tilde{\mathcal{O}}(\gamma_T \sqrt{T})$$

◇ $\gamma_s = \max_{A \subset \mathcal{X}, |A|=s} \mathcal{I}([y(x)]_{x \in A}; [\hat{f}(x)]_{x \in A})$

◇ Mutual information: $\mathcal{I}([y(x)]_{x \in A}; [\hat{f}(x)]_{x \in A})$

◇ Mutual information is closely related to the effective dimension of the kernel

◇ Matérn: $\gamma_T = \mathcal{O}\left(T^{\frac{d}{2\nu+d}} (\log(T))^{\frac{2\nu}{2\nu+d}}\right)$,
Squared Exponential: $\gamma_T = \mathcal{O}\left((\log(T))^{d+1}\right)$ [Srinivas et al., 2010, Vakili et al., 2021]

Setting Up Our Theorem

◇ Assumption 3: quality of the approximate standard deviation

$$\frac{1}{\underline{a}}\sigma_t(x) - \epsilon \leq \tilde{\sigma}_t(x) \leq \bar{a}\sigma_t(x) + \epsilon$$

◇ Assumption 4: quality of the approximate prediction

$$|\tilde{\mu}_t(x) - \mu_t(x)| \leq c\sigma_t(x)$$

◇ We show that this conditions are satisfied with proper parameters m_t and M

◇ The additive error in $\tilde{\sigma}_t(x)$ in particular makes the analysis challenging

Regret Bound for S-GP-TS

Theorem: S-GP-TS with $\alpha_t = 2\tilde{u}_t(1/(t^2))$, Under Assumptions 1,2,3 and 4, satisfies

$$R(T; f) = \mathcal{O}(\underline{a}\bar{a}BR\sqrt{d\gamma_T(\gamma_{TB} + \log(T))T\log(T)} + \underline{a}\epsilon TBR\sqrt{d(\gamma_{TB} + \log(T))\log(T)})$$

◇ $\tilde{u}_t(\delta)$, is a confidence interval width multiplier

$$\tilde{u}_t(\delta) = \underline{a}_t\left(\mathcal{B} + R\sqrt{2(\gamma_{tB} + 1 + \log(1/\delta))} + c_t\right)$$

◇ That is with probability at least $1 - \delta$,

$$|f(x) - \tilde{\mu}_t(x)| \leq \tilde{u}_t(\tilde{\sigma}_t(x) + \epsilon_t)$$

Regret Bound for S-GP-TS

Theorem Under assumptions 1 and 2, with parameters given in the table, S-GP-TS offers

$$R(T, B; f) = O(B\sqrt{\gamma_T \gamma_{TB} T \log(T)})$$

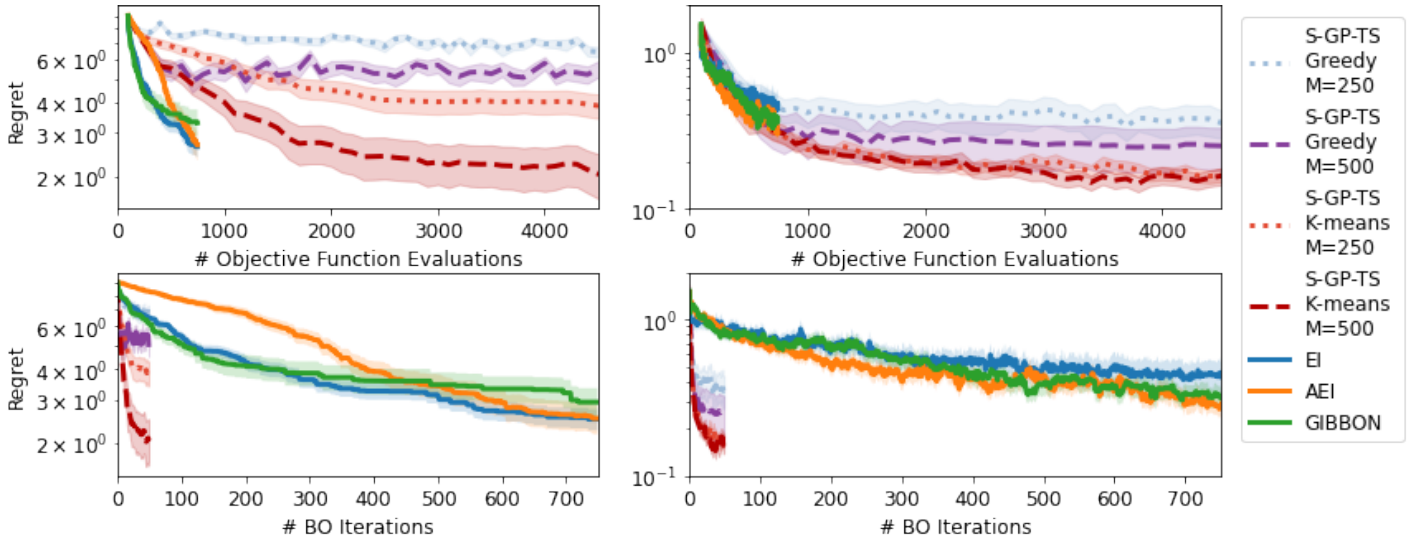
		Inducing points	Inducing features
Matérn	Condition	$m_t \sim T^{\frac{2d}{2\nu - d}}$, $M \sim T^{\frac{(2\nu+d)d}{2(2\nu-d)\nu}}$	$m_t \sim T^{\frac{d}{2\nu}}$, $M \sim T^{\frac{(2\nu+d)d}{4\nu^2}}$
	Cost	$O\left(BN_T T^{\frac{4\nu^2+d^2}{2(2\nu-d)\nu}} + BT^2 \min\{T^{\frac{4d}{2\nu-d}}, T^2\}\right)$	$O\left(BN_T T^{\frac{(2\nu+d)^2-2\nu d}{4\nu^2}} + BT^{\frac{2\nu+d}{\nu}}\right)$
SE	Condition	$m_t, M \sim (\log(T))^d$	$m_t, M \sim (\log(T))^d$
	Cost	$O\left(BN_T T \log^d(T) + BT^2 \log^{2d}(T)\right)$	$O\left(BN_T T \log^d(T) + BT^2 \log^{2d}(T)\right)$

$\diamond$ with $B = 1$, the same regret bound as exact GP-TS is recovered

Experiments

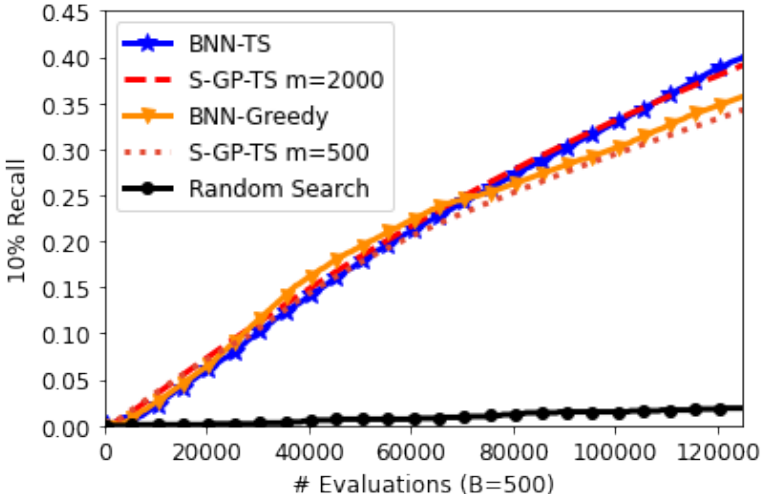
- ◇ Experiments on benchmark functions: Shekel and Hartmann
- ◇ Experiments on a high throughput molecular screening problem
- ◇ Our implementation is based on *gpflow* and *gpflux* for modeling and *trieste* for BO

Experiments on Benchmark Functions



◇ Shekel (4D, left) and Hartmann (6D, right)

Experiments on Molecular Screening



◇ S-GP-TS performs comparable to the established baseline of Bayesian NNs

References

S. R. Chowdhury and A. Gopalan. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, pages 844–853, 2017.

M. Phan, Y. Abbasi Yadkori, and J. Domke. Thompson sampling and approximate inference. In *Advances in Neural Information Processing Systems 32*, pages 8804–8813. Curran Associates, Inc., 2019.

N. Srinivas, A. Krause, S. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: no regret and experimental design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 1015–1022. Omnipress, 2010.

M. K. Titsias. Variational Learning of Inducing Variables inSparse Gaussian Processes. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 567–574, 2009.

S. Vakili, K. Khezeli, and V. Picheny. On information gain and regret

bounds in gaussian process bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 82–90. PMLR, 2021.

J. T. Wilson, V. Borovitskiy, A. Terenin, P. Mostowsky, and M. P. Deisenroth. Efficiently sampling functions from gaussian process posteriors. *arXiv preprint arXiv:2002.09309*, 2020.