Contextual Gaussian Process Bandit Optimization

Andreas Krause and Cheng Soon Ong
Department of Computer Science, ETH Zurich, Switzerland

ETH Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Contributions

- An efficient algorithm, CGP-UCB, for the contextual GP bandit problem
- Flexibly combining kernels over contexts and actions
- Generic approach for deriving regret bounds for composite kernel functions
- Evaluate CGP-UCB on automated vaccine design and sensor management

Contextual Bandits [cf., Auer '02; Langford & Zhang '08]

Play a game for T rounds:

- Receive *context* $\mathbf{z}_t \in Z$
- Choose an *action* $\mathbf{s}_t \in S$
- Receive a payoff $y_t = f(\mathbf{s}_t, \mathbf{z}_t) + \epsilon_t$ (f unknown).

Cumulative regret for context specific action

- Incur *contextual regret* $r_t = \sup_{\mathbf{s}' \in S} f(\mathbf{s}', \mathbf{z}_t) - f(\mathbf{s}_t, \mathbf{z}_t)$
- After T rounds, the *cumulative contextual regret* is $R_T = \sum_{t=1}^T r_t$.
- *Context-specific best action is a demanding benchmark.*

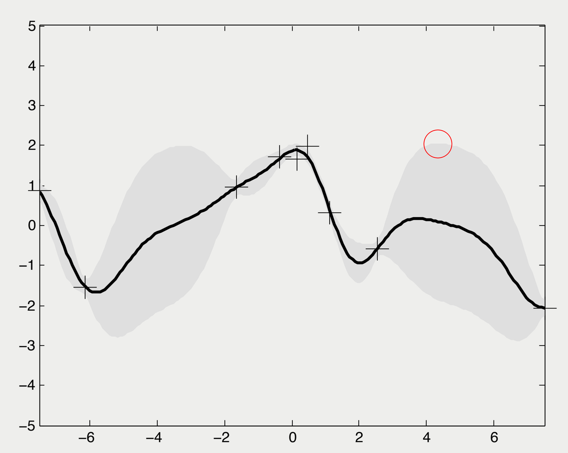
Gaussian Processes (GP)

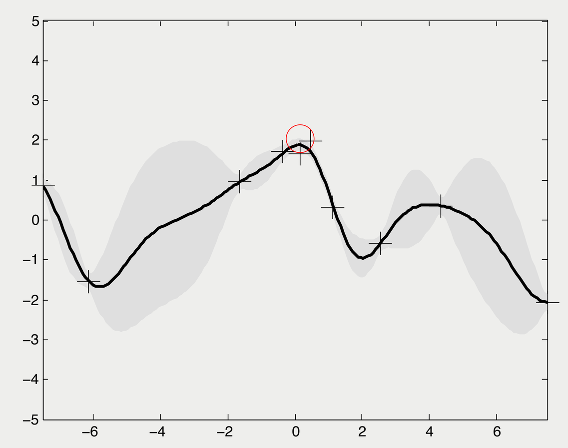
- Model payoff function using GPs: $f \sim GP(\mu, k)$
- observations $\mathbf{y}_T = [y_1 \dots y_T]^T$ at inputs $A_T = \{\mathbf{x}_1, \dots, \mathbf{x}_T\}$
- $y_t = f(\mathbf{x}_t) + \epsilon_t$ with i.i.d. Gaussian noise $\epsilon_t \sim N(0, \sigma^2)$
- Posterior distribution over f is a GP with

$$\begin{aligned} \text{mean} \quad & \mu_T(\mathbf{x}) = \mathbf{k}_T(\mathbf{x})^T(\mathbf{K}_T + \sigma^2\mathbf{I})^{-1}\mathbf{y}_T, \\ \text{covariance} \quad & k_T(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - \mathbf{k}_T(\mathbf{x})^T(\mathbf{K}_T + \sigma^2\mathbf{I})^{-1}\mathbf{k}_T(\mathbf{x}'), \\ \text{variance} \quad & \sigma_T^2(\mathbf{x}) = k_T(\mathbf{x}, \mathbf{x}), \end{aligned}$$

where $\mathbf{k}_T(\mathbf{x}) = [k(\mathbf{x}_1, \mathbf{x}) \dots k(\mathbf{x}_T, \mathbf{x})]^T$ and $\mathbf{K}_T$ is the kernel matrix.

GP-UCB [Srinivas, Krause, Kakade, Seeger ICML 2010]





Context free upper confidence bound algorithm (GP-UCB)

At round t, GP-UCB picks action $\mathbf{s}_t = \mathbf{x}_t$ such that

$$\mathbf{s}_t = \operatorname*{argmax}_{\mathbf{s} \in S} \mu_{t-1}(\mathbf{s}) + \beta_t^{1/2}\sigma_{t-1}(\mathbf{s}),$$

with appropriate β_t. Trades *exploration* (high σ) and *exploitation* (high μ).

Maximum information gain bounds regret

The (context-free) regret R_T of GP-UCB is bounded by $\mathcal{O}^*(\sqrt{T\beta_T\gamma_T})$, where γ_T is defined as the maximum information gain:

$$\gamma_T := \max_{A \subseteq S : |A| = T} \mathrm{I}(y_A; f), \qquad \text{where} \qquad \mathrm{I}(y_A; f) = \mathrm{H}(y_A) - \mathrm{H}(y_A | f)$$

quantifies the reduction in uncertainty about f achieved by revealing y_A.

Bounds for Kernels

Bounds on γ_T exist for linear, squared exponential and Matérn kernels.

Contextual Upper Confidence Bound Algorithm (CGP-UCB)

$$\mathbf{s}_t = \operatorname*{argmax}_{\mathbf{s} \in S} \mu_{t-1}(\mathbf{s}, \mathbf{z}_t) + \beta_t^{1/2}\sigma_{t-1}(\mathbf{s}, \mathbf{z}_t)$$

where $\mu_{t-1}(\cdot)$ and $\sigma_{t-1}(\cdot)$ are the posterior mean and standard deviation of the GP over the joint set $X = S \times Z$ conditioned on the observations $(\mathbf{s}_1, \mathbf{z}_1, y_1), \dots, (\mathbf{s}_{t-1}, \mathbf{z}_{t-1}, y_{t-1})$.

Bounds on Contextual Regret

Let $\delta \in (0, 1)$. Suppose one of the following assumptions holds

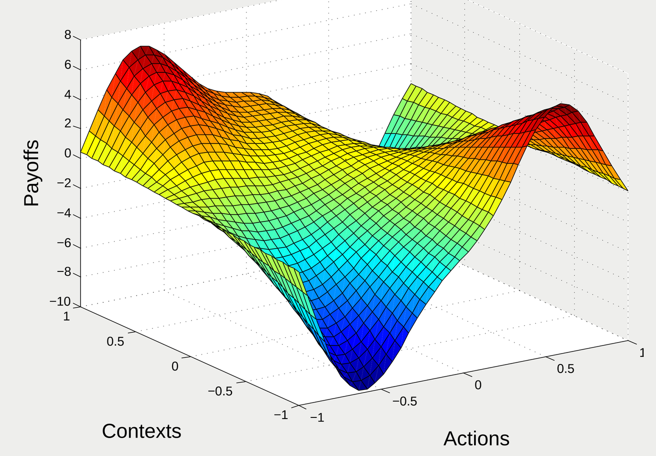
- X **is finite,** f is sampled from a known GP prior with known noise variance σ^2,
- X **is compact and convex,** $\subseteq [0, r]^d$, $d \in \mathbb{N}$, $r > 0$. Suppose f is sampled from a known GP prior with known noise variance σ^2, and that $k(\mathbf{x}, \mathbf{x}')$ has smooth derivatives,
- X **is arbitrary;** $\|f\|_k \le B$**.** The noise variables ϵ_t form an *arbitrary* martingale difference sequence (meaning that $\mathbb{E}[\varepsilon_t \,|\, \varepsilon_1, \dots, \varepsilon_{t-1}] = 0$ for all $t \in \mathbb{N}$), uniformly bounded by σ.

Then for appropriate choices of β_t, the *contextual regret* of CGP-UCB is bounded by $\mathcal{O}^*(\sqrt{T\gamma_T\beta_T})$ w.h.p. Precisely,

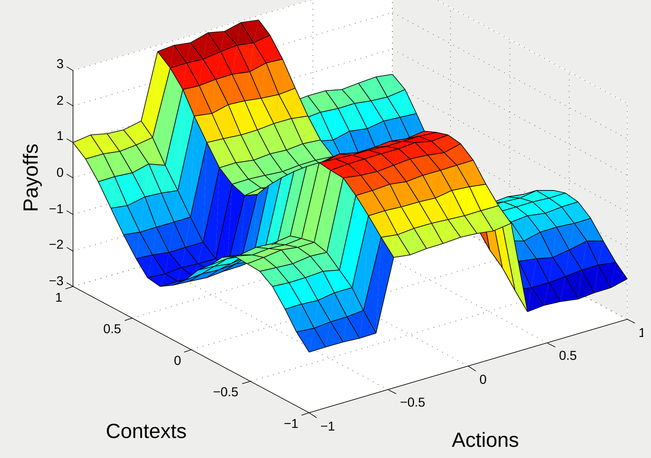
$$\Pr\left\{R_T \le \sqrt{C_1 T \beta_T \gamma_T} + 2 \quad \forall T \ge 1\right\} \ge 1 - \delta.$$

where $C_1 = 8/\log(1 + \sigma^{-2})$.

Composite Kernels



Product of squared exponential kernel and linear kernel



Additive combination of payoff that smoothly depends on context, and exhibits clusters of actions.

Product kernel

- $k = k_S \otimes k_Z$, where $(k_S \otimes k_Z)((\mathbf{s}, \mathbf{z}), (\mathbf{s}', \mathbf{z}')) = k_Z(\mathbf{z}, \mathbf{z}')k_S(\mathbf{s}, \mathbf{s}')$
- Two context-action pairs are similar (large correlation) if the contexts are similar and actions are similar

Additive kernel

- $(k_S \oplus k_Z)((\mathbf{s}, \mathbf{z}), (\mathbf{s}', \mathbf{z}')) = k_Z(\mathbf{z}, \mathbf{z}') + k_S(\mathbf{s}, \mathbf{s}')$
- Generative model: first sample a function $f_S(\mathbf{s}, \mathbf{z})$ that is constant along $\mathbf{z}$, and varies along $\mathbf{s}$ with regularity as expressed by k_s; then sample a function $f_z(\mathbf{s}, \mathbf{z})$, which varies along $\mathbf{z}$ and is constant along $\mathbf{s}$;

$$f = f_s + f_z.$$

Bounds for Composite Kernels

Maximum information gain for a GP with kernel k on set V

$$\gamma(T; k; V) = \max_{A \subseteq V, |A| \le T} \frac{1}{2} \log \left| \mathbf{I} + \sigma^{-2}[k(\mathbf{v}, \mathbf{v}')]_{\mathbf{v}, \mathbf{v}' \in A} \right|,$$

Product kernel

Let k_Z be a kernel function on Z with rank at most d. Then

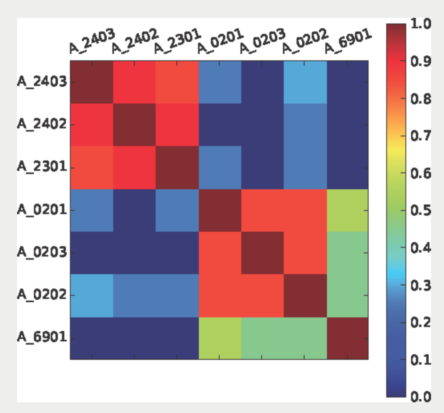
$$\gamma(T; k_S \otimes k_Z; X) \le d\gamma(T; k_S; S) + d \log T.$$

Additive kernel

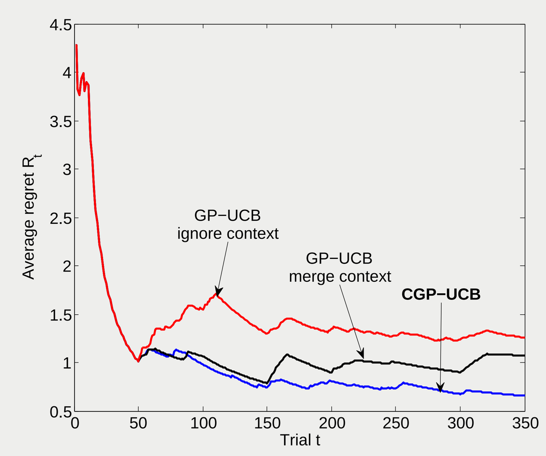
Let k_S and k_Z be kernel functions on S and Z respectively. Then

$$\gamma(T; k_S \oplus k_Z; X) \le \gamma(T; k_S; S) + \gamma(T; k_Z; Z) + 2 \log T.$$

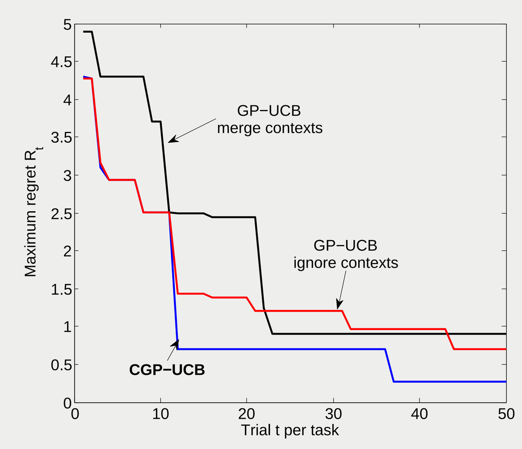
Multi-task Learning (Vaccine Design)



Context similarity using inter task predictions.



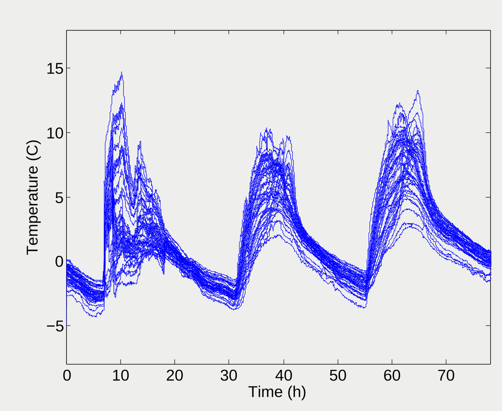
average regret of CGP-UCB



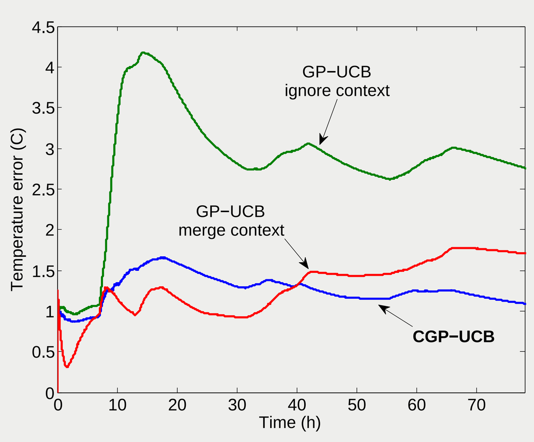
maximum regret of CGP-UCB

Task Discover peptide sequences binding to MHC molecules
Context Features encoding the MHC alleles
Action Choose a stimulus (the vaccine) $\mathbf{s} \in S$ that maximizes an observed response (binding affinity).
Kernels Use a finite inter-task covariance kernel $\mathbf{K}_Z$ with rank m_Z to model the similarity of different experiments, and a Gaussian kernel $k_S(\mathbf{s}, \mathbf{s}')$ to model the experimental parameters.

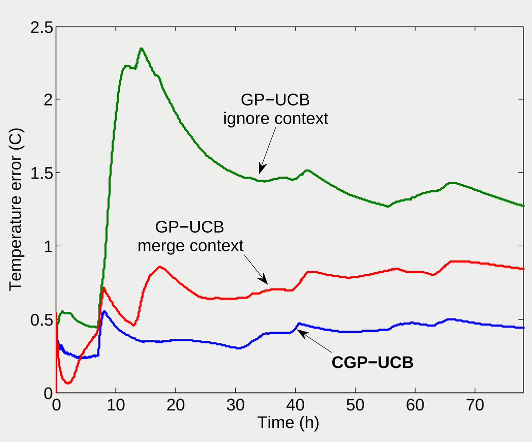
Learning to Monitor Sensor Networks



Temperature data from a network of 46 sensors at Intel Research.



CGP-UCB using average temperature



CGP-UCB using minimum temperature

Task Given a sensor network, monitor maximum temperatures in building
Context Time of day
Action Pick 5 sensors to activate
Kernels Joint spatio-temporal covariance function using the Matérn kernel