Gaussian Process Optimization with Mutual Information

Emile Contal[1] Vianney Perchet[2] Nicolas Vayatis[1]

[1]CMLA – Ecole Normale Suprieure de Cachan & CNRS, France
[2]LPMA – Université Paris Diderot & CNRS, France

June 22, 2014

Problem Statement

Sequential Optimization

Let $f : \mathcal{X} \to \mathbb{R}$ where $\mathcal{X} \subset \mathbb{R}^d$ is compact and convex.
We consider the problem of finding the maximum of f denoted by:

$$f(x^\star) = \max_{x \in \mathcal{X}} f(x)\,,$$

via sequential queries $f(x_1), f(x_2), \ldots$

Noisy Observations

At iteration T we choose x_{T+1} using the previous noisy observations $Y_T = \{y_1, \ldots, y_T\}$, where $\forall t \leq T$:

$$y_t = f(x_t) + \epsilon_t \text{ and } \epsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \eta^2)\,.$$

Gaussian Processes

Definition

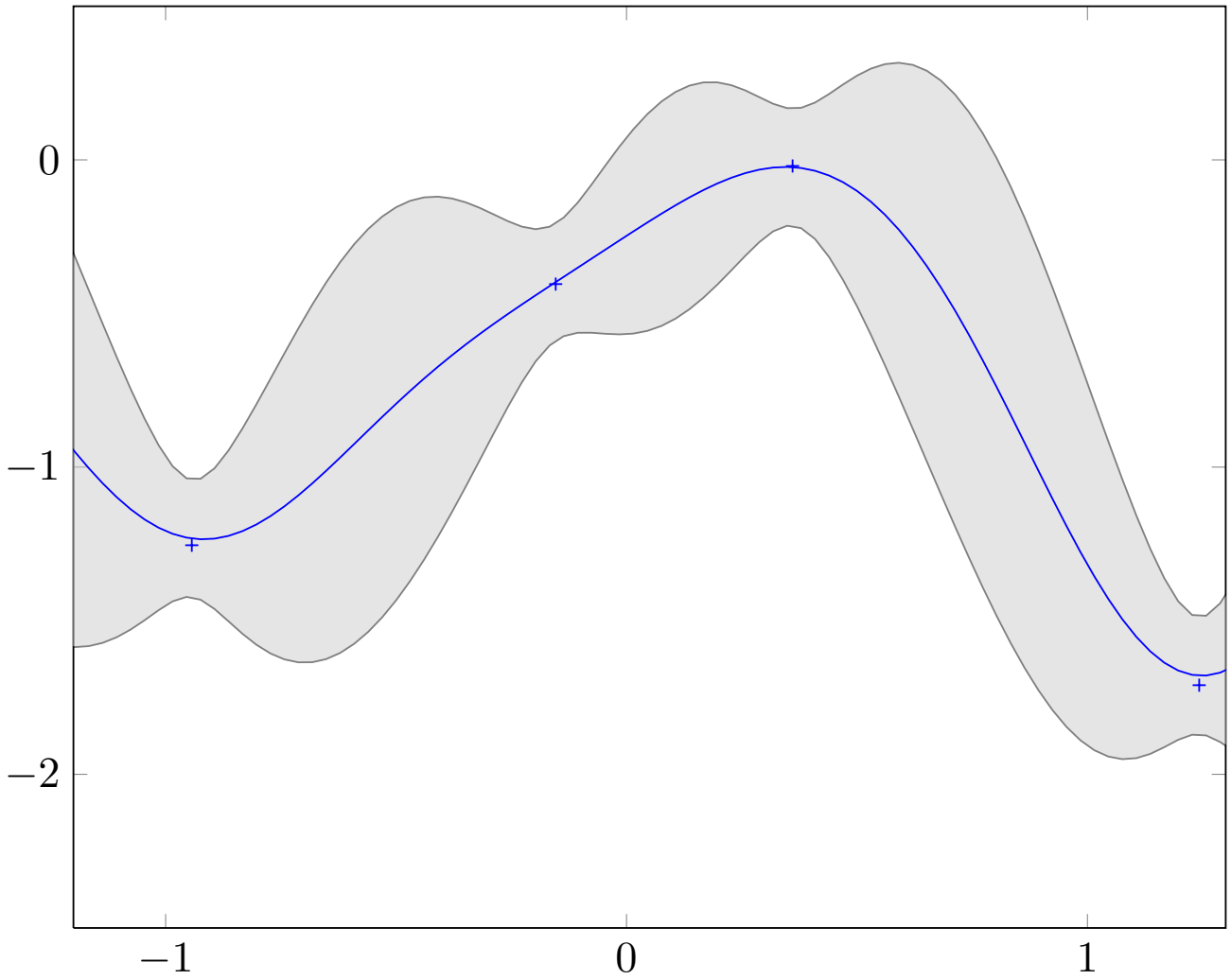
$f \sim \mathcal{GP}(m, k)$ with mean function $m : \mathcal{X} \to \mathbb{R}$ and kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$, when for all $x_1, \dots, x_n \in \mathcal{X}$ we have:

$$\big(f(x_1), \dots, f(x_n)\big) \sim \mathcal{N}\big([m(x_i)]_i, [k(x_i, x_j)]_{i,j}\big) .$$

Bayesian Inference

Given Y_T, the posterior distribution $\Pr[f \mid Y_T]$ is a GP with mean μ_{T+1} (prediction) and variance σ^2_{T+1} (uncertainty) computed by Bayesian inference.

Gaussian Processes



Objective

Cumulative Regret

The efficiency of a policy is measured via the cumulative regret:

$$R_T = \sum_{t<T} f(x^\star) - f(x_t)\,.$$

Goal

The cumulative regret is unknown in practice. Our aim is to obtain upper bounds on R_T with high probability.

Related Work

Dani al. 2008

In the general linear optimization problem, we have a lower bound on the cumulative regret of $\Omega(d\sqrt{T})$.

Srinivas et al. 2012

For the GP-UCB algorithm with linear GP, we have an upper bound on the cumulative regret of $\mathcal{O}(d\sqrt{T})$ with high probability.

Mutual Information – An Important Ingredient

Information Gain

The information gain on f at X_T is the mutual information between f and Y_T. For a GP distribution with $\mathbf{K}_T$ the kernel matrix of X_T:

$$I_T(X_T) = \frac{1}{2} \log \det(\mathbf{I} + \eta^{-2}\mathbf{K}_T)\,.$$

We define $\gamma_T = \max_{|X|=T} I_T(X)$ the maximum information gain by a sequence of T queries points.

Empirical Lower Bound

For GPs with bounded variance, we have: [Srinivas et al. 2012]

$$\widehat{\gamma}_T = \sum_{t=1}^{T} \sigma_t^2(x_t) \leq C\gamma_T \text{ where } C = \frac{2}{\log(1+\eta^{-2})}$$

GP-MI – A Novel Algorithm for Sequential Optimization

$\widehat{\gamma}_0 \leftarrow 0$

for $t = 1, 2, \ldots$ **do**

Compute μ_t and σ_t^2 using Bayesian inference

$$\phi_t(x) \leftarrow \sqrt{\alpha}\left(\sqrt{\sigma_t^2(x) + \widehat{\gamma}_{t-1}} - \sqrt{\widehat{\gamma}_{t-1}}\right)$$

$x_t \leftarrow \operatorname{argmax}_{x \in \mathcal{X}} \mu_t(x) + \phi_t(x)$

$\widehat{\gamma}_t \leftarrow \widehat{\gamma}_{t-1} + \sigma_t^2(x_t)$

Query at x_t and observe y_t

end

Main Result – Regret bounds for GP-MI

For all $\delta > 0$ and $T > 1$, set $\alpha = \log \frac{2}{\delta}$. The cumulative regret R_T incurred by the GP-MI algorithm on f distributed as a GP perturbed by independent Gaussian noise with variance η^2 satisfies the following bounds:

$$\Pr\left[R_T \leq 5\sqrt{\alpha C \gamma_T} + 4\sqrt{\alpha}\right] \geq 1 - \delta\,,$$

where $C = \frac{2}{\log(1+\eta^{-2})}$.

Application to Specific Kernels

- For linear kernel: $R_T = \mathcal{O}(\sqrt{d \log T})$
- For RBF kernel: $R_T = \mathcal{O}\big(\sqrt{(\log T)^{d+1}}\big)$
- For Matérn kernel: $R_T = \mathcal{O}\big(\sqrt{T^a \log T}\big)$,
  where $a = \frac{d(d+1)}{2\nu + d(d+1)} < 1$ and ν is the Matérn parameter.

High Probabilistic Bounds for R_T

Gaussian Martingale

The sequence $M_T = R_T - \sum_{t=1}^{T} \left(\mu_t(x^\star) - \mu_t(x_t)\right)$ is a Gaussian martingale with respect to Y_{T-1}.

Concentration Inequalities [Becu et al. 2008]

For all $\delta > 0$ and $T > 1$, with $y = 8(C\gamma_T + 1)$ we have:

$$\Pr\left[M_T \leq \sqrt{2\alpha y} + \sqrt{\frac{2\alpha}{y}\sum_{t=1}^{T}\sigma_t^2(x^\star)}\right] \geq 1-\delta$$

$$\Pr\left[R_T \leq \sqrt{2\alpha y} + \sqrt{\frac{2\alpha}{y}\sum_{t=1}^{T}\sigma_t^2(x^\star)} + \sum_{t=1}^{T}\left(\mu_t(x^\star) - \mu_t(x_t)\right)\right] \geq 1-\delta\,.$$

Regret Bounds for the GP-MI Algorithm

Inequality for the GP-MI Algorithm

Using the function ϕ_t as defined in GP-MI, we have:

$$\sum_{t=1}^{T} \left(\mu_t(x^\star) - \mu_t(x_t)\right) \leq \sum_{t=1}^{T} \left(\phi_t(x_t) - \phi_t(x^\star)\right)$$

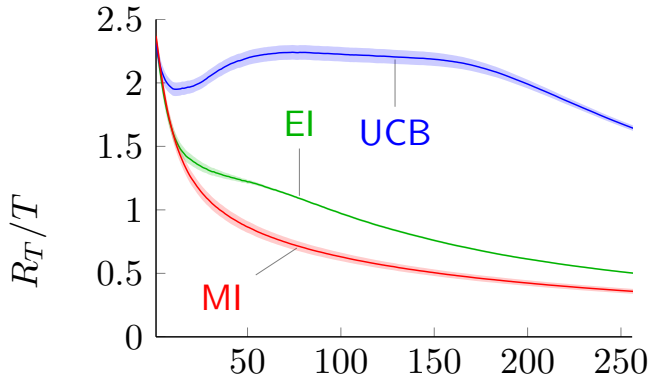
$$\leq \sqrt{\alpha C \gamma_T} - \sqrt{\frac{2\alpha}{y} \sum_{t=1}^{T} \sigma_t^2(x^\star)}\,.$$

Regret bounds

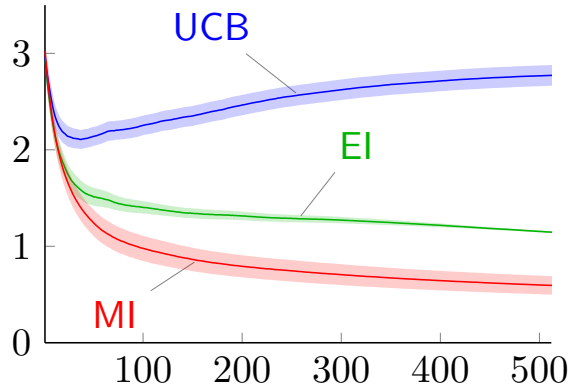
Plugging this inequality in the previous concentration bound for R_T:

$$\Pr\left[R_T \leq 5\sqrt{\alpha C \gamma_T} + 4\sqrt{\alpha}\right] \geq 1 - \delta\,.$$

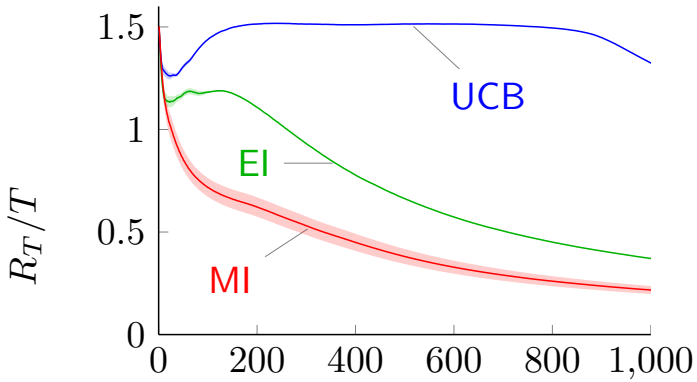
Empirical Average Regret



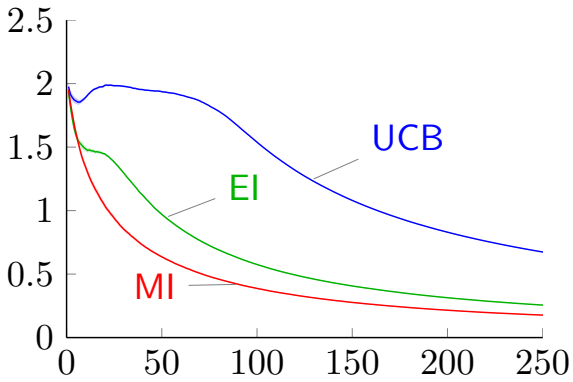
(a) Generated GP ($d = 2$)



(b) Generated GP ($d = 4$)

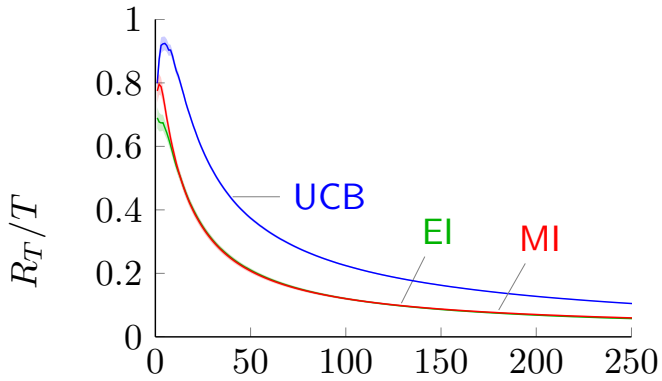


(c) Gaussian mixture

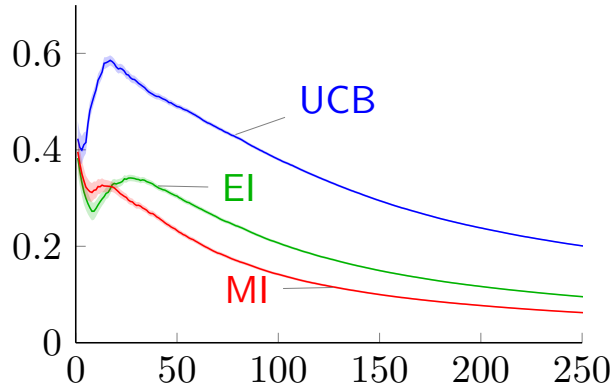


(d) Himmelblau

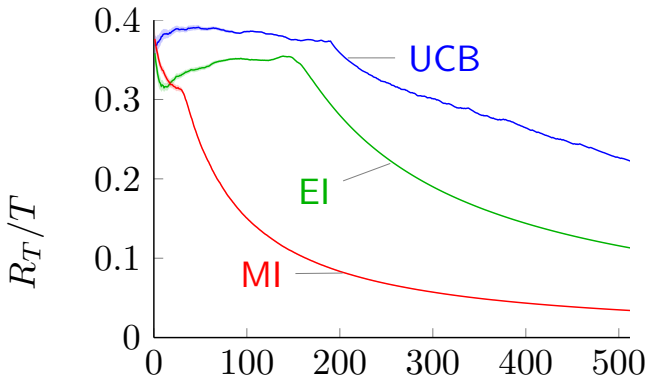
Empirical Average Regret



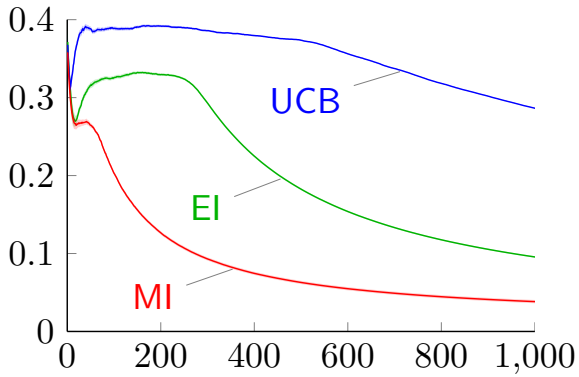
(e) Branin



(f) Goldstein



(g) Tsunamis



(h) Mackey-Glass

Implementation

Exact Inference for Gaussian likelihood

Numerical complexity in $\mathcal{O}(T^2)$ using the Cholesky sequential updates of the covariance matrix.

Algorithms for non-Gaussian likelihood

For other likelihood functions (e.g. Laplacian or Student's t), one can use the EP algorithm or Monte Carlo sampling.

Open Questions and Discussion

- Theoretical guarantees for simple regret
- Empirical performance for simple regret
- Kernel learning procedure
- Calibration of δ