

贝叶斯优化

黄政宇

北京大学北京国际数学研究中心

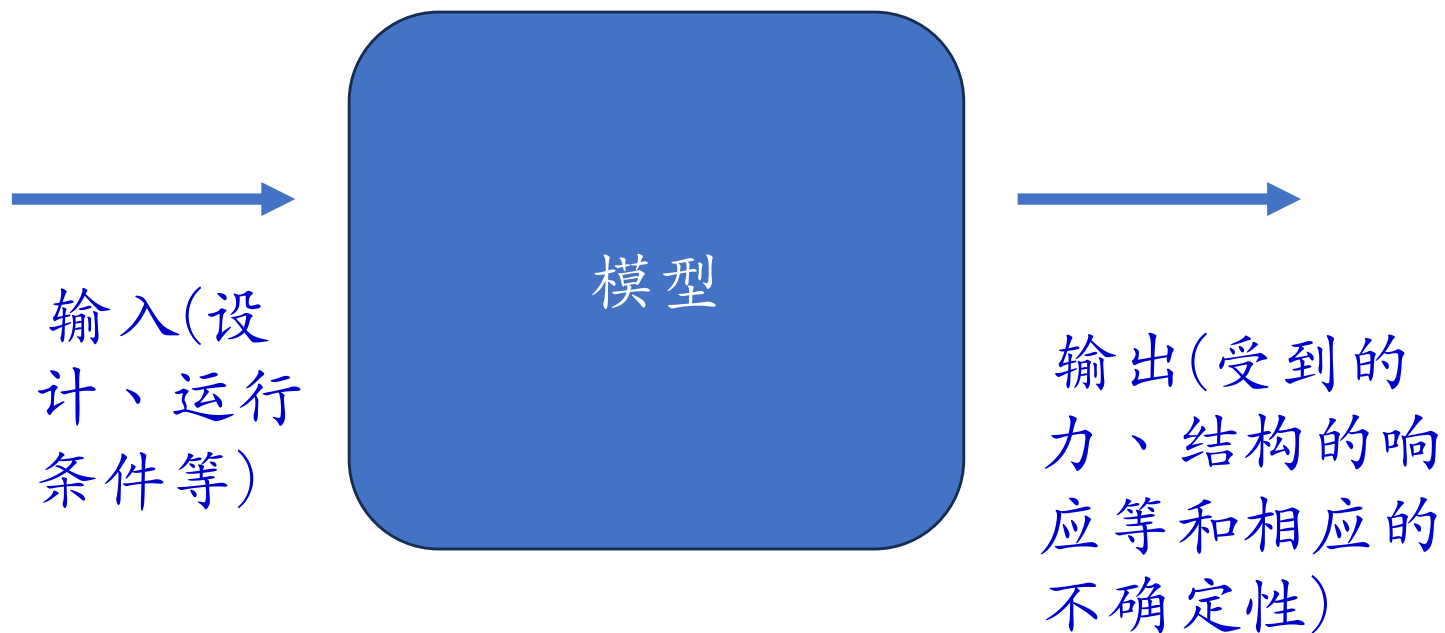
北京大学国际机器学习研究中心



贝叶斯优化

➤ 目标函数

$$\max f(x)$$

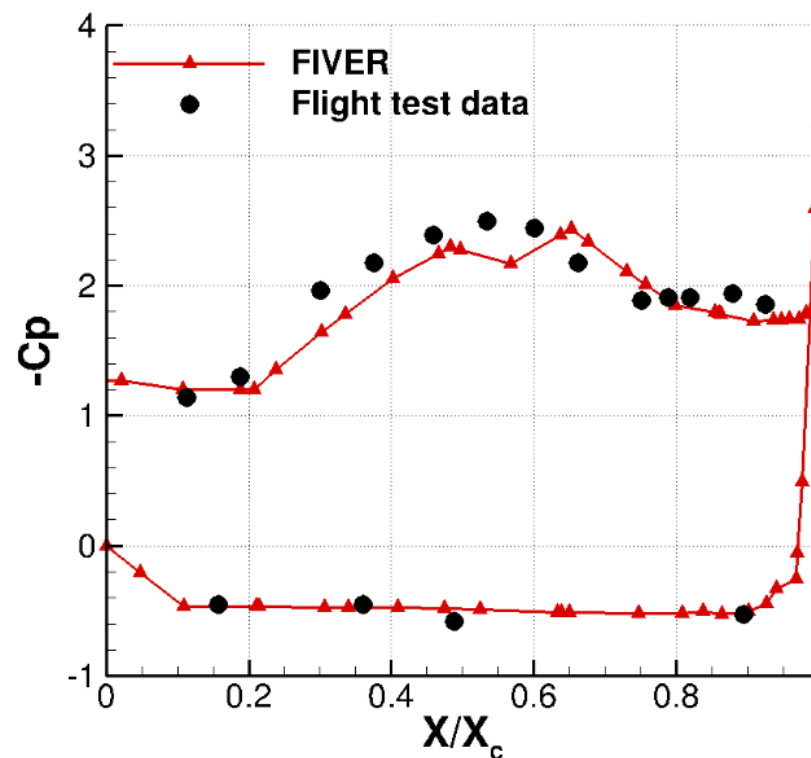
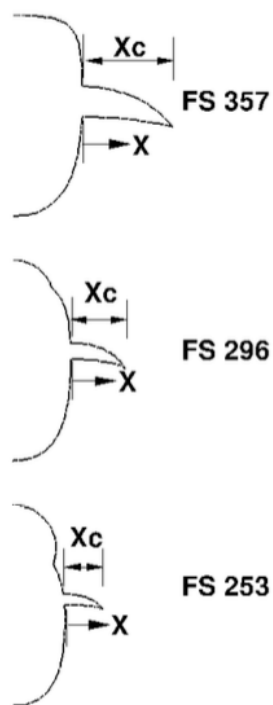
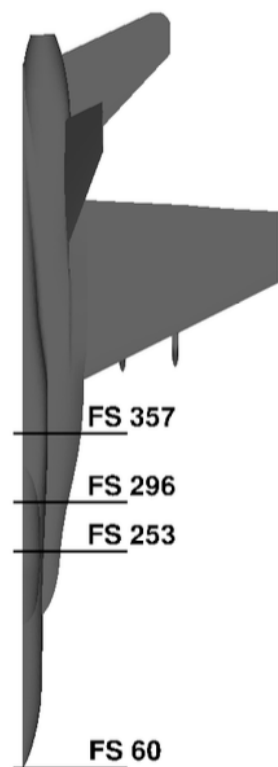




贝叶斯优化

➤ 对不同工况反复求解优化设计

$M_\infty = 0.243, \text{AoA} = 30.3^\circ$



4天 1000CPU's



贝叶斯优化

➤ 挑战

每次计算 $f(x)$ 计算量很大，函数 $f(x)$ 的评估被认为是“昂贵的”，这意味着可以执行的评估次数是有限的，通常只限于几百次。

$f(x)$ 像由黑盒，函数 f 缺乏诸如凹性或线性等已知的特殊结构，也难以得到一阶或二阶导数的信息，这些结构会使得利用相关技术来优化函数变得容易并提高效率。我们总结这一点，说 f 是一个“黑箱”。

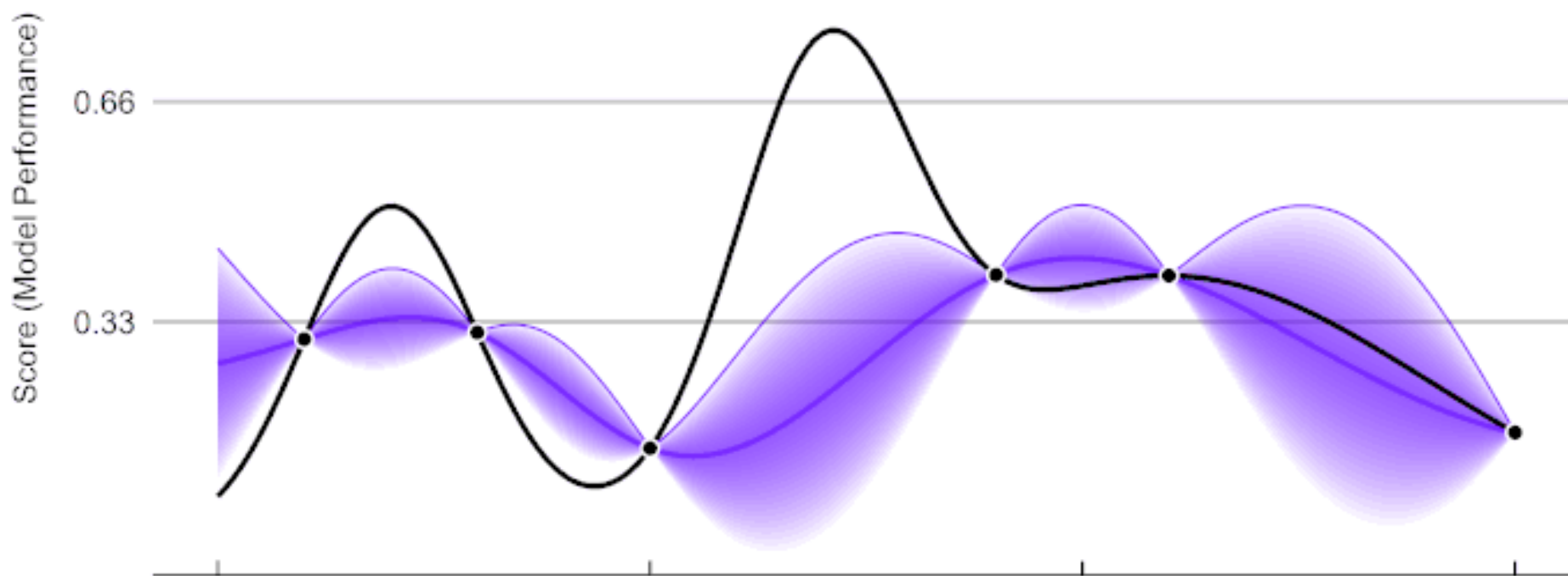
关注点在于寻找全局最优解，而非局部最优解



贝叶斯优化

➤ 贝叶斯优化

ParBayesianOptimization in Action (Round 1)





贝叶斯优化

➤ 贝叶斯优化

给定先验分布 $f(x) \sim GP(0, \kappa(x, x'))$

给定事先选定的 n_0 点 $X = \{x_1, x_2, \dots, x_{n_0}\}$ ，以及相应的值 $y = \{f(x_1), f(x_2), \dots, f(x_{n_0})\}$

依次添加新的点，根据

计算后验分布 $f|X, y$

根据后验分布，以及给定的采样函数(acquisition function)，选取新的点 x_{n+1} ，计算 $f(x_{n+1})$



采样函数

➤ 概率改进(Probability of improvement)采样函数

$$\text{定义 } f_n^* = \max_{i \leq n} f(x_i)$$

当我们选了新点 x ，我们得到的值为 $f(x)$

$$\text{定义效用函数 } u(x) = \begin{cases} 1 & f(x) \geq f_n^* \\ 0 & f(x) < f_n^* \end{cases}$$

我们希望选取新的点 x_{n+1} 最大化

$$\mathbb{E}[u(x) | x_{n+1} = x] = \int_{f_n^*}^{+\infty} \mathcal{N}(f; \mu_n(x), \sigma_n(x)^2) dx$$

$$\text{最后返回 } f_N^* = \max_{m \leq N} f(x_m)$$



采样函数

期望改进 (expected improvement) 采样函数

$$\text{定义 } f_n^* = \max_{i \leq n} f(x_i)$$

当我们选了新点 x ，我们得到的极大值为 $f(x)$ ($f(x) \geq f_n^*$)
或 f_n^* ($f_n^* \geq f(x)$)

最大值增量为 $[f(x) - f_n^*]^+ := \max\{0, f(x) - f_n^*\}$

对于后验分布 $f(x) \sim \mathcal{N}(\mu_n(x), \sigma_n^2(x))$ ，最大化

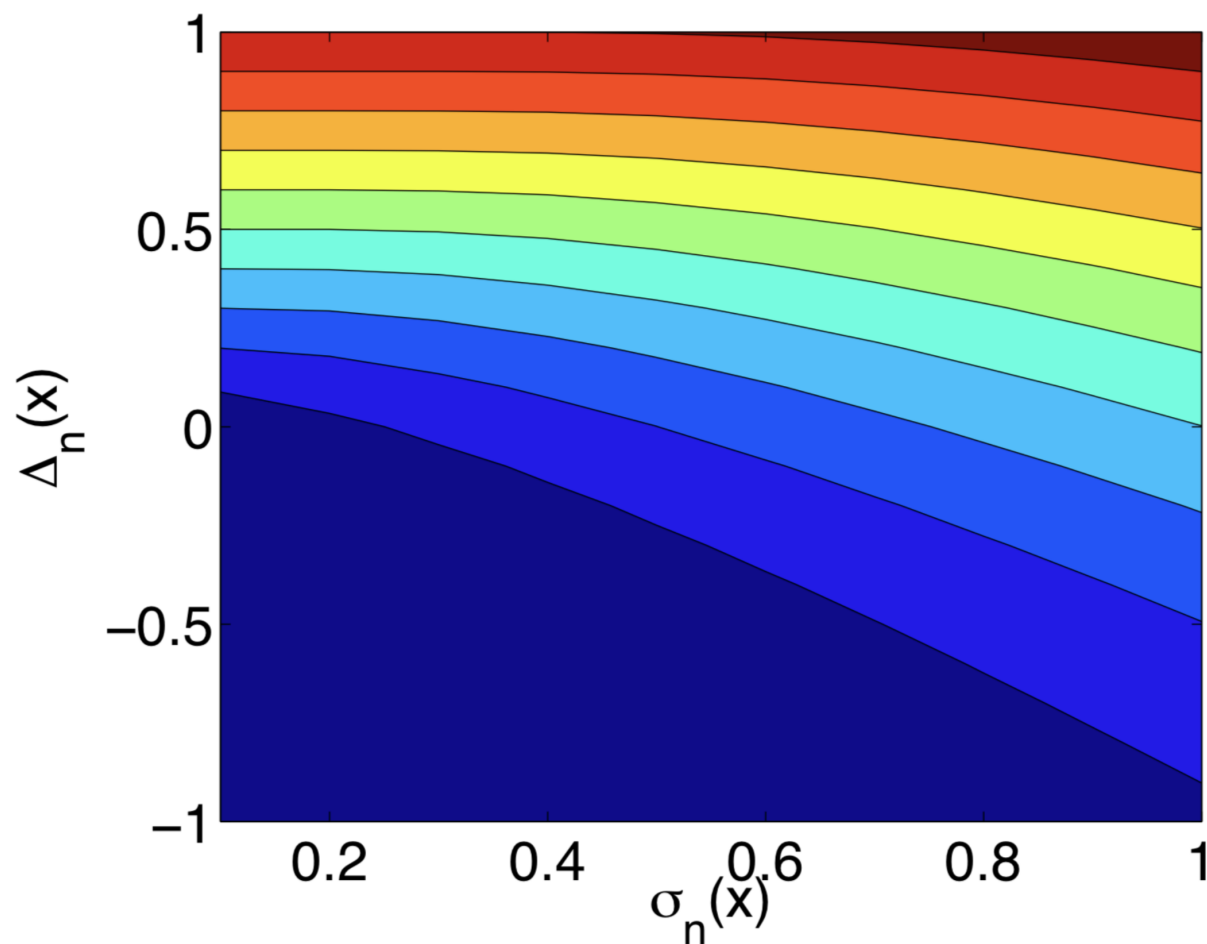
$$\begin{aligned} a_{EI}(x) &= \mathbb{E}[[f(x) - f_n^*]^+] \\ &= \Delta_n(x) \Phi\left(\frac{\Delta_n(x)}{\sigma_n(x)}\right) + \sigma_n \varphi\left(\frac{\Delta_n(x)}{\sigma_n(x)}\right) \end{aligned}$$

其中 $\Delta_n(x) = \mu_n(x) - f_n^*$ 。



采样函数

期望改进 (expected improvement) 采样函数



高期望 $\Delta_n(x)$
高不确定性 $\sigma_n(x)$

探索与利用的权 (exploration vs. exploitation tradeoff)



采样函数

➤ 信息增益(information gain)和实验设计(experimental design)

信息增益通常指在观察到新数据点后，对目标函数 $f(x)$ 的认识增加的程度。它通常通过互信息来量化，

$$I(y; f|X) = H(y|X) - H(y|X, f)$$

这表示观察到新数据后，模型对函数不确定性减少的程度。

假设每次观测有噪音： $y = f(x) + \epsilon$ $\epsilon \sim \mathcal{N}(0, \sigma^2)$

$$\begin{aligned} I(y; f|X) &= \frac{1}{2} \log |2\pi e(K + \sigma^2 I)| - \frac{1}{2} \log |2\pi e \sigma^2 I| \\ &= \frac{1}{2} \log |\sigma^{-2} K + I| \end{aligned}$$



采样函数

➤ 信息增益(information gain)和实验设计(experimental design)

假设每次观测有噪音： $y = f(x) + \epsilon$ $\epsilon \sim \mathcal{N}(0, \sigma^2)$

$$\begin{aligned} I(y; f|X) &= \frac{1}{2} \log |2\pi e(K + \sigma^2 I)| - \frac{1}{2} \log |2\pi e \sigma^2 I| \\ &= \frac{1}{2} \log |\sigma^{-2} K + I| \end{aligned}$$

选取 $X = \{x_1, x_2, \dots, x_n\}$ 最大化 $\operatorname{argmax}_X I(y; f|X)$,
NP-困难, 贪心法选取 X : 给定 $X = \{x_1, x_2, \dots, x_n\}$,
选取 x_{n+1} 最大化

$$\frac{1}{2} \log |\sigma^{-2} K + I|$$

$$\Rightarrow x_{n+1} = \operatorname{argmax} \sigma(x)$$



采样函数

➤ 上置信界 (upper confidence bound) 采样函数

$$a(x) = \mu(x) + \beta\sigma(x)$$

$\mu(x)$ 尽量小

$\sigma(x)$ 不确定性尽量大

$$x_{n+1} = \operatorname{argmax} \mu(x) + \beta\sigma(x)$$



采样函数

收敛性

考虑有限空间 D 上的优化问题，给点 $\delta \in (0,1)$ 和 $\beta_i = 2 \log\left(\frac{|D|i^2\pi^2}{6\delta^2}\right)$ ，使用上置信界 (upper confidence bound) 采样函数的贝叶斯优化满足

$$P\left(R_n \leq \sqrt{\frac{8n\beta_n\gamma_n}{\log(1+\sigma^2)}}\right) \geq 1 - \delta$$

其中观测误差满足 $\epsilon \sim \mathcal{N}(0, \sigma^2)$ ， $R_n = \sum_{i=1}^n r_i$ 是累计遗憾(cumulative regret)，即时遗憾 (regret) $r_i = f(x^*) - f(x_i)$ ， $\gamma_n = \max_{|X|=n} I(y; f|X)$ 是 n 轮后最大信息增益。



采样函数

➤ 上置信界 (upper confidence bound) 采样函数

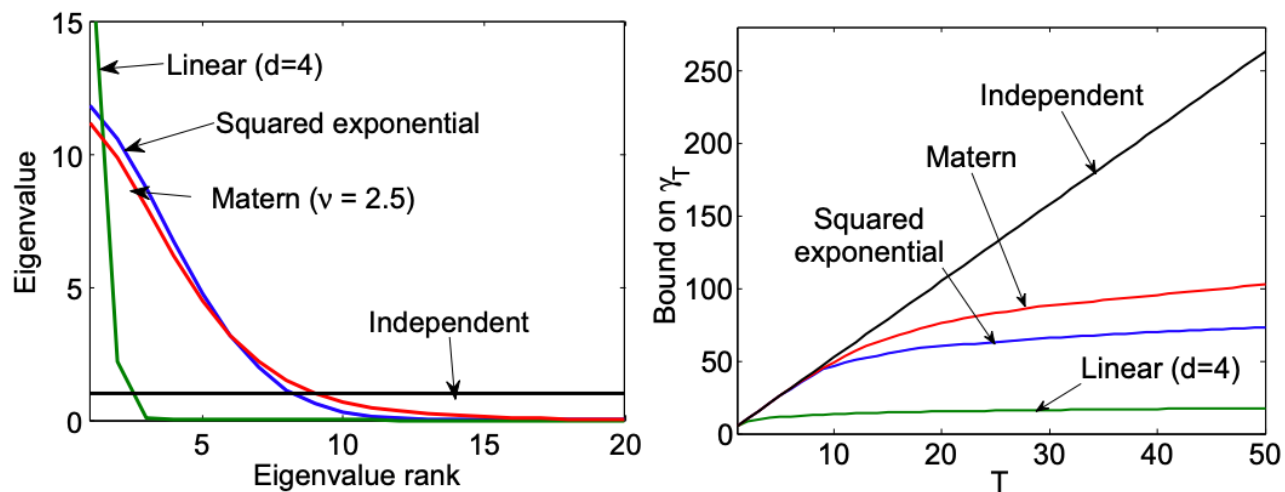


Figure 3. Spectral decay (left) and information gain bound (right) for independent (diagonal), linear, squared exponential and Matérn kernels ($\nu = 2.5$.) with equal trace.



采样函数

上述方法认为我们只愿意将之前评估过的点作为最终解返回。这一假设在评估无噪声且我们极度规避风险的情况下是合理的。但如果我们能够容忍一定的风险，则可能愿意报告带有一定不确定性的解。此外，如果评估存在噪声，那么输出的最终解必然具有不确定性，因为我们几乎不可能对其进行无限次的评估。

$$\text{输出 } \mu_n^* = \max_x \mu_n(x)$$



采样函数

➤ 知识梯度 (knowledge gradient) 采样函数

$$\text{定义 } \mu_n^* = \max_x \mu_n(x)$$

当我们选了新点 x_{n+1} ，我们得到的值为 $f(x)$ ，对于后验分布 $f(x) \sim \mathcal{N}(\mu_{n+1}(x), \sigma_{n+1}^2(x))$ ，那么 $\mu_{n+1}^* = \max_x \mu_{n+1}(x)$

$$a_{KG}(x) = \mathbb{E}_{f(x_{n+1})} [\mu_{n+1}^* - \mu_n^* | x_{n+1} = x]$$

随机近似：

$$a_{KG}(x) \approx \mu_{n+1}^*(x; y_{n+1}) - \mu_n^* \quad y_{n+1} \sim \mathcal{N}(\mu_n(x), \sigma_n^2(x))$$



采样函数

熵搜索(entropy search)采样函数

熵函数 $H(\rho) = -\int \rho(x) \log \rho(x) dx$, 越小越确定

是在取了 n 个点, 确定了极值点后 x_n^* , 我们希望极值点的熵尽量小 $H(\rho_n(x_n^*))$, 其中 $x_n^* = \operatorname{argmax}_x \mu_n(x)$ 。

$$a_{ES}(x)$$

$$= H(\rho_n(x_n^*)) - \mathbb{E}_{f(x_{n+1})} H(\rho_{n+1}(x_{n+1}^* | x_{n+1} = x, f(x_{n+1})))$$

$$\mathbb{E}_{f(x_{n+1})} H(\rho_{n+1}(x_{n+1}^* | x_{n+1} = x, f(x_{n+1})))$$

$$= \int \mathcal{N}(y; \mu_n(x), \sigma_n^2(x)) H(\rho_{n+1}(x_{n+1}^* | x_{n+1} = x, y)) dy$$



参考文献

➤ 参考文献

Frazier, Peter I. "A tutorial on Bayesian optimization." arXiv preprint arXiv:1807.02811 (2018).

Srinivas, Niranjan, Andreas Krause, Sham M. Kakade, and Matthias Seeger. "Gaussian process optimization in the bandit setting: No regret and experimental design." arXiv preprint arXiv:0912.3995 (2009).