Bayesian bandits

Finite arms / actions

Mean reward vector $\mu = \begin{bmatrix} \mu(1) \\ \vdots \\ \mu(K) \end{bmatrix} \sim P$ (prior)

Thompson Sampling

- Observe history $H_{t-1} = H = \{(a_1, r_1), \ldots (a_{t-1}, r_{t-1})\}$
- Sample mean vector from posterior $\mu_t \sim P_H(\mu)$ ← posterior
- Choose best arm acc to μ_t, $a_t = \arg\max_a \mu_t(a)$
- Add a_t, r_t to $H_{t-1} \rightarrow H_t$.

Bayesian Regret $O(\sqrt{KT \log T})$

a_t and a^* have same dist given H_{t-1} ✓

$E[UCB(a_t, H_t) | H_{t-1}] = E[UCB(a^*, H_{t-1}) | H_{t-1}]$ ✓

a^* : $\mu \sim P$ $\quad a^* = \arg\max_a \mu(a)$ $\quad \Pr(a^* | H_{t-1})$

a_t : $P_{H_{t-1}}(\mu)$ $\quad \mu_t \sim P_{H_t}(\mu)$ $\quad a_t = \arg\max_a \mu_t(a)$ $\quad \Pr(a_t | H_{t-1})$

$$\Pr(a_t | H_{t-1}) = \Pr(\mu_t(a_t) \geq \mu_t(a) \ \forall a \ | H_{t-1})$$
$$= \Pr(\mu(a_t) \geq \mu(a) \ \forall a \ | H_{t-1})$$
$$= \Pr(a^* | H_{t-1})$$

Continuous arms (correlated) - Gaussian Process prior

reward $\mu(x) \sim GP(m(x), k(x, x'))$ $\quad \forall x \in \mathcal{X}$

is continuous

covariance function = $\begin{cases} x^T x' & \text{linear (Bayesian)} \\ \exp\{-\frac{\|x - x'\|^2}{2r^2}\} & \text{RBF or squared kernel} \\ \text{Matern kernel} \end{cases}$

$r = \mu(x) + \varepsilon \quad \varepsilon \sim N(0, \sigma^2)$

$=$ ind.

$\mu \sim GP(0, K)$

Observations $(x_1, r_1) \dots (x_T, r_T)$

Posterior is also $GP(m_T(x), k_T(x, x'))$

$$m_T(x) = \underset{1\times T}{k_T^T(x)} \; \underset{T\times T}{(K_T + \sigma^2 I)^{-1}} \; \underset{T\times 1}{\vec{r}_T} \qquad \vec{r}_T = [r_1 \dots r_T]^T$$

$$K_T = [k(x_i, x_j)]_{x_i, x_j \in \text{observations}}$$

$$k_T(x) = [k(x, x_1) \dots k(x, x_T)]^T$$

posterior covariance $k_T(x, x') = k(x, x') - k_T^T(x)(K_T + \sigma^2 I)^{-1} k_T(x')$

$\sigma_T^2(x) = k_T(x, x)$ posterior variance $\equiv$ uncertainty

---

Experimental Design (Bayesian with GP prior)

Information gain $I(r_A, \mu) = \frac{1}{2} \log |I + \sigma^{-2} K_A|$

A - subset of samples from X $\quad I(r_A, \mu_A)$ $\quad K_A = [k(x, x')]_{x, x' \in A}$

$\max_{|A| \le T} \underbrace{I(r_A, \mu)}_{F(A)}$ is NP-hard (even for GP)

$F(A)$ is a submodular function Krause-Guestrin '05

$X, Y, X \subseteq Y$. for every $x \notin Y$

$$F(X \cup x) - F(X) \geq F(Y \cup x) - F(Y)$$

$A_t \cup x \quad A_t \quad A_{t+1} \cup x \quad A_{t+1}$

Diminishing returns

Any submodular function can be approximately (upto a constant) optimized using an efficient greedy procedure. Nemhauser 1978

monotonicity

$$F(A_T) \geq \left(1 - \frac{1}{e}\right) \max_{|A| \leq T} F(A)$$

Greedy algorithm:

$$x_t = \arg\max_{x \in \mathcal{X}} F(A_{t-1} \cup x) \qquad A_{t-1} = \{x_1, \dots, x_{t-1}\}$$

$$= \arg\max_{x} \sigma_{t-1}(x)$$

GP Bandits (exp design doesn't exploit, so doesn't optimize cum regret)

GP-UCB

$$x_t = \arg\max_{x} \; m_{t-1}(x) + \sqrt{\beta_t}\,\sigma_{t-1}(x)$$

↑ exploit ↑ explore

$$\beta_t = 2 \log \frac{|\mathcal{X}|\, t^2 \pi^2}{6\delta} \qquad |\mathcal{X}| \text{ finite}$$

$K(\cdot,\cdot) \leq 1$

Theorem: Regret $R_T = \sum_{t=1}^{T} r(x^*) - r(x_t) = O\left(\sqrt{T \beta_T \gamma_T}\right) \; \forall T$ w.p. $\geq 1-\delta$

max info gain $\gamma_T = \max_{|A| = T} I(r_A, \mu)$

Regret	Linear	RBF	Matern (ν)
$\sqrt{T \beta_T \gamma_T}$			
β_T	———	continuous: $d \log \frac{T}{\delta}$ or finite: $\log \frac{\lvert\mathcal{X}\rvert T}{\delta}$	———
γ_T	$d \log T$	$(\log T)^{d+1}$	$T^{\frac{d(d+1)}{2\nu + d(d+1)}} \log T$

Proof sketch: finite $\mathcal{X}$

w.p. $\geq 1-\delta$ $\quad |\mu(x) - m_{t-1}(x)| \leq \sqrt{\beta_t}\,\sigma_{t-1}(x) \quad \forall x$ $\quad \mu \sim GP(m_{t-1}(x), \sigma_{t-1}(x))$ conditioned on H_{t-1}

$$r_t = \mu(x^*) - \mu(x_t) \overset{\text{choice}}{\leq} \underbrace{m_{t-1}(x^*) + \sqrt{\beta_t}\,\sigma_{t-1}(x^*)} - \mu(x_t)$$

$$\overset{UCB}{\leq} m_{t-1}(x_t) + \sqrt{\beta_t}\,\sigma_{t-1}(x_t) - \mu(x_t)$$

$$\overset{conc.}{\leq} 2\sqrt{\beta_t}\,\sigma_{t-1}(x_t)$$

$$\sum_{t=1}^{T} R_t \leq 2\sum_{t=1}^{T} \sqrt{\beta_t}\,\sigma_{t-1}(x_t) \qquad = O\left(\sqrt{T\beta_T \gamma_T}\right) \quad \text{w.p.} \geq 1-\delta$$

$$\leq 2\sqrt{\beta_T} \sum_{t=1}^{T} \sigma_{t-1}(x_t) \cdot 1$$

$$\leq 2\sqrt{\beta_T}\,\sqrt{T}\,\sqrt{\sum_{t=1}^{T} \sigma_{t-1}^2(x_t)} \qquad \text{IF}$$

$$IF(\mathcal{A}, \mu) = \frac{1}{2}\sum_{t=1}^{T} \log\left(1 + \frac{\sigma_{t-1}^2(x)}{\sigma^2}\right) \rightsquigarrow \sigma^{-2} K_{\mathcal{A}}$$

$$\sum_{t=1}^{T} \sigma_{t-1}^2(x) = \sigma^2 \cdot \sum_{t=1}^{T} \frac{\sigma_{t-1}^2(x)}{\sigma^2} = O\left(\sigma^2 \sum_{t} \log\left(1 + \frac{\sigma_{t-1}^2(x)}{\sigma^2}\right)\right)$$

$$z \leq O(\log(1+z))$$

$$e^{c^2} \leq 1 + z$$

Continuous setting

$\mathcal{X} \subseteq \mathbb{R}^d$ compact, convex assumptions on kernel ($\equiv$ Lipschitz)

follows discretization

RKHS version

μ smoothness controlled by kernel $\|\mu\|_k \leq B$

- low RKHS norm induced by kernel k.

$\mu \in \mathcal{H}_k(\mathcal{X})$ - RKHS(K) - Reproducing Kernel Hilbert Space

$\langle \cdot, \cdot \rangle_k$ obeys $\langle \mu, k(x) \rangle_k = \mu(x)$

RKHS norm $\langle \mu, \mu \rangle_k = \|\mu\|_k^2$

Thm: $\mu \in$ RKHS(k) $\|\mu\|_k^2 \leq B$

$\beta_t = 2B + c\gamma_t \log 1/\delta$ w.p. $\geq 1-\delta$ $R_T = O(\sqrt{T\gamma_T \beta_T})$

Doesn't assume $f \in \mathcal{H}_k$ — uniform for all $f \in \mathcal{H}_k$ & $\|f\|_k^2 \le B$

Proof sketch: Freedman's inequality

$$Z_t = \|m_t - \mu\|^2_{K_T}$$

- 3 terms

same as HW linear problem

GP - Thompson sampling.

$$\mu_t \sim GP(m_t, K_t)$$

$$x_t = \arg\max_{x \in \mathcal{X}} \mu_t(x)$$

$$\mu_t(x) + \sqrt{\beta_t}\,\sigma_t(x)$$

$$\sqrt{K_t(x,x)}$$

$$1$$