An Analysis of Ensemble Sampling

Presenter: Hao Liang
217019008@link.cuhk.edu.cn

The Chinese University of Hong Kong, Shenzhen, China

March 1, 2022

Original Paper: Qin, C., Wen, Z., Lu, X., & Van Roy, B. (2022). An Analysis of Ensemble Sampling. arXiv preprint arXiv:2203.01303.

Outline

Thompson Sampling in Online Decision Making

- Thompson sampling (TS) is an effective heuristic for trading off between exploration and exploitation in online decision making problems
  - bandit: MAB, linear bandit, contextual bandit
  - RL
- Mechanism
  - Maintain a posterior distribution of models (initialized with a prior)
  - Sample models from the posterior distribution as statistically plausible models
  - Choose greedy/optimal action w.r.t. the statistically plausible models
- Limitations
  - Requires conjugacy properties, e.g. Beta/Bernoulli, Gaussian/Gaussian
  - Difficult to apply to complex models like neural net

From Thompson Sampling to Ensemble Sampling

- Approximate TS algorithms
  - Laplace approximation: limited to unimodal distribution
  - MCMC: computationally expensive for complex models
  - Ensemble sampling (ES)
  - Hypermodel: generalization/variation of ES
- ES as a practical approximation to TS
  - Fast incremental update/low computational cost
  - Applicable to neural net
- Applications of variations of ES
  - DRL [Osband et al., 2016, 2018, 2019]
  - Online recommendation [Lu et al., 2018, Hao et al., 2020, Zhu and Van Roy, 2021]
  - MARL [Dimakopoulou and Van Roy, 2018]

Main contributions

- No rigorous theory of ES
- Lu and Van Roy [2017] provided the first regret bound of ES applied to the linear bandit, with a flaw in the analysis
- Contributions
  - The first rigorous regret analysis of ES for the linear bandit
  - A general Bayesian regret bound for any algorithm for the linear bandit

Outline

Linear Gaussian Bandit

- Reward at time t and is linear in action a with Gaussian noise

$$R_{t,a} = a^\top \theta + W_{t,a}$$

  - $a \in \mathcal{A} \subset \mathbb{R}^d$ with $K = |\mathcal{A}|$
  - $\theta \in \mathbb{R}^d$ is the model parameter
  - W_t is an i.i.d. sequence with $W_t \sim N\left(0, \sigma^2 I_K\right)$
- Bayesian framework

$$\theta \sim \mathbb{P}_1(\theta \in \cdot) = N(\mu_0, \Sigma_0)$$

- At each time t, the agent chooses A_t and only observes R_{t,A_t}
- Bandit/partial feedback: $R_{t,a}$ for $a \neq A_t$ not revealed

Bayesian Regret

- History at time t

$$\mathcal{H}_t = \left(A_1, R_{1,A_1}, \ldots, A_{t-1}, R_{t-1,A_{t-1}}\right).$$

- Given a model θ, the optimal action

$$A_* := \arg\max_{a \in \mathcal{A}} \mathbb{E}\left[R_{t,a} | \theta\right] = \arg\max_{a \in \mathcal{A}} a^\top \theta$$

- Frequentist regret

$$\begin{aligned}
\mathrm{Regret}(T, \theta) &:= \sum_{t=1}^{T} \mathbb{E}\left[R_{t,A_*} - R_{t,A_t} | \theta\right] \\
&= \sum_{t=1}^{T} \mathbb{E}\left[A_*^\top \theta - A_t^\top \theta | \theta\right].
\end{aligned}$$

- Bayesian regret

$$\mathrm{Regret}(T) := \mathbb{E}_{\theta \sim \mathbb{P}_1}[\mathrm{Regret}(T, \theta)].$$

Ensemble Sampling

- Without conjugacy properties, exact TS becomes computationally infeasible
- ES serves as a practical approximation to TS

Algorithm 1 `Thompson Sampling`

for $t \in [T]$ **do**
  Sample $\tilde{\theta}_t \sim \mathbb{P}(\theta \in \cdot|\mathcal{H}_t)$
  Execute $A_t \sim \arg\max_{a \in \mathcal{A}} a^\top \tilde{\theta}_t$
  Observe R_{t,A_t}
  Update $\mathbb{P}(\theta \in \cdot|\mathcal{H}_t) \longrightarrow \mathbb{P}(\theta \in \cdot|\mathcal{H}_{t+1})$

Algorithm 2 `Ensemble Sampling`

Sample: $\tilde{\theta}_{1,1}, \ldots, \tilde{\theta}_{1,M} \sim \mathbb{P}_1(\theta \in \cdot)$
for $t \in [T]$ **do**
  Sample $m \sim \mathrm{unif}\{1, \ldots, M\}$
  Execute $A_t \sim \arg\max_{a \in \mathcal{A}} a^\top \tilde{\theta}_{t,m}$
  Observe R_{t,A_t}
  Update $\tilde{\theta}_{t,1:M} \longrightarrow \tilde{\theta}_{t+1,1:M}$

Update Details

- The posterior distribution at time $t+1$ is still Gaussian
$$\Sigma_{t+1} = \left(\Sigma_t^{-1} + \frac{1}{\sigma^2} A_t A_t^\top\right)^{-1} \quad \text{and} \quad \mu_{t+1} = \Sigma_{t+1}\left(\Sigma_t^{-1}\mu_t + \frac{R_{t,A_t}}{\sigma^2} A_t\right)$$
- Satisfies conjugacy properties though
- ES updates each m-th model according to
$$\tilde{\theta}_{t+1,m} = \Sigma_{t+1}\left(\Sigma_t^{-1}\tilde{\theta}_{t,m} + \frac{R_{t,A_t} + \tilde{W}_{t,m}}{\sigma^2} A_t\right),$$
where each $\tilde{W}_t = \left(\tilde{W}_{t,1}, \ldots, \tilde{W}_{t,M}\right) \sim N\left(0, \sigma^2 I_M\right)$ is an independent random perturbation.

Outline

Regret Bound

- Regret bound for ES
- General regret bound for any algorithms
- From general regret to regret bound for ES

Regret Bound for ES

Theorem 1.

Algorithm 2 for linear bandit with prior $N(\mu_0, \Sigma_0)$ and M models satisfies

$$\mathrm{Regret}(T) \leqslant \underbrace{\iota\sqrt{dT\mathbb{H}(A_*)}}_{(a)} + \underbrace{\kappa T\sqrt{\frac{K\log(6TM)}{M}}}_{(b)}$$

$$= \tilde{\mathcal{O}}\left(\sqrt{dT\mathbb{H}(A_*)} + T\sqrt{K/M}\right),$$

where $\mathbb{H}(A_*)$ is the entropy of the optimal action A_* under the prior, and

$$\iota := \sqrt{2\left(\max_{a\in\mathcal{A}} a^\top \Sigma_0 a + \sigma^2\right)}$$

and

$$\kappa := 2\sqrt{(4\log K + 5)\max_{a\in\mathcal{A}} a^\top \Sigma_0 a + \max_{a\in\mathcal{A}} (a^\top \mu_0)^2 + \sigma^2} = \tilde{\mathcal{O}}(\sqrt{\log K}).$$

Comparison with the regret bound on TS

The regret bound for ES

$$\mathrm{Regret}(T) \leqslant \tilde{\mathcal{O}}\left(\underbrace{\sqrt{dT\mathbb{H}\left(A_*\right)}}_{(a)} + \underbrace{T\sqrt{K/M}}_{(b)}\right)$$

- Term (a) is exactly the regret bound achieved by TS [Russo and Van Roy, 2016]
- Term (b) accounts for posterior distribution mismatch
- As $M \to \infty$, term (b) converges to 0, and the regret bound reduces to that of TS
- When M is finite and satisfies $M = \Omega(KT/d)$, the regret bound matches TS

Notations

- For two discrete distributions $P = (p_1, \ldots, p_n)$ and $Q = (q_1, \ldots, q_n)$, the KL divergence and Hellinger distance between P and Q are defined as

$$\mathbf{d}_{\mathrm{KL}}(P\|Q) := \sum_{i\in[n]} p_i \log\left(p_i/q_i\right) \quad \text{and} \quad \mathbf{d}_{\mathrm{H}}(P\|Q) := \sqrt{\sum_{i\in[n]} \left(\sqrt{p_i} - \sqrt{q_i}\right)^2}$$

- They satisfy

$$\mathbf{d}_{\mathrm{H}}^2(P\|Q) \leqslant \min\left\{\mathbf{d}_{\mathrm{KL}}(P\|Q), \mathbf{d}_{\mathrm{KL}}(Q\|P)\right\}$$

Notations

- The Shannon entropy of X

$$\mathbb{H}(X) := -\sum_{x\in\mathcal{X}} \mathbb{P}(X=x)\log\mathbb{P}(X=x)$$

- The entropy of X conditional on $Y=y$

$$\mathbb{H}(X \mid Y=y) := -\sum_{x\in\mathcal{X}} \mathbb{P}(X=x \mid Y=y)\log\mathbb{P}(X=x \mid Y=y)$$

- The conditional entropy of X given Y

$$\mathbb{H}(X \mid Y) := \mathbb{E}_Y\left[-\sum_{x\in\mathcal{X}} \mathbb{P}(X=x \mid Y)\log\mathbb{P}(X=x \mid Y)\right]$$

- The mutual information between X and Y

$$\mathbb{I}(X;Y) := \mathbf{d}_{\mathrm{KL}}(P(X,Y)\|P(X)P(Y))$$

- The conditional mutual information between X and Y given Z

$$\mathbb{I}(X;Y \mid Z) := \mathbb{E}_Z\left[\mathbf{d}_{KL}(P(X,Y \mid Z)\|P(X \mid Z)P(X \mid Z))\right]$$

General Regret Bound

- Derive a general Bayesian regret bound for any learning algorithm
- It is of independent interest and might be used to analyze other bandit algorithms
- Use subscript t to denote conditioning on H_t,

$$\mathbb{P}_t(\cdot) := \mathbb{P}\left(\cdot \mid H_t\right) \quad \text{and} \quad \mathbb{E}_t[\cdot] := \mathbb{E}\left[\cdot \mid H_t\right]$$

- Define

$$\bar{p}_t(\cdot) := \mathbb{P}_t\left(A_t = \cdot\right) \quad \text{and} \quad p_t(\cdot) := \mathbb{P}_t\left(A_* = \cdot\right)$$

- Both $\bar{p}_t$ and p_t are specified by the algorithm
  - Under TS, $\bar{p}_t = p_t$
  - For approximate TS, $\bar{p}_t \approx p_t$

General Regret Bound

Theorem 2.

Any learning algorithm for linear bandit satisfies

$$\mathrm{Regret}(T) \leqslant \iota\sqrt{dT\mathbb{H}\left(A_*\right)} + \eta\sum_{t=1}^{T}\sqrt{\mathbb{E}\left[\mathbf{d}_{\mathrm{H}}^{2}\left(\bar{p}_t \| p_t\right)\right]},$$

where

$$\eta := 2\sqrt{\mathbb{E}\left[\max_{a\in\mathcal{A}}\left(a^{\top}\theta\right)^{2}\right] + \sigma^{2}}.$$

Note that the expectation is taken w.r.t. the prior over θ.

Discussion on the General Regret Bound

- The first term matches TS
- The second term quantifies the cumulative difference between $\bar{p}_t$ and p_t
- $\mathbf{d}_{\mathrm{H}}^2 \left(\bar{p}_t \| p_t\right)$
  - vanishes for TS
  - should be small for well approximated TS
- $\eta = 2\sqrt{\mathbb{E}\left[\max_{a\in\mathcal{A}} \left(a^\top\theta\right)^2\right] + \sigma^2}$ depends on the prior. For Gaussian prior,
  - $a^\top\theta$ is Gaussian r.v. for any $a \in \mathcal{A}$
  - the expectation of the maximum of K squares of Gaussian r.v.s is $\mathcal{O}(\log K)$
- Indeed, $\eta \leq \kappa = 2\sqrt{(4\log K + 5)\max_{a\in\mathcal{A}} a^\top\Sigma_0 a + \max_{a\in\mathcal{A}} \left(a^\top\mu_0\right)^2 + \sigma^2}$

Regret Bound in terms of KL divergence

- Analyzing the KL divergence are sometimes easier
- Recall

$$\mathbf{d}_{\mathrm{H}}^2(P\|Q) \leqslant \min\{\mathbf{d}_{\mathrm{KL}}(P\|Q), \mathbf{d}_{\mathrm{KL}}(Q\|P)\}$$

Corollary 3.

Under then setting of Theorem 2,

$$\mathrm{Regret}(T) \leqslant \iota\sqrt{dT\mathbb{H}(A_*)} + \eta\sum_{t=1}^{T}\sqrt{\mathbb{E}\left[\min\left\{\mathbf{d}_{\mathrm{KL}}\left(\bar{p}_t\|p_t\right), \mathbf{d}_{\mathrm{KL}}\left(p_t\|\bar{p}_t\right)\right\}\right]}$$

- Will show that for ES, $\mathbf{d}_{\mathrm{KL}}\left(\bar{p}_t\|p_t\right)$ can be bounded in terms of M

Proof Sketch of Theorem 2

- Step 1: rewrite cumulative regret

$$\text{Regret}(T) = \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{E}_t\left[R_{t+1,A_*} - R_{t+1,A_t}\right]\right]$$

- Step 2: regret decomposition as sum of "main regret" and "approximation error",

$$\mathbb{E}_t\left[R_{t,A_*} - R_{t+1,A_t}\right] = G_t + D_t$$

where

$$G_t \triangleq \sum_{a\in\mathcal{A}} \sqrt{\bar{p}_t(a)p_t(a)}\left(\mathbb{E}_t\left[R_{t,a} \mid A_* = a\right] - \mathbb{E}_t\left[R_{t+,a}\right]\right)$$

and

$$D_t \triangleq \sum_{a\in\mathcal{A}} \left(\sqrt{p_t(a)} - \sqrt{\bar{p}_t(a)}\right)\left(\sqrt{p_t(a)}\mathbb{E}_t\left[R_{t,a} \mid A_* = a\right] + \sqrt{\bar{p}_t(a)}\mathbb{E}_t\left[R_{t,a}\right]\right)$$

Proof Sketch of Theorem 2

- Step 3: Bound $\sum_{t=1}^{T} \mathbb{E}\left[G_t\right]$

$$G_t \leqslant \iota \sqrt{d \cdot \mathbb{I}_t\left(A_*; \left(A_t, R_{t,A_t}\right)\right)} \qquad \text{(information-theoretic)}$$

$$\Longrightarrow \sum_{t=1}^{T} \mathbb{E}\left[G_t\right] \leqslant \iota \sqrt{dT\mathbb{H}\left(A_*\right)} \qquad \text{(Cauchy-Schwartz inequality + chain rule)}$$

- Step 4: Bound $\sum_{t=1}^{T} \mathbb{E}\left[G_t\right]$

$$\mathbb{E}\left[D_t\right] \leqslant \eta \sqrt{\mathbb{E}\left[\mathbf{d}_{\mathrm{H}}^2\left(\bar{p}_t \| p_t\right)\right]} \qquad \text{(Cauchy-Schwartz inequality)}$$

Outline

Bound $\sum_{t=1}^{T} \sqrt{\mathbb{E}\left[\mathbf{d}_{\mathrm{H}}^{2}\left(\bar{p}_t \| p_t\right)\right]}$

Lemma 4.

Under ES, for all $t \in [T]$*,*

$$\mathbb{E}\left[\mathbf{d}_{KL}\left(\bar{p}_t \| p_t\right)\right] \leqslant \frac{K \log(6(t+1)M)}{M}.$$

- Plugging Lemma 4 into the general regret bound in Theorem 2, we achieve the regret bound for ES in Theorem 1

Outline

Proof of Lemma 4

- ES first uniformly samples $m \in [M]$, and then samples the action A_t corresponding to $\tilde{\theta}_{t,m}$ uniformly from the optimal action set

$$\tilde{\mathcal{A}}_{t,m} := \arg\max_{a \in \mathcal{A}} a^\top \tilde{\theta}_{t,m}$$

- Define the following approximation of $p_t(a)$

$$\hat{p}_t(a) := \frac{1}{M} \sum_{m=1}^{M} \frac{1}{\left|\tilde{\mathcal{A}}_{t,m}\right|} \mathbb{I}\left\{a \in \tilde{\mathcal{A}}_{t,m}\right\}$$

- History H_t does not include $\tilde{W}_t$, and

$$\tilde{\theta}_{t,m} = \Sigma_t \left(\Sigma_{t-1}^{-1} \tilde{\theta}_{t-1,m} + \frac{R_{t,A_t} + \tilde{W}_{t,m}}{\sigma^2} A_t \right)$$

- Given H_t, $\hat{p}_t(a)$ is still random

$$\bar{p}_t(a) = \mathbb{E}_t\left[\hat{p}_t(a)\right]$$

Proof of Lemma 4

- By convexity of KL divergence, the per-period approximation error

$$\mathbf{d}_{\mathrm{KL}}\left(\bar{p}_t \| p_t\right) = \mathbf{d}_{\mathrm{KL}}\left(\mathbb{E}_t\left[\hat{p}_t\right] \| p_t\right) \leqslant \mathbb{E}_t\left[\mathbf{d}_{\mathrm{KL}}\left(\hat{p}_t \| p_t\right)\right]$$

- Taking expectation on both sides

$$\mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\bar{p}_t \| p_t\right)\right] \leqslant \mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\hat{p}_t \| p_t\right)\right]$$

- $$\mathbb{E}\left[\mathbf{d}_{KL}\left(\hat{p}_t \| p_t\right)\right] = \int_0^\infty \mathbb{P}\left(\mathbf{d}_{KL}\left(\hat{p}_t \| p_t\right) > \epsilon\right) \mathrm{d}\epsilon$$

- Derive an upper bound on $\mathbb{P}\left(\mathbf{d}_{KL}\left(\hat{p}_t \| p_t\right) > \epsilon\right)$ for any $\epsilon > 0$

Proof of Lemma 4

- For simplicity, consider deterministic action sequence $a_{1:t} := (a_1, ..., a_t)$. Write
$$p_t^{a_{1:t-1}}(\cdot) := \mathbb{P}_t\left(A_* = \cdot \mid a_1, R_{1,a_1}, \ldots, a_{t-1}, R_{t,a_{t-1}}\right)$$
- Under deterministic action sequence $a_{1:t}$
$$\tilde{\theta}_{t,1}^{a_{0:t-1}}, \ldots, \tilde{\theta}_{t,M}^{a_{0:t-1}} | R_{1,a_1}, \ldots, R_{t,a_{t-1}} \sim \mathbb{P}_t(\theta \in \cdot)$$
- $\hat{p}_t^{a_{0:t-1}}$ is an empirical distribution for $p_t^{a_{0:t-1}}$

Fact 5 (Sanov's theorem).

Relate to stochastic action sequence

- Using Sanov's theorem

$$\mathbb{P}\left(\mathbf{d}_{\mathrm{KL}}\left(\hat{p}_t^{a_{0:t-1}} \| p_t^{a_{0:t-1}}\right) > \epsilon \mid \theta\right) \leqslant (M+1)^K e^{-M\epsilon}$$

- By applying the union bound over action counts, instead of action sequences

$$\mathbb{P}\left(\max_{a_{0:t-1} \in \mathcal{A}^t} \mathbf{d}_{\mathrm{KL}}\left(\hat{p}_t^{a_{0:t-1}} \| p_t^{a_{0:t-1}}\right) > \epsilon \mid \theta\right) \leqslant (t+1)^K (M+1)^K e^{-M\epsilon}$$

- Relate to the KL divergence associated with stochastic action sequence

$$\mathbb{P}\left(\mathbf{d}_{KL}\left(\hat{p}_t \| p_t\right) > \epsilon \mid \theta\right) \leqslant \mathbb{P}\left(\max_{a_{0:t-1} \in \mathcal{A}^t} \mathbf{d}_{KL}\left(\hat{p}_t^{a_{0:t-1}} \| p_t^{a_{0:t-1}}\right) > \epsilon \mid \theta\right)$$

Proof

- Fix t. For any $\epsilon > 0$,

$$\mathbb{P}\left(\mathbf{d}_{KL}\left(\hat{p}_t \| p_t\right) > \epsilon\right) = \mathbb{E}[\mathbb{P}\left(\mathbf{d}_{KL}\left(\hat{p}_t \| p_t\right) > \epsilon|\theta\right)|\theta] \leqslant (t+1)^K (M+1)^K e^{-M\epsilon}.$$

- For any threshold $\delta \geqslant 0$,

$$\begin{aligned}\mathbb{E}\left[\mathbf{d}_{\mathrm{KL}}\left(\hat{p}_t \| p_t\right)\right] &= \int_0^\infty \mathbb{P}\left(\mathbf{d}_{\mathrm{KL}}\left(\hat{p}_t \| p_t\right) > \epsilon\right) \mathrm{d}\epsilon \\ &\leqslant \delta + (t+1)^K (M+1)^K \int_\delta^\infty e^{-M\epsilon} \mathrm{d}\epsilon \\ &= \delta + \frac{(t+1)^K (M+1)^K e^{-M\delta}}{M}\end{aligned}$$

- Choosing the optimal $\delta^* = \frac{K[\log(t+1)+\log(M+1)]}{M}$

$$\mathbb{E}\left[\mathbf{d}_{KL}\left(\hat{p}_t \| p_t\right)\right] \leqslant \frac{K[\log(t+1) + \log(M+1)] + 1}{M} \leqslant \frac{K \log(6(t+1)M)}{M}$$