Thompson Sampling for Online Learning with Linear Experts

Aditya Gopalan
Technion, Haifa, Israel
aditya@ee.technion.ac.il

November 5, 2013

Abstract

In this note, we present a version of the Thompson sampling algorithm for the problem of online linear generalization with full information (i.e., the experts setting), studied by Kalai and Vempala, 2005. The algorithm uses a Gaussian prior and time-varying Gaussian likelihoods, and we show that it essentially reduces to Kalai and Vempala’s Follow-the-Perturbed-Leader strategy, with exponentially distributed noise replaced by Gaussian noise. This implies $\sqrt{T}$ regret bounds for Thompson sampling (with time-varying likelihood) for online learning with full information.

1 Setup

Consider the full-information linear generalization setting, similar to the one studied by Kalai and Vempala [1]. We can select, at each time $t \geq 1$, a decision $d_t$ from an action set $D \subset \mathbb{R}^n$. Following the $t$-th decision $d_t$, we get to observe $s_t \in S \subset \mathbb{R}^n$ and receive a reward of $\langle d_t, s_t \rangle$. The goal is to maximize the total reward $\sum_t \langle d_t, s_t \rangle$.

As shorthand we will write $S_t$ for the vector $s_1 + s_2 + \ldots + s_t$, and $x_i$ for the $i$-th coordinate of a vector $x$. Throughout, $I_n$ and $\mathbf{1}_n$ denote the identity matrix and all-ones vector in dimension $n$ respectively.

Consider the Thompson Sampling algorithm TSG($\epsilon$), with $\epsilon > 0$, and Gaussian prior and likelihood (Algorithm 1).

\begin{algorithm}
\caption{TSG($\epsilon$)}
\begin{algorithmic}
\FOR{$t = 1, 2, 3, \ldots$}
\STATE 1. Assume that $\{s_k\}_{k \leq t}$ are independent and identically distributed $N(\mu, \frac{1}{\sqrt{\epsilon(t-1)}}I_n)$ samples, where $\mu$ follows the prior distribution $\mu \sim N(0, \frac{1}{\epsilon}I_n)$. Draw $\theta_t \in \mathbb{R}^n$ from the posterior distribution $P[\mu | \{s_k\}_{k \leq t}]$.
\STATE 2. Play $d_t = \arg\max_{d \in D} \langle d, \theta_t \rangle$.
\END\FOR
\end{algorithmic}
\end{algorithm}
By standard results, upon observing iid standard normal samples $x_1, \ldots, x_{t-1}$ distributed as $N(\mu, \sigma^2)$, with nonrandom variance $\sigma^2$ and prior $\mu \sim N(\mu_0, \sigma_0^2)$, the posterior distribution of the mean $\mu$ is again Gaussian with mean $\frac{\sigma_0^2 x + \sigma^2 \mu_0}{\sigma_0^2 + \sigma^2}$ and variance $\left(\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2}\right)^{-1}$. In our case, at time $t \geq 2$,

$$
P \left[ \theta_t \mid \{s_k\}_{k < t} \right] \sim N \left( \frac{\frac{1}{e} + \frac{1}{e(t-1)^2}}{t-1} S_{t-1} + \frac{1}{t-1} e + e(t-1)^2 \right)
$$

$$
\Rightarrow \quad P \left[ \left( t - 1 + \frac{1}{t-1} \right) \theta_t \mid \{s_k\}_{k < t} \right] \sim N \left( S_{t-1}, e^{-1} \left( 1 + \frac{1}{(t-1)^2} \right) \right)
$$

Thus, the TSG algorithm perturbs the aggregate 'state' $S_{t-1}$ seen so far with Gaussian noise, and takes the best decision for this perturbed state. This is akin to the Follow-the-Perturbed-Leader (FPL) strategy developed by Kalai and Vempala [1], and we apply their techniques to provide regret bounds for TSG that hold over all sequences $s_1, s_2, \ldots$ in $S$. Our result involves the following parameters:

$$
D \triangleq \sup_{d, d' \in D} \| d - d' \|_1, \quad R \triangleq \sup_{d \in D, s \in S} | \langle d, s \rangle |, \quad A_1 \triangleq \sup_{s \in S} \| s \|_1, \quad A_2 \triangleq \sup_{s \in S} \| s \|_2.
$$

As usual, for a sequence of states $s_1, s_2, \ldots, s_T$, we define the regret $R^A(T)$ of a strategy $A$ to be the difference between the reward earned by $A$ on the sequence and the reward earned by the best fixed decision in hindsight:

$$
R^A(T) \triangleq \sup_{d \in D} \sum_{t=1}^{T} \langle d, s_t \rangle - \sum_{t=1}^{T} \langle d^A_t, s_t \rangle.
$$

**Theorem 1.** The expected regret of TSG($\epsilon$) satisfies

$$
E \left[ R^{\text{TSG}(\epsilon)}(T) \right] \leq \sqrt{\epsilon} R A_2 K_{2,n} T + \frac{\epsilon R A_2^2 T}{2} + \frac{2 D K_{\infty,n}}{\sqrt{\epsilon}},
$$

where $K_{2,n}$ and $K_{\infty,n}$ are positive constants that depend only on $n$.

Note: Setting $\epsilon = \frac{1}{T}$ implies an expected regret of $O(\sqrt{T})$.  

**Proof.** Let us introduce the notation $M(x) \triangleq \arg \max_{d \in D} \langle d, x \rangle$. TSG chooses the decision $M(S_{t-1} + p_t)$ at time $t$, where $p_t \sim N(0, e^{-1}(1 + q_t))$, $q_t = \frac{1}{(t-1)^2}$.

First, an application of Lemma 3.1 in [1] gives that for any state sequence $s_1, s_2, \ldots, T > 0$ and vectors $p_0 = 0, p_1, \ldots, p_T$,

$$
\langle M(S_T), S_T \rangle \leq \sum_{t=1}^{T} \langle M(S_t + p_t), s_t \rangle + D \sum_{t=1}^{T} \| p_t - p_{t-1} \|_\infty.
$$

(1)

Next, observe that the expected reward is unchanged if for each $t > 1$, $p_t = p_t \sqrt{1 + q_t}$. For
such a noise sequence,

\[ \|p_t - p_{t-1}\|_\infty = \|p_1\|_\infty \cdot \sqrt{1 + q_t - \sqrt{1 + q_{t-1}}} \]

\[ \leq \|p_1\|_\infty \cdot \left| \left( \sqrt{1 + q_t} \right)^2 - \left( \sqrt{1 + q_{t-1}} \right)^2 \right| \]

\[ = \|p_1\|_\infty \cdot |q_t - q_{t-1}| \]

\[ \Rightarrow \sum_{t=2}^T \|p_t - p_{t-1}\|_\infty \leq \|p_1\|_\infty \sum_{t=2}^T |q_t - q_{t-1}| \]

\[ \leq \|p_1\|_\infty \sum_{t=2}^T \left( \frac{1}{(t-1)^2} - \frac{1}{t^2} \right) \]

\[ \leq \|p_1\|_\infty \cdot (2) \tag{2} \]

TSG earns reward \( \langle M(S_{t-1} + p_t), s_t \rangle \) at each time \( t \), and the best possible reward in hindsight over the entire time horizon \( 1, 2, \ldots, T \) is \( \langle M(S_T), S_T \rangle \), so in order to bound the regret of TSG using (1), it remains to bound the expectation of the difference \( \langle M(S_t + p_t), s_t \rangle - \langle M(S_{t-1} + p_t), s_t \rangle \). Let \( e_t^{-1} \triangleq e^{-1}(1 + q_t) \), and let \( d\nu_a(\cdot) \) be Gaussian measure on \( \mathbb{R}^n \) with mean 0 and variance \( a^{-1}I_n \). Observe that

\[
\mathbb{E} [\langle M(S_{t-1} + p_t), s_t \rangle] = \int_{x \in \mathbb{R}^n} \langle M(S_{t-1} + x), s_t \rangle \ d\nu_{e_t}(x)
\]

\[ = \sum_{y \in \mathbb{R}^n} \langle M(S_{t-1} + s_t + y), s_t \rangle \ d\nu_{e_t}(y + s_t) \]

\[ = \sum_{y \in \mathbb{R}^n} \langle M(s_t + y), s_t \rangle \ d\nu_{e_t}(y + s_t) \]

\[ = \sum_{y \in \mathbb{R}^n} \langle M(s_t + y), s_t \rangle e^{\frac{n}{2}\left(\|y\|^2 - \|y + s_t\|^2\right)} \ d\nu_{e_t}(y). \]

Thus, we can write

\[
\mathbb{E}[\langle M(S_t + p_t), s_t \rangle - \langle M(S_{t-1} + p_t), s_t \rangle]
\]

\[ = \sum_{z \in \mathbb{R}^n} \langle M(S_t + e_t^{\frac{1}{2}}z), s_t \rangle \left[ 1 - e^{\frac{n}{2}\left(\|z\|^2 - \|z + s_t\|^2\right)} \right] \ d\nu_{e_t}(z) \]

\[ = \sum_{z \in \mathbb{R}^n} \langle M(S_t + e_t^{\frac{1}{2}}z), s_t \rangle \left[ 1 - e^{\frac{n}{2}\left(-2\langle z, s_t \sqrt{e_t} \rangle - \|s_t\|^2\right)} \right] \ d\nu_{e_t}(z) \]

\[ \leq \sum_{z \in \mathbb{R}^n} \langle M(S_t + e_t^{\frac{1}{2}}z), s_t \rangle \left[ 1 - e^{-\sqrt{e_t} \|s_t\| \|z\| - \frac{e_t A^2}{2}} \right] \ d\nu_{e_t}(z) \]

(Cauchy-Schwarz, and assuming that \( (d, s) \geq 0 \ \forall d \in D, s \in S \))

\[ \leq \sum_{z \in \mathbb{R}^n} \left[ \sqrt{e_t} \|s_t\| \|z\| + \frac{e_t A^2}{2} \right] \ d\nu_{e_t}(z) \]

(since \( 1 - e^{-x} \leq x \))

\[ \leq \sqrt{e_t} RA_2 K_2 n + \frac{e_t R A_2^2}{2} \leq \sqrt{e_t} RA_2 K_2 n + \frac{e R A_2^2}{2}, \]
where \( K_{p,n} \triangleq \int_{z \in \mathbb{R}^n} \|z\|_p \, d\nu_1(z) \) for \( p \geq 1 \). Combining the above with (1) and (2) and summing over \( 1, 2, \ldots, T \) gives

\[
E[(M(S_T), S_T)] - E\left[ \sum_{t=1}^T \langle d_t^{SG}, s_t \rangle \right] \leq \sqrt{2\epsilon R A_2 K_{2,n} T} + \frac{\epsilon R A_2^2 T}{2} + \frac{2D K_{80,n}}{\sqrt{\epsilon}};
\]

completing the proof. \( \square \)

References

[1] A. T. Kalai and S. Vempala, “Efficient algorithms for online decision problems,” J. Comput. Syst. Sci., vol. 71, no. 3, pp. 291–307, 2005.