In this section, we consider a more general problem setting.

An agent applies a sequence of actions $x_1, x_2, x_3, \cdots$ to a system, selecting from a set $\mathcal{X}$. $\mathcal{X}$ could be either finite or infinite. After applying x_t, the agent observes y_t, which is generated from a conditional probability measure $q_\theta(y \mid x_t)$. The agent then collects a reward $r_t = r(y_t)$, where r is a known function. The agent is initially uncertain about the value of θ, and he represents his uncertainty using a prior distribution p.

We first list the Greedy Algorithm and TS algorithm and then explain.

Greedy $(\mathcal{X}, p, q, r)$	Thompson Sampling $(\mathcal{X}, p, q, r)$
for $t = 1, 2, \cdots$, do	for $t = 1, 2, \cdots$ do
# estimate θ from prior	# sample prior
$\hat{\theta} = \mathbb{E}_p[\theta]$	$\hat{\theta} \sim p$
# select and apply action	# select and apply action
$x_t = \underset{x \in \mathcal{X}}{\operatorname{argmax}}\ \mathbb{E}_{q_{\hat{\theta}}}[r(y_t) \mid x_t = x]$	$x_t = \underset{x \in \mathcal{X}}{\operatorname{argmax}}\ \mathbb{E}_{q_{\hat{\theta}}}[r(y_t) \mid x_t = x]$
Apply action x_t and get y_t.	
# update prior	# update prior
$p = \mathbb{P}_{p,q}(\theta \in \cdot \mid x_t, y_t)$	$p = \mathbb{P}_{p,q}(\theta \in \cdot \mid x_t, y_t)$

For Greedy.

* $\hat{\theta}$ is the optimal estimator from θ with. MSE. $\Rightarrow \mathbb{E}_p[\theta]$

* Suppose $y \in \mathcal{Y}$, then by going through all possible outcomes, and figure out which action, on average, rewards the most,

$$\mathbb{E}_{q_{\hat{\theta}}}[r(y_t) \mid x_t = x] = \sum_{y \in \mathcal{Y}} q_{\hat{\theta}}(y \mid x) \cdot r(y)$$

* The prior distribution p is updated based on observation y_t. If θ is sampled from a finite set, apply Bayes Rule., the estimation of p is updated follow: (for each value of u)

$$P(\theta = u \mid x_t, y_t) = \frac{\boxed{p(u)}\, q_u(y_t \mid x_t)}{\sum_v p(v)\, q_v(y_t \mid x_t)}.$$

old prior

Example.1. (Independent Travel Times)

Background:

An agent commutes from home to work every morning. She would like to minimise the travel time. How can she learn efficiently and minimize the expected travel time?

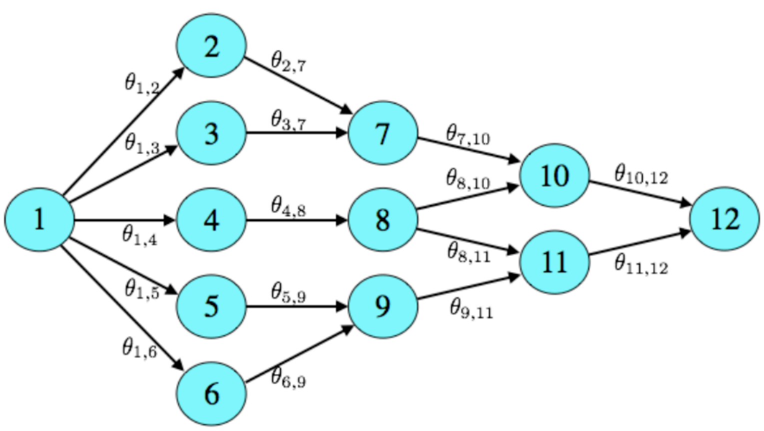

Figure 1.1: Shortest path problem.

Let $G=(V,E)$ be the graph representing the route.

$V=[N]$, $E=\{(i,j) \mid i,j \in V, \text{ edge}(i,j) \text{ in graph}\}$.

Vertex 1 is the source

Vertex N is the destination.

An action is a sequence of distinct edges from 1 to N, or a path.

After applying action x_t, the travel time at edge $e \in x_t$ is $y_{t,e}$ and $y_{t,e}$ is independently sampled from a distribution with mean θ_e.

The cost (negative of the reward) is $\sum_{e \in x_t} y_{t,e}$.

Consider a prior for θ_e to be Log-Gaussian with μ_e and σ_e^2, i.e. $\ln(\theta_e) \sim N(\mu_e, \sigma_e^2)$, then.

$$f_\theta(\theta) = \frac{1}{x} \cdot \frac{1}{\sqrt{2\pi}\sigma_e} \cdot \exp\left(-\frac{(\ln x - \mu_e)^2}{2\sigma^2}\right)$$

$$\theta \implies \mathbb{E}[\theta] = \exp\left(\mu_e + \frac{\sigma^2}{2}\right)$$

$$\mathrm{Var}[\theta] = \left[\exp(\sigma^2) - 1\right]\left[\overbrace{\exp(2\mu + \sigma^2)}^{\mathbb{E}[\theta]^2}\right].$$

In the problem setting, we assume $y_{t,e} \mid \theta$ is independent across $e \in E$, so, $\mathbb{E}[y_{t,e} \mid \theta_e] = \theta_e \leftarrow$ problem setting.

and $y_{t,e} \mid \theta_e \sim$ log-Gaussian with parameter $\left(\ln\theta_e - \frac{\hat{\sigma}^2}{2}, \hat{\sigma}^2\right)$

where $\hat{\sigma}$ is known.

Conjugacy property inspires the following update when $y_{t,e}$ is observed.

$$(\mu_e, \sigma_e^2) \leftarrow \left(\frac{\frac{1}{\sigma_e^2}\mu_e + \frac{1}{\hat{\sigma}^2}\left(\ln(y_{t,e}) + \frac{\hat{\sigma}^2}{2}\right)}{\frac{1}{\sigma_e^2} + \frac{1}{\hat{\sigma}^2}}, \frac{1}{\frac{1}{\sigma_e^2} + \frac{1}{\hat{\sigma}^2}} \right)$$

Suppose the agent knows the distance of each edge d_e, but is not sure about the travel time $y_{t,e}$. Initially, it may be a good idea to let $y_{t,e} \propto d_e$, or even more brutally let $\mathbb{E}[y_{t,e}] = d_e$. So initially, we have the averaged travel time

$$\mathbb{E}[y_{t,e}] = \exp(\mu_e + \frac{\sigma^2}{2}) = d_e$$

$$\Rightarrow \mu_e = \ln(d_e) - \frac{\sigma_e^2}{2}$$

However, from a single d_e, it's impossible to determin μ_e and σ_e^2 simultaneously. Any value can be possible.

Note that, $Var[y_{t,e}] = (\exp(\sigma_e^2) - 1) \cdot d_e^2$.

After that, Greedy algorithm as TS algorithm can be applied as follows:

At the beginning of each round, the agent has (μ_e, σ_e^2) for each edge from previous trials.

① * For Greedy Algorithm, $\hat{\theta}_e = \mathbb{E}_p[\theta_e] = \exp(\mu_e + \frac{\sigma_e^2}{2})$

* For Thompson Sampling, $\hat{\theta}_e$ is drawn from a log-Gaussian with mean (μ_e, σ_e^2)

② Then, each algorithm select its path to maximize

$$\mathbb{E}_{q_{\hat{\theta}}}[r(y_t) \mid x_t = x] = -\sum_{e \in x_t} \hat{\theta}_e .$$

This can be solved efficienty using Dijkstra for example.

③ Apply x_t.

④ Update μ_e and σ_e for each involved edge.

Example 2. (Correlated Travel Time)

We modify the shortest path problem.

We change the observation time model to

$$y_{t,e} = \underline{\xi_{t,e}} \cdot \underline{\eta_t} \cdot U_{t,\underline{l(e)}} \cdot \theta_e$$

$\xi_{t,e}$: the idiosyncratic factor with edge e.

η_t: a factor common to all edges.

$l(e)$ indicates whether edge e locates in the lower half of the binomial bridge. $l(e) \in \{0, 1\}$.

$U_{t,1}$: edge $e \in$ Lower half

$U_{t,0}$: $e \in$ upper half.

We let $\xi_{t,e}$, η_t, $U_{t,0}$, $U_{t,1}$ to be independent log-Gaussian with parameter $(-\tilde{\sigma}^2/6, \tilde{\sigma}^2/3)$. Their distributions are known, but θ_e also subjects to log-Gaussian (μ_e, σ_e^2) with unknown μ_e and σ_e^2.

Given $\xi_{t,e}$, η_t, $U_{t,0}$, $U_{t,1}$, the marginal distribution $y_{t,e} | \theta$ is identical to Example 1.

- Common factor η_t is used to model a global factor, say weather.
- $U_{t,0}$, $U_{t,1}$ refect events affecting only half of the path. Say, a bridge locates in the middle of the path.

Conjugate properties benefits the updating process.

Let $\phi_e = \ln(\theta_e)$, and $z_{t,e} = \begin{cases} \ln(y_{t,e}) & \text{if } e \in \mathcal{X}_t \\ 0 & \text{o.w.} \end{cases}$

$z_t = \begin{bmatrix} z_{t,1} \\ z_{t,2}, \\ \vdots \\ z_{t,N} \end{bmatrix}$

We then formulate a covariance matrix $\tilde{\Sigma}_{e,e'} \in R^{|\mathcal{X}_t| \times |\mathcal{X}_t|}$

$$\tilde{\Sigma}_{e,e'} = \begin{cases} \tilde{\sigma}^2 & \text{if } e = e', \\ \frac{2}{B}\tilde{\sigma}^2 & \text{if } e \neq e' \text{ but } l(e) = l(e') \\ \tilde{\sigma}^2/3 & \text{o.w.} \end{cases} \quad \text{for } e, e' \in \mathcal{X}_t$$

$$\tilde{C}_{e,e'} = \begin{cases} \tilde{\Sigma}^{-1}_{e,e'} & \text{if } e, e' \in \mathcal{X}_t \quad \text{for all } e, e' \in E \\ 0 & \text{o.w.} \end{cases}$$

Then $(\mu, \Sigma) = \left((\Sigma^{-1} + \tilde{C})^{-1} (\Sigma^{-1}\mu + \tilde{C} \cdot z_t), (\Sigma^{-1} + \tilde{C})^{-1} \right)$

Again, we can use Greedy as TS.