# Exploration in Online Decision Making
# (A whirlwind tour w/ everything but MDPs)

Daniel Russo

*Columbia University*

*Dan.Joseph.Russo@gmail.com*

# Outline: Part I

1. Briefly discuss classical bandit problems
2. Use the shortest path problem to teach TS
   - Emphasize flexible modeling of problem features
   - Discuss a range of issues like
     - Prior distribution specification
     - Approximate posterior sampling
     - Non-stationarity
     - Constraints, caution, and context
3. Discuss shortcomings and alternatives

Material drawn from

*A Tutorial on Thompson Sampling - Russo, Van Roy, Kazerouni, Osband, and Wen.*

*Learning to optimize via information-directed sampling – Russo and Van Roy.*

# Outline: Part 2 (Next week)

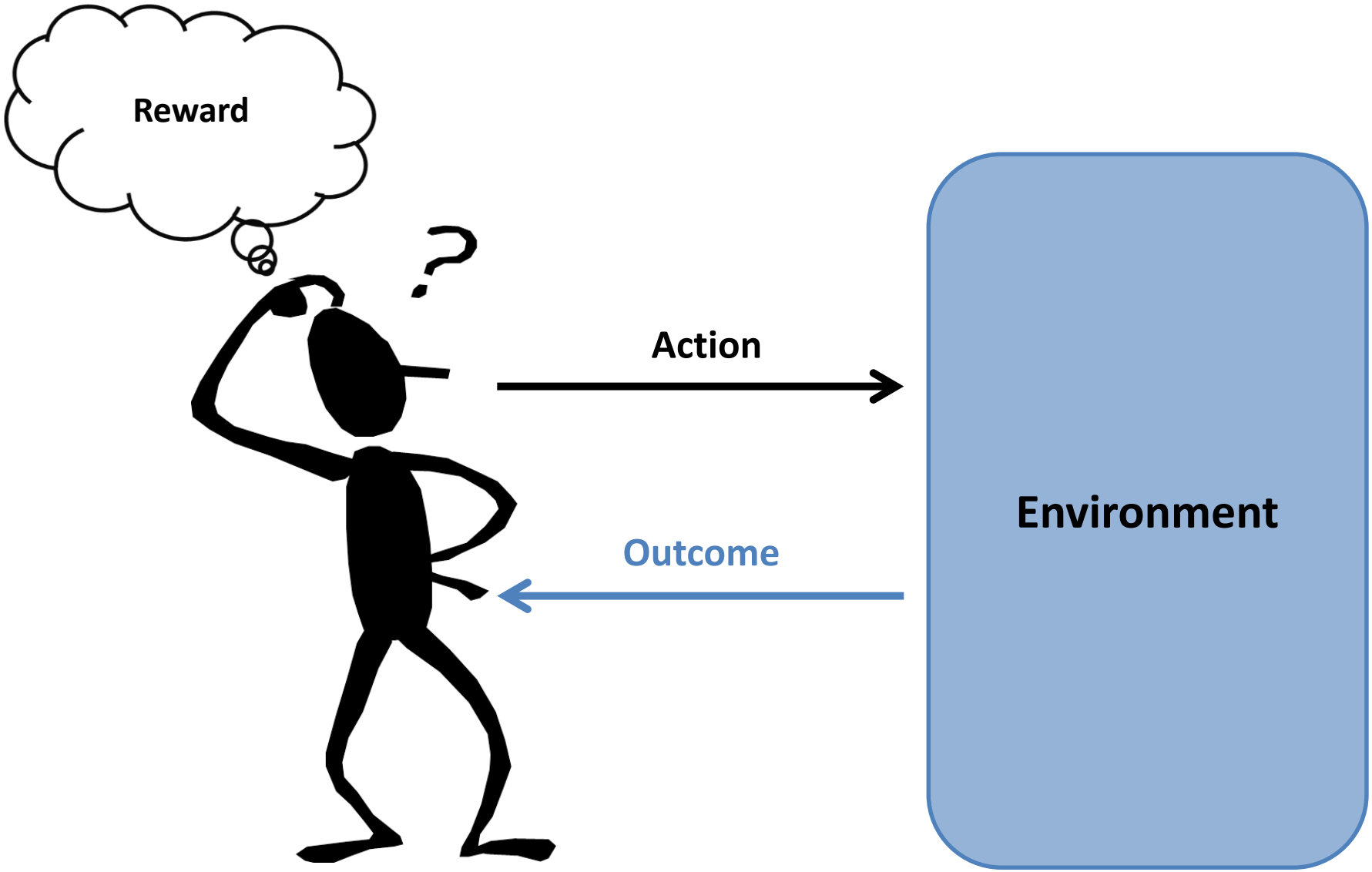
- Introduction to regret analysis.
- Focus on the case of a online linear optimization with "bandit feedback" and Gaussian observation noise.
- Give a regret analysis that applies to TS and UCB.

Material drawn from

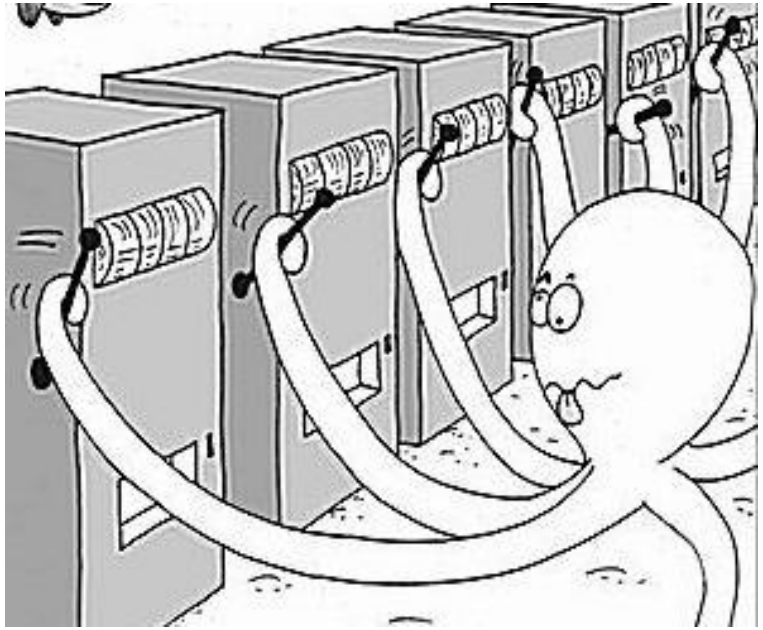
- *Russo and Van Roy: Learning to optimize via posterior sampling*
- *Dani, Hayes and Kakade: Stochastic Linear Optimization under Bandit Feedback*
- *Rusmevichientong and Tsitsiklis: Linearly parameterized bandits*

# Interactive Machine Learning:

## Intelligent information gathering

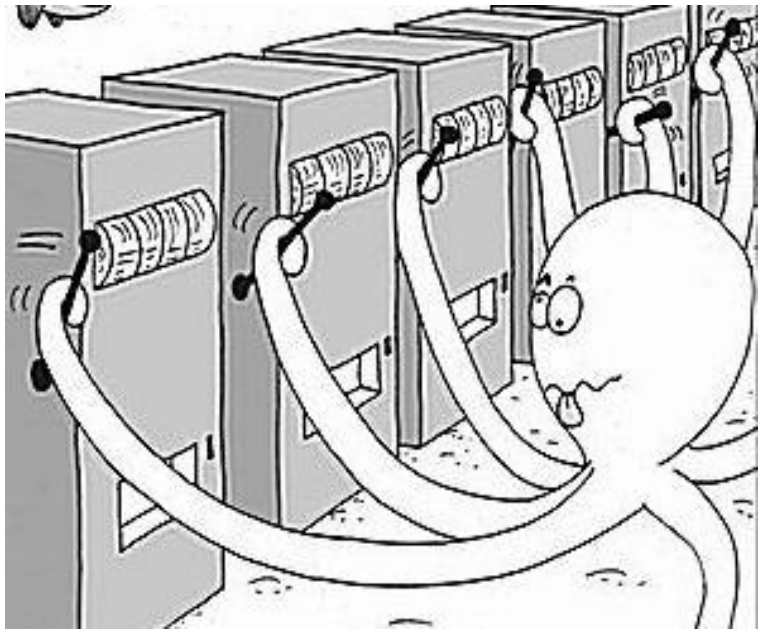


# The Multi-armed Bandit Problem



- A sequential learning and experimentation problem
- Crystalizes the **exploration/exploitation** tradeoff

# The Multi-armed Bandit Problem



- A sequential learning and experimentation problem
- Crystalizes the exploration/exploitation tradeoff
- Initial motivation: **clinical trials**

# Website Optimization


PLAY ANIMATION


PLAY ANIMATION

- Choose ad to show to User 1
- Observe click?
- Choose ad to show to User 2
- Observe click?
- …..

# Broad Motivation

- The information revolution is spawning systems that:
  - Make rapid decisions
  - Generate huge volumes of data
- Allows for small scale, adaptive, experiments

# Website Optimization: A Simple MAB problem

- 3 advertisements
- Unknown click probability: $\theta_1, \ldots, \theta_3 \in [0,1]$
- Choose adaptive algorithm displaying ads
- Goal: Maximize cumulative number of clicks.

# Greedy Algorithms

- Always play the arm with highest estimated success rate.

What is wrong with this?

*This algorithm requires point estimation*

- *a procedure for predicting the mean reward of an action given past data.*

# $\epsilon$-Greedy Algorithm

- With probability $1 - \epsilon$ play the arm with highest estimated success rate.
- With Probability $\epsilon$, pick an arm uniformly at random.

Why is this wasteful?

*This algorithm requires point estimation*

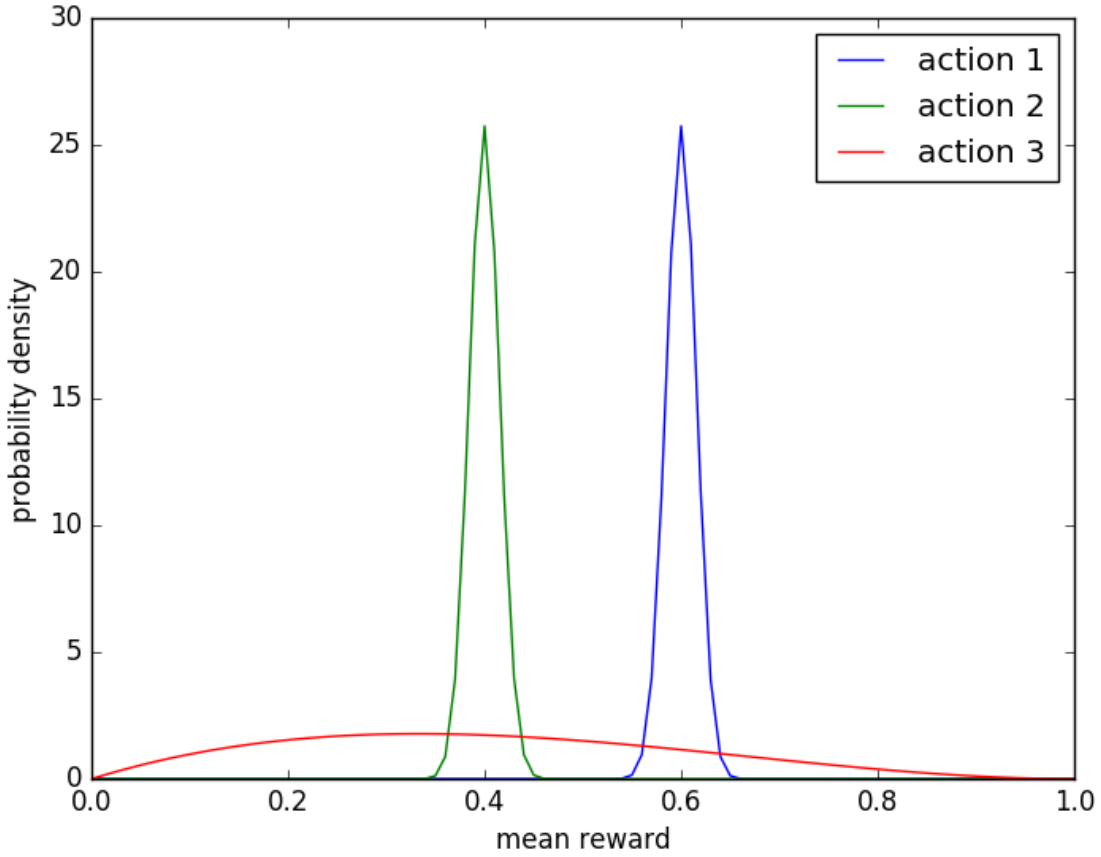
- *a procedure for predicting the mean reward of an action given past data.*

# An example

- Historical data on 3 actions
  - Played (1000,1000, 5) times respectively
  - Observed (600,400, 2) successes respectively.

- Synthesize observations with an independent uniform prior on each arm.

# Posterior Beliefs



# Comments

- Greedy is likely to play action 1 forever, even though there is a reasonable chance action 3 is better.
- $\epsilon$—Greedy fails to write off bad actions
  - Effectively wastes effort measuring action 2, and regardless of how convincing evidence against arm to is.

# Improved algorithmic design principles

- Continue to play actions that are plausibly optimal.
- Gradually write off actions as that are very unlikely to be optimal.

*This requires inference*

- *procedures assessing the uncertainty in estimated mean rewards.*

# Beta-Bernoulli Bandit

- A $k$ armed bandit with binary rewards
- Success probabilities $\theta = (\theta_1, \ldots \theta_k)$ are unknown but fixed over time.

$$p(r_t = 1 | x_t = i, \theta) = \theta_i$$

- Begin with a Beta prior with parameters $\alpha = (\alpha_1, \ldots, \alpha_k)$ and $\beta = (\beta_1, \ldots \beta_k)$.

$$p(\theta_k) = \frac{\Gamma(\alpha_k + \beta_k)}{\Gamma(\alpha_k)\Gamma(\beta_k)} \theta_k^{\alpha_k - 1} (1 - \theta_k)^{\beta_k - 1}$$

# Beta-Bernoulli Bandit

- Note, Beta(1,1)=Uniform(0,1)
- Posterior distributions are also Beta distributed, with simple update rule

$$(\alpha_k, \beta_k) = \begin{cases} (\alpha_k, \beta_k) & if\ x_t \neq k \\ (\alpha_k, \beta_k) + (r_t, 1 - r_t) & if\ x_t = k \end{cases}$$

- Posterior mean is $\alpha_k/(\alpha_k + \beta_k)$.

# Greedy

- For every period
  - Compute posterior means $(\mu_1, \ldots, \mu_K)$
  - $\mu_k = \alpha_k/(\alpha_k + \beta_k)$
  - Play $x = \text{argmax}_\text{k}\ \mu_k$
  - Observe reward and update $(\alpha_x, \beta_x)$

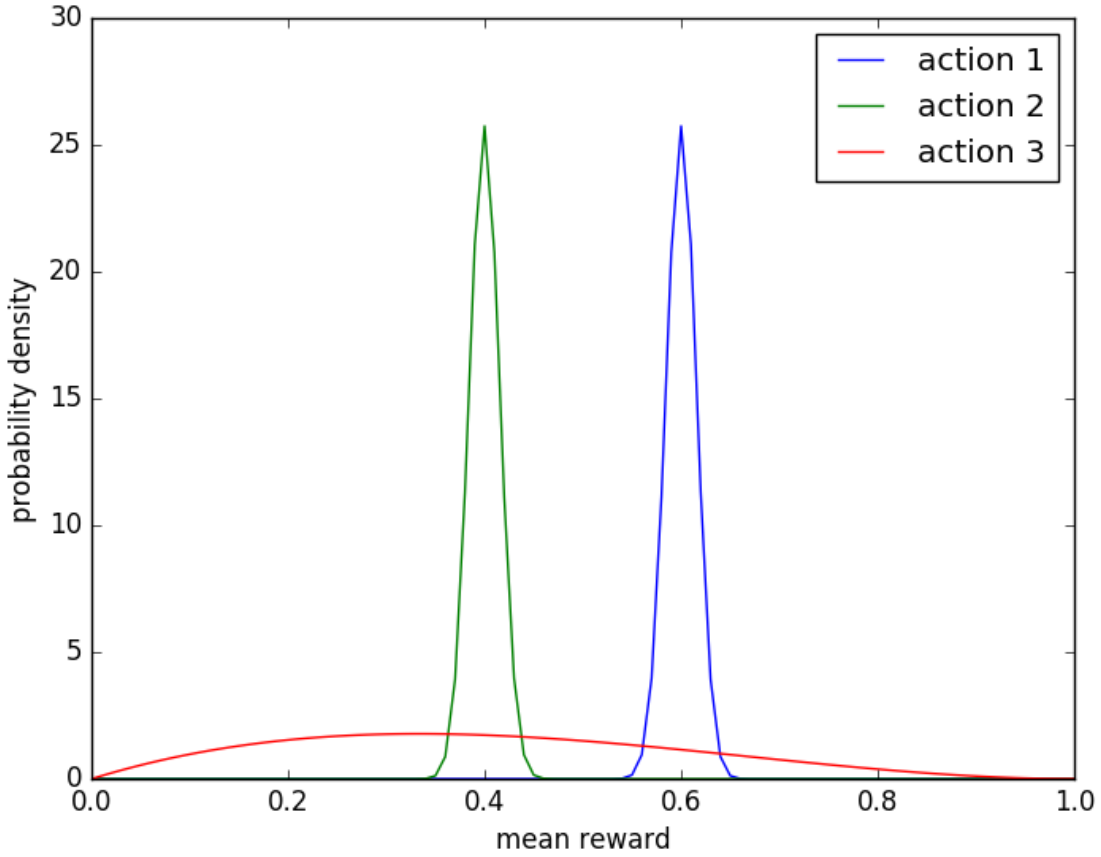
# Bayesian UCB

- For every period
  - Compute upper confidence bounds $(U_1, \ldots, U_K)$
    - $P_{\theta_k \sim Beta(\alpha_k, \beta_k)}(\theta_k \geq U_k) \leq \text{threshold}$
  - Play $x = \text{argmax}_\text{k}\ \text{U}_k$
  - Observe reward and update $(\alpha_x, \beta_x)$

# Thompson Sampling

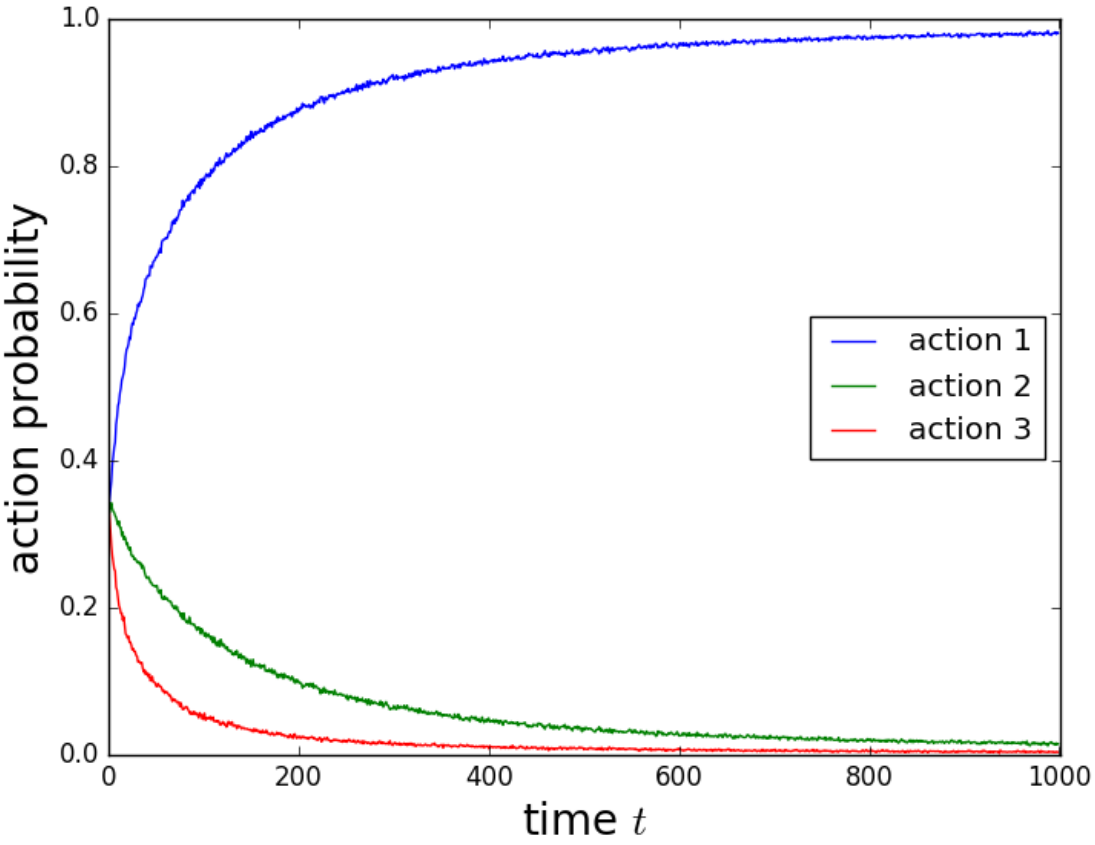
- For every period
  - Draw random samples $(\hat{\theta}_1, \ldots, \hat{\theta}_K)$
    - $\theta_k \sim Beta(\alpha_k, \beta_k)$
  - Play $x = \text{argmax}_{\text{k}}\, \hat{\theta}_k$
  - Observe reward and update $(\alpha_x, \beta_x)$

# What do TS and UCB do here?



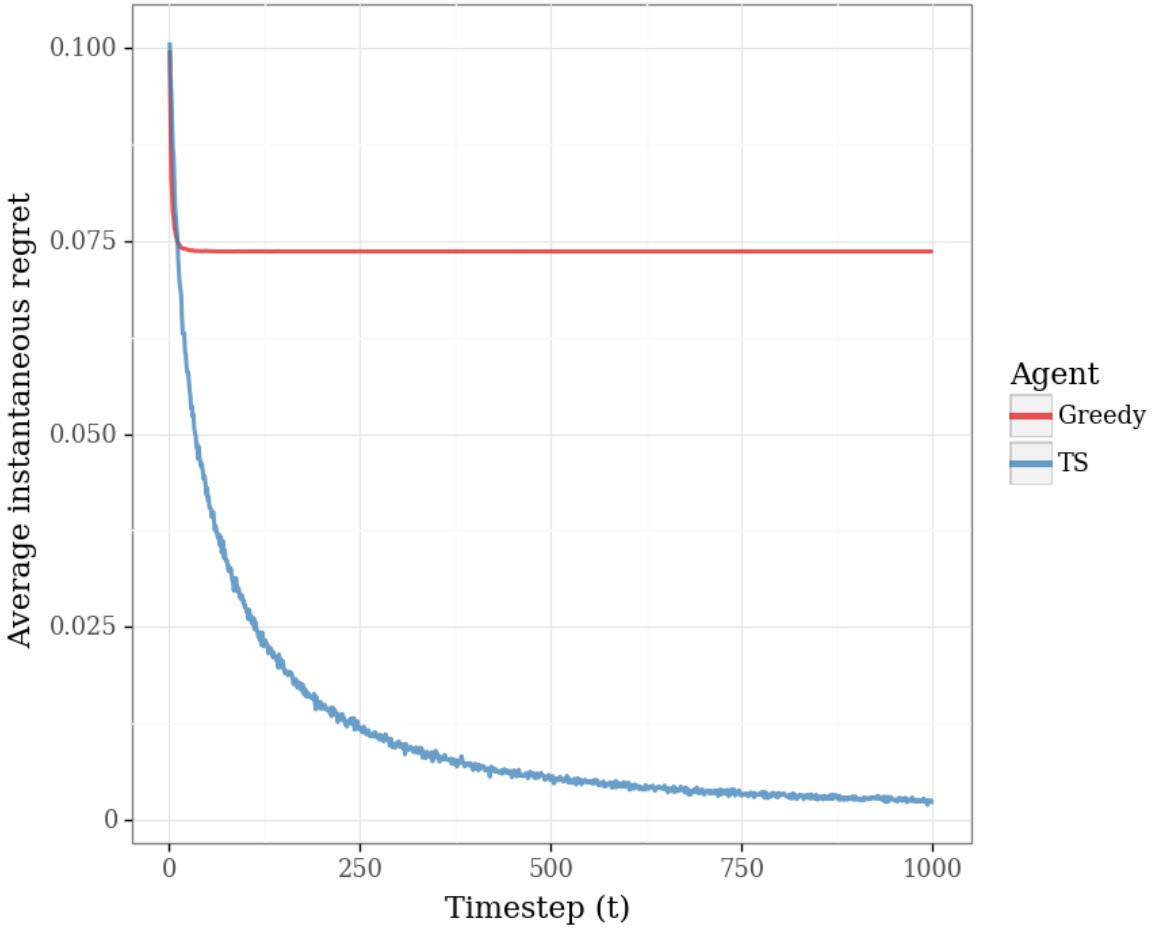
# A simulation of TS

- Fixed problem instance $\theta = (.9, .8, .7)$



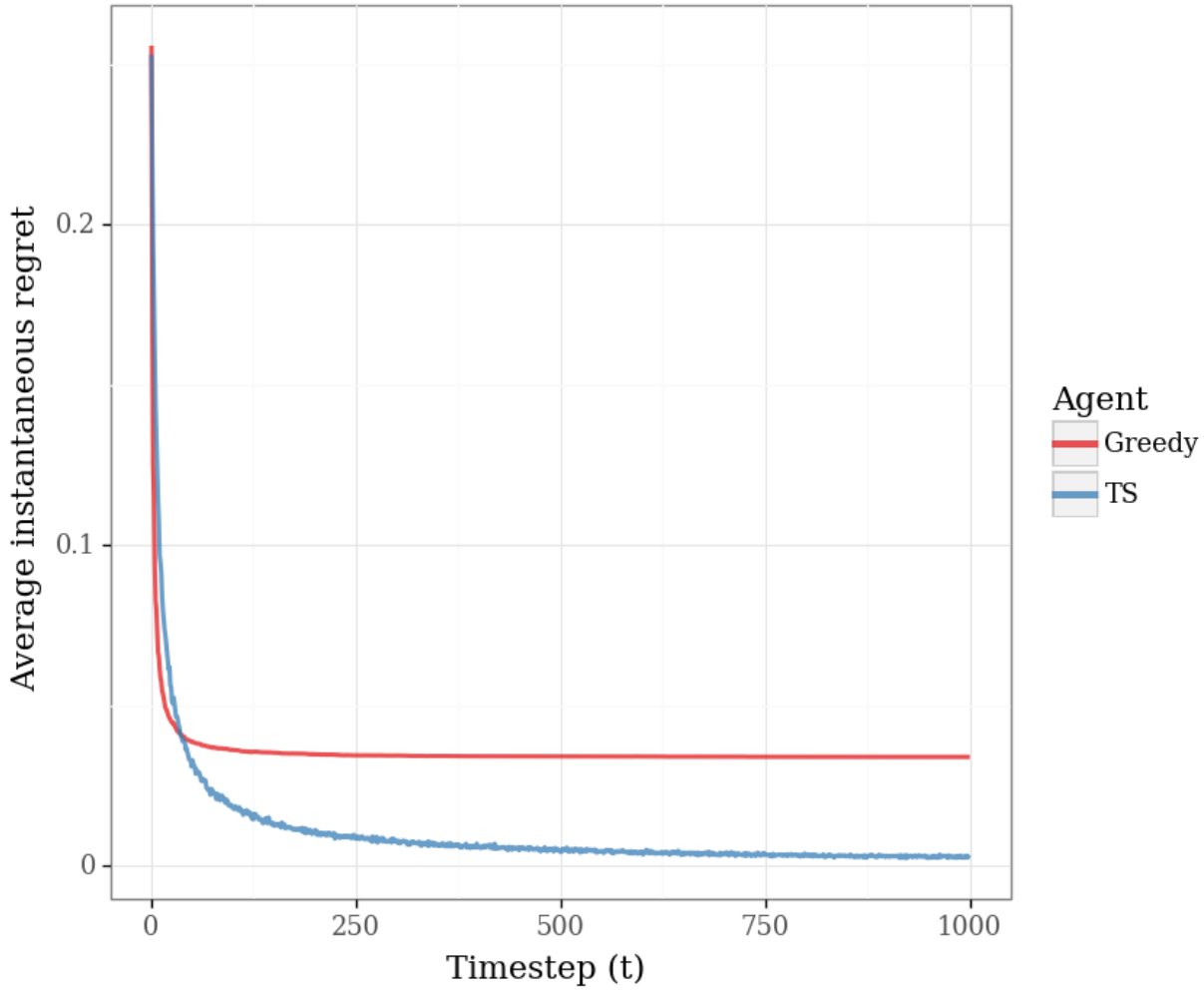
# A simulation of TS

- Fixed problem instance $\theta = (.9, .8, .7)$



# A simulation of TS

- Random instance $\theta_i \sim \text{Beta}(1,1)$



# Prior Distribution Specification

How I think about this:

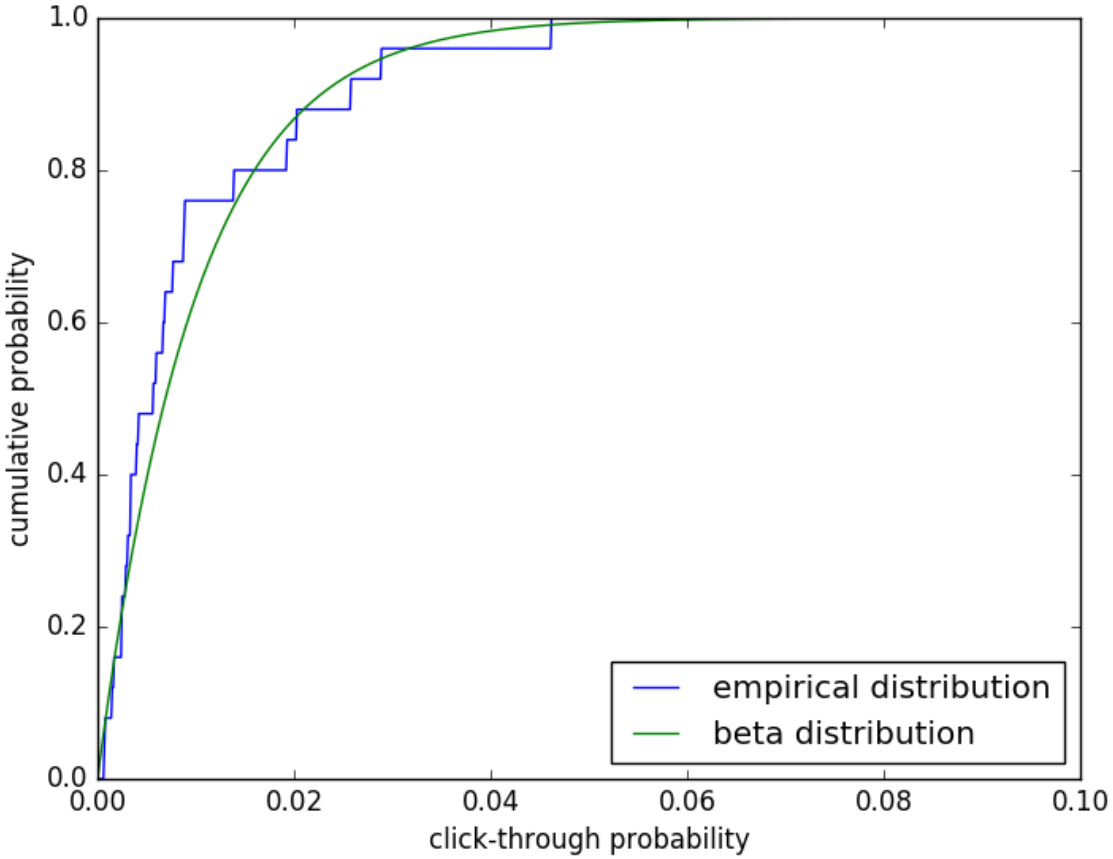
- No algorithm minimizes $\mathbb{E}[\text{Total\_regret}|\theta]$ for all possible instances $\theta$.
  - E.g. an algorithm that always plays arm 1 is optimal when $\theta_1 \geq \theta_2, \ldots, \theta_1 \geq \theta_k$ but is terrible otherwise.

- A prior directs the algorithm that certain instances are more likely than others, and to prioritize good performance on those instances.

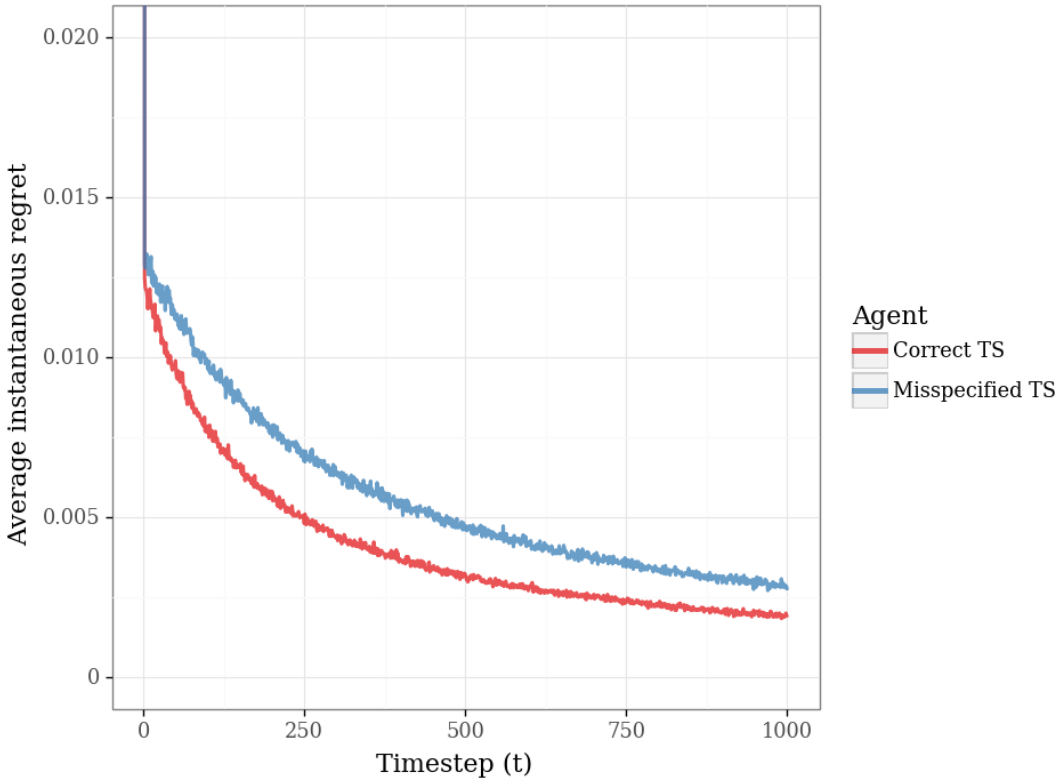
# Empirical Prior Distribution Specification

- We want to identify the best of $K$ banner ads
- Have historical data from previous products
- For each ad $k$ we can identify the past products with similar stylistic features, and use that to construct an informed prior.

# Empirical Prior Distribution Specification



# The value of a thoughtful prior

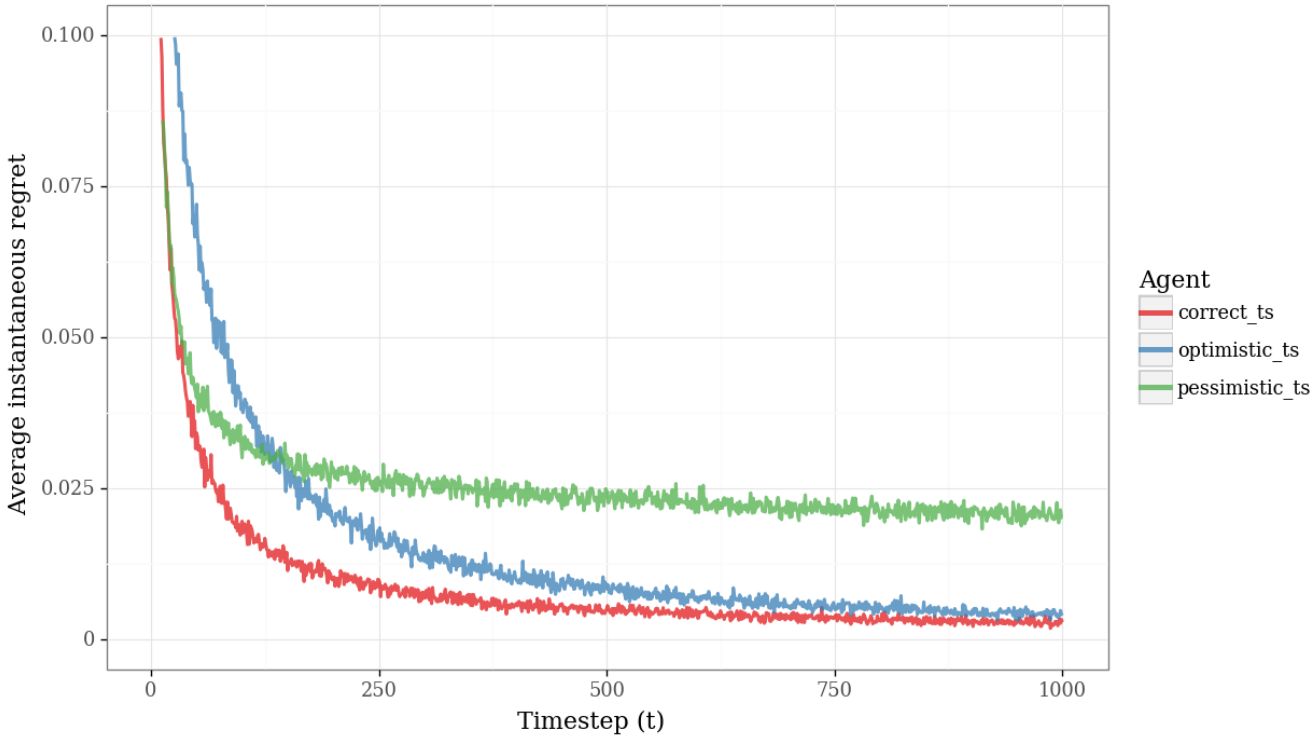



- Mispecified TS has prior $\alpha = (1,1,1)$ & $\beta = (100,100,100)$
- Correct_TS has prior $\alpha = (1,1,1)$ & $\beta = (50,100,200)$

# Prior Robustness and Optimistic Priors

- The effect of the prior distribution usually washes out once a lot of data has been collected.
- The impact in bandit problems is more subtle
- An agent who believes an action is very likely to be bad is, naturally, unlikely to try that action.
- Overly “optimistic” priors usually lead to fairly efficient learning.
- There is still limited theory establishing this.

# Prior Robustness and Optimistic Priors




- correct_ts has prior $\alpha = (1,1,1)$ & $\beta = (1,1,1)$
- optimistic_ts has prior $\alpha = (10,10,10)$ & $\beta = (1,1,1)$
- pessimistic_ts has prior $\alpha = (1,1,1)$ & $\beta = (10,10,10)$

# Recap so far

- Looked at a simple bandit problem.
- Introduces TS+UCB
- Understood their potential advantage over $\epsilon$-greedy
- Discusses priors specification.

# Classical Bandit Problems

- Small number of actions
- Informationally decoupled actions
- Observations = rewards
- No long run influence. (*no credit assignment*)

- **How to address more complicated settings?**

# Example: personalizing movie recommendations for a new user

- Action space is large and complex.
- Complex link between actions/observations.
- Substantial prior knowledge:
  - Which movies are similar?
  - Which movies are popular?
- Delayed consequences.

# General Thompson Sampling

**Summary on TS**

- Optimize a *perturbed estimate of the objective*
  - Add noise in proportion to uncertainty
- Often generates sophisticated exploration.
- A general paradigm

# General Thompson Sampling

**Summary on TS**

- Optimize a *perturbed estimate of the objective*
  - Add noise in proportion to uncertainty
- Often generates sophisticated exploration.
- A general paradigm

Misleading view in the literature: TS is "optimal," is the best algorithm empirically, and performs much better than UCB.

My view: TS is a simple way to generate fairly sophisticated exploration while still enabling rich and flexible modeling.

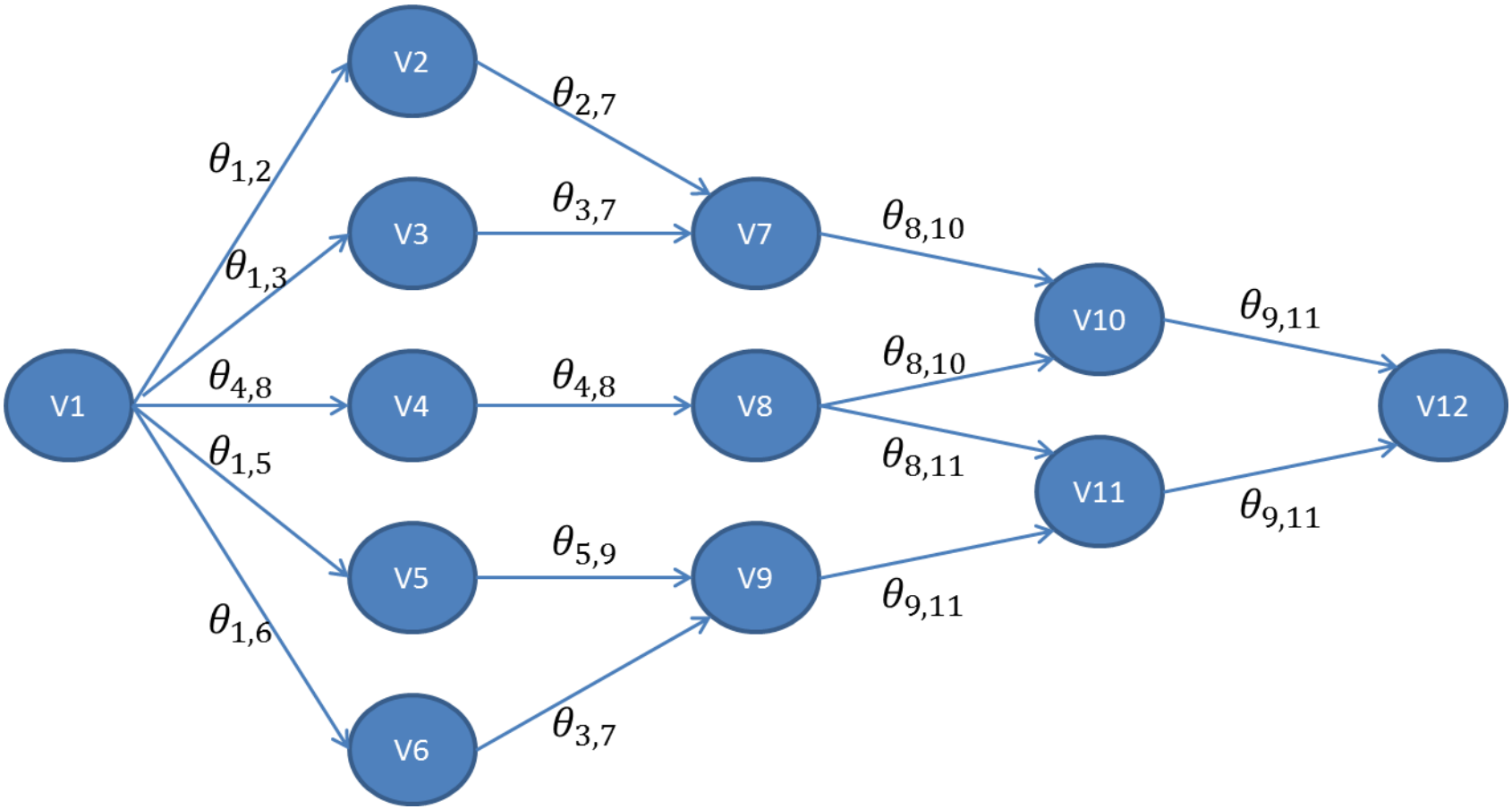
# Part I:Thompson Sampling

- Use the online shortest path problem to understand the Thompson sampling algorithm.

  1. Why is the problem challenging?
  2. How TS works in this setting.
  3. Touch on a theoretical guarantee.

- Thompson (1933), Scott (2010), Chappelle and Li (2011), Agrawal and Goyal (2012)

# Online Shortest Path Problem



# Shortest Path Problem

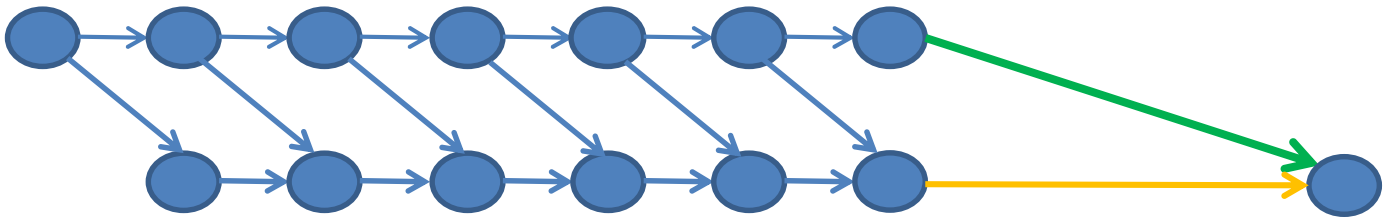
The number of paths can be exponential in the number of edges.

**Associated Challenges**

1. Computational
   - Natural algorithms optimize a surrogate objective in each time-step.
   - Optimizing this surrogate objective may be intractable.

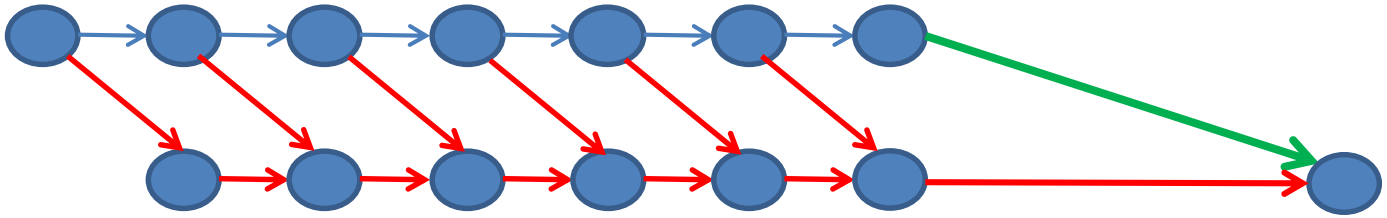
2. Statistical
   - Many natural algorithms only explore locally.
   - Time to learn may scale with the number of paths.

# Dithering (i.e. $\epsilon$ –greedy ) for Shortest Path



- Short back-roads, marked blue.
- Two long highways, marked green and orange.
- We think green *might* be much faster than orange

# Dithering (i.e. $\epsilon$ −greedy ) for Shortest Path



- Time to learn scales with the **number of paths** (exponential in number of edges)

# Bayesian Shortest Path

- Begin with a prior over mean travel times $\boldsymbol{\theta}$.
- Observe realized travel times on traversed edges.
- Track posterior beliefs.
  - (Require posterior-samples)

# Conjugate Example

## Log-Normal Distribution

- $log(\theta_e) \sim N(\mu_e, \sigma_e^2)$
- Conditioned on $\theta_e$, realized travel times along edge $e$ have mean $\theta_e$ and are lognormally distributed.
- Simple update rule for posterior parameters

# Conjugate Example

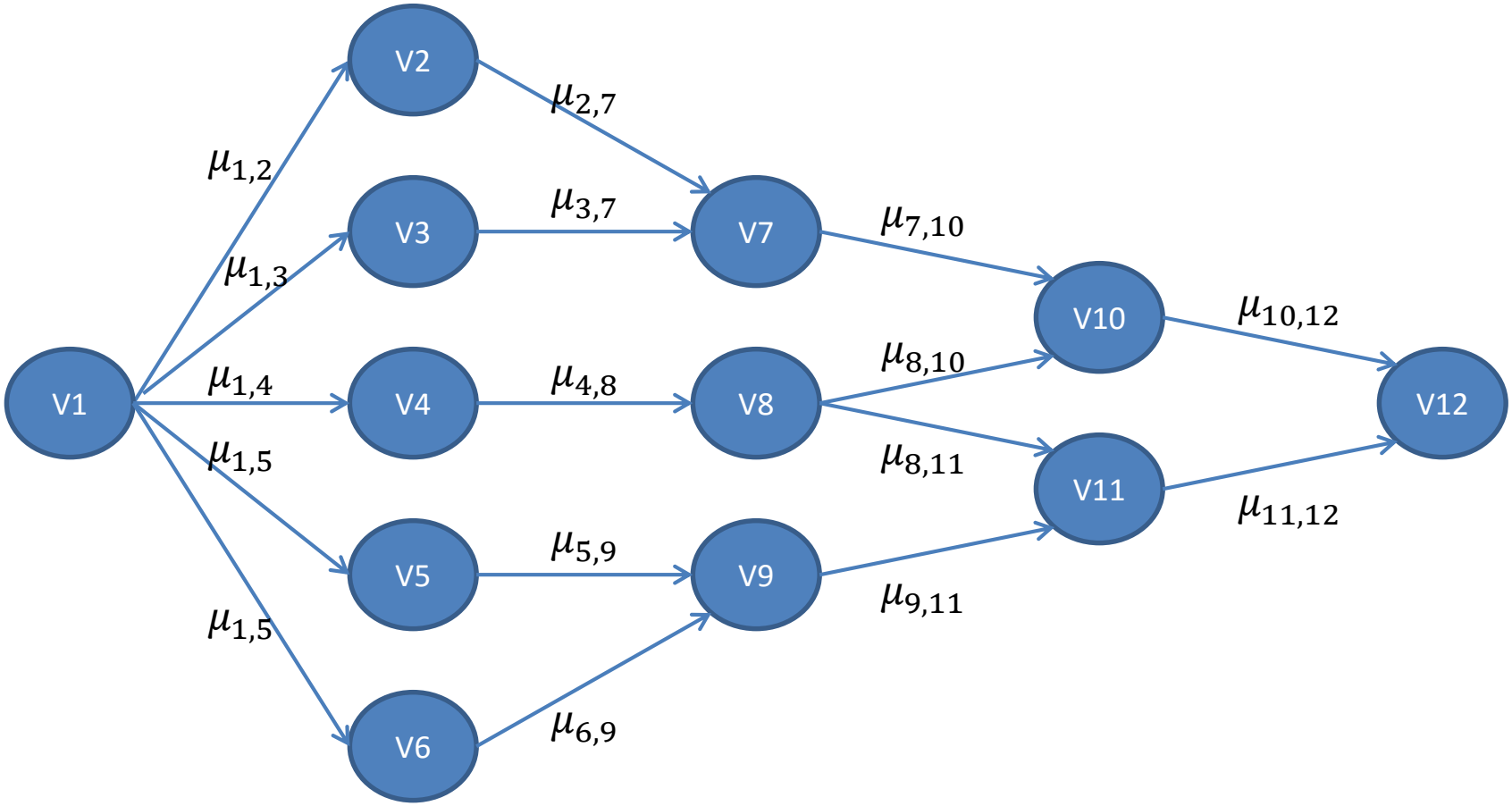
## Log-Normal Distribution

- $log(\theta_e) \sim N(\mu_e, \sigma_e^2)$
- Conditioned on $\theta_e$, realized travel times along edge $e$ have mean $\theta_e$ and are lognormally distributed.
- Simple update rule for posterior parameters

## An Informed Prior

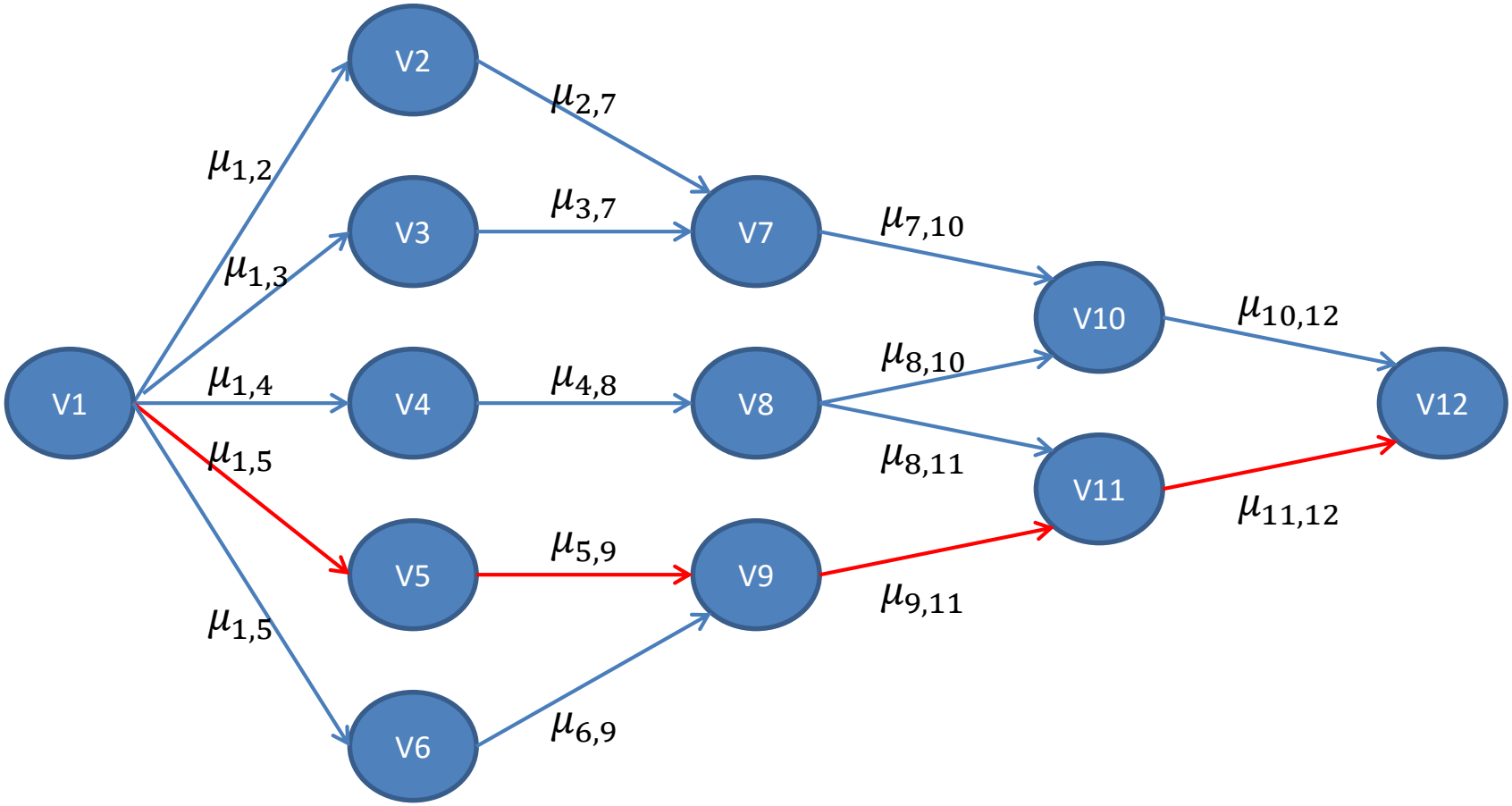
- With known travel distances for each edge, one can pick $(\mu_e, \sigma_e^2)$ so
  - $\mathbb{E}[\theta_e] = d_e$
  - $Var(\theta_e) \propto d_e^2$

# Greedy for Shortest Path



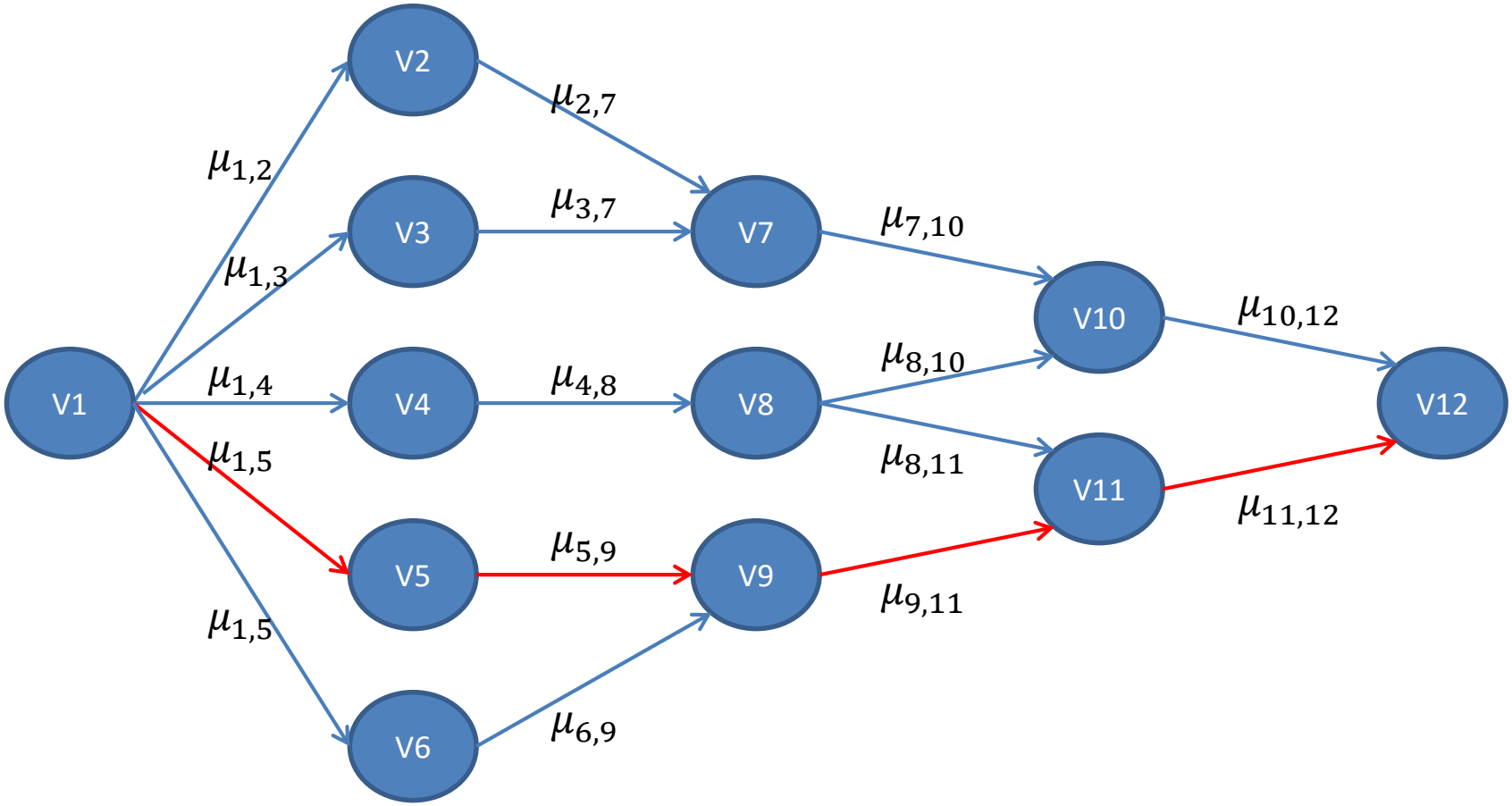

1. Set: $\boldsymbol{\mu}$ to be the posterior mean of $\boldsymbol{\theta}$

# Greedy for Shortest Path



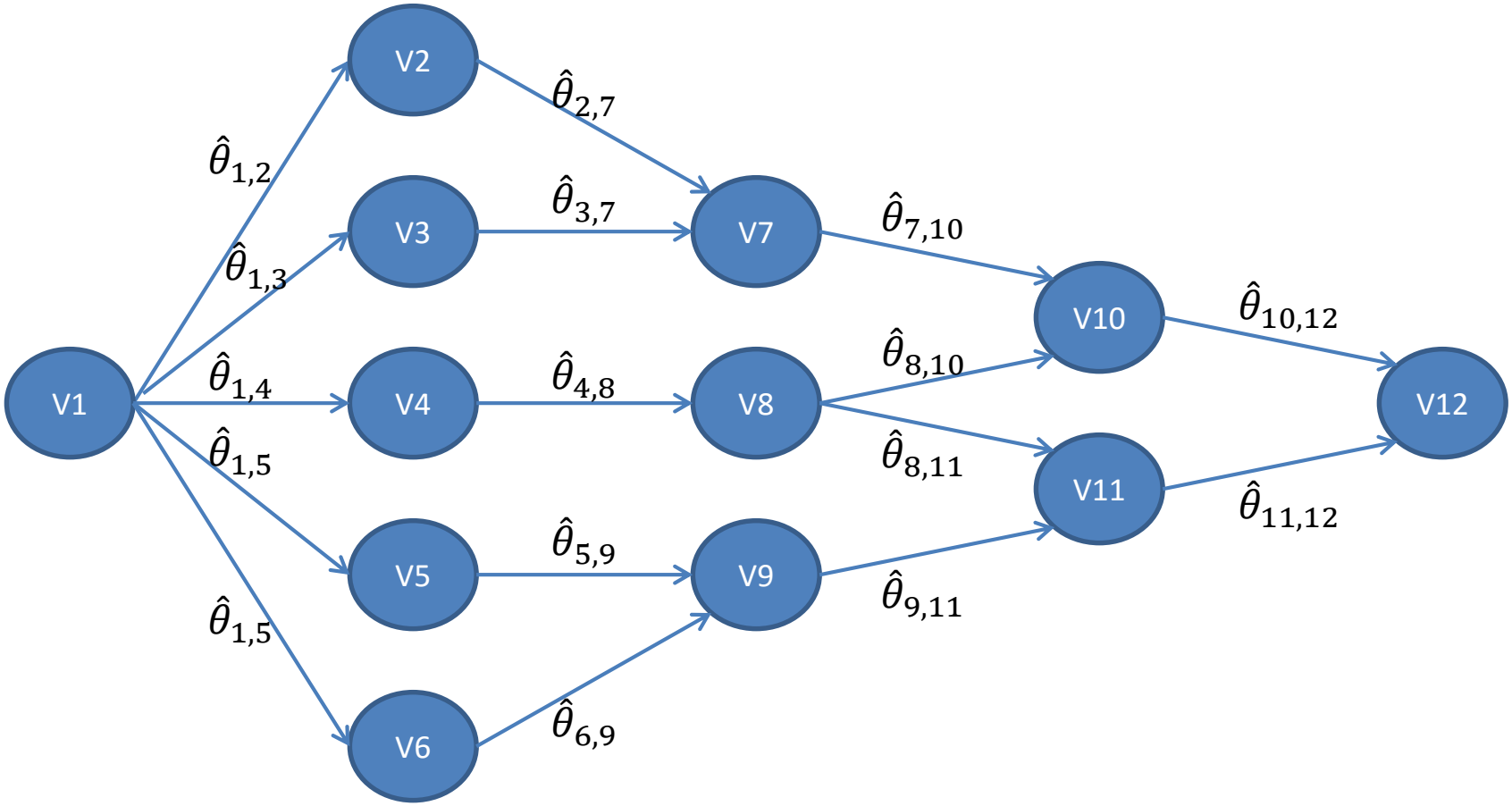

1. Set: $\boldsymbol{\mu}$ to be the posterior mean of $\boldsymbol{\theta}$
2. Follow the shortest path under $\boldsymbol{\mu}$

# Greedy for Shortest Path



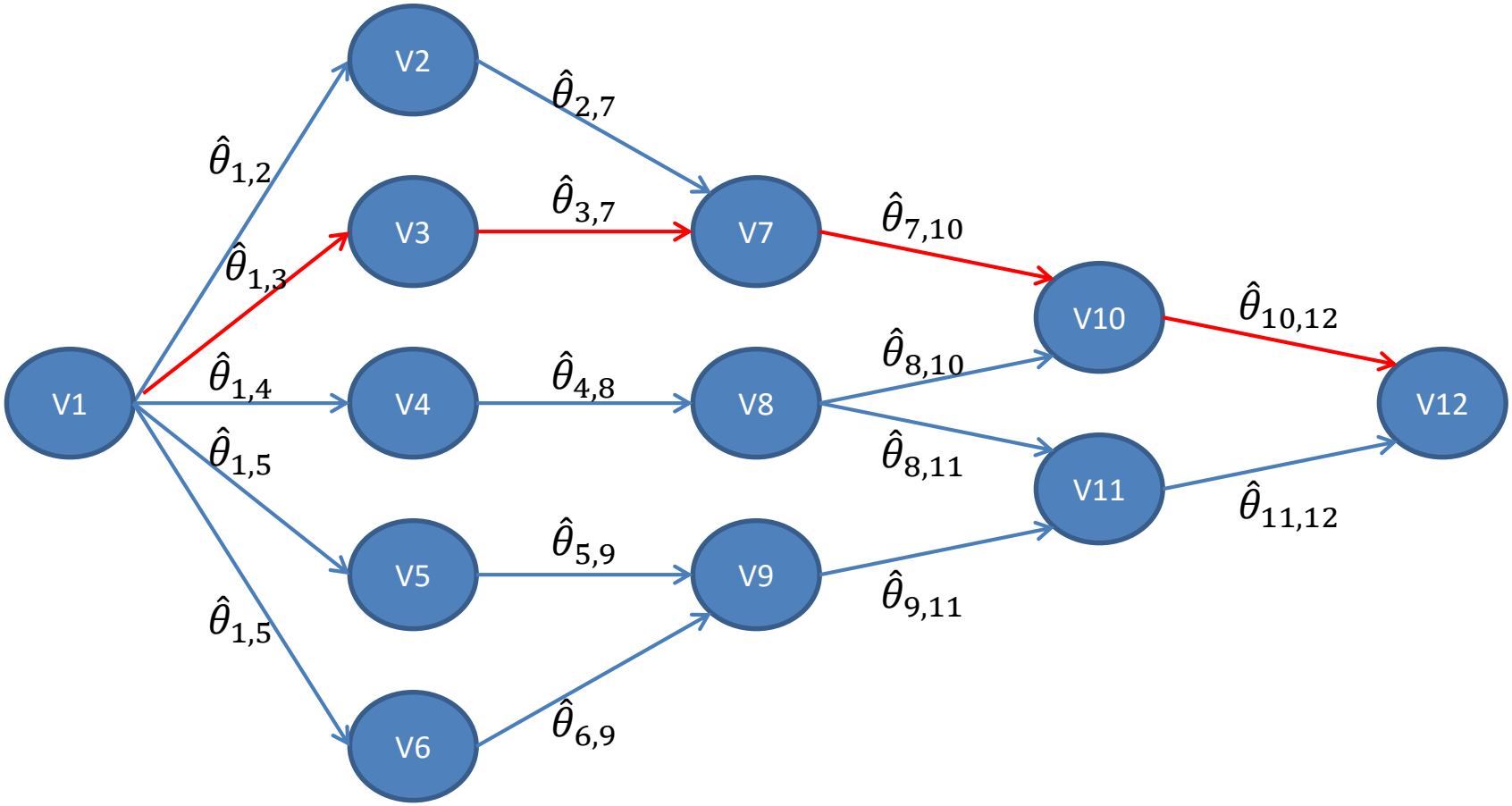

1. Set: $\boldsymbol{\mu}$ to be the posterior mean of $\boldsymbol{\theta}$
2. Follow the shortest path under $\boldsymbol{\mu}$
3. Update beliefs

# Thompson Sampling for Shortest Path



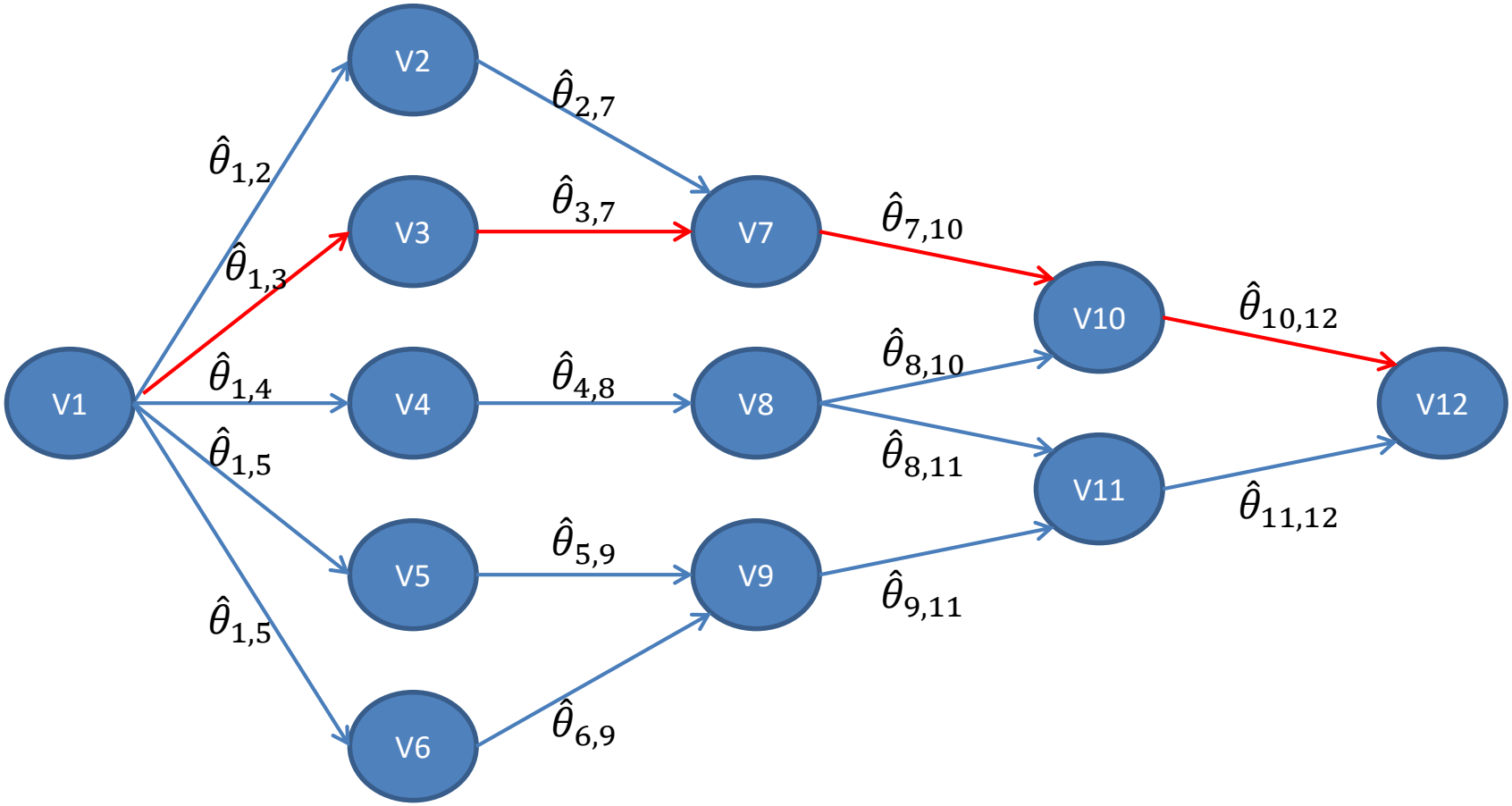

1. Sample from posterior: $\widehat{\boldsymbol{\theta}} \sim \pi_t(d\boldsymbol{\theta})$

# Thompson Sampling for Shortest Path



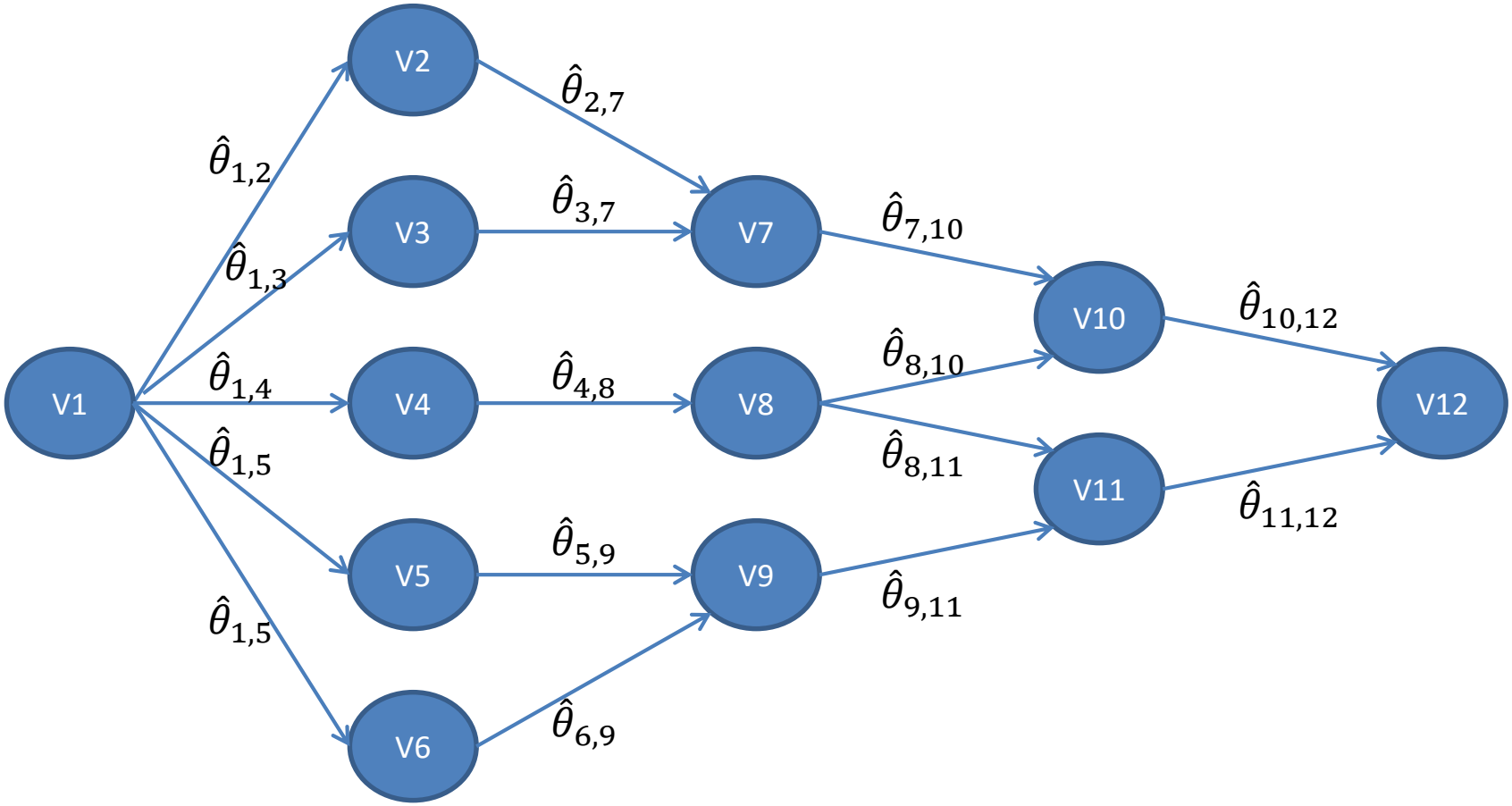

1. Sample from posterior: $\widehat{\boldsymbol{\theta}} \sim \pi_t(d\boldsymbol{\theta})$
2. Follow shortest path under sampled weights

# Thompson Sampling for Shortest Path



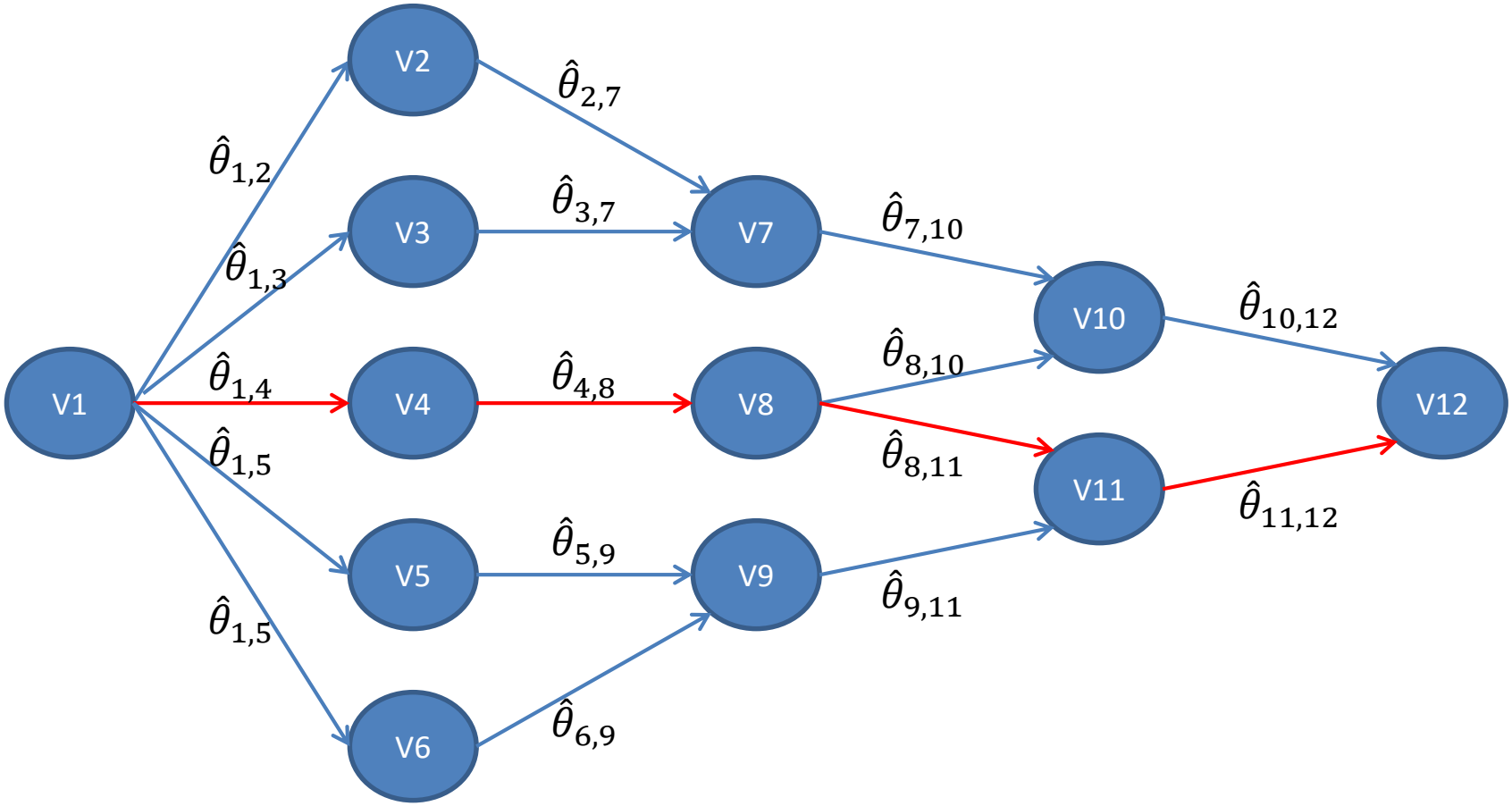

1. Sample from posterior: $\widehat{\boldsymbol{\theta}} \sim \pi_t(d\boldsymbol{\theta})$
2. Follow shortest path under sampled weights
3. Update beliefs

# Thompson Sampling for Shortest Path



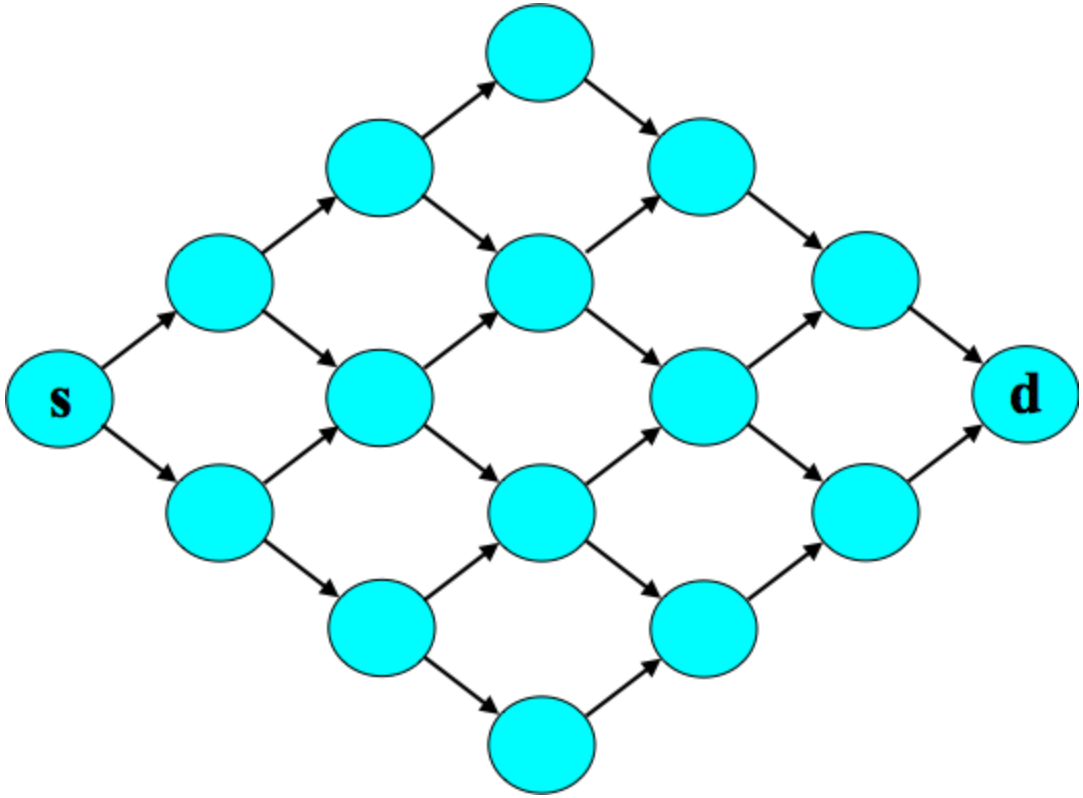

1. Sample from posterior: $\widehat{\boldsymbol{\theta}} \sim \pi_{t+1}(d\boldsymbol{\theta})$

# Thompson Sampling for Shortest Path



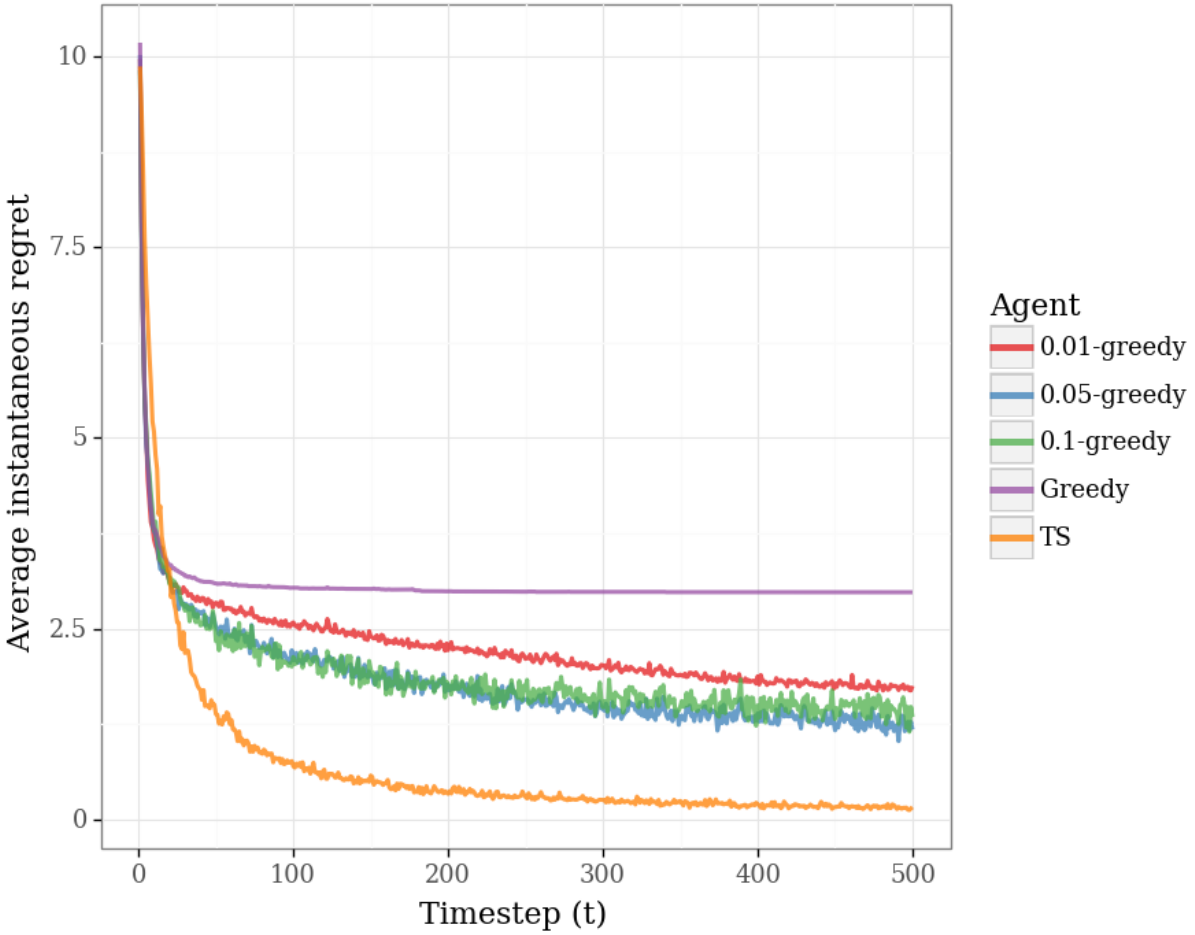

1. Sample from posterior: $\widehat{\boldsymbol{\theta}} \sim \pi_{t+1}(d\boldsymbol{\theta})$
2. Follow shortest path under sampled weights
3. Update beliefs

# Binomial Bridge




- Twenty rather than six stages
- 184,757 paths

# Shortest Path Simulation



# Why does this work?

Let $x^*(\boldsymbol{\theta}) \in \mathcal{X}$ denote the shortest path under $\boldsymbol{\theta}$

**Posterior sampling definition:**

- Sample $\widehat{\boldsymbol{\theta}}_t \sim \pi_t(d\boldsymbol{\theta})$
- Play $x^*\left(\widehat{\boldsymbol{\theta}}_t\right)$

**Probability matching definition:**

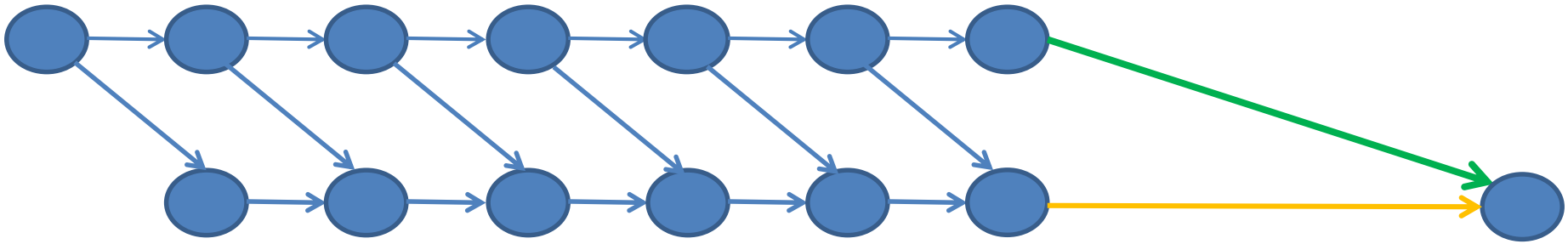
- Play $x$ with probability $\mathbb{P}_{\boldsymbol{\theta} \sim \pi_t}(x^*(\theta) = a)$

# Why does this work?

*Sample a path according to the posterior probability it's the shortest path.*

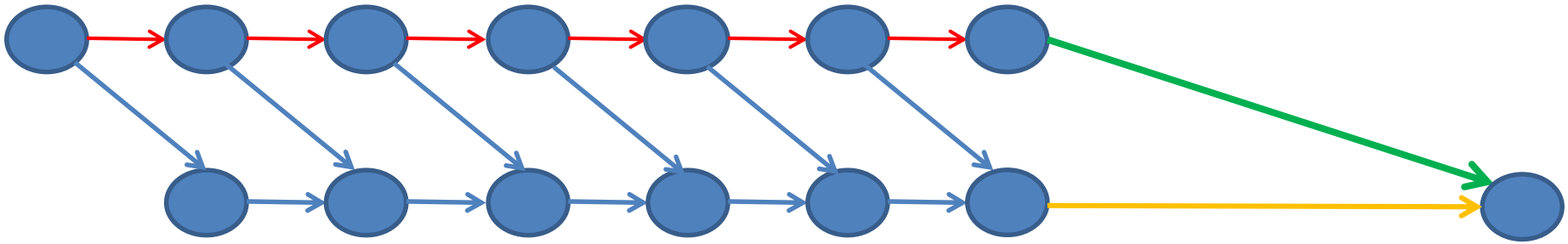
1. Continue to explore all edges that could plausibly be in the shortest path.
2. Don't waste effort exploring edges that are very unlikely to be in the shortest path.

# Thompson Sampling vs Dithering



- Short back-roads, marked blue.
- Two long highways, marked green and orange.
- We think green might be much faster than orange

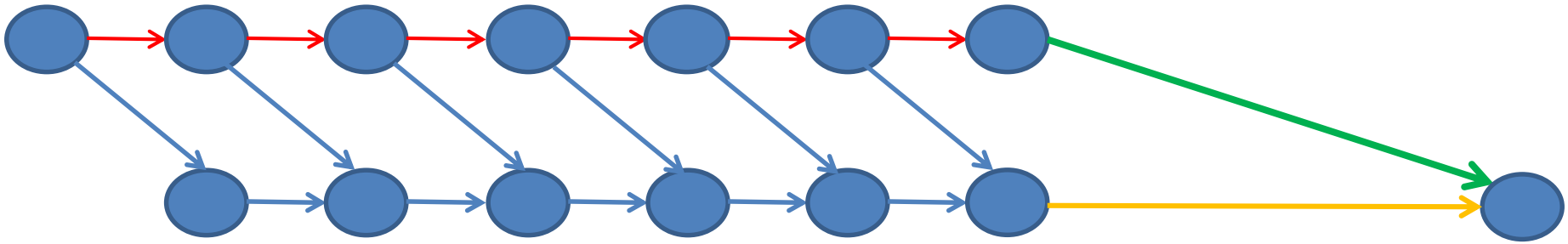
# Thompson Sampling vs Dithering



- Short back-roads, marked blue.
- Two long highways, marked green and orange.
- We think green might be much faster than orange

  **TS navigates to, and samples, the green edge**

# Thompson Sampling vs Dithering



- Short back-roads, marked blue.
- Two long highways, marked green and orange.
- We think green might be much faster than orange

  **TS navigates to, and samples, the green edge**

  **Performs “Deep exploration”**

# The practice of TS

- A richer model of edge delays
- Posterior approximations
- Non-stationary environments
- Constraints, caution, and context

# Extension: Correlated Travel Times

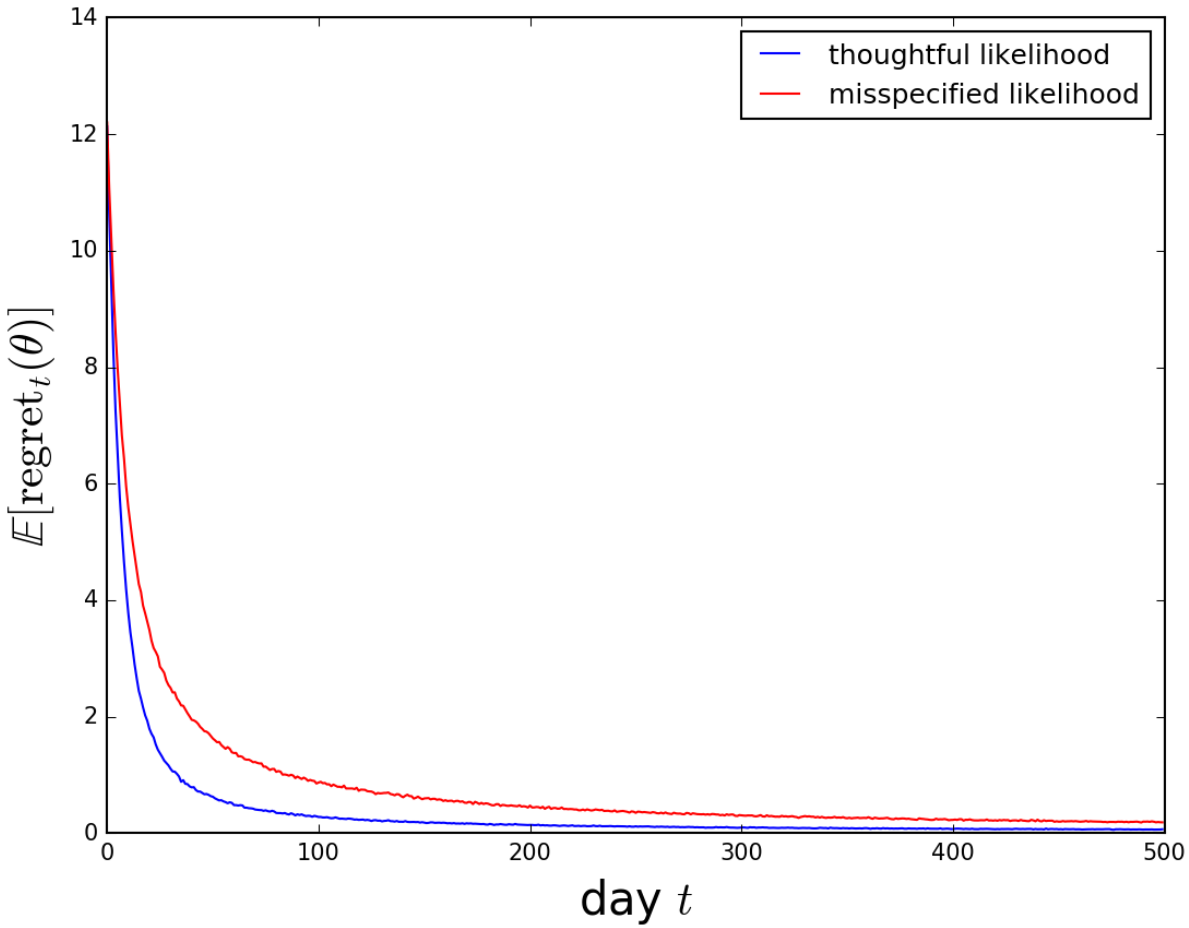
- Graph can be broken up into regions
  - For simplicity, uptown and downtown

- Delays on an edge are influenced by
  - Shocks associated with that edge
  - Shocks to the whole system
  - Shocks to the region containing the current edge

# Simulation trial

- For each edge e
  - $travel\ time = (idosyncratic\ shock) \times (region\ shock) \times (system\ shock) \times (mean\ travel\ time)$
- Shocks are lognormal with known parameters.
- Simple update rule for posterior parameters

# Benefits of modeling correlation



# A Path Recommendation Problem: (A non-conjugate example)

- Route recommendation service suggests paths
- Users give binary ratings
- Probabilities reflect quality of path

$$y_t|\theta \sim \begin{cases} 1 & \text{with probability } \frac{1}{1+\exp\left(\sum_{e\in x_t}\theta_e - M\right)} \\ 0 & \text{otherwise.} \end{cases}$$

# A Path Recommendation Problem

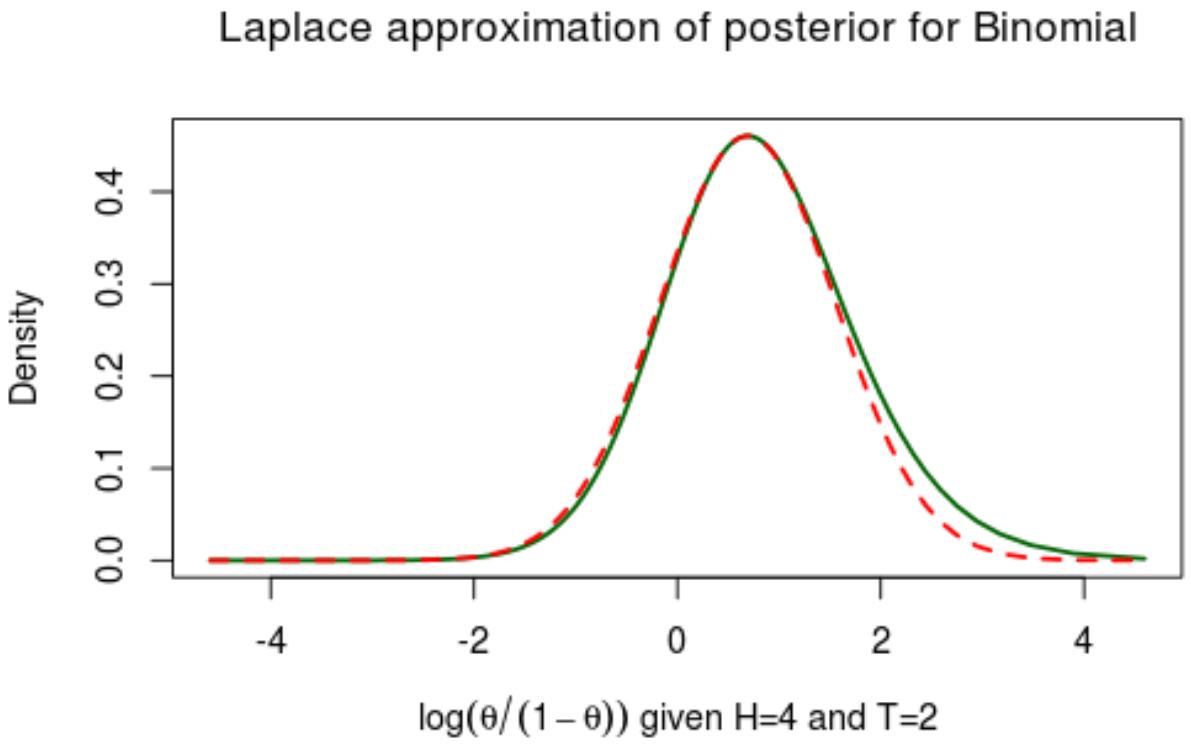
- Computing MAP estimates is straightforward
- No closed form posterior.
- How do we apply Thompson sampling?

# Posterior Approximations / Sampling

1. Gibbs Sampling / Metropolis Hastings
2. Laplace Approximation
3. Langevin Monte Carlo
4. Bootstrap Sampling
5. Ensemble methods

# Laplace Approximation

*Approximate the posterior by a Gaussian centered at its mode*



# Laplace Approximation

*Approximate the posterior by a Gaussian centered at its mode*

The log posterior density

$$g(\theta) \propto \log \pi_0(\theta) + \sum_1^n \log(p(y_i|\theta))$$

is concave.

- Taylor expansion around mode $\hat{\theta}$:

$$g(\theta) \propto g(\hat{\theta}) - (\hat{\theta} - \theta)^T \hat{C} (\hat{\theta} - \theta) + o(\|\theta - \hat{\theta}\|^2)$$

where $\hat{C}$ is the Hessian of g at $\hat{\theta}$.

- Leads to approximate posterior $N(\hat{\theta}, \hat{C})$

# Langevin MCMC

*Construct a Markov chain by running gradient ascent + noise*

The log posterior density

$$g(\theta) \propto \log \pi_0(\theta) + \sum_1^n \log(p(y_i|\theta))$$

is concave.

Under some technical conditions, $\pi_n(\theta) \propto e^{g(\theta)}$ is the unique stationary distribution of the Langevin diffusion

$$d\theta_t = \nabla g(\theta_t) + \sqrt{2} dB_t$$

where $B_t$ is standard Brownian motion.

# Langevin MCMC

*Construct a Markov chain by running gradient ascent + noise*

Simulate a Euler discretaton of the Langevin Diffusion

$$\theta_{t+1} = \theta_t + \epsilon \nabla g(\theta_t) + \sqrt{2\epsilon} dB_t$$

There is theory showing this mixes rapidly

- (e.g. when $\nabla^2 g(\theta) \preccurlyeq -LI$)

I have found it is helpful to initialize at the MAP estimate, and 'precondition' by the inverse Hessian at the MAP estimate.

# Bootstrap Sampling

*Subsample training data with replacement. Train as usual.*

Standard bootstrap

1. $H_n = \{(x_1, y_1) \ldots (x_n, y_n)\}$
2. Sample hypothetical history with replacement
   - $\widehat{H}_n = \{(\hat{x}_1, \hat{y}_1), \ldots, (\hat{x}_n, \hat{y}_n)\}$
3. Construct MAP estimate on $\widehat{H}_n$.

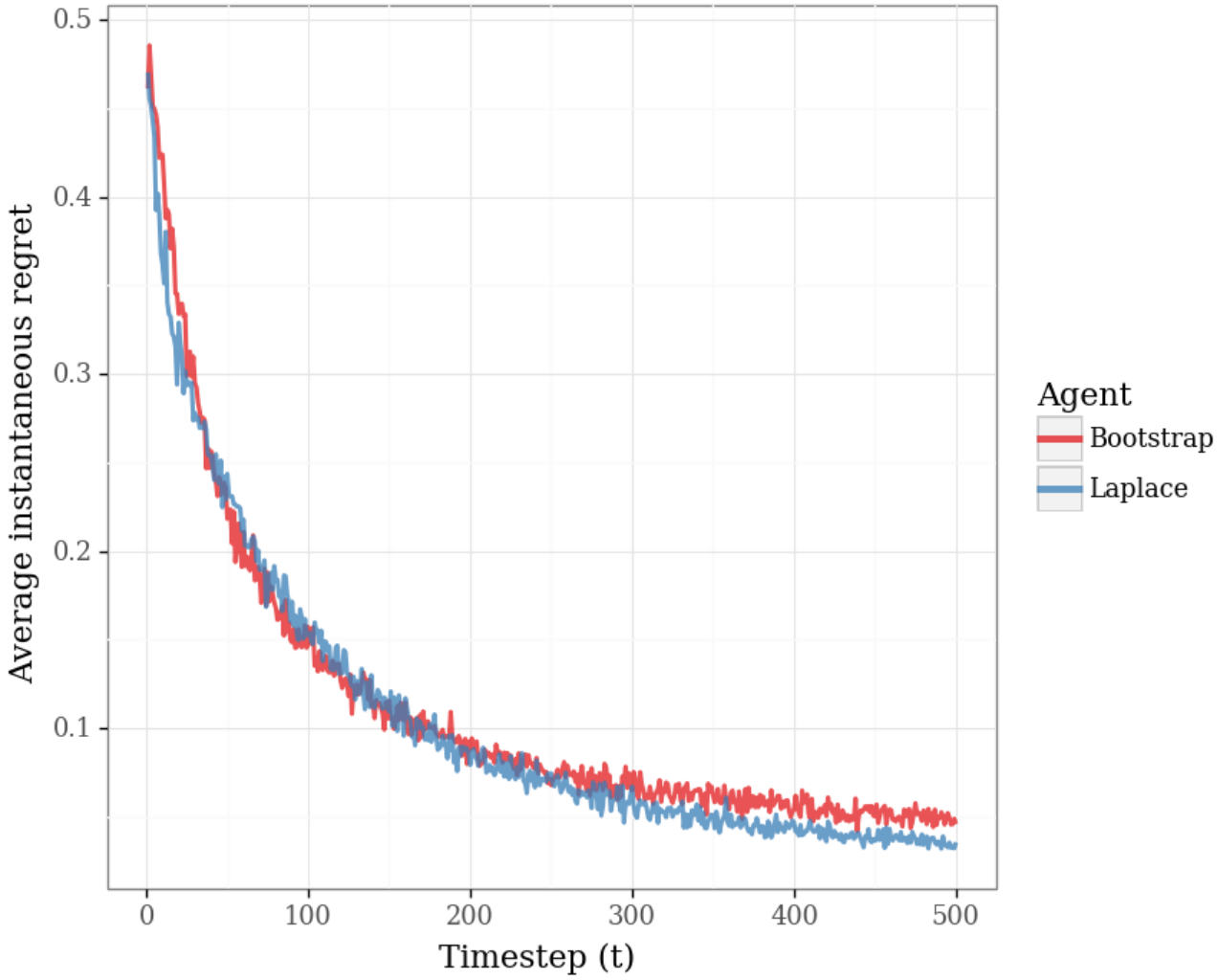
# Bootstrap Sampling

*Subsample training data with replacement. Train as usual.*

The tutorial covers a nonstandard bootstrap.

- Injects some additional uncertainty by sampling from prior.

# Simulating Path Recommendation



# Non-stationarity

**First observation:**

- If the environment changes very rapidly, it is not worth exploring.

**Second observation:**

- Slowly changing environments can be addressed by running TS while gradually "forgetting" the past
  1. w/ a sliding window.
  2. w/ geometric down-weighting of the past.
  3. w/ more sophisticated Bayesian filtering techniques.

# Constraints, Caution, and Context

Beyond the shortest path problem, can be written the form

1. Sample $\widehat{\boldsymbol{\theta}}_t \sim \pi_t(d\boldsymbol{\theta})$
2. Play $\max_{x \in \mathcal{X}_t} \mathbb{E}[r_t | x_t = x, \hat{\theta}_t]$

Here $r_t$ denotes the reward at time $t$, and $x_t$ denotes the action.

# Constraints, Caution, and Context

**Observation:**

It is easy to apply TS in a problem with arbitrary changing action sets: $\mathcal{X}_1, \mathcal{X}_2, \mathcal{X}_3$ …

1. Observe $\mathcal{X}_t$
2. Sample $\widehat{\boldsymbol{\theta}}_t \sim \pi_t(d\boldsymbol{\theta})$
3. Play $\max_{x \in \mathcal{X}_t} \mathbb{E}[r_t | x_t = x, \hat{\theta}_t]$

# Constraints, Caution, and Context

Constrained action sets provide substantial modeling flexibility.

1. Routes are inherently constrained by announced road closures.

2. We enforce impose constraints to provide caution against very poor performance.
   - $\mathcal{X}_t = \{x | \mathbb{E}[r_t | x_t = x, \mathcal{F}_{t-1}] \geq \underline{r}\}$
   - The set of actions with posterior mean above $\underline{r}$

3. We observe contextual information before acting.
   - e.g. a weather report

# Constraints, Caution, and Context

3. We observe contextual information before acting.
   - e.g. a weather report

- Let $z_t$ be the weather report at time $t$.
- Write $x_t = (chosen\ path, weather\ report)$
- $\mathcal{X}_t$ is the set of paths with weather report $z_t$
- $\max_{x \in \mathcal{X}_t} \mathbb{E}[r_t | x_t = x, \theta]$ gives the best path givne the weather report $z_t$ and parameter $\theta$.

# Theoretical Guarantees?

- I've emphasized the ability of TS to accommodate general modeling and rich forms of prior knowledge.
- I've argued  prior knowledge improves performance.

- Can we say something formal?

# Example of a Theoretical Guarantee

- Normalize so travel times are in $[0,1]$.
- Let $x^*(\theta) \in \mathcal{X}$ denote the shortest path under $\theta$

Russo and Van Roy, *A Information Theoretic Analysis of Thompson Sampling*, JMLR 2016

# Example of a Theoretical Guarantee

- Normalize so travel times are in $[0,1]$.
- Let $x^*(\theta) \in \mathcal{A}$ denote the shortest path under $\theta$

$$\mathbb{E}[\text{Regret}(T)] \leq \sqrt{\frac{1}{2}\,\text{Entropy}\big(x^*(\theta)\big)(\#edges)T}$$

- Note that $\text{Entropy}\big(x^*(\theta)\big) \leq \log|\mathcal{X}|$.

Russo and Van Roy, *A Information Theoretic Analysis of Thompson Sampling*, JMLR 2016

# Information-Theoretic Analysis

**Proof idea:**

- Posterior-entropy of $x^*$ quantifies uncertainty
- Show that in ***every*** period

$$\mathbb{E}[\text{regret}]^2 \leq .5\ (\#edges)\mathbb{E}[\text{entropy reduction}]$$

Russo and Van Roy, *A Information Theoretic Analysis of Thompson Sampling*, JMLR 2016

We'll cover a different analysis in class.

# Recap so far

- Understood TS in the context of the shortest path problem.
- Discussed a range of practical issues
  - Correlated feedback
  - Approximate posterior sampling
  - Prior specification.
  - Non-stationarity
  - Constraints and context
- Made note of one theoretical guarantee.

# Thompson Sampling

**Summary on TS**

- Optimize a *perturbed estimate of the objective*
    - Add noise in proportion to uncertainty
- Often generates sophisticated exploration.
- A general paradigm