Efficient Rare-Event Simulation for Many-Server Loss Systems

Jose Blanchet
IEOR Department
Columbia University

Many-Server Loss System

- Loss system: GI/GI/s/0 no waiting room → customers are lost if all servers are busy…

- Assume s = # of servers large

- Focus of this talk:
  - Computing loss probabilities / overflow events
  - Conditional distribution at the time of a loss
  - Also discuss systems with time varying / Markov modulated arrivals

Many-Server Loss System: The Model

Many-Server Loss System

Application 1: Long Distance Lines

- A local company sets up long-distance call lines
- "Customers" are the employees (can be over 5000 in big companies)
- "Service times" are the call holding times

How many call lines should be set up to guarantee a loss probability of less than, say 0.1%?

Application 2: Telecommunication Switches

- Digital switches provide connections among phone calls, internet etc.
- Switch holds a buffer capacity; packets beyond the capacity are rejected
- What is the value of buffer capacity to achieve a loss probability typically in the order of 10^{-9}?

Application 3: Insurance Portfolio

- A life insurance company sells insurance contracts to policyholders
- Policyholders pay regular (or lump-sum) premium to the company; in return, the company pays benefit to policyholders in the contingent event (e.g. death)
- "Customers" are the policyholders
- "Service times" are the times to contingent event (or the tenor of contract, if shorter)
- Large insurance companies have millions of policyholders
- The cash flow of insurance company is a functional of the statuses of customers in the system:

$$\begin{aligned} net\ cash\ position &= net\ discounted\ premium\ received \\ &+ net\ discounted\ benefit\ paid \end{aligned}$$

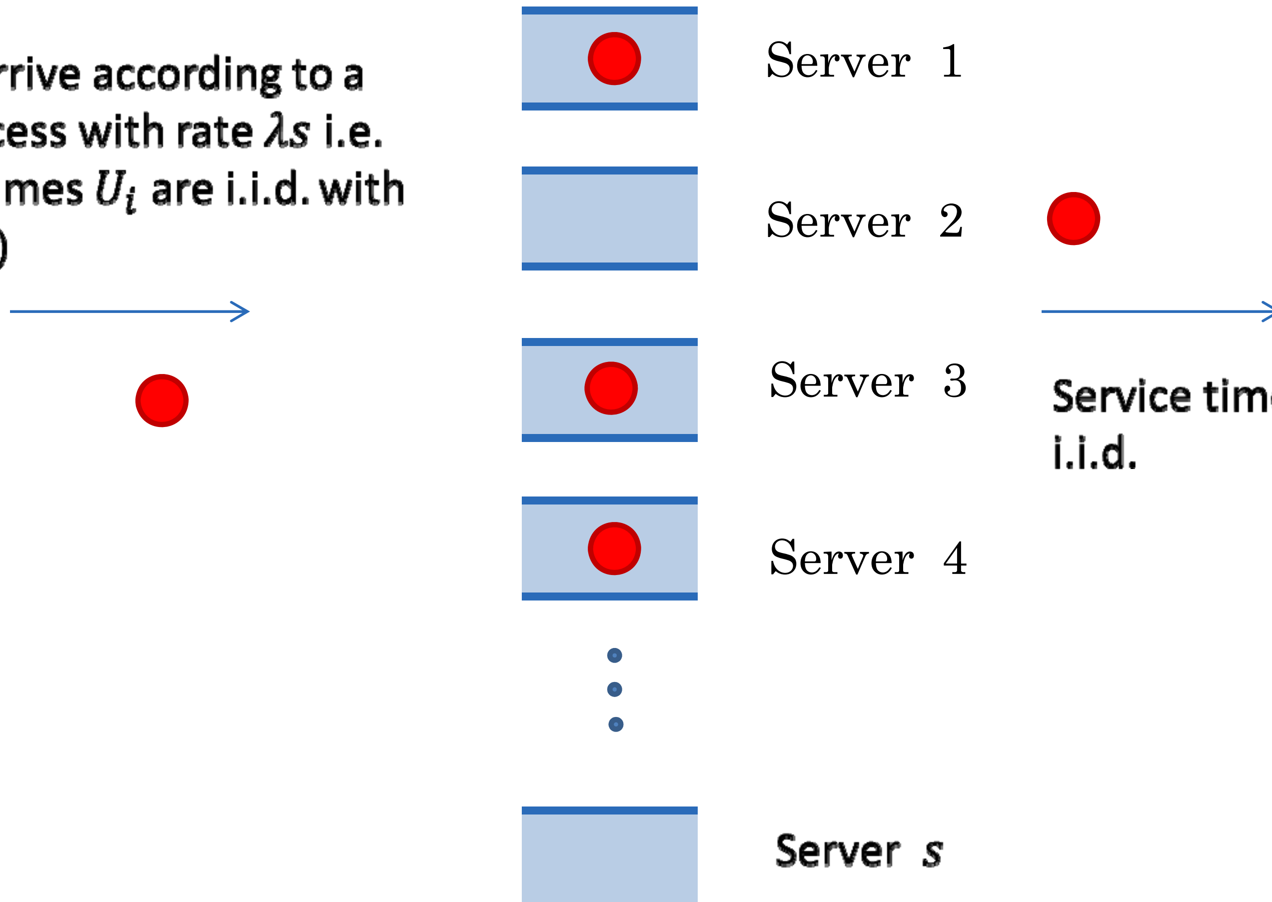
- What is the probability that the insurance company suffers from insolvency?

Important Features of Many-Server Loss System

- Many servers! (order of 10^2 to 10^3 depending on context)
- Customers arrive frequently i.e. heavy traffic
- Stable system
- Loss event is rare (order of 10^{-3} to 10^{-9})
- Other features: time-varying, limited waiting room capacity etc.

The Basic Model

Customers arrive according to a renewal process with rate λs i.e. interarrival times U_i are i.i.d. with mean $1/(\lambda s)$



Service times V_i are i.i.d.

The Basic Model

Notes:

- State-space of the process (if insisting on being Markov) at a time t is high-dimensional (measure-valued). It consists of:
  - Number of customers
  - Residual service time for each present customer
  - Age of the process since last arrival
- One convenient way of encoding the state:

$$Y(t) = \big(Q(t,y), B(t)\big) \in \mathcal{D} \times \mathbb{R}_+$$

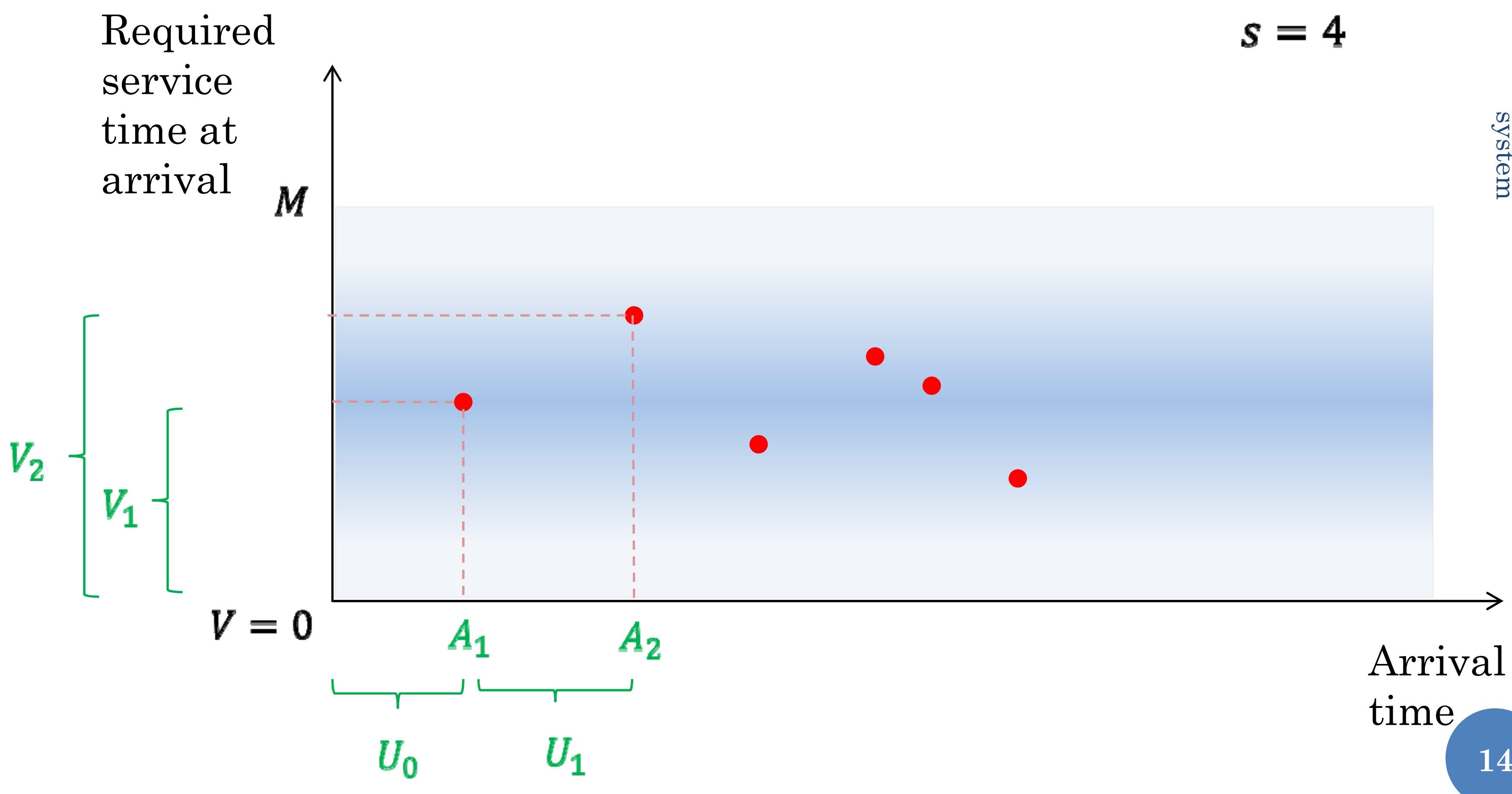
$Q(t,y)$: number of customers at time t who have residual service times larger than y
$B(t)$: age of process since last arrival

- Traffic intensity $\rho := \lambda EV < 1 \Rightarrow$ stable system
- Technical assumptions:
  - Interarrival times U_i possess exponential moments i.e. $Ee^{\theta U_i} < \infty$ for some θ in a neighborhood of 0
  - Service times V_i have bounded support

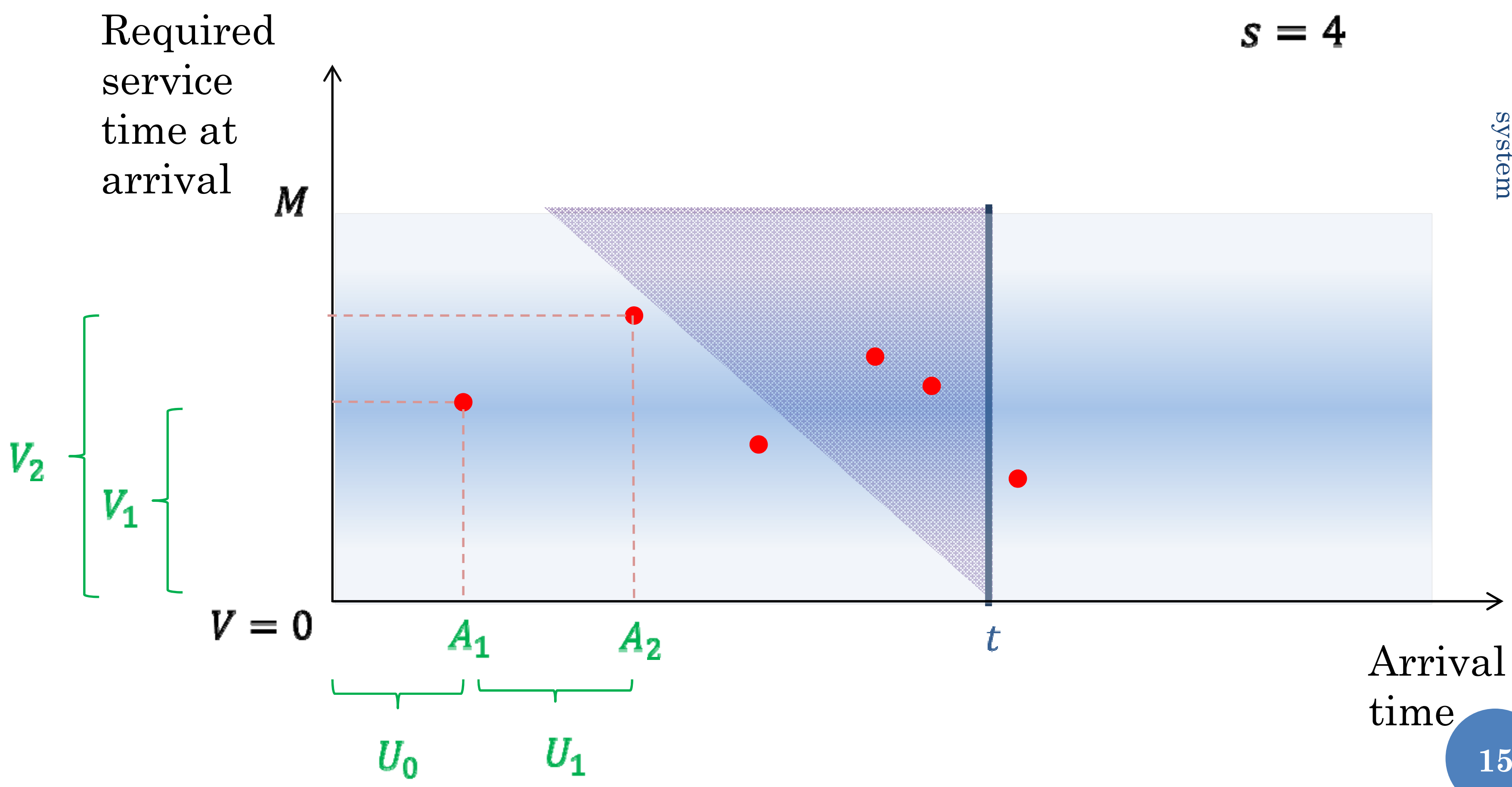
Main Goal of the Talk

- Provide an "optimal" importance sampling algorithm to estimate the steady-state loss probability
- Main Motivations:
  - Analytical solution not available except Poisson arrival
  - Typical order of magnitude $\approx 10^{-3}$ to 10^{-9} $\Rightarrow$ crude Monte Carlo is inefficient, if not infeasible
- More motivations:
  - Since our simulation is pathwise, other quantities of interest can be simulated e.g. conditional expectation of functional of the statuses of customers before loss happens
  - The algorithm can be generalized to a range of more complicated models

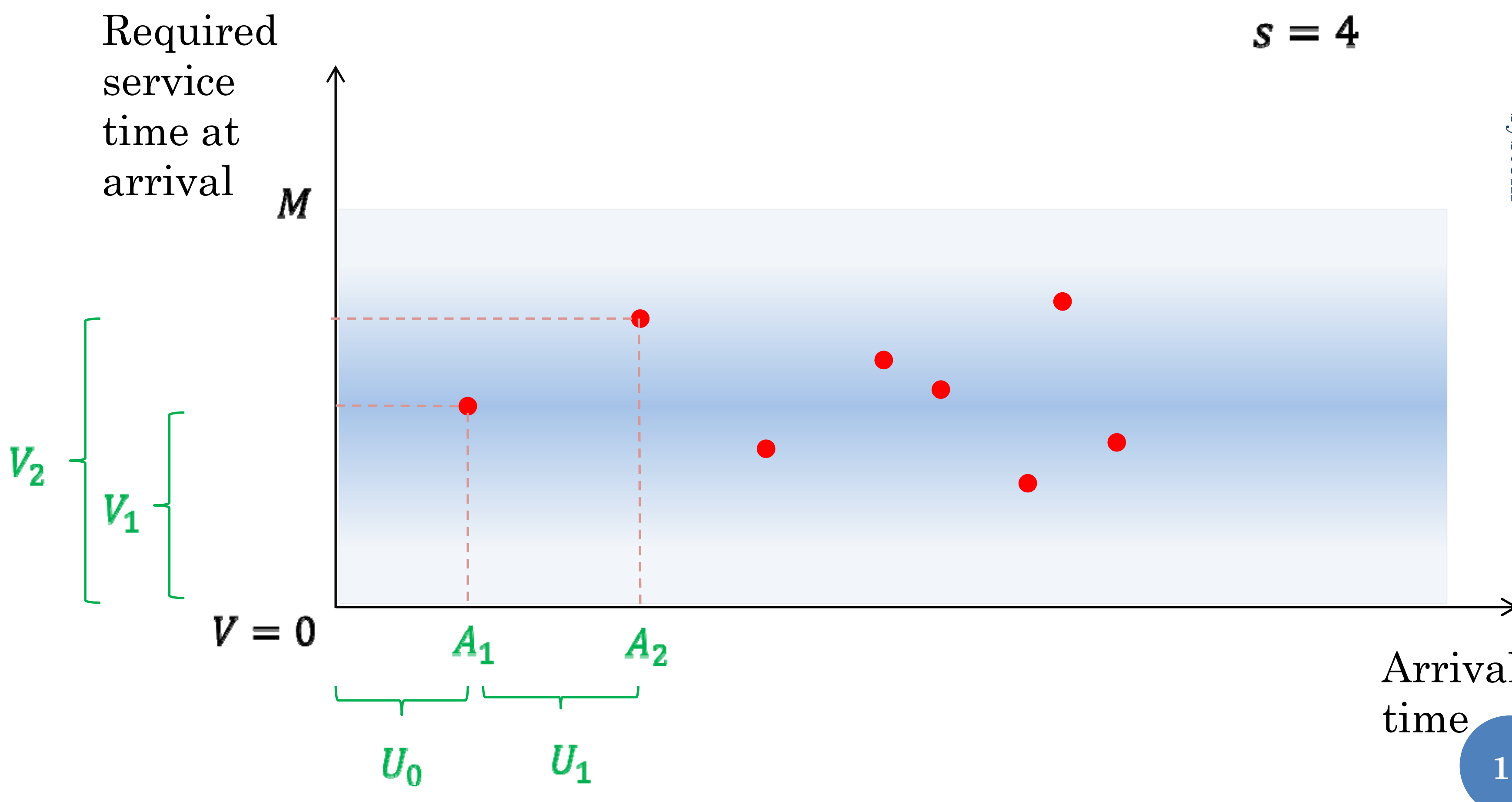
Crude Monte Carlo Scheme



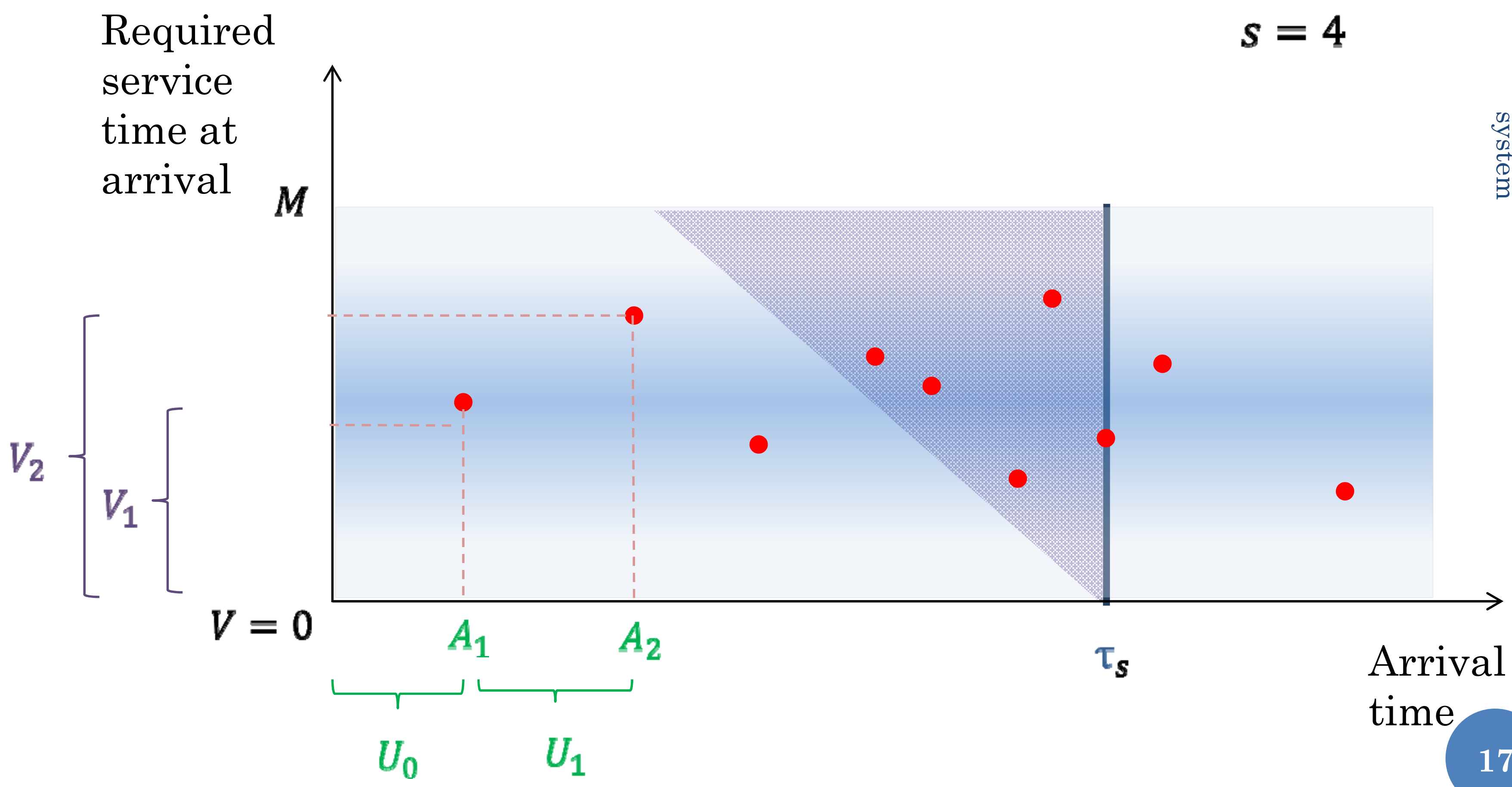
Crude Monte Carlo Scheme



CRUDE MONTE CARLO SCHEME



Crude Monte Carlo Scheme



Crude Monte Carlo Scheme

- Run the process for a long time

- $$\hat{P}(loss) = \frac{\#\ loss}{\lambda s \times total\ time\ units\ simulated}$$

A Numerical Example

Parameters/Assumptions:

$s = 100, \lambda = 1, EV_i = 0.5$
Poisson arrival with rate $\lambda s = 100$
Service time $V_i \sim U[0,1]$

Loss probability calculated from Erlang's loss formula	1.630319×10^{-10}
Running time to simulate 1000 time units using crude Monte Carlo	$5.16\ seconds$
Approximate number of customers that can be simulated in this time	$1000 \times 100 = 10^5$
Approximate time to simulate one loss event	$\frac{1/(1.63031 \times 10^{-10})}{10^5} \times 5.16 = 3.66\ days$
Approximate time to simulate 100 loss events	$366\ days$

Our Algorithm…

Organization of the Talk

1. Introduce notions in rare-event simulation
2. Explain in detail our importance sampling scheme
3. Algorithmic efficiency
4. Generalizations e.g. renewal arrivals, Markov-modulation, time-varying system

LITERATURE REVIEW

- Central Limit Theorems / diffusion approximation: Iglehart (1965), Halfin and Whitt (1981), Reed (2007)…
- Rare event analysis / large deviations:
  - most papers are on queues with single server / several servers, e.g. Asmussen (1982), Anantharam (1988), Sadowsky (1991, 1993), Frater et al. (1989, 1990, 1991, 1994), Glasserman and Kou (1995), Dupuis et al. (2007), Lehtonen and Nyrhinen (1992), Chang et al. (1993, 1994)
  - Many-server queues under heavy traffic: Glynn (1995), Szechtman and Glynn (2002), Ridder (2009)

Fundamental Challenge of Rare-Event Simulation

- Suppose one wants to estimate $P(A_n) \searrow 0$ as $n \nearrow \infty$
- Crude Monte Carlo estimator i.e.

$$\frac{1}{N}\sum_{l=1}^{N} \mathbb{1}(A_{l,n})$$

gives unbiased estimate with variance

$$\frac{1}{N^2} P(A_n)(1 - P(A_n))$$

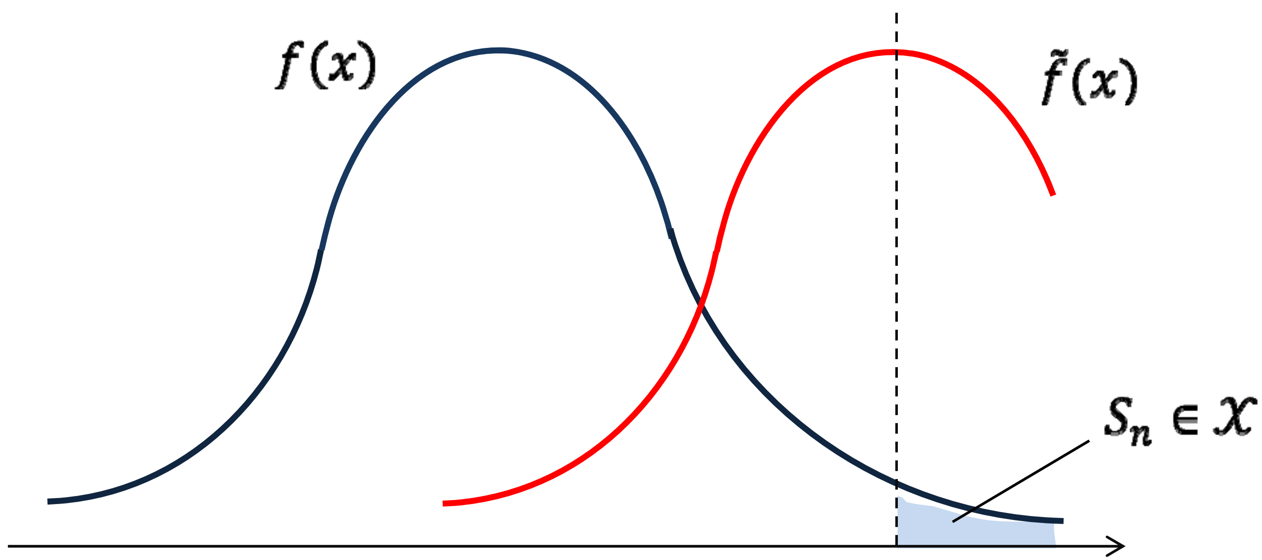
- Relative error (coefficient of variation) defined by the ratio of standard deviation to mean gives

$$\sqrt{\frac{1 - P(A_n)}{NP(A_n)}}$$

- $N \sim 1/P(A_n)$ number of samples is required to retain reasonable level of relative error
- If $P(A_n)$ is exponentially small in n, number of samples required also blows up exponentially in n

IMPORTANCE SAMPLING

- For illustration let $A_n = \{S_n \in \mathcal{X}\}$ where S_n has density $f(\cdot)$
- Instead of sampling from density $f(\cdot)$, we sample from $\tilde{f}(\cdot)$
- Likelihood ratio $L(S_n) = \frac{f(X_n)}{}$



Importance Sampling

Definition 1: An estimator is called strongly efficient if its relative error is bounded in n i.e.

$$\frac{\tilde{E}\left(L(S_n)\mathbb{1}(S_n \in \mathcal{X})\right)^2}{P(S_n \in \mathcal{X})^2} < C$$

for all n.

Definition 2: An estimator is called asymptotically optimal, or logarithmically efficient, if

$$\limsup_{n \nearrow \infty} \frac{\log \tilde{E}\left(L(S_n)\mathbb{1}(S_n \in \mathcal{X})\right)^2}{\log P(S_n \in \mathcal{X})} = 2$$

Note:

1. If $P(S_n \in \mathcal{X}) \to 0$ exponentially in n, then Definition 2 means that the second moment of the estimator decays in twice the exponential rate as $P(S_n \in \mathcal{X})$
2. A zero-variance sampler has a density

$$\frac{f(x)\mathbb{1}(x \in \mathcal{X})}{P(S_n \in \mathcal{X})}$$

A Simple (Simplest) Example…

- Consider $P(S_n > an)$ where $S_n = \sum_{i=1}^{n} X_i$, X_i are i.i.d. r.v.'s with $EX_i = 0$ and $\psi(\theta) = \log Ee^{\theta X_i} < \infty$ for all $\theta \in \mathbb{R}$, and $a > 0$
- By Law of Large Numbers, $P(S_n > an) \to 0$ as $n \nearrow 0$
- Consider the importance sampling scheme where the probability distribution of each X_i is tilted along its exponential family so that $\tilde{E}X_i = a$ i.e. $d\tilde{P} = e^{\theta^* X_i - \psi(\theta^*)} dP$ where θ^* is the solution to $\psi'(\theta) = a$
- Cramer's Theorem:

$$P(S_n > an) = \tilde{E}\left[e^{-\theta^* S_n + n\psi(\theta^*)}; S_n > an\right]$$
$$= e^{-\theta^* an + n\psi(\theta^*)} \tilde{E}\left[e^{\theta^*(an - S_n)}; S_n > an\right] \approx e^{-nI(a)}$$

where $I(a) = \theta^* a - \psi(\theta^*)$ is called the rate function in large deviations theory

NOTES FROM THE EXAMPLE

- The proof of large deviations suggests a natural importance sampling scheme
- This scheme can be shown to be asymptotically optimal:

$$\tilde{E}(L\mathbb{1}(S_n > an))^2 = E[L; S_n > an]$$
$$= E\left[e^{-\theta^* S_n + n\psi(\theta^*)}; S_n > an\right]$$
$$= e^{-nI(a)} E[e^{\theta^*(an - S_n)}; S_n > an] \approx e^{-2nI(a)}$$

- The importance sampling scheme mimics the zero-variance sampler in the sense that

$$P(X_1 \in B_1, \ldots, X_n \in B_n | S_n > an) \to \tilde{P}(X_1 \in B_1) \cdots \tilde{P}(X_n \in B_n)$$

for all Borel sets $B_1, \ldots, B_n$

Large Deviations and Importance Sampling

- Contrary to central limit theorems where information on moments is enough, large deviations typically depend on the behavior of the moment generating function
- Gartner-Ellis Theorem as a generalization of Cramer's Theorem: Under regularity conditions, suppose a random object S_n possesses logarithmic moment generating function $\psi_n(\theta) = \log E e^{\theta S_n}$ such that $\psi_\infty(\theta) := \lim_{n\to\infty} \frac{1}{n}\psi_n(\theta)$ (Gartner-Ellis limit) exists on a sufficiently large enough interval of θ, then

$$\lim_{n\to\infty} \frac{1}{n} \log P(S_n \in \mathcal{X}) = -\inf_{a\in\mathcal{X}} I(a)$$

where $I(a) = \sup_{\theta\in\mathbb{R}} \langle\theta, a\rangle - \psi(\theta)$ is the rate function

Large Deviations and Importance Sampling

To find an optimal importance sampler for large deviations event…

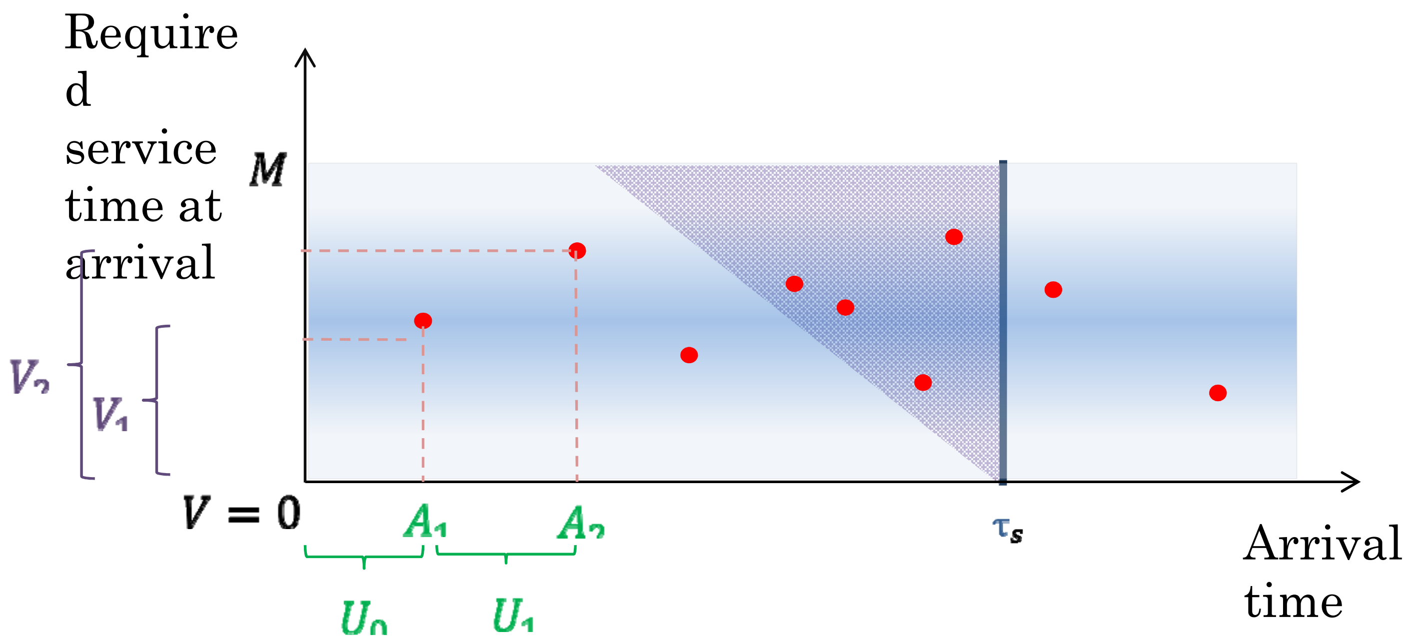
- Formulate Gartner-Ellis limit of the random object
- Decode from the limit the contributions of more "elementary" objects that lead to the rare event
- In many cases (but not all), the naturally suggested sampler is asymptotically optimal

Back to Our Problem…

- Do we know the Gartner-Ellis limit of the random object i.e. the steady-state loss distribution?
- Problem reformulation

Crude Monte Carlo Revisited

- Run the process for a long time
- $\hat{P}(loss) = \frac{\# \, loss}{\lambda s \times total\ time\ units\ simulated}$



A Regenerative View

- Suppose $A \in \mathcal{D} \times \mathbb{R}_+$ is a regenerative set of the system
- Kac's formula:

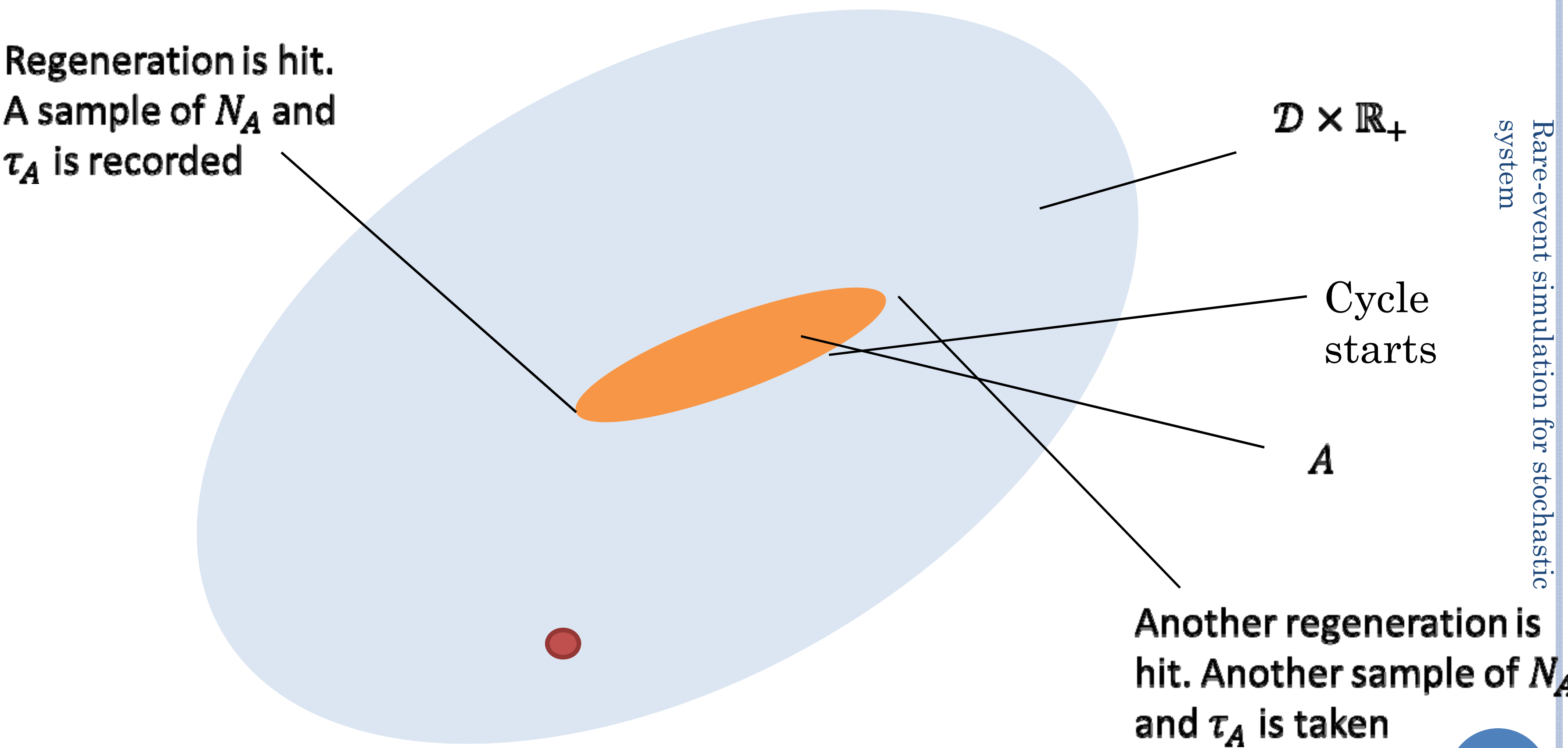
$$P_\pi(loss) = \frac{E_A N_A}{\lambda s E_A \tau_A}$$

Notations:

  - N_A = number of loss before reaching A again
  - τ_A = time units to reach A again
  - $E_A[\cdot]$ = expectation with initial state in steady-state conditional on being in A

- If we choose A to be a “good” set i.e. it does not take exponential amount of time to reach, then crude Monte Carlo is equivalent to using sample mean for both numerator and denominator
- Mixing guaranteed by finite support assumption on service time

A Regenerative View

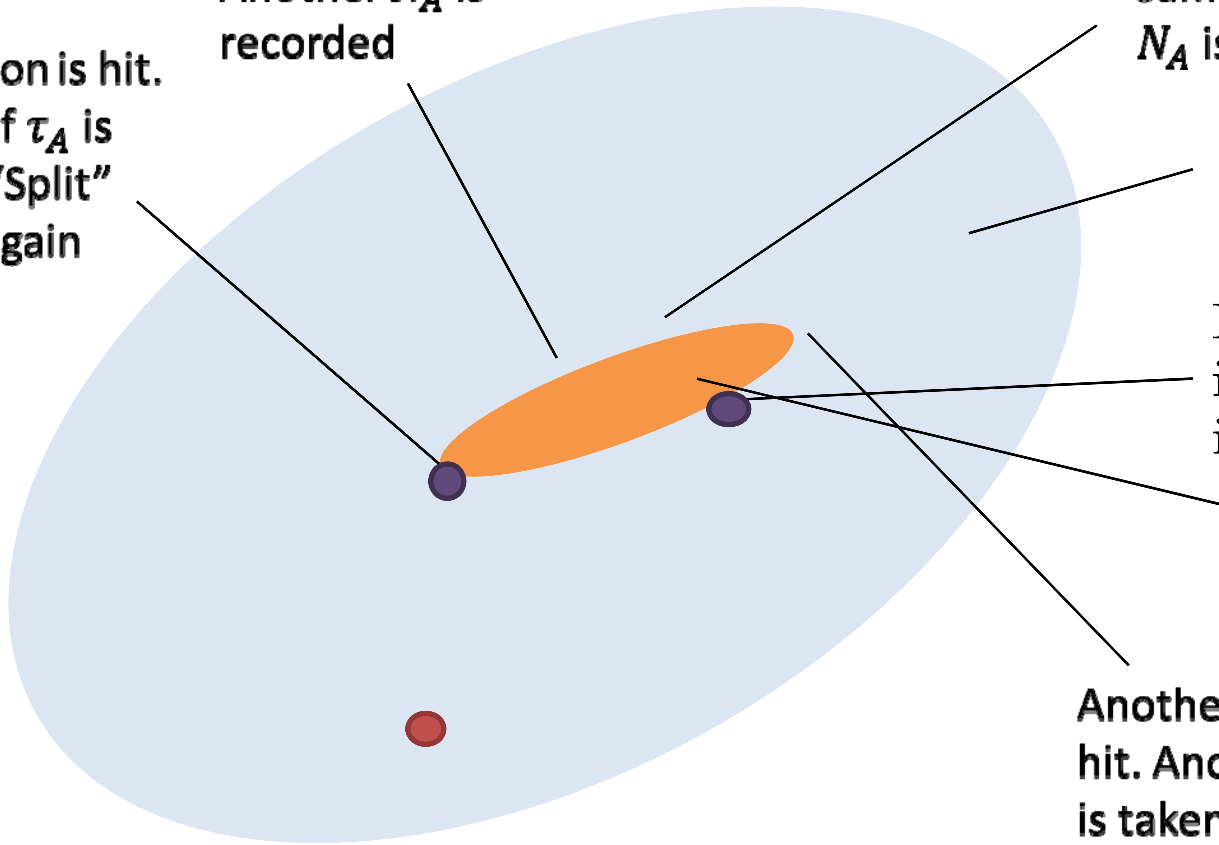


“Splitting” Algorithm

- Do importance sampling on the numerator
- Run crude Monte Carlo; every time A is hit, “split” the path into two: one continue with original evolution; another is applied with importance sampling to get a sample of N_A. Then continue with the original path

$$\hat{P}(loss) = \frac{weighted\ sum\ of\ N_A}{\lambda s \times total\ time\ units}$$

"SPLITTING" ALGORITHM



What is a good choice of set A?

- A should be occupied frequently (but not too frequently!) in steady-state
- When s is large, one can use Central Limit Theorem to approximate the "central limit" region of $\big(Q(t,y), B(t)\big) \in \mathcal{D} \times \mathbb{R}_+$
- (Pang and Whitt (2008)) Suppose there is no capacity constraint i.e. number of servers is infinite (but arrival rate is still λs), then

$$\frac{Q(\infty, y) - \lambda s \int_y^\infty \bar{F}(u)du}{\sqrt{s}} \approx Z(y)$$

where $Z(y)$ is a Gaussian process with

$$Var\, Z(y) = \lambda c^2 \int_y^\infty \bar{F}(u)^2 du + \lambda \int_y^\infty F(u)\bar{F}(u)du$$

- We can choose

$$A = \left\{ Q(t,y) \in \left(\lambda s \int_y^\infty \bar{F}(u)du - sd\big(Z(y)\big)\sqrt{s}\, , \lambda s \int_y^\infty \bar{F}(u)du + sd\big(Z(y)\big)\sqrt{s} \right), t \in \{\Delta, 2\Delta, \ldots\} \right\}$$

Key Questions

- Our problem becomes estimating $E_r N_A$ for some $r \in A$
- Do we have information from large deviations theory (i.e. Gartner-Ellis limit)?
- How does the rare event i.e. loss happen?
- Does the intuition give an asymptotically optimal estimator (or more)?

A Simpler Problem

- Consider a simpler problem in which Gartner-Ellis limit can be computed:
  - A "coupled" system that has no capacity constraint i.e. number of servers is infinite
  - Fix a time horizon t and initial state, say 0
- What is the probability that there are more than s customers in the system at time t?

SOLUTION TO THE SIMPLER PROBLEM

- This is mathematically

$$P(Q(t) > s)$$

where

$$Q(t) = \sum_{i=1}^{N(t)} \mathbb{1}(V_i > t - A_i)$$

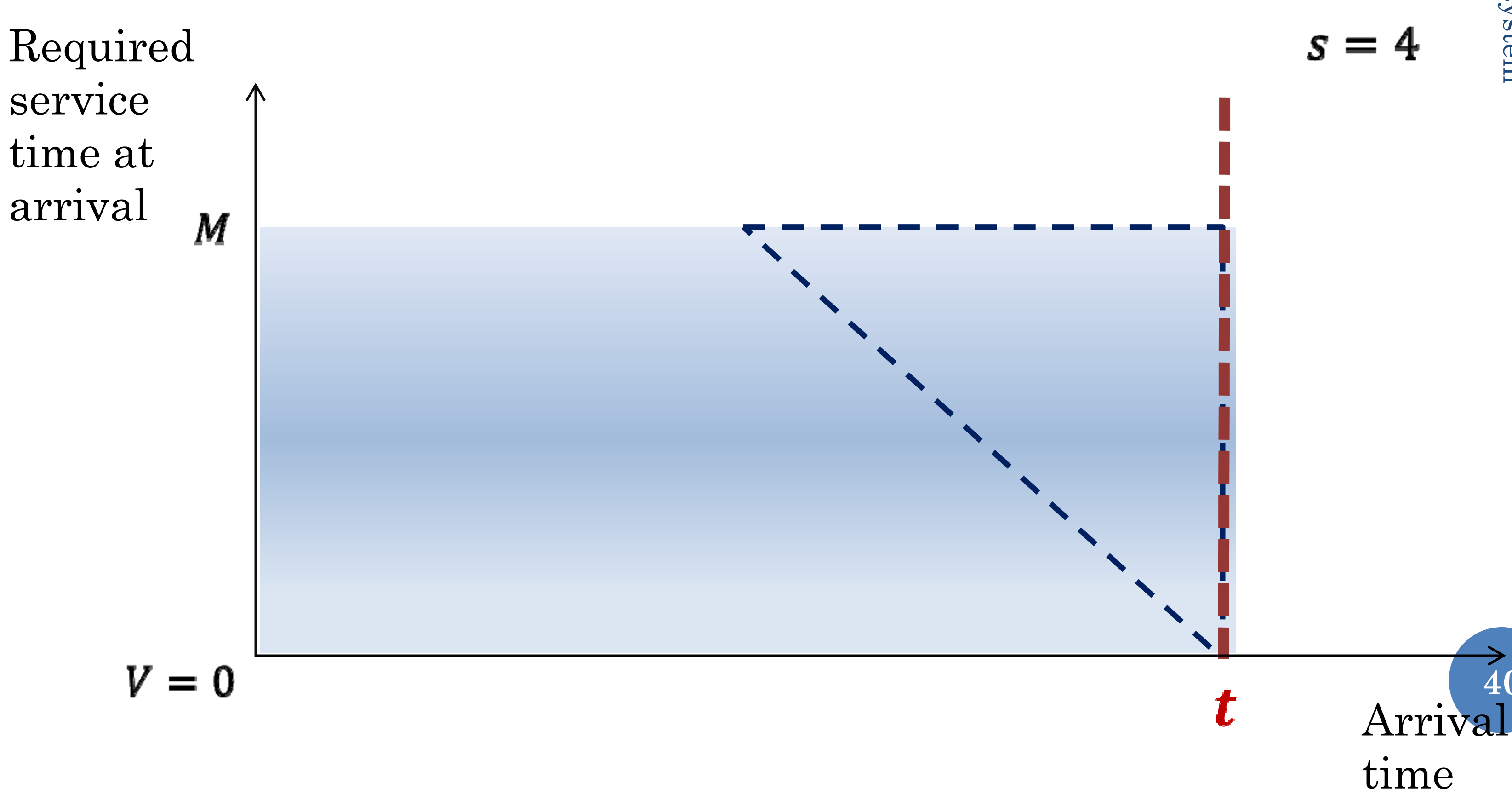
is the number of customers in the system at time t

- Gartner-Ellis limit

$$\psi_\infty(\theta) = \int_0^t \psi_N\left(\log\left(e^\theta \bar{F}(t-u) + F(t-u)\right)\right) du$$

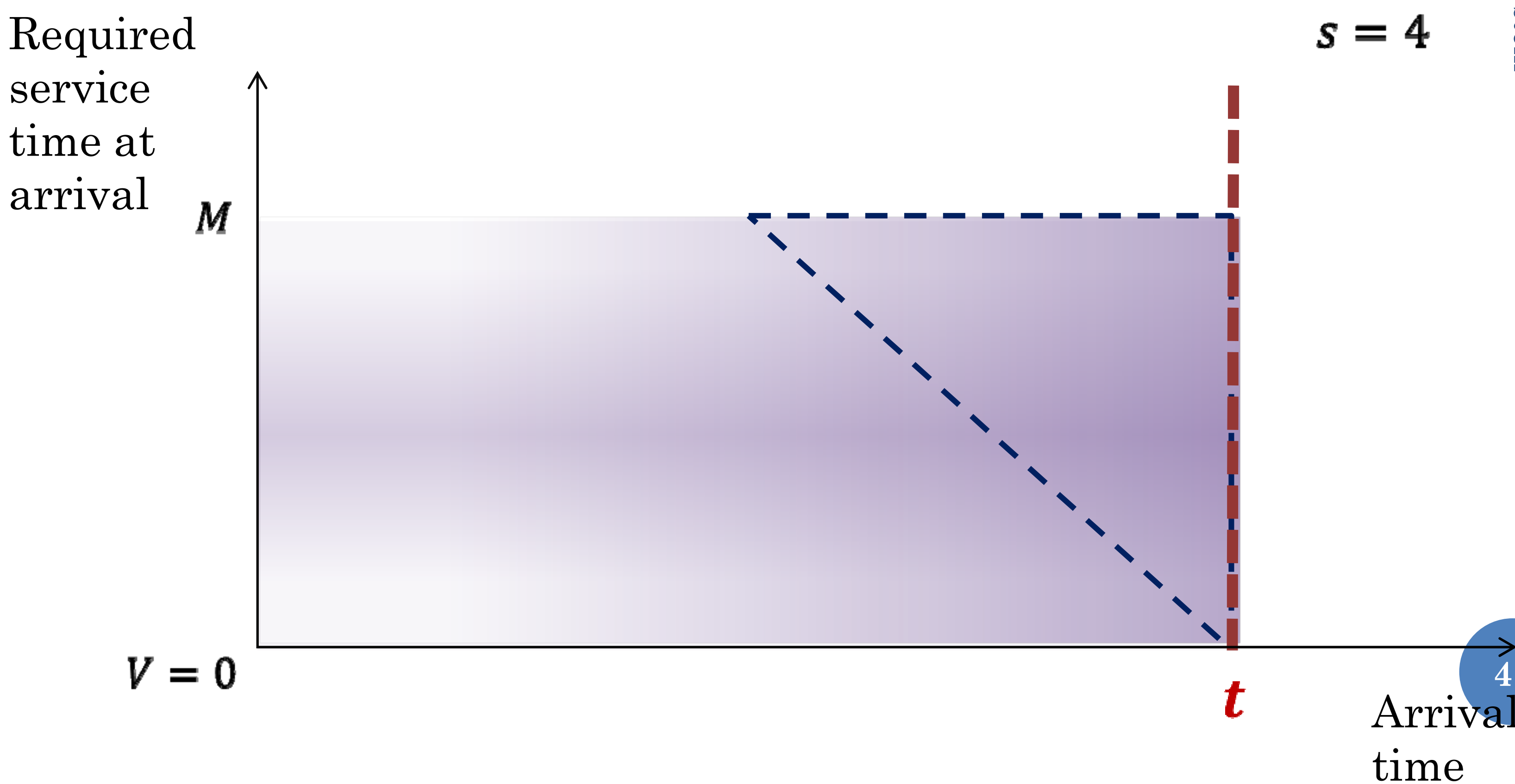
where $\psi_N(\cdot)$ is the infinitesimal logarithmic moment generating function of the arrival process

Importance Sampling for the Simpler Problem



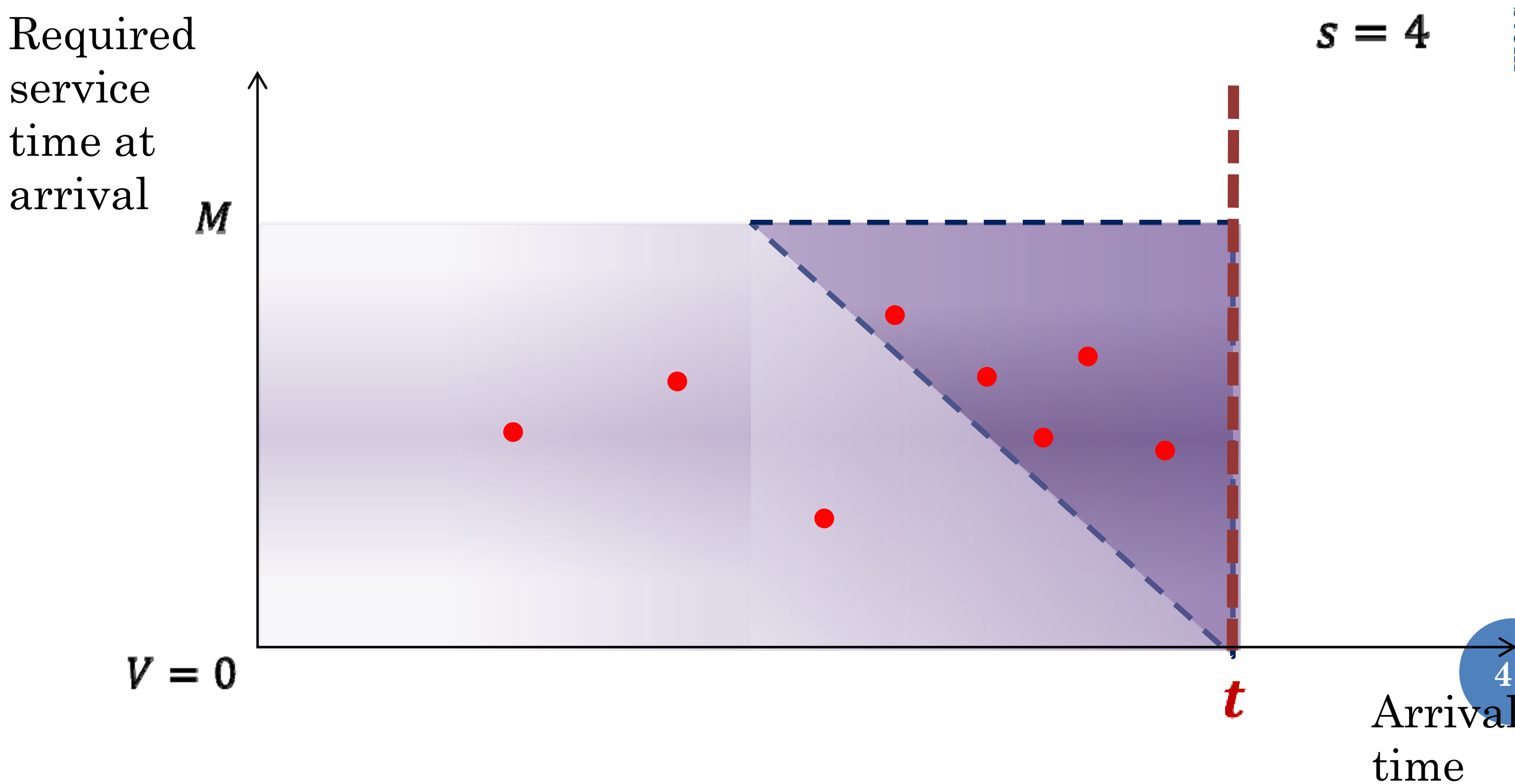
Importance Sampling for the Simpler Problem

1. Arrival rate is accelerated towards t by tilting the interarrival times U_i by $\psi_N(\log(e^{\theta^*}\bar{F}(t-A_i)+F(t-A_i))$



Importance Sampling for the Simpler Problem

1. Arrival rate is accelerated towards t by tilting the interarrival times U_i by $\psi_N(\log(e^{\theta^*}\bar{F}(t-A_i)+F(t-A_i))$
2. Service time density is increased by a factor of e^{θ^*} inside the triangle



INTUITION FOR OUR PROBLEM

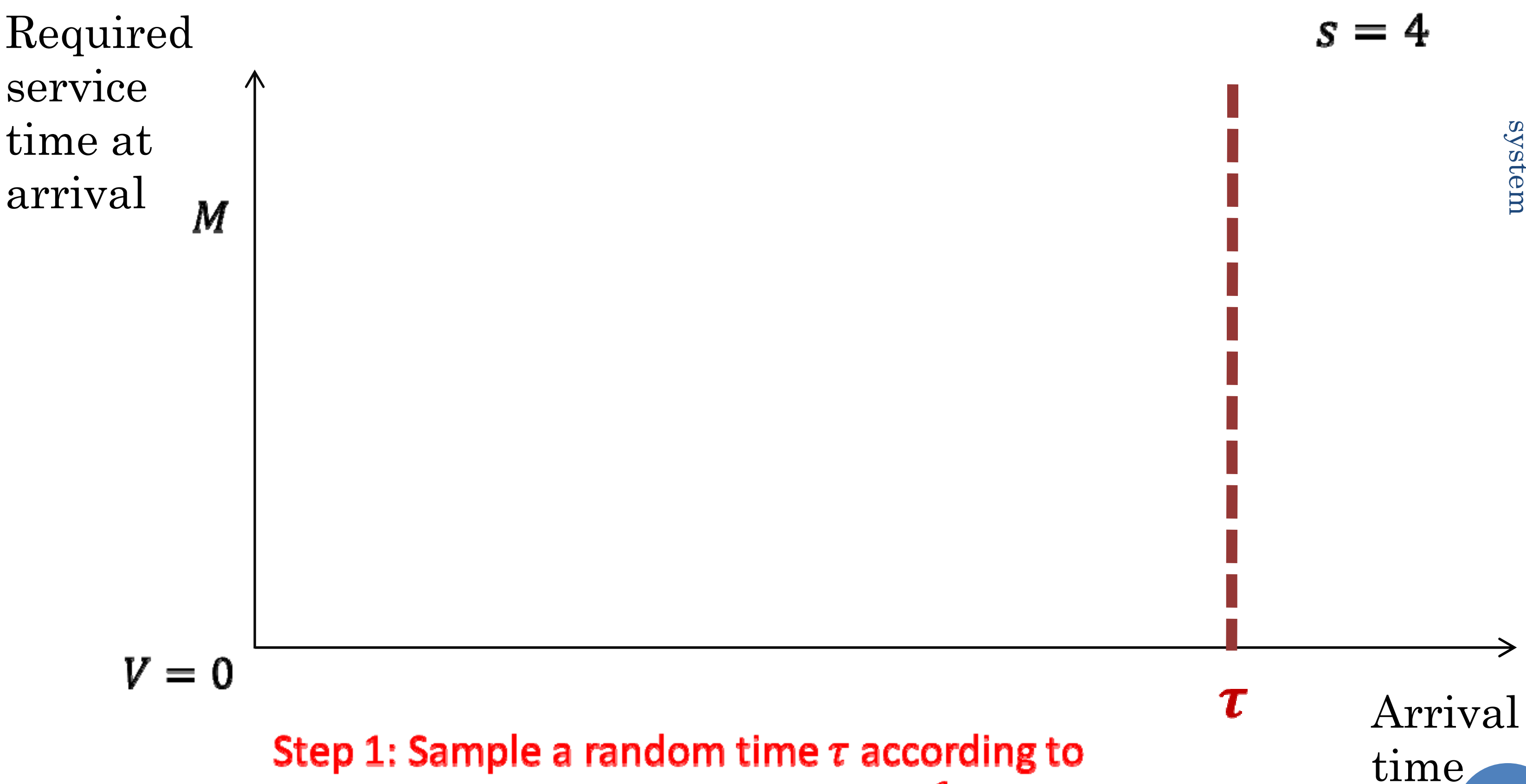
- Given an initial state $r \in A$,

$$E_r N_A = E_r[N_A; \tau_s < \tau_A] \approx P_r(\tau_s < \tau_A)$$
$$\approx \sum_t P_r(\tau_s = t < \tau_A) \approx \sum_t P_r(Q(t) > s) \approx P_r(Q(t^*) > s)$$
$$\approx e^{-sI_{t^*}}$$

Idea 1: Before the first loss, the system acts the same as if there are infinite number of servers

Idea 2: Since time horizon for loss is not fixed, we shall randomize the time horizon (also preventing blowing up variance due to non-optimal sample paths)

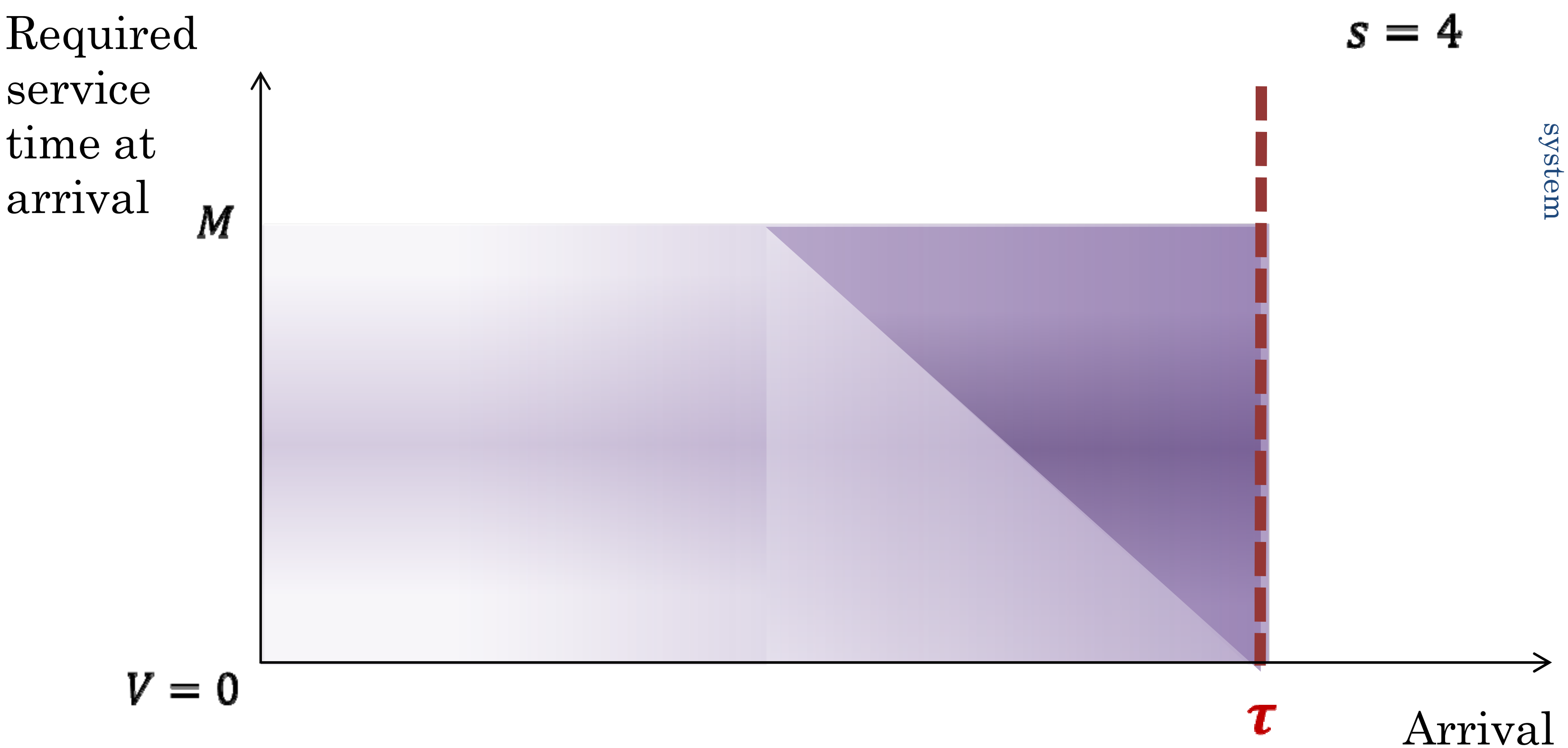
Algorithm for $E_r N_A$



Step 1: Sample a random time τ according to

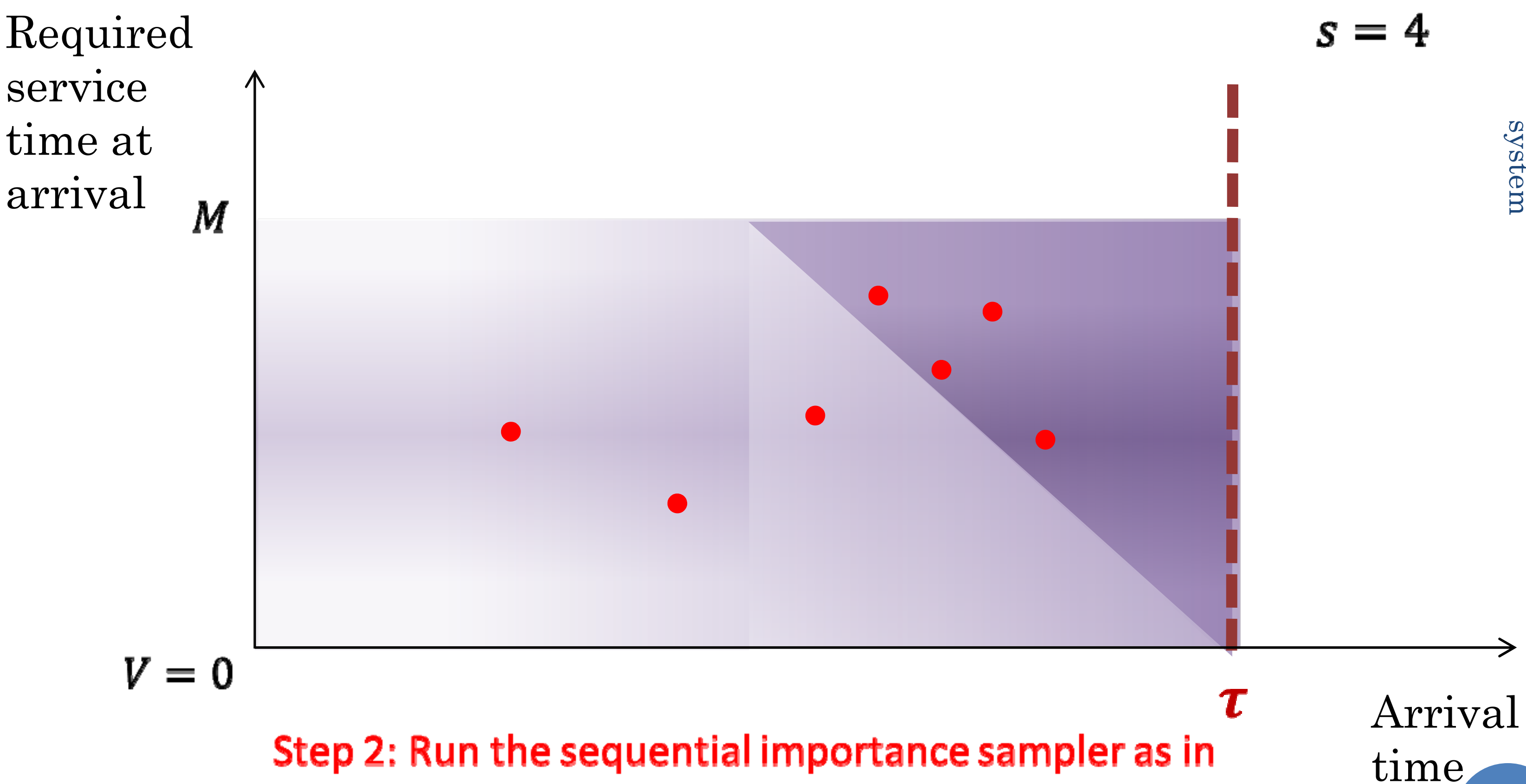
$$P(\tau = T + k\delta) = \frac{6}{\pi^2(k+1)^2}$$

Algorithm for $E_r N_A$



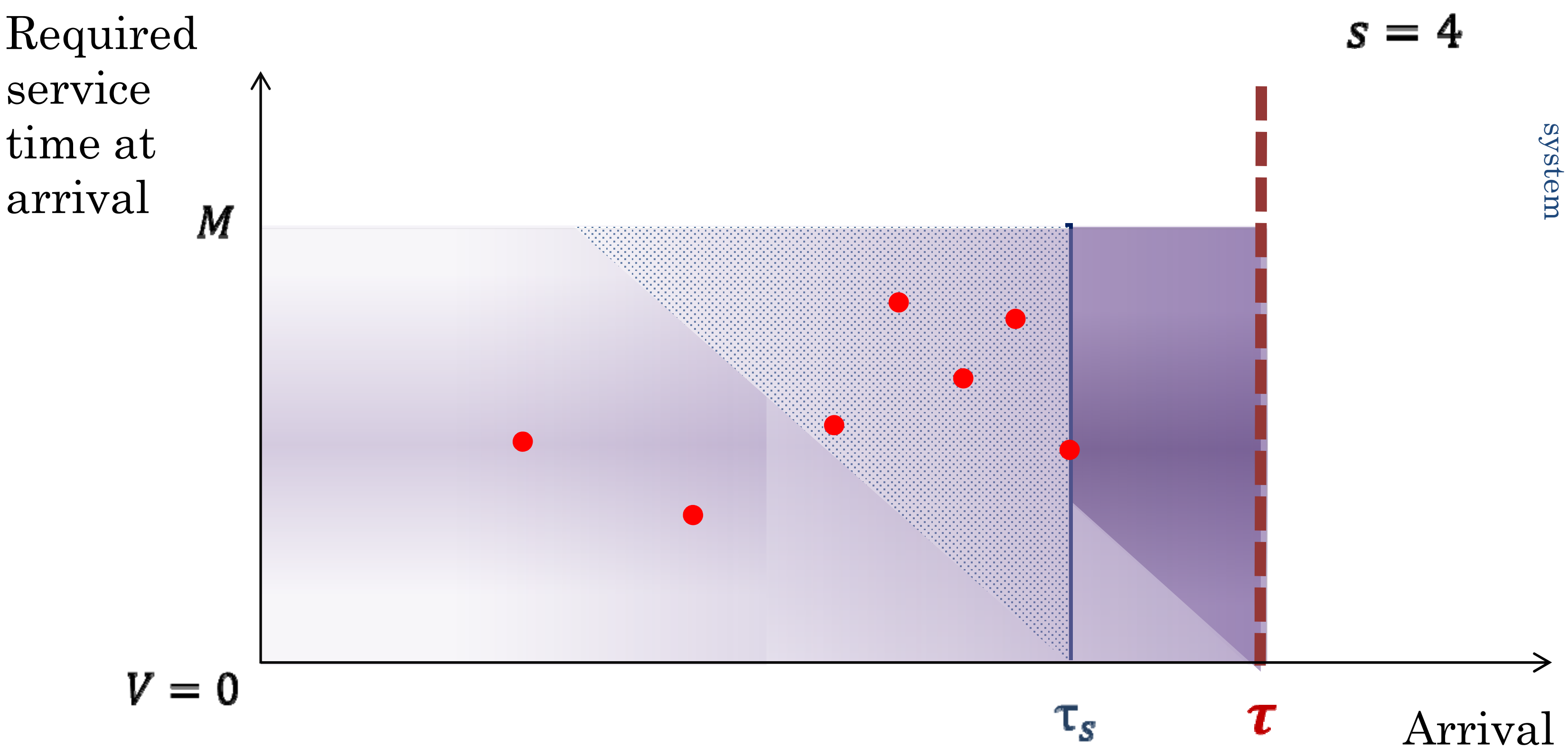
Step 2: Run the sequential importance sampler as in the simpler problem, pretending infinite number of servers and time horizon of the realized τ, until $\tau_s \wedge \tau_A$

Algorithm for $E_r N_A$



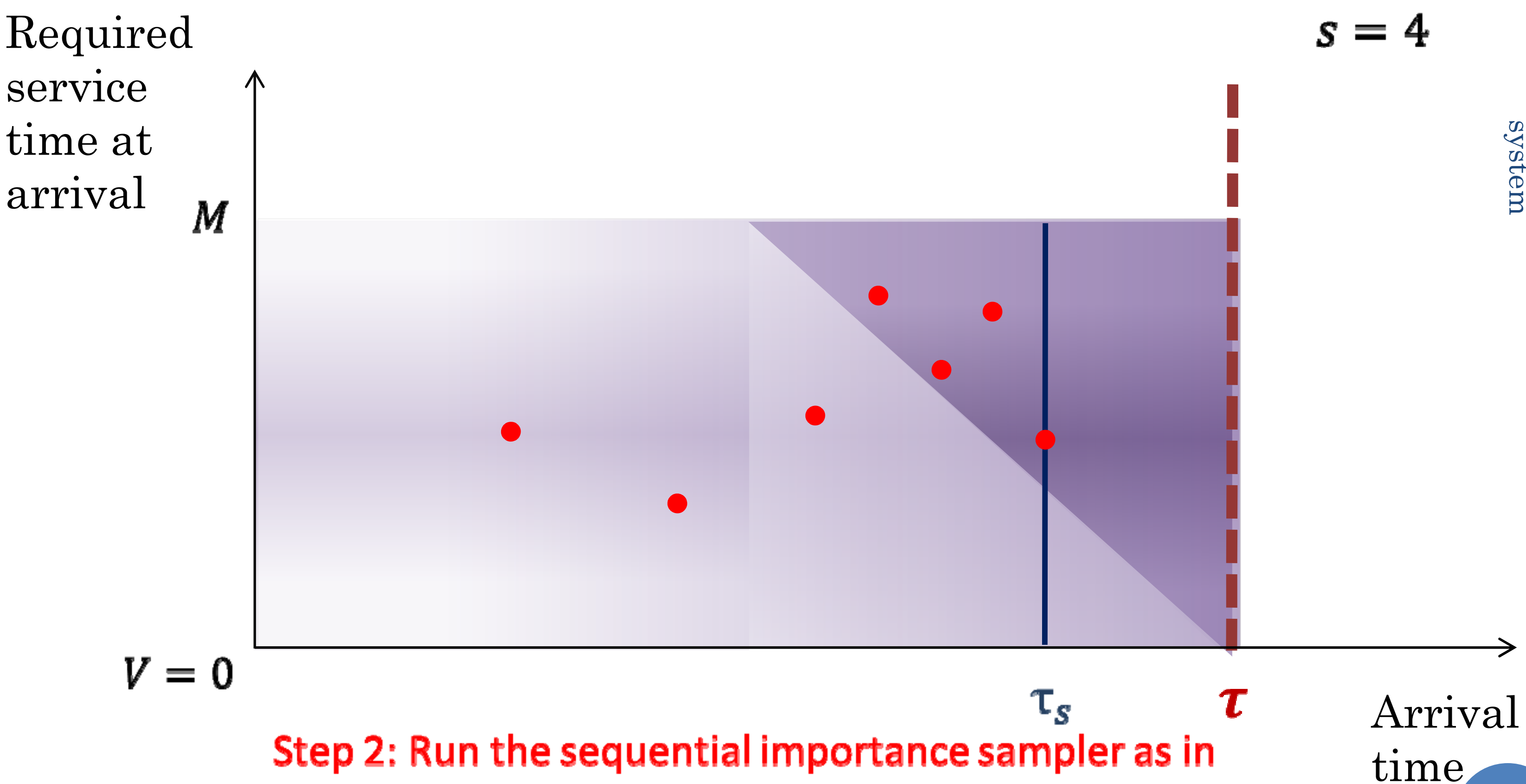
Step 2: Run the sequential importance sampler as in the simpler problem, pretending infinite number of servers and time horizon of the realized τ, until $\tau_s \wedge \tau_A$

Algorithm for $E_r N_A$



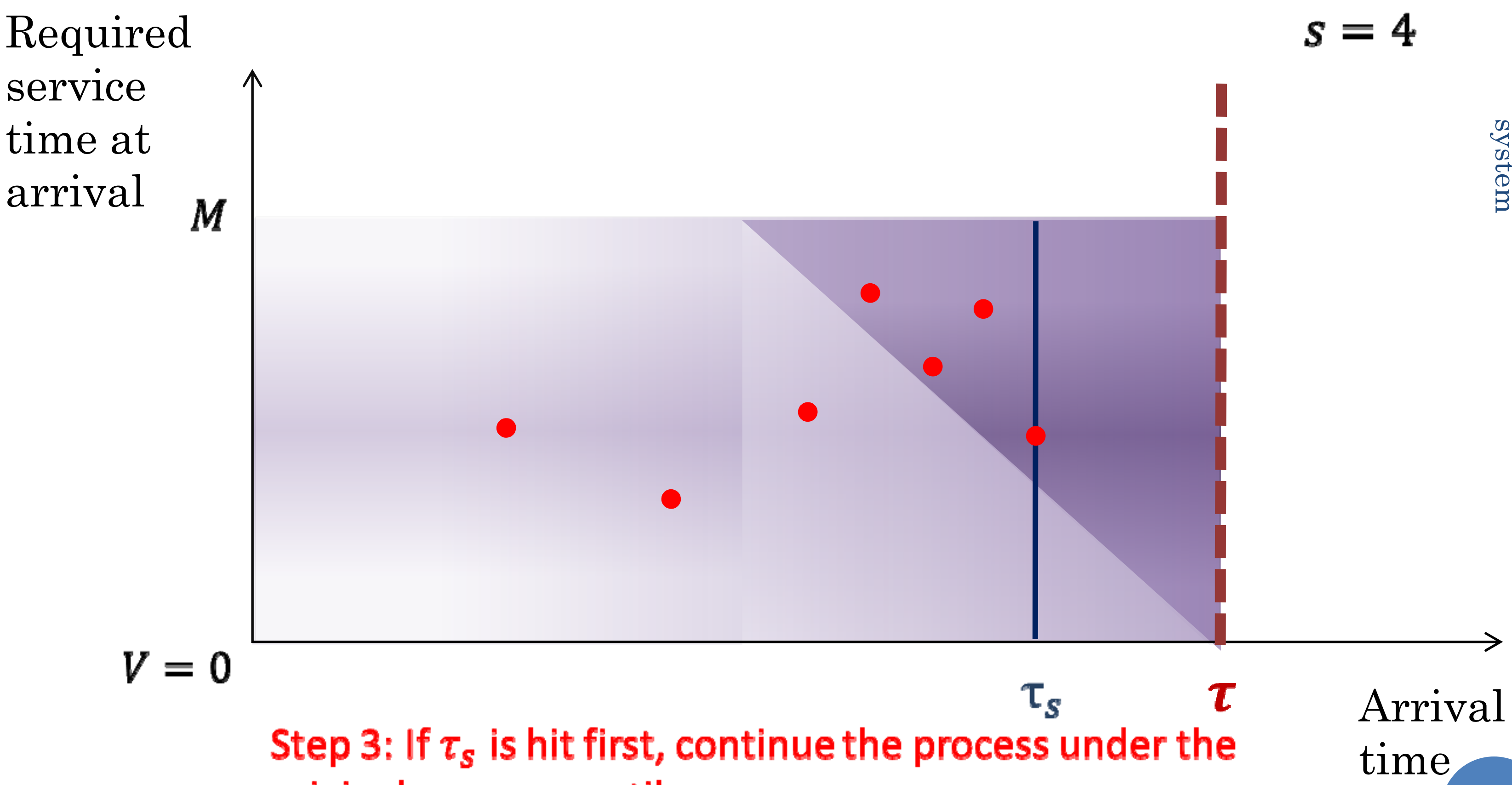
Step 2: Run the sequential importance sampler as in the simpler problem, pretending infinite number of servers and time horizon of the realized τ, until $\tau_s \wedge \tau_A$

Algorithm for $E_r N_A$



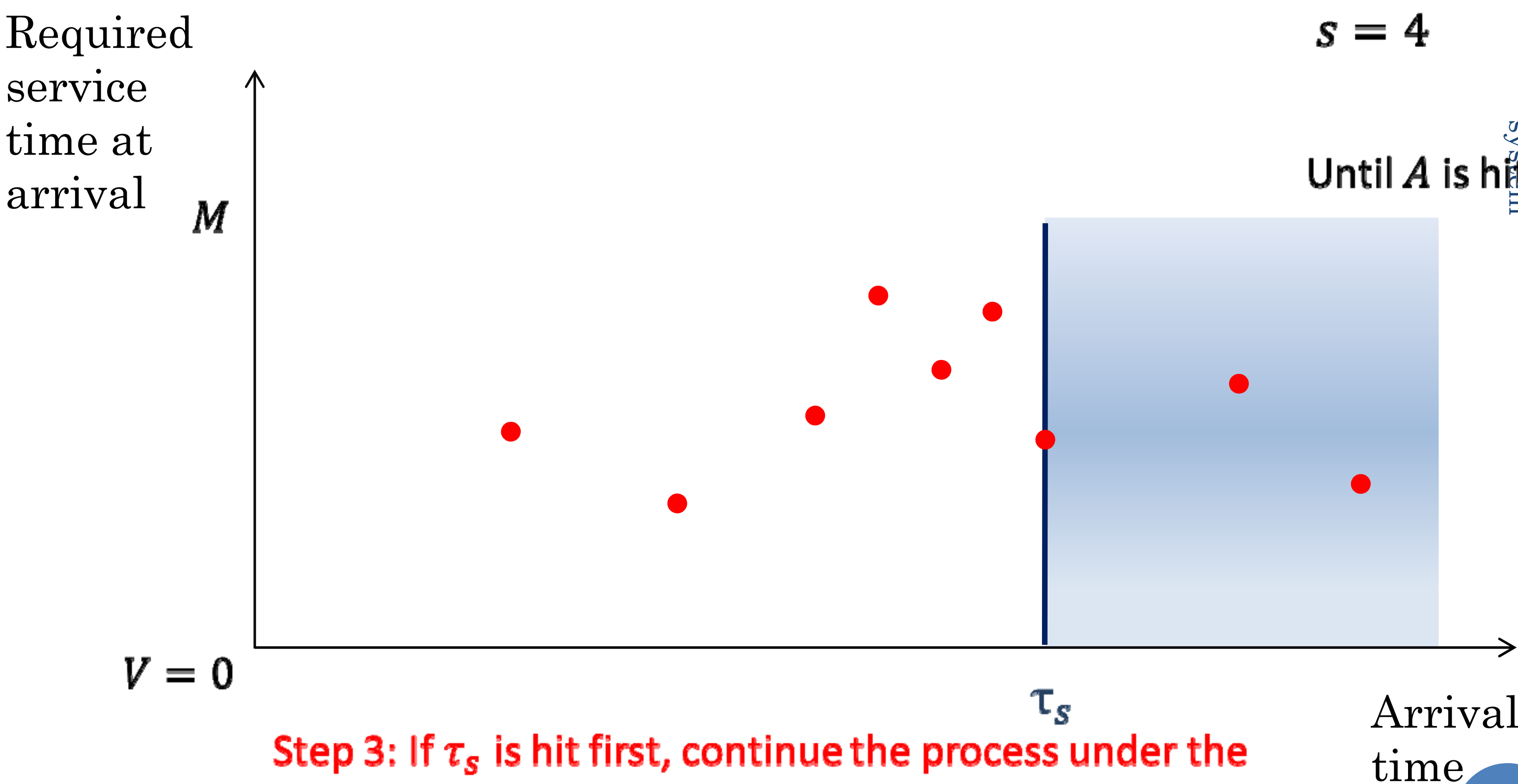
Step 2: Run the sequential importance sampler as in the simpler problem, pretending infinite number of servers and time horizon of the realized τ, until $\tau_s \wedge \tau_A$

Algorithm for $E_r N_A$



Step 3: If τ_s is hit first, continue the process under the original measure until τ_A

Algorithm for $E_r N_A$



Step 3: If τ_s is hit first, continue the process under the original measure until τ_A

Asymptotic Optimality

The likelihood ratio for the algorithm under $\tau_s < \tau_A$ is

$$L(Y_u, 0 \le u \le \tau_s) = \frac{1}{\sum_t P(\tau = t) {L_t}^{-1}(Y_u, 0 \le u \le \tau_s)}$$

where L_t is the likelihood ratio conditional on $\tau = t$ given by

$$\exp\left\{ s \sum_{i=1}^{N(\tau_s)-1} \psi_N\left(\log\left(e^{\theta_t}\bar{F}(t - A_i) + F(t - A_i)\right)\right) U_i - \theta_t \sum_{i=1}^{N(\tau_s)-1} \mathbb{1}(V_i > t - A_i) \right\}$$

for $t \ge \tau_s$ and

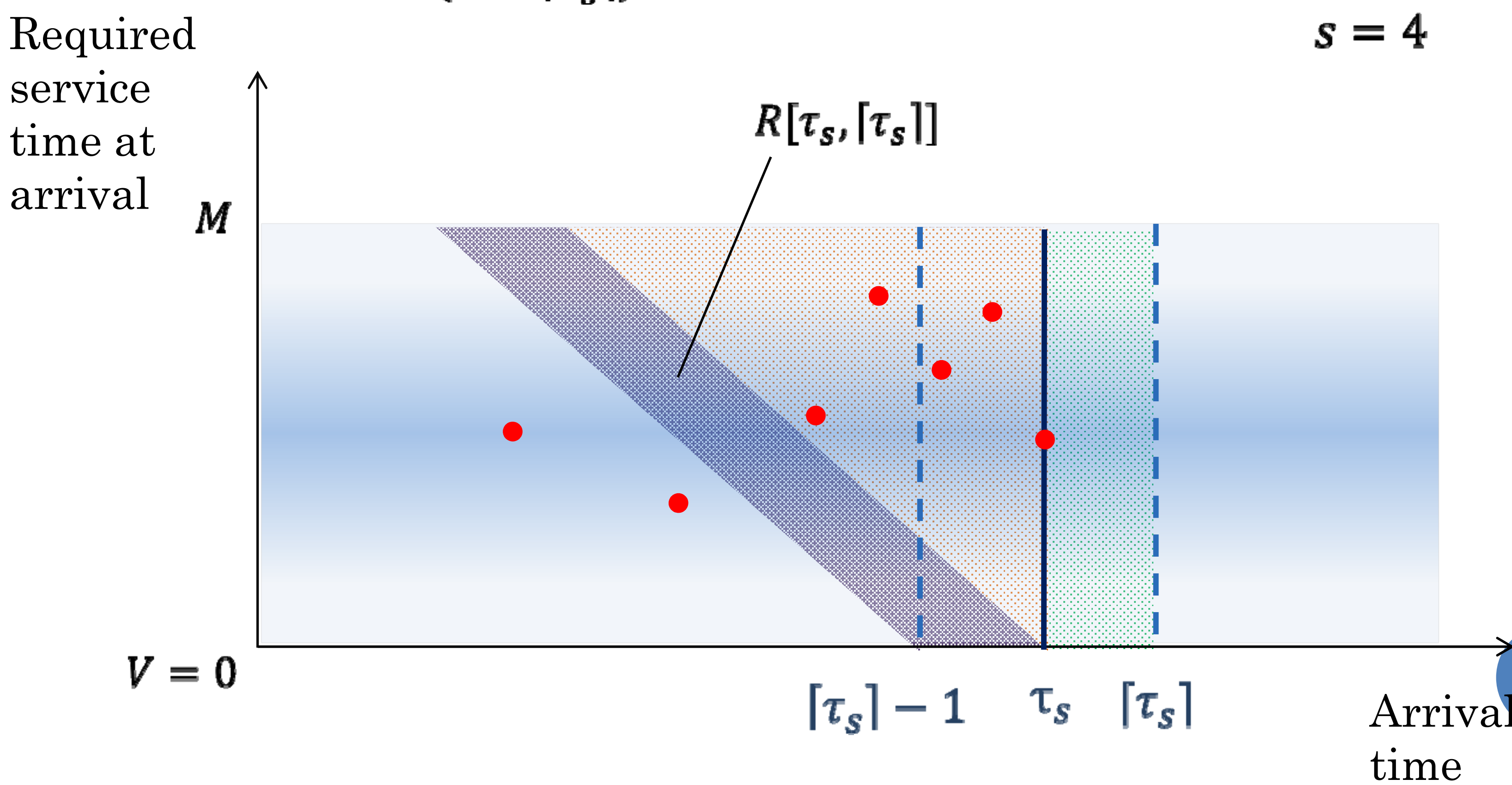
$$\exp\left\{ s \sum_{i=1}^{N(t)} \psi_N\left(\log\left(e^{\theta_t}\bar{F}(t - A_i) + F(t - A_i)\right)\right) U_i - \theta_t \sum_{i=1}^{N(t)} \mathbb{1}(V_i > t - A_i) \right\}$$

for $t < \tau_s$

Asymptotic Optimality

- The likelihood ratio is bounded from above by

$$\frac{L_{\lceil\tau_s\rceil}}{P(\tau = \lceil\tau_s\rceil)} \leq c\lceil\tau_s\rceil^2 e^{-sI_{t^*}+\theta_{\lceil\tau_s\rceil}R[\tau_s,\lceil\tau_s\rceil]}$$



Asymptotic Optimality

- Second moment of likelihood ratio is

$$\tilde{E}[N_A^2 L^2; \tau_s < \tau_A] = E[N_A^2 L; \tau_s < \tau_A]$$
$$\leq c e^{-sI_{t^*}} E[e^{\theta_{\lceil \tau_s \rceil} R[\tau_s, \lceil \tau_s \rceil]}; \tau_s < \tau_A]$$

- When τ is sampled at scale $O\left(\frac{1}{s}\right)$,
  - For the case of Poisson arrival, given τ_s, $R[\tau_s, \lceil \tau_s \rceil] \sim Binomial(s, p)$ where p is the ratio of purple area to trapezoid
  - For general case, condition on the arrival times of the contributing customers
- One can get a logarithmic bound of $e^{-2sI_{t^*}}$

Asymptotic Optimality

Theorem: We have

$$\lim_{s \nearrow \infty} \frac{1}{s} \log P(loss) = -I_{t^*}$$

where $t^* = \text{argmin}\, I_t$ and

$$\lim_{s \nearrow \infty} \frac{1}{s} \log \tilde{E}[N_A^2 L^2; \tau_s < \tau_A] = -2I_{t^*}$$

Hence the algorithm is asymptotically optimal.

<u>Sketch of Proof:</u>

- Lower bound $\lim_{s \nearrow \infty} \frac{1}{s} \log P(loss) \geq -I_{t^*}$ is established by explicitly identifying the optimal sample path
- For upper bound,

$$-2I_{t^*} \leq \lim_{s \nearrow \infty} \frac{1}{s} \log P(loss)^2 \leq \lim_{s \nearrow \infty} \frac{1}{s} \log \tilde{E}[N_A^2 L^2; \tau_s < \tau_A] \leq -2I_{t^*}$$

SIMPLIFICATION AND EXTENSIONS

For Poisson arrival,

- A faster algorithm can be obtained by, after sampling τ, generating $Q(t)$ using tilted measure and then sampling the customers exploiting the Poisson random measure description
- It is interesting to note that the seemingly more powerful idea of conditionally sampling $Y_u, 0 \leq u \leq t | Q(t)$ (instead of exponential tilting) will blow up the second moment of the likelihood ratio at a neighborhood of the time of first loss, due to "discontinuity" of the likelihood ratio at τ_s
- However, this works for a discrete version of the process (Blanchet, Glynn and Lam (2009))

Simplification and Extensions

- The case of Markov-modulated arrivals can be simulated by restricting set A to the optimal Markov state i.e. the state that gives the highest arrival rate
- The case of both Markov-modulated arrivals and (possibly correlated) service times can be simulated by augmenting the state-space to identify the residual service times of customers who enter at each Markov state
- Time-varying arrivals (in this case we are interested in loss during an interval instead of the steady-state) can be simulated using exactly the same methodology (with truncation of τ)