Adaptive Discretization in Model-Based RL

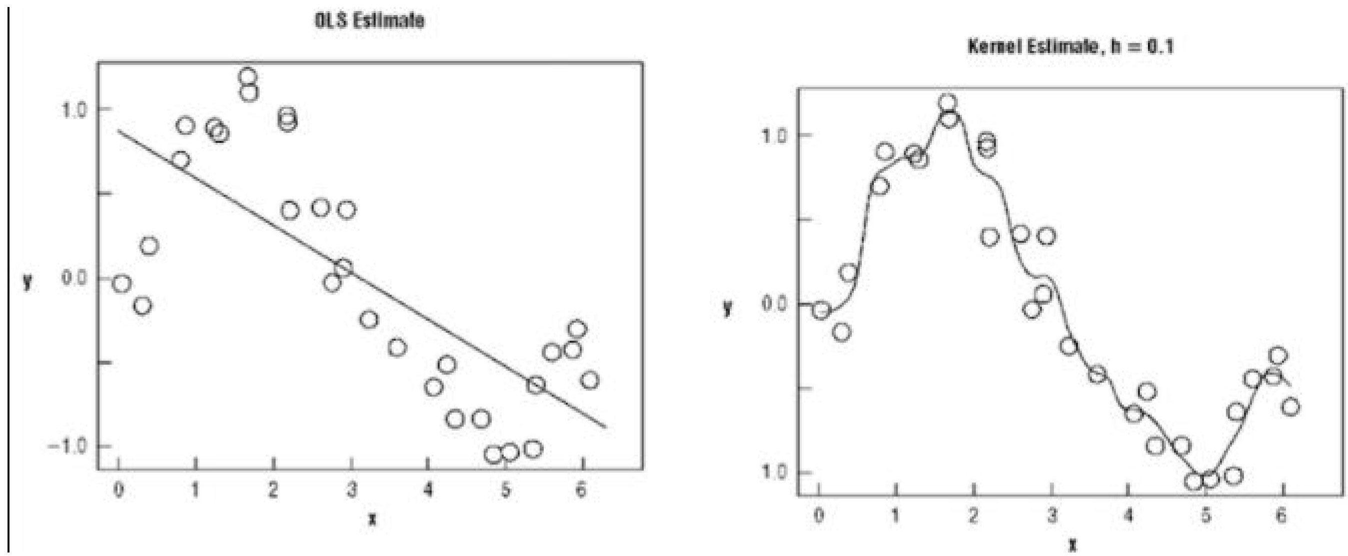
Robert Ferrando

11/23/2021

Nonparametric RL

- In class, we discussed several examples of **parametric RL** – making assumptions about the structure of the value functions
    - e.g. linear Bellman completeness
- Nonparametric RL make few such parametric assumptions; i.e., the learned function can take any form
    - Common assumption: Q-function is Lipschitz continuous

Nonparametric vs. Parametric Regression



Modified from https://www.section.io/engineering-education/parametric-vs-nonparametric/fit.png.

Why Nonparametric RL?

- Don’t know what Q should “look like”
- Large and/or continuous state-action spaces
- E.g., My research – optimal trading in energy markets

Model-Based vs. Model Free

- Model-based: store values of reward function and transition dynamics
- Model-free: only store values of Q function – do not need to "learn" the system
- Previous work on adaptive discretization for model-free RL has been published

Comparison of Methods

Algorithm	Regret	Time Complexity	Space Complexity
AdaMB (Alg. 1) ($d_{\mathcal{S}} > 2$)	$H^{1+\frac{1}{d+1}} K^{1-\frac{1}{d+d_{\mathcal{S}}}}$	$HK^{1+\frac{d_{\mathcal{S}}}{d+d_{\mathcal{S}}}}$	HK
($d_{\mathcal{S}} \leq 2$)	$H^{1+\frac{1}{d+1}} K^{1-\frac{1}{d+d_{\mathcal{S}}+2}}$	$HK^{1+\frac{d_{\mathcal{S}}}{d+d_{\mathcal{S}}+2}}$	$HK^{1-\frac{2}{d+d_{\mathcal{S}}+2}}$
Adaptive Q-Learning [37]	$H^{5/2} K^{1-\frac{1}{d+2}}$	$HK \log_d(K)$	$HK^{1-\frac{2}{d+2}}$
Kernel UCBVI [11]	$H^3 \quad K^{1-\frac{1}{2d+1}}$	HAK^2	HK
Net-Based Q-Learning [41]	$H^{5/2} K^{1-\frac{1}{d+2}}$	HK^2	HK
Lower-Bounds [39]	$H \quad K^{1-\frac{1}{d+2}}$	N/A	N/A

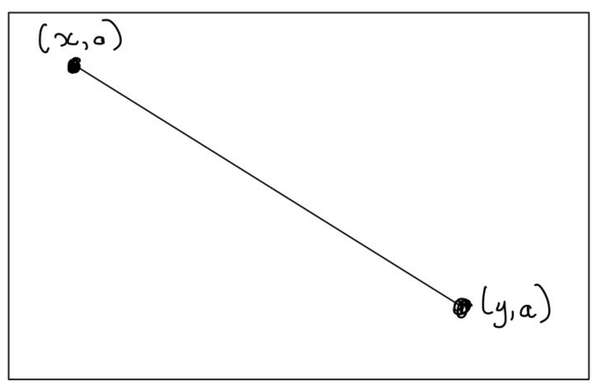
Ada-MB: model-based – discussed in paper

Kernel-UCBVI: model-based

Adaptive Q-learning: model-free

Problem Setting

- $\mathcal{S}$: state space
  - bounded metric space with metric $\mathcal{D}_S$
- $\mathcal{A}$: action space
  - bounded metric space with metric $\mathcal{D}_A$
- $\mathcal{S} \times \mathcal{A}$: state-action space
  - has product metric $\mathcal{D}$



Notation

- $d_{\mathcal{S}}$: dimension of $\mathcal{S}$
- $d_{\mathcal{A}}$: dimension of $\mathcal{A}$
- $d := d_{\mathcal{S}} + d_{\mathcal{A}}$
- $r_h(x, a)$: reward at time step h for $(x, a) \in \mathcal{S} \times \mathcal{A}$
- $T_h(x' \mid x, a)$: prob. of transitioning from state $x \to x'$ via action a

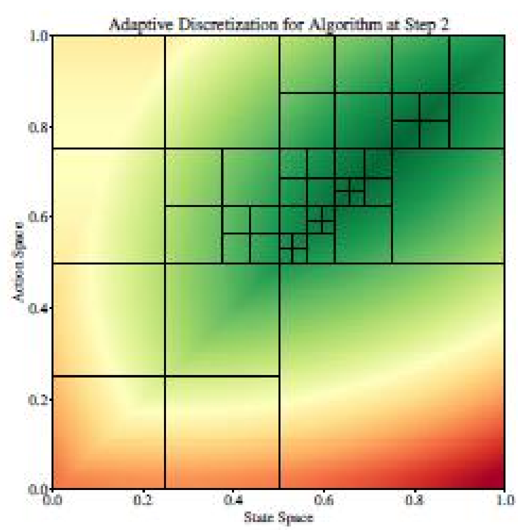
Assumptions

- Reward estimate: Lipschitz with constant L_R
- Transition estimate: Lipschitz with constant L_T
- Q_h^*, V_h^*: Lipschitz with constant L_V
- Through sampling, we have access to metrics $\mathcal{D}_S$, $\mathcal{D}_A$, $\mathcal{D}$

The Algorithm

Why Adaptive Discretization?

- Consider a continuous state-action space – discretize to yield empirical results
- Fixed discretization may be unnecessarily refined, especially in areas of the state-action space which are not profitable
- Adaptive discretization helps save memory and time, leading to a more efficient RL algorithm



Partitioning in practice

Ada-MB Algorithm

Ada-MB: Adaptive discretization for model-based RL

What we know:

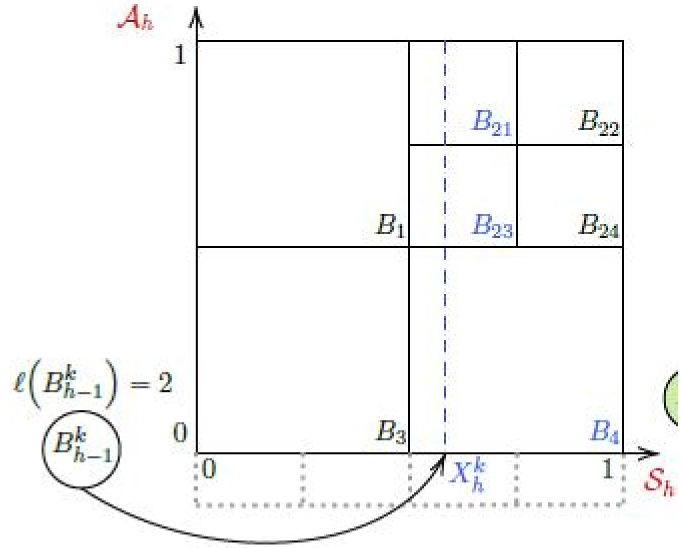
- $\mathcal{S}$: state space
- $\mathcal{A}$: action space
- $\mathcal{D}$: metric on state-action space
- H: time horizon
- K: # of episodes
- δ: probability of “bad events”

Ada-MB Algorithm

The following process occurs for each episode:

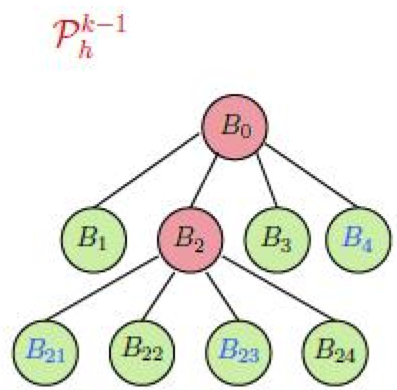
- Start with partitioning $\mathcal{S} \times \mathcal{A}$ into a collection of “balls”
- Receive starting state X_1^k
- For each time step:
  - Figure out which balls X_h^k belongs to
  - Which of those balls is most profitable? Call it B_h^k
  - Play action associated with greedy ball (center, or sample uniformly)
  - Update count, transition, reward associated with B_h^k
  - Partition if warranted

Ada-MB Algorithm



Partitioning the State-Action Space

- Define a **splitting rule** (*more on next slide*): if a ball is played enough times, split it
  - want to give the learner lots of options to play in profitable regions of $\mathcal{S} \times \mathcal{A}$
- Splitting is done "dyadically": consider the following tree



- If a ball at level $\ell(B)$ is chosen to be split, partition into 2^d smaller balls of diameter $2^{-(\ell(B)+1)}$

Splitting Rule

Partition B when $n_h^k(B) + 1 > n_+(B)$

- Splitting threshold $n_+(B) = \phi 2^{\gamma \ell(B)}$:
  - $d_{\mathcal{S}} > 2$: $n_+(B) = \phi 2^{d_{\mathcal{S}} \ell(B)}$
  - $d_{\mathcal{S}} \leq 2$: $n_+(B) = \phi 2^{(d_{\mathcal{S}}+2) \ell(B)}$
    - $\phi = H^{(d+d_{\mathcal{S}})/(d+1)}$ – reduce H-dependence of regret bound

Updating Counts

- $n_h^k(B)$: number of times ball B has been played by time step h of episode k

Reward

- $\hat{\mathbf{r}}_h^k(B)$: empirical reward from playing B
- $\overline{\mathbf{r}}_h^k(B)$: empirical reward from playing B *and its ancestors*

Transition

- $\hat{\mathbf{T}}_h^k(\cdot \mid B)$: empirical probability of transitioning from B
- $\overline{\mathbf{T}}_h^k(\cdot \mid B)$: empirical probability of transitioning from B *and its ancestors*

Bonus Terms

We want to maintain an *optimistic* estimate of the value functions – introduction of bonus terms allows us to do so (much like in multi-armed bandits.)

$$\textsc{Rucb}_h^k(B) = \sqrt{\frac{8\log(2HK^2/\delta)}{\sum_{B' \supseteq B} n_h^k(B')}} + 4L_r\mathcal{D}(B)$$

$$\textsc{Tucb}_h^k(B) = \begin{cases} L_V\left((5L_T+4)\mathcal{D}(B) + 4\sqrt{\frac{\log(HK^2/\delta)}{\sum_{B' \subseteq B} n_h^k(B')}} + c\left(\sum_{B' \subseteq B} n_h^k(B')\right)^{-1/d_{\mathcal{S}}}\right) & \text{if } d_{\mathcal{S}} > 2 \\ L_V\left((5L_T+6)\mathcal{D}(B) + 4\sqrt{\frac{\log(HK^2/\delta)}{\sum_{B' \supseteq B} n_h^k(B')}} + c\sqrt{\frac{2^{d_{\mathcal{S}}\ell(B)}}{\sum_{B' \supseteq B} n_h^k(B')}}\right) & \text{if } d_{\mathcal{S}} \leq 2 \end{cases}$$

Value Function Estimates

Q Estimate

$$\overline{\mathbf{Q}}_h^k(B) = \begin{cases} \overline{\mathbf{r}}_H^k(B) + \mathrm{RUCB}_H^k(B) & h = H \\ \overline{\mathbf{r}}_h^k(B) + \mathrm{RUCB}_h^k(B) + & \\ \mathbb{E}_{A \sim \overline{\mathbf{T}}_h^k(\cdot \mid B)}\left[\overline{V}_{h+1}^{k-1}(A)\right] + \mathrm{TUCB}_h^k(B) & h < H. \end{cases}$$

V Estimate

$$\overline{\mathbf{V}}_h^k(x) = \min_{A' \in \mathcal{S}(\mathcal{P}_h^k)} \left(\tilde{\mathbf{V}}_h^k(A') + L_V \mathcal{D}_{\mathcal{S}}(x, \tilde{x}(A'))\right)$$

$$\tilde{\mathbf{V}}_h^k(A) = \min\left\{\tilde{\mathbf{V}}_h^{k-1}(A), \max_{B \in P_h^k \text{ s.t. } A \subseteq \mathcal{S}(B)} \overline{\mathbf{Q}}_h^k(B)\right\}$$

Main Result

Worst-Case Regret

The main result is the following:

Theorem. With probability at least $1-\delta$, the regret of Ada-MB for any sequence of starting states $\{X_1^k\}_{k=1}^K$, is upper bounded as follows:

$$R(K) \lesssim \begin{cases} LH^{1+\frac{1}{d+1}} K^{\frac{d+d_{\mathcal{S}}-1}{d+d_{\mathcal{S}}}}, & d_{\mathcal{S}} > 2 \\ LH^{1+\frac{1}{d+1}} K^{\frac{d+d_{\mathcal{S}}+1}{d+d_{\mathcal{S}}+2}}, & d_{\mathcal{S}} \leq 2. \end{cases}$$

Proof Sketch

We consider three pieces:

1. Concentration and clean events
   - Error bound of reward estimate
   - Error bound of transition estimate
   - Optimism of Q and V estimates
2. Regret decomposition – bounding pieces
3. Bounds on partition size, bonus terms

The ultimate regret bound follows by algebraic manipulation using the pieces/lemmas above.

Concentration and Clean Events

Reward Estimate

Lemma. With probability at least $1-\delta$, for all h, k, $B \in \mathcal{P}_h^k$:

$$|\overline{\mathbf{r}}_h^k(B) - r_h(x,a)| \leq \mathrm{RUCB}_h^k(B).$$

Proof Sketch:

1. Rewrite using definition of $\overline{\mathbf{r}}$.
2. Subtract and add $r_h(X_h^{k'}, A_h^{k'})$.
3. Triangle inequality.
4. First term: Azuma-Hoeffding.
5. Second term: Lipschitz continuity of r.

Transition Estimate

Lemma. With probability at least $1 - 2\delta$, for all h, k, $B \in \mathcal{P}_h^k$ with $(x, a) \in B$:

$$d_W\left(\overline{\mathbf{T}}_h^k(\cdot \mid B), T_h^k(\cdot \mid x, a)\right) \leq \frac{1}{L_V} \mathrm{TUCB}_h^k(B).$$

d_W: Wasserstein metric

Optimism

Lemma. With probability at least $1 - 3\delta$, for all k, h, $\mathcal{P}_h^k$:

$$\begin{cases} \forall B \in \mathcal{P}_h^k, (x, a) \in B\text{:} & \overline{\mathbf{Q}}_h^k(B) \geq Q_h^*(x, a) \\ \forall A \in \mathcal{S}(\mathcal{P}_h^k), x \in A\text{:} & \tilde{\mathbf{V}}_h^k(A) \geq V_h^*(x) \\ \forall x \in \mathcal{S}\text{:} & \overline{\mathbf{V}}_h^k(x) \geq V_h^*(x). \end{cases}$$

Proof:

1. Forward induction on k, backward induction on H.
2. Use previous Lemmas on reward and transition estimate in the induction step.

Note: When a ball B splits, it retains any optimistic properties of its ancestors.

Regret Decomposition

Regret Decomposition

Lemma. The *expected* ($\star$) regret for Ada-MB can be decomposed as:

$$
\begin{aligned}
\mathbb{E}[R(K)] \lesssim & \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\left[\tilde{\mathbf{V}}_h^{k-1}(\mathcal{S}(\mathcal{P}_h^{k-1}, X_h^k) - \tilde{\mathbf{V}}_h^{k}(\mathcal{S}(\mathcal{P}_h^{k}, X_h^k))\right] \\
& + \sum_{h=1}^{H}\sum_{k=1}^{K}\mathbb{E}\left[2RUCB_h^k(B_h^k)\right] \\
& + \sum_{h=1}^{H}\sum_{k=1}^{K}\mathbb{E}\left[2TUCB_h^k(B_h^k)\right] \\
& + \sum_{k=1}^{K}\sum_{h=1}^{H}L_V\mathbb{E}\left[\mathcal{D}(B_h^k)\right].
\end{aligned}
$$

Regret Decomposition

Note: ($\star$) It suffices to bound the *expected* regret to achieve a bound on the regret – consider the “worst-case scenario” for each term under expectation.

The ultimate regret bound will follow from the expected regret bound via Azuma-Hoeffding.

Partition Size, Bonus Terms

Bounds on Partition Size

Lemma.

- Partition: $\mathcal{P}_h^k$ for any k, h, with splitting threshold $n_+(\ell)$.
- Penalty vector: $\{a_\ell\}_{\ell \in \mathbb{N}_0}$
  - $a_{\ell+1} \geq a_\ell \geq 0$
  - $\forall \ell \in \mathbb{N}_0$: $2a_{\ell+1}/a_\ell \leq n_+(\ell)/n_+(\ell-1)$
- $\ell^* := \inf\{\ell \,|\, 2^{d(\ell-1)} n_+(\ell-1) \geq k\}$

Then

$$\sum_{\ell=0}^{\infty} \sum_{B \in \mathcal{P}_h^k; \ell(B)=\ell} a_\ell \leq 2^{d\ell^*} \alpha_{\ell^*}$$

Proof of Lemma

1. Let x_ℓ be the number of active balls at level ℓ. Solve the LP:

$$\begin{aligned}
\max \ & \sum_{\ell=0}^{\infty} a_\ell x_\ell \\
\text{s.t.} \ & \sum_{\ell} 2^{-\ell d} x_\ell \leq 1 \\
& \sum_{\ell} n_+(\ell - 1) 2^{-d} x_\ell \leq k \\
& \forall \ell\colon x_\ell \geq 0.
\end{aligned}$$

Proof cont.

1. Use the weak duality theorem: optimal value is $\leq \alpha + \beta$ when: $2^{-\ell d}\alpha + n_+(\ell - 1)2^{-d}\beta \geq \alpha_\ell$.

2. Set: $\hat{\alpha} = \frac{2^{d\ell^*} a_{\ell*}}{2}$, $\hat{\beta} = \frac{2^d a_{\ell*}}{2n_+(\ell^*-1)}$.

- Satisfy constraints – check.
- $\hat{\alpha} + \hat{\beta}$ gives desired value of objective – check.

Worst-Case Partition Size

Corollary. For any h, k:

- $\ell^* \leq \frac{1}{d+\gamma} \log_2(k/\phi) + 2$
- $|\mathcal{P}_h^k| \leq 4^d (k/\phi)^{d/(d+\gamma)}$.

Proof Sketch.

1. Take $\alpha_\ell = 1$ for all ℓ as an upper bound.
2. ℓ^* definition $\Rightarrow 2^{d(\ell^*-2)} n_+(\ell^* + 2) \leq k$
3. Use definition of splitting rule $\Rightarrow$ two results in corollary.

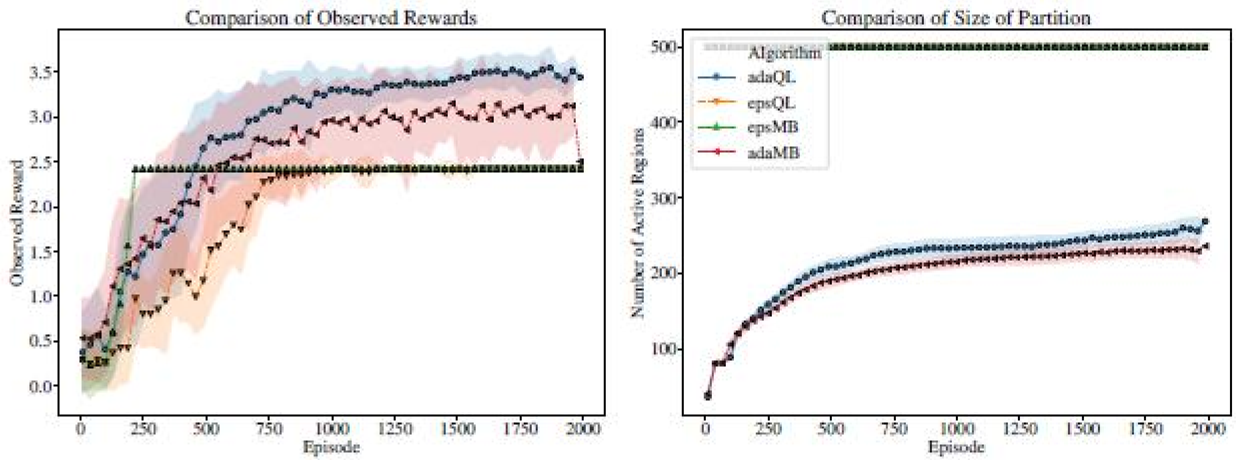
This corollary is used to achieve bounds in regret expansion.

Experiments

Oil Discovery

- Setup: similar to “Grid World”
  - agent surveys 1D map – survey function gives prob. of striking oil
  - cost depends on distance traveled

Results

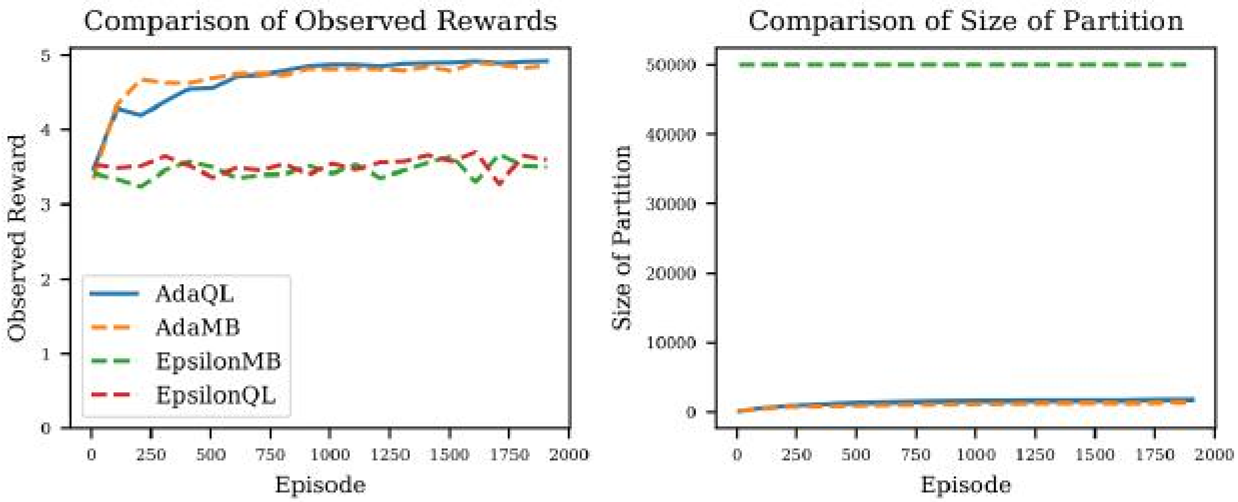


- Model-free QL has longer run-time than model-based
- Adaptive algorithms have smaller partition size
  - reduces unnecessary exploration
  - learner achieves optimal policy faster

Ambulance Routing

- k ambulances over H hours
- want to minimize cost of transportation, response time to emergencies
- 911 call drawn from prob. distribution
- action - dispatch one ambulance, reposition others

Results



- adaptive algorithms give smaller partition
- model free a bit more efficient, *but* has finer partition

Further Work

- Model based vs. model free in continuous settings
  - How to eliminate dependence of MB on dimension of $\mathcal{S} \times \mathcal{A}$?
- Gap-dependent analysis –> can model-based outperform model-free?

Thank you!