Safety in Sequential Decision-making

Mohammad Ghavamzadeh
Outline

Safety: Problem Formulation

Different Approaches to Safety

Risk-sensitive Decision-making (optional)

Safety w.r.t. Undesirable Situations (optional)
Outline

Safety: Problem Formulation

Different Approaches to Safety
  Model-free Approach
  Model-based Approach
  Online Approach

Risk-sensitive Decision-making (optional)

Safety w.r.t. Undesirable Situations (optional)
Safety: Problem Formulation

Disaster, Bankruptcy, Death...

Data → Machine Learning Algorithm → \( \pi \) → Manager
Safety: Problem Formulation

Data → Machine Learning Algorithm → $\pi$ → Disaster, Bankruptcy, Death...

Baseline performance $V_b$ → Confidence level $1 - \delta$

M. Ghavamzadeh – Safety in Sequential Decision-making
Different Approaches to Safety

Safety: Problem Formulation

Different Approaches to Safety

Model-free Approach
Model-based Approach
Online Approach

Risk-sensitive Decision-making (optional)

Safety w.r.t. Undesirable Situations (optional)
1. Model-free Approach
Offline Setting

Different Approaches to Safety

- Baseline performance $V_b$
- Confidence level $1 - \delta$

Behavior policy $\pi_b$ to Batch of Data

Decision $\pi$ to Manager

Disaster, Bankruptcy, Death...
Model-free Approach

Different Approaches to Safety

- baseline performance $V_b$
- Confidence level $1 - \delta$
- Disaster, Bankruptcy, Death...

behavior policy $\pi_b$ → Batch of Data → Compute Directly from the Batch of Data → $\pi$ → MANAGER
Model-free Approach

Different Approaches to Safety

behavior policy $\pi_b$ → Batch of Data → Compute Directly from the Batch of Data → $\pi$ → Disaster, Bankruptcy, Death...

1. historical data $D$ → $V_{\pi}$ → $V_{\pi} \geq V_b$
2. new policy $\pi$ → $V_b$
3. baseline performance $V_b$
4. confidence level $1 - \delta$

Yes / No
Model-free Approach

Different Approaches to Safety

behavior policy $\pi_b$ \rightarrow \text{Batch of Data} \rightarrow \text{Compute Directly from the Batch of Data} \rightarrow \pi$

baseline performance $V_b$ \rightarrow \text{Disaster, Bankruptcy, Death...}

Confidence level $1 - \delta$

historical data $D$ \rightarrow

new policy $\pi$ \rightarrow $V_\pi \geq V_b$

baseline performance $V_b$

confidence level $1 - \delta$

Risk Plot

M. Ghavamzadeh – Safety in Sequential Decision-making
Publications


2. P. Thomas, G. Theocharous, and MGH. "High Confidence Policy Improvement". *ICML-2015*.

Other Publications


2. Model-based Approach
Offline Setting

Different Approaches to Safety

Baseline performance $V_b$
Confidence level $1 - \delta$

Disaster, Bankruptcy, Death...

Behavior policy $\pi_b$ → Batch of Data

Manager

M. Ghavamzadeh – Safety in Sequential Decision-making
Model-based Approach

Different Approaches to Safety

behavior policy $\pi_b$ → Batch of Data → Build a Simulator of the System → Compute Directly from the Simulator → $\pi$ → MANAGER

$V_b$, baseline performance

1 - $\delta$, Confidence level

Disaster, Bankruptcy, Death...
Main Question: Given the simulator and the error in building it, how to compute a policy that is guaranteed (with a given confidence level) to perform at least as well as a baseline???
Publications

3. Online Approach
Online Approach

- Data traffic
- %α
- Our policy
- π

- Company's policy
- % (1 - α)
- π₀

- Manager

Different Approaches to Safety

M. Ghavamzadeh – Safety in Sequential Decision-making
Online Approach

Data traffic:
- $\%\alpha$ of the traffic handled by our policy $\pi$.
- $(1-\alpha)$ of the traffic handled by the company's policy $\pi_b$.

Manager's thoughts:
- Loss of handling $\%\alpha$ of the traffic by $\pi$ instead of $\pi_b$.

M. Ghavamzadeh – Safety in Sequential Decision-making
Online Approach

- Data traffic
  - $\% \alpha$
  - $\% (1 - \alpha)$

- Our policy: $\pi$
- Company's policy: $\pi_b$

- Loss of handling $\% \alpha$ of the traffic by $\pi$ instead of $\pi_b$?

- If loss $\geq \epsilon$
  - Panic button
Online Approach

Different Approaches to Safety

- Our policy
  - %α
  - π
  - data traffic
    - %α
    - % (1 - α)
  - Our policy
  - πb
  - Company's policy
    - %α
  - Company's policy
  - MANAGER

- loss of handling %α of the traffic by π instead of πb?

if loss ≥ ε
  - panic
if loss < ε
  - %α

M. Ghavamzadeh – Safety in Sequential Decision-making
Publications


Other Publications

Outline

Safety: Problem Formulation

Different Approaches to Safety
  Model-free Approach
  Model-based Approach
  Online Approach

Risk-sensitive Decision-making (optional)

Safety w.r.t. Undesirable Situations (optional)
Model-free Approach
Model-free Approach

Different Approaches to Safety  Model-free Approach

```
behavior policy \rightarrow \pi_b \rightarrow \text{Batch of Data} \rightarrow \text{Compute Directly from the Batch of Data} \rightarrow \pi
```

Disaster, Bankruptcy, Death...

```
historical data \rightarrow \mathcal{D} \rightarrow V_{\pi} \geq V_{b} \rightarrow \text{Risk Plot}
```

Yes / No

```
1 - \delta \rightarrow V_{b}
```

M. Ghavamzadeh – Safety in Sequential Decision-making
Problem Definition

new policy $\pi$ has been computed using a finite number of samples

- Is it \textit{safe} to deploy policy $\pi$?
- Is $\pi$ guaranteed to perform at least as well as $\pi_b$?
- Is the performance of $\pi$ guaranteed to be at least $V_b$?
new policy $\pi$ has been computed using a finite number of samples

- Is it \textit{safe} to deploy policy $\pi$?
- Is $\pi$ guaranteed to perform at least as well as $\pi_b$?
- Is the performance of $\pi$ guaranteed to be at least $V_b$?

\textbf{an important problem in many different fields including marketing, health, and finance}
High-Confidence Off-Policy Evaluation
Problem Formulation

- **System Trajectory**

  \[ \tau = \{x_1, a_1, r_1, x_2, a_2, r_2, \ldots, s_T, a_T, r_T \} \quad r_t = r(x_t, a_t) \]

- **Return of a Trajectory** (assume rewards are in \([0,1]\))

  \[ D(\tau) = \sum_{t=1}^{T} \gamma^{t-1} r_t \in [0, 1/(1-\gamma)] \]

- **Policy Performance**

  \[ V^\pi = \mathbb{E}[D(\tau)] \in [0, 1/(1-\gamma)] \quad \text{if} \quad a_t \sim \pi(\cdot | x_t) \]
Problem Formulation

- **Historical Data**

$$\mathcal{D} = \left\{ (\tau_i, \pi_i) \right\}_{i=1}^{n}$$

$\tau_i$ has been generated by $\pi_i$
Problem Formulation

- **Historical Data**
  \[ D = \left\{ (\tau_i, \pi_i) \right\}_{i=1}^{n} \]
  \( \tau_i \) has been generated by \( \pi_i \)

- **Behavior Policies** \( \pi_1, \ldots, \pi_n \)

  **Target Policy** \( \pi \)
Problem Formulation

► **Historical Data**

\[ D = \left\{ (\tau_i, \pi_i) \right\}_{i=1}^{n} \]

\( \tau_i \) has been generated by \( \pi_i \)

► **Behavior Policies** \( \pi_1, \ldots, \pi_n \)

► **Target Policy** \( \pi \)

► **Baseline Performance** \( V_b \)
Problem Formulation

- **Historical Data**
  \[ \mathcal{D} = \{ (\tau_i, \pi_i) \}_{i=1}^{n} \]
  \( \tau_i \) has been generated by \( \pi_i \)

- **Behavior Policies** \( \pi_1, \ldots, \pi_n \)
  **Target Policy** \( \pi \)

- **Baseline Performance** \( V_b \)

---

Risk Plot

[Diagram showing decision-making process]

- historical data \( \mathcal{D} \)
- new policy \( \pi \)
- baseline performance \( V_b \)
- confidence level \( 1 - \delta \)
- decision \( V^\pi \geq V_b \)
Weighted Importance Return

For any \((\tau_i, \pi_i) \in \mathcal{D}\)

\[
\hat{D}(\tau_i, \pi, \pi_i) = D(\tau_i) \frac{\operatorname{Pr}(\tau_i \mid \pi)}{\operatorname{Pr}(\tau_i \mid \pi_i)} = D(\tau_i) \prod_{t=1}^{T} \frac{\pi(\alpha_t \mid x_t)}{\pi_i(\alpha_t \mid x_t)}
\]
Weighted Importance Return

For any \((\tau_i, \pi_i) \in D\)

\[
\hat{D}(\tau_i, \pi, \pi_i) = D(\tau_i) \frac{\Pr(\tau_i | \pi)}{\Pr(\tau_i | \pi_i)} = D(\tau_i) \left( \prod_{t=1}^{T} \frac{\pi(a_t | x_t)}{\pi_i(a_t | x_t)} \right)
\]

For each \(\pi_i\), \(\hat{D}(\tau_i, \pi, \pi_i)\) is a random variable

(by generating a trajectory \(\tau_i\))
Weighted Importance Return

$\hat{D}(\tau_i, \pi, \pi_i)$ is a random variable such that

- $\mathbb{E}[\hat{D}(\tau_i, \pi, \pi_i)] \leq V^\pi$
- if $\forall x, a \ni \pi_i(a|x) = 0 \implies \pi(a|x) = 0$ then $\mathbb{E}[\hat{D}(\tau_i, \pi, \pi_i)] = V^\pi$
- $\hat{D}(\tau_i, \pi, \pi_i) \geq 0$ and $\mathbb{E}[\hat{D}(\tau_i, \pi, \pi_i)] \in [0, 1/1 - \gamma]$
- $\hat{D}(\tau_i, \pi, \pi_i)$ may have a very large upper-bound (very long tail)
Weighted Importance Return

\( \hat{D}(\tau_i, \pi, \pi_i) \) is a random variable such that

- \( \mathbb{E}[\hat{D}(\tau_i, \pi, \pi_i)] \leq V^\pi \)
- if \( \forall x, a \ni \pi_i(a | x) = 0 \implies \pi(a | x) = 0 \) then \( \mathbb{E}[\hat{D}(\tau_i, \pi, \pi_i)] = V^\pi \)
- \( \hat{D}(\tau_i, \pi, \pi_i) \geq 0 \) and \( \mathbb{E}[\hat{D}(\tau_i, \pi, \pi_i)] \in [0, 1/1 - \gamma] \)
- \( \hat{D}(\tau_i, \pi, \pi_i) \) may have a very large upper-bound

![Probability Density](image)

- two policies in the Mountain Car problem
- \( T = 20 \)
- PDF is estimated from 100,000 trajectories
- sample mean = 0.191
- maximum observed WIR = 316
- upper-bound on WIR = \( 10^{9.4} \)
Problem Formulation

Given the data set $\mathcal{D} = \{(\tau_i, \pi_i)\}_{i=1}^{n}$, we have

- $n$ independent bounded random variables

  \[ X_1 = \hat{D}(\tau_1, \pi, \pi_1) \quad X_2 = \hat{D}(\tau_2, \pi, \pi_2) \quad \ldots \quad X_n = \hat{D}(\tau_n, \pi, \pi_n) \]

- for all $i \in \{1, \ldots, n\}$,

  \[ \Pr(X_i \in [0, b_i]) = 1 \quad \text{and} \quad \mathbb{E}[X_i] = \mu \leq V^\pi \]
Problem Formulation

Given the data set $D = \{(\tau_i, \pi_i)\}_{i=1}^{n}$, we have

- $n$ independent bounded random variables

$$X_1 = \hat{D}(\tau_1, \pi, \pi_1) \quad X_2 = \hat{D}(\tau_2, \pi, \pi_2) \quad \ldots \quad X_n = \hat{D}(\tau_n, \pi, \pi_n)$$

- for all $i \in \{1, \ldots, n\}$,

$$\Pr(X_i \in [0, b_i]) = 1 \quad \text{and} \quad \mathbb{E}[X_i] = \mu \leq V^\pi$$

**Main Objective:** to derive a **tight** high probability (w.p. $\geq 1 - \delta$) lower-bound $\mu_-$ on $\mu$

*(note that $b_i$'s can be very large)*
Problem Formulation

\[ \text{if } \mu_- \geq V_b, \text{ then so does } V^\pi (w.p. \geq 1 - \delta) \]

otherwise

Yes

no answer

---

historical data \[ \xrightarrow{D} \]
new policy \[ \xrightarrow{\pi} V_b \]
baseline performance \[ \xrightarrow{V_b} 1 - \delta \]
confidence level

\[ V^\pi \geq V_b \]

Risk Plot

M. Ghavamzadeh – Safety in Sequential Decision-making
Problem Formulation

if $\mu \geq V_b$, then so does $V^\pi$ (w.p. $\geq 1 - \delta$)

otherwise

Yes

no answer

how to derive a **tight** lower-bound on $\mu$???

(note that $b_i$'s can be very large)
Concentration Inequalities

**Chernoff-Hoeffding (CH)**

\[
\mu \geq \frac{1}{n} \sum_{i=1}^{n} X_i - b \sqrt{\frac{\log(1/\delta)}{2n}} \quad \text{w.p. } \geq 1 - \delta
\]

**Maurer-Pontil empirical Bernstein (MPeB)** (Maurer & Pontil, 2009)

\[
\mu \geq \frac{1}{n} \sum_{i=1}^{n} X_i - \frac{7b \log(2/\delta)}{3(n-1)} - \frac{1}{n} \sqrt{\frac{\log(2/\delta)}{n-1} \sum_{i,j=1}^{n} (X_i - X_j)^2} \quad \text{w.p. } \geq 1 - \delta
\]

**Anderson-Massart (AM)** (Anderson, 1969; Massart, 1990)

\[
\mu \geq z_n - \sum_{i=0}^{n-1} (z_{i+1} - z_i) \min \left\{ 1, \frac{i}{n} + \sqrt{\frac{\log(2/\delta)}{2n}} \right\} \quad \text{w.p. } \geq 1 - \delta
\]

\[z_1 \leq z_2 \leq \ldots \leq z_n\] are sorted samples of the random variables \(X_1, \ldots, X_n\)
Concentration Inequalities

Chernoff-Hoeffding (CH)

\[ \mu \geq \frac{1}{n} \sum_{i=1}^{n} X_i - b \sqrt{\frac{\log(1/\delta)}{2n}} \quad \text{w.p. } \geq 1 - \delta \]

Maurer-Pontil empirical Bernstein (MPeB) (Maurer & Pontil, 2009)

\[ \mu \geq \frac{1}{n} \sum_{i=1}^{n} X_i - \frac{7b \log(2/\delta)}{3(n-1)} - \frac{1}{n} \sqrt{\frac{\log(2/\delta)}{n-1} \sum_{i,j=1}^{n} (X_i - X_j)^2} \quad \text{w.p. } \geq 1 - \delta \]


\[ \mu \geq z_n - \sum_{i=0}^{n-1} (z_{i+1} - z_i) \min \left\{ 1, \frac{i}{n} + \sqrt{\frac{\log(2/\delta)}{2n}} \right\} \quad \text{w.p. } \geq 1 - \delta \]

\[ z_1 \leq z_2 \leq \ldots \leq z_n \text{ are sorted samples of the random variables } X_1, \ldots, X_n \]
Concentration Inequalities

**Chernoff-Hoeffding (CH)**

\[
\mu \geq \frac{1}{n} \sum_{i=1}^{n} X_i - b \sqrt{\frac{\log(1/\delta)}{2n}} \quad \text{w.p.} \geq 1 - \delta
\]

(for i.d. and i.i.d.)

**Maurer-Pontil empirical Bernstein (MPeB)**

\[
\mu \geq \frac{1}{n} \sum_{i=1}^{n} X_i - \frac{7b \log(2/\delta)}{3(n-1)} - \frac{1}{n} \sqrt{\frac{\log(2/\delta)}{n-1} \sum_{i,j=1}^{n} (X_i - X_j)^2} \quad \text{w.p.} \geq 1 - \delta
\]

(for i.d. and i.i.d.)

**Anderson-Massart (AM)**

\[
\mu \geq z_n - \sum_{i=0}^{n-1} (z_{i+1} - z_i) \min \left\{ 1, \frac{i}{n} + \sqrt{\frac{\log(2/\delta)}{2n}} \right\} \quad \text{w.p.} \geq 1 - \delta
\]

(only for i.i.d.)

\[z_1 \leq z_2 \leq \ldots \leq z_n\] are sorted samples of the random variables \(X_1, \ldots, X_n\)
Concentration Inequalities

Maurer-Pontil empirical Bernstein (MPeB) (Maurer & Pontil, 2009)

$$\mu \geq \frac{1}{n} \sum_{i=1}^{n} X_i - \frac{7b \log(2/\delta)}{3(n-1)} - \frac{1}{n} \sqrt{\frac{\log(2/\delta)}{n-1} \sum_{i,j=1}^{n} (X_i - X_j)^2} \quad \text{w.p.} \geq 1 - \delta$$
Concentration Inequalities

**Maurer-Pontil empirical Bernstein (MPeB) (Maurer & Pontil, 2009)**

\[ \mu \geq \frac{1}{n} \sum_{i=1}^{n} X_i - \frac{7b \log(2/\delta)}{3(n-1)} - \frac{1}{n} \sqrt{\frac{\log(2/\delta)}{n-1} \sum_{i,j=1}^{n} (X_i - X_j)^2} \quad \text{w.p.} \geq 1 - \delta \]

**New MPeB bound:** by collapsing the tails of the distributions of the random variables and then bounding the means of the new distributions **a from of Winsorization (Wilcox & Keselman, 2003)**
New MPeB Bound

Theorem

Let $X_1, \ldots, X_n$ be $n$ independent bounded \textit{(in $[0, b_i]$)} random variables with the same mean $\mathbb{E}[X_i] = \mu$. Let $Y_1, \ldots, Y_n$ such that $Y_i = \min\{X_i, c_i\}$ be bounded versions of $X_1, \ldots, X_n$ with fixed thresholds $c_i > 0$. Then \textit{w.p.} $\geq 1 - \delta$, we have

$$
\mu \geq \frac{1}{\left( \sum_{i=1}^{n} \frac{1}{c_i} \right)^{-1} \sum_{i=1}^{n} \frac{Y_i}{c_i}} - \frac{7n \log(2/\delta)}{3(n - 1)} - \left( \sum_{i=1}^{n} \frac{1}{c_i} \right)^{-1} \left( \frac{\log(2/\delta)}{n - 1} \sum_{i, j=1}^{n} \left( \frac{Y_i}{c_i} - \frac{Y_j}{c_j} \right)^2 \right).$
$$

- The first term is the empirical mean.
- The second term goes to zero as $1/n$.
- The third term goes to zero as $1/\sqrt{n}$.
New MPeB Bound

**Theorem**

Let $X_1, \ldots, X_n$ be $n$ independent bounded (in $[0, b_i]$) random variables with the same mean $\mathbb{E}[X_i] = \mu$. Let $Y_1, \ldots, Y_n$ such that $Y_i = \min\{X_i, c_i\}$ be bounded versions of $X_1, \ldots, X_n$ with fixed thresholds $c_i > 0$. Then $w.p. \geq 1 - \delta$, we have

$$\mu \geq \left( \sum_{i=1}^{n} \frac{1}{c_i} \right)^{-1} \sum_{i=1}^{n} \frac{Y_i}{c_i} - \left( \sum_{i=1}^{n} \frac{1}{c_i} \right)^{-1} \frac{7n \log(2/\delta)}{3(n - 1)}$$

The term in braces is the empirical mean. The term in the denominator goes to zero as $1/n$. The term in the last underbrace goes to zero as $1/\sqrt{n}$.

**how to select $c_i$'s independent of the realization of $X_i$'s??**
Selecting the Thresholds $c_i$'s

- dividing the data $\mathcal{D}$ into two parts $\mathcal{D}_{\text{pre}}$ and $\mathcal{D}_{\text{post}}$
  
  \hspace{1cm} (1/20 in $\mathcal{D}_{\text{pre}}$, 19/20 in $\mathcal{D}_{\text{post}}$)

- calculating the optimal values of $c_i$ by maximizing $\mu_-$ using $\mathcal{D}_{\text{pre}}$
  \hspace{1cm} (computing $c_i^\ast$'s)

- calculating $\mu_-$ with $c_i^\ast$'s using $\mathcal{D}_{\text{post}}$

- If $\mu_- \geq V_b$, return Yes, else return no answer
Experimental Results
Results - Mountain Car Problem

- two policies in the Mountain Car problem
- $T = 20$
- PDF is estimated from 100,000 trajectories
- sample mean = 0.191
- maximum observed WIR = 316
- upper-bound on WIR = $10^{9.4}$

95% confidence lower-bounds

<table>
<thead>
<tr>
<th></th>
<th>New MPeB</th>
<th>CH</th>
<th>MPeB</th>
<th>AM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_-$</td>
<td>0.154</td>
<td>-5,831,000</td>
<td>-129,703</td>
<td>0.055</td>
</tr>
</tbody>
</table>
Results - Personalized Ad Recommendation (simulated data)

- target policy = a RL policy
- performance of behavior policy = 0.0765
- performance of target policy = 0.086
- trajectories of size T=20
- 95% confidence lower-bound

2M traj (behavior)
95% lower-bound = 0.077
5M traj (behavior)
95% lower-bound = 0.0803
5M traj (behavior) + 1M traj (target)
95% lower-bound = 0.081
Improving the Lower-Bound

MPeB lower-bound
general - no assumption on the data

One-sided Student’s t-Test
assumes $\hat{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ is normally distributed (is true as $n \to \infty$ - CLT)

Bias Corrected and accelerated (BCa) Bootstrap: (Efron, 1987)
use bootstrapping to estimate the true distribution of $\hat{X}$ and then use it to produce a lower-bound
Improving the Lower-Bound

heavy upper-tail Gamma distribution

95% confidence lower-bound on the mean - 100,000 trials
Outline

Safety: Problem Formulation

Different Approaches to Safety
  Model-free Approach
  Model-based Approach
  Online Approach

Risk-sensitive Decision-making *(optional)*

Safety w.r.t. Undesirable Situations *(optional)*
Model-based Approach
Model-based Approach

Main Question: Given the simulator and the error in building it, how to compute a policy that is guaranteed \(\text{with a given confidence level}\) to perform at least as well as a baseline??
Safe Policy Improvement by Minimizing Robust Baseline Regret

Problem Definition *(Model Uncertainty)*

- True Dynamics

- Simulator

- Error Function

\[ \forall (x, a) \in X \times A : \| P^*(\cdot|x, a) - \hat{P}(\cdot|x, a) \|_1 \leq e(x, a) \]
Problem Definition *(Model Uncertainty)*

- **True Dynamics**

- **Simulator**

- **Error Function**

  \[ \forall (x, a) \in \mathcal{X} \times \mathcal{A} : \| P^*(\cdot | x, a) - \hat{P}(\cdot | x, a) \|_1 \leq e(x, a) \]

- **Uncertainty Set**

  \[ \Xi(\hat{P}, e) = \left\{ \xi : \mathcal{X} \times \mathcal{A} \to \Delta^\mathcal{X} : \right\}

  \[ \| \xi(\cdot | x, a) - \hat{P}(\cdot | x, a) \|_1 \leq e(x, a), \forall (x, a) \in \mathcal{X} \times \mathcal{A} \right\} \]
Problem Definition *(Safety)*

- Baseline Policy *(deterministic)*
- Return of a Policy $\pi$ in world $\xi$
- True Return of a Policy $\pi$
- True Optimal Policy
- Policy $\pi$ is safe if

$$\pi^* \in \arg \max_{\pi} V(\pi, P^*)$$

$$V(\pi, P^*) \geq V(\pi_b, P^*)$$
Problem Formulation 1

Solving the Simulator

\[ \pi_{\text{sim}} \in \arg \max_{\pi} V(\pi, \widehat{P}) \]
Problem Formulation I

Solving the Simulator

\[ \pi_{\text{sim}} \in \arg \max_{\pi} V(\pi, \hat{P}) \]

▶ no guarantee that \( \pi_{\text{sim}} \) is \textit{safe}

Theorem: Bound on Performance Loss of \( \pi_{\text{sim}} \)

\[ \Phi(\pi_{\text{sim}}) \triangleq V(\pi^*, P^*) - V(\pi_{\text{sim}}, P^*) \leq \frac{2\gamma R_{\text{max}}}{(1 - \gamma)^2} \| e \|_\infty \]
Problem Formulation II

Solving the Robust MDP

\[ \pi_0 \in \arg\max_{\pi} \min_{\xi \in \Xi} V(\pi, \xi) \] (1)
Problem Formulation II

Solving the Robust MDP

\[ \pi_0 \in \arg \max_{\pi} \min_{\xi \in \Xi} V(\pi, \xi) \] (1)

\[ \pi_R = \begin{cases} 
\pi_0 & \text{if } \min_{\xi \in \Xi} V(\pi_0, \xi) > \max_{\xi \in \Xi} V(\pi_b, \xi), \\
\pi_b & \text{otherwise.}
\end{cases} \]
Problem Formulation II

Solving the Robust MDP

\[
\pi_0 \in \arg\max_{\pi} \min_{\xi \in \Xi} V(\pi, \xi)
\]  

(1)

\[
\pi_R = \begin{cases} 
\pi_0 & \text{if } \min_{\xi \in \Xi} V(\pi_0, \xi) > \max_{\xi \in \Xi} V(\pi_b, \xi), \\
\pi_b & \text{otherwise.}
\end{cases}
\]

\[\pi_R\] is guaranteed to be \textit{safe}

Theorem: Bound on Performance Loss of \(\pi_R\)

\[
\Phi(\pi_R) \leq \min \left\{ \frac{2\gamma R_{\max}}{(1 - \gamma)^2} \left( \| e_{\pi^*} \|_{1, u_{\pi^*}}^* + \| e_{\pi_b} \|_{1, u_{\pi_b}^*} \right), \Phi(\pi_b) \right\}
\]
Our Proposed Formulation

Robust Policy Improvement

\[ \pi^*_S \in \arg \max_{\pi} \min_{\xi \in \Xi} \left( V(\pi, \xi) - V(\pi_b, \xi) \right) \]  \hspace{1cm} (2)
Our Proposed Formulation

Robust Policy Improvement

\[ \pi_S \in \arg \max_{\pi} \min_{\xi \in \Xi} \left( V(\pi, \xi) - V(\pi_b, \xi) \right) \]  \hspace{1cm} (2)

- \( \pi_S \) is guaranteed to be safe

- \( \pi_S \) can outperform \( \pi_R \) by an arbitrarily large margin
Comparison with the Robust Solution

Solving the Robust MDP

\[ \pi_0 \in \arg \max_{\pi} \min_{\xi \in \Xi} V(\pi, \xi) \]

\[ \pi_R = \begin{cases} 
\pi_0 & \text{if } \min_{\xi \in \Xi} V(\pi_0, \xi) > \max_{\xi \in \Xi} V(\pi_b, \xi), \\
\pi_b & \text{otherwise.} 
\end{cases} \]

Robust Policy Improvement

\[ \pi_S \in \arg \max_{\pi} \min_{\xi \in \Xi} \left( V(\pi, \xi) - V(\pi_b, \xi) \right) \]
Comparison with the Robust Solution

Solving the Robust MDP

$$\pi_0 \in \arg \max_{\pi} \min_{\xi \in \Xi} V(\pi, \xi)$$

$$\pi_R = \begin{cases} 
\pi_0 & \text{if } \min_{\xi \in \Xi} V(\pi_0, \xi) > \max_{\xi \in \Xi} V(\pi_b, \xi), \\
\pi_b & \text{otherwise.}
\end{cases}$$

Robust Policy Improvement

$$\pi_S \in \arg \max_{\pi} \min_{\xi \in \Xi} \left( V(\pi, \xi) - V(\pi_b, \xi) \right)$$

Robust Policy Improvement

$$\max_{\pi} \min_{\xi} \left( V(\pi, \xi) - V(\pi_b, \xi) \right) \geq \max_{\pi} \min_{\xi} V(\pi, \xi) - \max_{\xi} V(\pi_b, \xi)$$
Properties of Robust Policy Improvement Formulation

1. Policy Class

<table>
<thead>
<tr>
<th>Theorem</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimization problem (2) may not have solution in the space of deterministic policies and ignoring this may cause huge loss.</td>
</tr>
</tbody>
</table>
Properties of Robust Policy Improvement Formulation

1. Policy Class

Theorem
Optimization problem (2) may not have solution in the space of deterministic policies and ignoring this may cause huge loss.

2. Performance Bound

Theorem
\[
\Phi(\pi_S) \leq \min \left\{ \frac{2\gamma R_{\max}}{(1-\gamma)^2} \left( ||e_{\pi^*1,u_{\pi^*b}}|| + ||e_{\pi_b1,u_{\pi_b^*}}|| \right) , \Phi(\pi_b) \right\}
\]
Properties of Robust Policy Improvement Formulation

1. Policy Class

Theorem
Optimization problem (2) may not have solution in the space of deterministic policies and ignoring this may cause huge loss.

2. Performance Bound

Theorem
\[ \Phi(\pi_S) \leq \min \left\{ \frac{2\gamma R_{\text{max}}}{(1 - \gamma)^2} \left( \| e_{\pi^*}^{1,u_{\pi^*}} \| + \| e_{\pi_b}^{1,u_{\pi_b}} \| \right), \Phi(\pi_b) \right\} \]

3. Computational Complexity

Theorem
Optimization problem (2) is NP-hard.
Solutions to Robust Policy Improvement Problem

1. Exact Solution with Extra Information

**Assumption:** Markov chain induced by $\pi_b$ is known
Solutions to Robust Policy Improvement Problem

1. Exact Solution with Extra Information

Assumption: Markov chain induced by $\pi_b$ is known

$$\forall x \in X, \quad \tilde{P}(\cdot | x, \pi_b(x)) = P^*(\cdot | x, \pi_b(x))$$
Solutions to Robust Policy Improvement Problem

1. Exact Solution with Extra Information

**Assumption:** Markov chain induced by \( \pi_b \) is known

\[
\forall x \in \mathcal{X}, \quad \tilde{P}(\cdot | x, \pi_b(x)) = P^*(\cdot | x, \pi_b(x))
\]

\[
\forall x \in \mathcal{X}, \quad e(x, \pi_b(x)) = 0
\]
Solutions to Robust Policy Improvement Problem

1. Exact Solution with Extra Information

Assumption: Markov chain induced by $\pi_b$ is known

$$\forall x \in \mathcal{X}, \quad \hat{P}(\cdot | x, \pi_b(x)) = P^*(\cdot | x, \pi_b(x))$$

$$\forall x \in \mathcal{X}, \quad e(x, \pi_b(x)) = 0$$

$$\forall \xi \in \Xi(\hat{P}, e), \quad \forall x \in \mathcal{X}, \quad \xi(\cdot | x, \pi_b(x)) = \hat{P}(\cdot | x, \pi_b(x))$$
Solutions to Robust Policy Improvement Problem

1. Exact Solution with Extra Information

Assumption: Markov chain induced by $\pi_b$ is known

\[ \forall x \in \mathcal{X}, \quad \hat{P}(\cdot | x, \pi_b(x)) = P^*(\cdot | x, \pi_b(x)) \]

\[ \forall x \in \mathcal{X}, \quad e(x, \pi_b(x)) = 0 \]

\[ \forall \xi \in \Xi(\hat{P}, e), \quad \forall x \in \mathcal{X}, \quad \xi(\cdot | x, \pi_b(x)) = \hat{P}(\cdot | x, \pi_b(x)) \]

\[ \forall \xi \in \Xi(\hat{P}, e), \quad V(\pi_b, \xi) \text{ is fixed} \]
Solutions to Robust Policy Improvement Problem

1. Exact Solution with Extra Information

**Assumption:** Markov chain induced by $\pi_b$ is known

\[
\forall x \in \mathcal{X}, \quad \hat{P}(\cdot | x, \pi_b(x)) = P^*(\cdot | x, \pi_b(x))
\]

\[
\forall x \in \mathcal{X}, \quad e(x, \pi_b(x)) = 0
\]

\[
\forall \xi \in \Xi(\hat{P}, e), \ \forall x \in \mathcal{X}, \quad \xi(\cdot | x, \pi_b(x)) = \hat{P}(\cdot | x, \pi_b(x))
\]

\[
\forall \xi \in \Xi(\hat{P}, e), \quad V(\pi_b, \xi) \text{ is fixed}
\]

\[
\arg\max_{\pi \in \Pi_R} \arg\min_{\xi \in \Xi} \left( V(\pi, \xi) - V(\pi_b, \xi) \right) \rightarrow \arg\max_{\pi \in \Pi_R} \arg\min_{\xi \in \Xi} V(\pi, \xi)
\]

Robust MDP

M. Ghavamzadeh – Safety in Sequential Decision-making
Solutions to Robust Policy Improvement Problem

1. Exact Solution with Extra Information

Assumption: Markov chain induced by $\pi_b$ is known

$$\forall x \in \mathcal{X}, \quad \hat{P} \left( \cdot \mid x, \pi_b(x) \right) = P^* \left( \cdot \mid x, \pi_b(x) \right)$$

$$\forall x \in \mathcal{X}, \quad e(x, \pi_b(x)) = 0$$

$$\forall \xi \in \Xi(\hat{P}, e), \forall x \in \mathcal{X}, \quad \xi \left( \cdot \mid x, \pi_b(x) \right) = \hat{P} \left( \cdot \mid x, \pi_b(x) \right)$$

$$\forall \xi \in \Xi(\hat{P}, e), \quad V(\pi_b, \xi) \text{ is fixed}$$

$$\arg \max_{\pi \in \Pi_R} \min_{\xi \in \Xi} \left( V(\pi, \xi) - V(\pi_b, \xi) \right) \rightarrow \arg \max_{\pi \in \Pi_R} \min_{\xi \in \Xi} V(\pi, \xi) \quad \text{Robust MDP}$$
Solutions to Robust Policy Improvement Problem

2. A Heuristic Solution

**Assumption:** Simulator is accurate for all the actions suggested by $\pi_b$

$$\forall x \in \mathcal{X}, \quad \hat{P}(\cdot | x, \pi_b(x)) \approx P^*(\cdot | x, \pi_b(x))$$
Solutions to Robust Policy Improvement Problem

2. A Heuristic Solution

Assumption: Simulator is accurate for all the actions suggested by $\pi_b$

$$\forall x \in X, \quad \hat{P}(\cdot | x, \pi_b(x)) \approx P^*(\cdot | x, \pi_b(x))$$

solve the robust MDP

$$\arg\max_{\pi \in \Pi_R} \min_{\xi \in \Xi(\hat{P}, e)} V(\pi, \xi)$$

where in the uncertainty set $\Xi(\hat{P}, e)$, we force

$$\forall x \in X, \quad \hat{P}(\cdot | x, \pi_b(x)) = P^*(\cdot | x, \pi_b(x))$$
Experimental Results
Grid Problem

interaction with the system

<table>
<thead>
<tr>
<th>overall satisfaction</th>
<th>$x_{1,1}$</th>
<th>$x_{1,2}$</th>
<th>$x_{1,3}$</th>
<th>$\ldots$</th>
<th>$x_{1,11}$</th>
<th>$x_{1,12}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$x_{2,1}$</td>
<td>$x_{2,2}$</td>
<td>$x_{2,3}$</td>
<td>$\ldots$</td>
<td>$x_{2,11}$</td>
<td>$x_{2,12}$</td>
</tr>
<tr>
<td></td>
<td>$x_{3,1}$</td>
<td>$x_{3,2}$</td>
<td>$x_{3,3}$</td>
<td>$\ldots$</td>
<td>$x_{3,11}$</td>
<td>$x_{3,12}$</td>
</tr>
</tbody>
</table>

reward
-1 1 2 $\cdots$ 4 5

Actions: left right up down
Results - Grid Problem

**baseline policy**: optimal ignoring the row part of the state

**simulator**: built by samples from random policy

![Graph showing performance improvement over baseline with varying number of samples.](image)

- **EXP**: solving simulator
- **RWA**: reward-adjusted MDP
- **ROB**: robust MDP
- **RBC**: robust policy improvement
## Energy Arbitrage Problem

### Problem Description

**State:** charge level, capacity, price

**Action:** amount of energy to buy or sell

**Dynamics:**
- capacity degrades when energy is stored or retrieved
- model for charge level and capacity are *known*
- model for price is *unknown*
- price is discretized to 10 separate levels

![Histogram of Price State Index](image)
Results - Energy Arbitrage Problem

**Baseline policy**: optimal for the model in which price is discretized to 3 levels

---

**EXP**: solving simulator

**ROB**: robust MDP

**RBC**: robust policy improvement
Summary
Safe Policy Improvement

**Error Function**

\[ \forall (x, a) \in \mathcal{X} \times \mathcal{A} : \| P^*(\cdot | x, a) - \hat{P}(\cdot | x, a) \|_1 \leq e(x, a) \]

**Uncertainty Set**

\[ \Xi(\hat{P}, e) = \left\{ \xi : \mathcal{X} \times \mathcal{A} \rightarrow \Delta^\mathcal{X} : \| \xi(\cdot | x, a) - \hat{P}(\cdot | x, a) \|_1 \leq e(x, a), \forall (x, a) \in \mathcal{X} \times \mathcal{A} \right\} \]

**Robust Policy Improvement**

\[ \pi_S \in \arg \max_{\pi} \min_{\xi \in \Xi} \left( V(\pi, \xi) - V(\pi_b, \xi) \right) \]  \hspace{1cm} (3)

- \( \pi_S \) is guaranteed to be **safe**
- \( \pi_S \) is less conservative than the other solutions
- (3) may not have solution in the space of deterministic policies
- Optimization problem (3) is NP-hard
- Under an assumption, (3) can be solved with a polynomial algorithm
Outline

Safety: Problem Formulation

Different Approaches to Safety
  Model-free Approach
  Model-based Approach
  Online Approach

Risk-sensitive Decision-making (optional)

Safety w.r.t. Undesirable Situations (optional)
Online Approach
Online Approach

Different Approaches to Safety  Online Approach

loss of handling $\%\alpha$ of the traffic by $\pi$ instead of $\pi_b$?

if loss $\geq \varepsilon$

if loss $< \varepsilon$

Our policy

Company's policy

data traffic

$\pi$}

$\pi_b$

$\%(1 - \alpha)$

M. Ghavamzadeh – Safety in Sequential Decision-making
Conservative Contextual Linear Bandits

Contextual Linear Bandit

At each round $t$,

- Learner selects an action $a_t \in \mathcal{A}_t$ ($\mathcal{A}_t$ possibly infinite set)
- $a_t$ is associated with the feature vector $\phi_{a_t}^t \in \mathbb{R}^d$
Contextual Linear Bandit

At each round $t$,

- Learner selects an action $a_t \in A_t$ ($A_t$ possibly infinite set)
  - $a_t$ is associated with the feature vector $\phi_{a_t}^t \in \mathbb{R}^d$

- Learner observes a random reward
  \[ Y_t = \langle \theta^*, \phi_{a_t}^t \rangle + \eta_t \]
  - $\langle \theta^*, \phi_{a_t}^t \rangle = \mathbb{E}[Y_t] = r_{a_t}^t$ (expected reward)
  - $\eta_t$ is conditionally $\sigma$-sub-Gaussian noise
  - $\theta^* \in \mathbb{R}^d$ is the unknown parameter
Contextual Linear Bandit

At each round $t$,

- Learner selects an action $a_t \in A_t$ ($A_t$ possibly infinite set)
  - $a_t$ is associated with the feature vector $\phi_{a_t}^t \in \mathbb{R}^d$

- Learner observes a random reward
  \[
  Y_t = \langle \theta^*, \phi_{a_t}^t \rangle + \eta_t
  \]
  - $\langle \theta^*, \phi_{a_t}^t \rangle = \mathbb{E}[Y_t] = r_{a_t}^t$ (expected reward)
  - $\eta_t$ is conditionally $\sigma$-sub-Gaussian noise
  - $\theta^* \in \mathbb{R}^d$ is the unknown parameter

Assumption: $\|\theta^*\|_2 \leq B$ and $\|\phi_{a}^t\|_2 \leq D$ and $r_{a}^t \in [0, 1]$
Contextual Linear Bandit (*Main Objective*)

**Optimal Action**

$$a_t^* = \arg \max_{a \in A_t} \langle \theta^*, \phi^t_a \rangle$$

**Minimizing (pseudo)-Regret**

$$R_T = \sum_{t=1}^{T} \langle \theta^*, \phi^t_{a_t^*} \rangle - \sum_{t=1}^{T} \langle \theta^*, \phi^t_{a_t} \rangle$$
Conservative Contextual Linear Bandit

same as contextual linear bandit except
Conservative Contextual Linear Bandit

same as contextual linear bandit except

- there exists a baseline policy $\pi_b$ that at each round $t$
  - selects action $b_t \in A_t$
  - observes expected reward $r_{b_t}^t = \langle \theta^*, \phi_{b_t}^t \rangle$ ($r_{b_t}^t$ is known)
Conservative Contextual Linear Bandit

same as contextual linear bandit except

- there exists a baseline policy $\pi_b$ that at each round $t$
  - selects action $b_t \in A_t$
  - observes expected reward $r^t_{b_t} = \langle \theta^*, \phi^t_{b_t} \rangle$ ($r^t_{b_t}$ is known)

- **Performance Constraint:** At each round $t$, $\alpha \in (0, 1)$

\[
\sum_{i=1}^{t} r^i_{b_i} - \sum_{i=1}^{t} r^i_{a_i} \leq \alpha \sum_{i=1}^{t} r^i_{b_i}
\]

*cumulative loss*
Conservative Contextual Linear Bandit

same as contextual linear bandit except

- there exists a baseline policy $\pi_b$ that at each round $t$
  - selects action $b_t \in A_t$
  - observes expected reward $r_{b_t}^t = \langle \theta^*, \phi_{b_t}^t \rangle$ ($r_{b_t}^t$ is known)

- Performance Constraint: At each round $t$, $\alpha \in (0, 1)$

\[
\sum_{i=1}^{t} r_{b_i}^i - \sum_{i=1}^{t} r_{a_i}^i \leq \alpha \sum_{i=1}^{t} r_{b_i}^i, \quad \sum_{i=1}^{t} r_{a_i}^i \geq (1-\alpha) \sum_{i=1}^{t} r_{b_i}^i
\]

small $\alpha$ more conservative, large $\alpha$ less conservative
A Conservative Contextual Linear Bandit Algorithm

Given \( \alpha \) and \( r_{b_t}^t \), it should

minimize (pseudo)-regret

\[
R_T = \sum_{t=1}^{T} \langle \theta^*, \phi_{a_t^*}^t \rangle - \sum_{t=1}^{T} \langle \theta^*, \phi_{a_t}^t \rangle
\]

satisfy performance constraint

\[
\sum_{i=1}^{t} r_{a_i}^i \geq (1 - \alpha) \sum_{i=1}^{t} r_{b_i}^i
\]
A Conservative Contextual Linear Bandit Algorithm

At each round $t$, the CLUCB algorithm
A Conservative Contextual Linear Bandit Algorithm

At each round $t$, the CLUCB algorithm

- uses the previous observations and builds a confidence set $C_t$
  \[w.h.p. \text{ contains } \theta^*\]
A Conservative Contextual Linear Bandit Algorithm

At each round $t$, the CLUCB algorithm

- uses the previous observations and builds a **confidence set** $C_t$ *(w.h.p. contains $\theta^*$)*

- computes the **optimistic action**

$$a'_t = \arg\max_{a \in A_t} \max_{\theta \in C_t} \langle \theta, \phi^t_a \rangle$$
A Conservative Contextual Linear Bandit Algorithm

At each round $t$, the CLUCB algorithm

- uses the previous observations and builds a **confidence set** $C_t$ (w.h.p. contains $\theta^*$)

- computes the **optimistic action**

$$a'_t = \arg\max_{a \in A_t} \max_{\theta \in C_t} \langle \theta, \phi^t_a \rangle$$

- plays the **optimistic action** ($a_t = a'_t$), only if

$$\sum_{i \in S_{t-1}^\circ} r_{b_i}^i + \min_{\theta \in C_t} \left( \langle \theta, \phi^t_{a'_t} \rangle + \sum_{i \in S_{t-1}} \langle \theta, \phi^i_{a'_t} \rangle \right) \geq (1 - \alpha) \sum_{i=1}^t r_{b_i}^i$$

plays the **baseline action** ($a_t = b_t$), otherwise.
Construction of Confidence Sets (Abbasi-Yadkori et al., 2011)

At each round $t$,

given the observed data $\{(\phi_{a_i}^i, Y_i)\}_{i=1}^{S_t}$, CLUCB updates the confidence set as

$C_{t+1} = \{\theta \in \mathbb{R}^d : \|\theta - \hat{\theta}\|_{V_t} \leq \beta_{t+1}\}$

$\hat{\theta}_t = (\Phi_t \Phi_t^T + \lambda I)^{-1} \Phi_t Y_t$

$V_t = \lambda I + \Phi_t \Phi_t^T$

$\beta_{t+1} = \sigma \sqrt{d \log \left( \frac{1 + (|S_t| + 1)D^2/\lambda}{\delta} \right)} + \sqrt{\lambda B}$
Construction of Confidence Sets (Abbasi-Yadkori et al., 2011)

At each round $t$,

given the observed data $\{(\phi_{a_i}^t, Y_i)\}_{i=1}^{S_t}$. CLUCB updates the confidence set as

$$(S_t = \text{set of rounds we play optimistic})$$

$$C_{t+1} = \{\theta \in \mathbb{R}^d : \|\theta - \tilde{\theta}\|_{V_t} \leq \beta_{t+1}\}$$

\[
\begin{align*}
\tilde{\theta}_t &= (\Phi_t\Phi_t^T + \lambda I)^{-1}\Phi_t Y_t \\
V_t &= \lambda I + \Phi_t\Phi_t^T \\
\beta_{t+1} &= \sigma \sqrt{d \log \left(\frac{1+(S_t + 1)D^2/\lambda}{\delta}\right)} + \sqrt{\lambda B}
\end{align*}
\]

**Proposition:** For any $C_t$ and $\delta > 0$, we have $\mathbb{P}[\theta^* \in C_t, \forall t \in \mathbb{N}] \geq 1 - \delta$. 
Regret Analysis

**Assumption**
There exists $0 \leq \Delta_l \leq \Delta_h$ and $0 < r_l$ such that at each round $t$,

$$\Delta_l \leq \Delta_{b_t}^t = r_{a_{t}^*}^t - r_{b_t}^t \leq \Delta_h \quad \text{and} \quad r_l \leq r_{b_t}^t.$$

**Proposition**
The regret of CLUCB can be decomposed into two terms as follows:

$$R_T(\text{CLUCB}) \leq R_{S_T}(\text{LUCB}) + \left| S_T^c \right| \Delta_h$$
Regret Analysis

Theorem

With probability at least $1 - \delta$, CLUCB satisfies the performance constraint for all $t \in \mathbb{N}$ and has the regret bound

$$R_T(\text{CLUCB}) = O \left( \frac{d\sqrt{T} \log \left( \frac{DT}{\lambda \delta} \right)}{R_{ST}(\text{LUCB})} + \frac{K \Delta_i}{\alpha r_i} \right)$$

where

$$K = 1 + 114d^2 \frac{(B \sqrt{\lambda} + \sigma)^2}{\Delta_l + \alpha r_i} \left[ \log \left( \frac{62d(B \sqrt{\lambda} + \sigma)}{\sqrt{\delta(\Delta_l + \alpha r_i)}} \right) \right]^2$$
Experimental Results

\[ d = 4, \quad \lambda = 1, \quad \delta = 0.001, \quad \theta^* \sim \mathcal{N}(0, I_4), \quad \eta_t \sim \mathcal{N}(0, 1) \]

\# of arms = 100, baseline = 10\textsuperscript{th} best arm, averaged over 1,000 runs

CLUCB starts to play optimistically more quickly for larger values of \( \alpha \)
Experimental Results

\[ d = 4, \quad \lambda = 1, \quad \delta = 0.001, \quad \theta^* \sim \mathcal{N}(0, I_4), \quad \eta_t \sim \mathcal{N}(0, 1) \]
\[ \# \text{ of arms} = 100, \quad \text{baseline} = 10'\text{th best arm}, \quad \text{averaged over 1,000 runs} \]

![Graph 1](image1.png)

**per-step regret**

![Graph 2](image2.png)

**per-step regret at \( t = 4 \times 10^4 \)**

---

**Performance of CLUCB converges to that of LUCB more quickly for larger values of \( \alpha \)**
Summary
Conservative Contextual Linear Bandit

Given $\alpha$ and $r_{b_t}^t$, minimize (pseudo)-regret

$$R_T = \sum_{t=1}^{T} \langle \theta^*, \phi_{a_t}^t \rangle - \sum_{t=1}^{T} \langle \theta^*, \phi_{a_t}^t \rangle$$

s.t. the performance constraint

$$\forall t \in \{1, \ldots, T\} \quad \sum_{i=1}^{t} r_{a_i}^i - \sum_{i=1}^{t} r_{b_i}^i \leq \alpha \sum_{i=1}^{t} r_{b_i}^i$$
Outline

Safety: Problem Formulation

Different Approaches to Safety
  Model-free Approach
  Model-based Approach
  Online Approach

Risk-sensitive Decision-making (optional)

Safety w.r.t. Undesirable Situations (optional)
Risk-sensitive Decision-making *(optional)*

Risk-sensitive Decision-making *(brief overview)*
Risk-sensitive Decision-making

Policy $\pi$

Trajectory 1

Trajectory 2

Trajectory 3

Trajectory 4

Return
Risk-sensitive Decision-making

**Policy** $\pi$

- Trajectory 1
- Trajectory 2
- Trajectory 3
- Trajectory 4

![Graph showing probability distribution and return trajectories](image-url)
Risk-sensitive Decision-making

\[
\max_{\pi} \text{Mean}(D^{\pi})
\]
Risk-sensitive Decision-making

\[
\max_{\pi} \text{Mean}(D^\pi)
\]
\[
\text{s.t.} \quad \text{var}(D^\pi) \leq \beta
\]
Risk-sensitive Decision-making

\[
\max_{\pi} \text{Mean}(D^\pi) \quad \text{s.t.} \quad \text{var}(D^\pi) \leq \beta
\]

\[
\max_{\pi} \text{Mean}(D^\pi) \quad \text{s.t.} \quad \text{CVaR}_\alpha(D^\pi) \geq \beta
\]
Publications


2. Y. Chow and MGH. “Algorithms for CVaR Optimization in MDPs“. **NIPS-2014**.


8. B. Liu and MGH. “A Block Coordinate Ascent Algorithm for Mean-Variance Optimization“. **submitted**.
Tutorial on
Risk-averse Decision-making & Control

Marek Petrik and Mohammad Ghavamzadeh

AAAI-2017
Outline

Safety w.r.t. Undesirable Situations *(optional)*

Safety: Problem Formulation

Different Approaches to Safety
  Model-free Approach
  Model-based Approach
  Online Approach

Risk-sensitive Decision-making *(optional)*

Safety w.r.t. Undesirable Situations *(optional)*
Safety w.r.t. Undesirable Situations (optional)

Safety w.r.t. Undesirable Situations (brief overview)
Safe RL

CMDP Formulation

\[
\min_{\pi} \mathbb{E} \left[ \sum_{t=0}^{T-1} c(x_t, a_t) \mid x_0, \pi \right], \quad \text{s.t.} \quad \mathbb{E} \left[ \sum_{t=0}^{T-1} d(x_t) \mid x_0, \pi \right] \leq d_0
\]

Safe Policy Iteration (SPI)

1. finding the Lyapunov function

\[
\max_{\varepsilon: \mathcal{X} \to \mathbb{R}^+} ||\varepsilon||_1, \quad \text{s.t.} \quad \mathcal{T}_{d+\varepsilon}^\pi [L_k](x) = L_k(x), \forall x \in \mathcal{X}, \quad L_k(x_0) \leq d_0
\]

\[
L_k(x) = V_{d+\varepsilon}^\pi(x), \forall x \in \mathcal{X}
\]

2. policy evaluation

\[
V_k = V_c^\pi_k
\]

3. policy improvement

\[
\pi_{k+1} \in \arg \min_{\pi \in \mathcal{P}_{L_k}(x)} \mathcal{T}_c^\pi [V_k]
\]

\[
\mathcal{P}_{L_k}(x) = \{ \pi(\cdot|x) \mid \mathcal{T}_d^\pi [L_k](x) \leq L_k(x) \}
\]

(a) all \( \pi_k \)'s are safe, (b) \( \pi_{k+1} \) is no worse than \( \pi_k \), (c) SPI converges
| SPI / Large Penalty | Optimal / Small Penalty |
Thank you!!

Mohammad Ghavamzadeh
ghavamza@gmail.com     OR
mohammad.ghavamzadeh@inria.fr