Tutorial: Sample Complexity Lower Bounds

The non-i.i.d. case

Alessio Russo

September 2024

Boston University

Overview i

1. Introduction
2. Change of Measure (recap)
3. Detecting a change in a stream of data as quickly as possible
4. Best Policy Identification: Tabular Markov Decision Processes
5. Best Policy Identification: Linear Markov Decision Processes
6. Conclusions

Introduction

The non-i.i.d. case

- In general much harder to deal with compared to the i.i.d. case.
- For Markovian models it is possible to say something sometimes.
- Extending results to partially observable models is extremely challenging [Fuh03] (and still an open question in almost every case afaik).

The non-i.i.d. case

- In general much harder to deal with compared to the i.i.d. case.
- For Markovian models it is possible to say something sometimes.
- Extending results to partially observable models is extremely challenging [Fuh03] (and still an open question in almost every case afaik).

The non-i.i.d. case

- In general much harder to deal with compared to the i.i.d. case.
- For Markovian models it is possible to say something sometimes.
- Extending results to partially observable models is extremely challenging [Fuh03] (and still an open question in almost every case afaik).

Change of Measure (recap)

Change of Measure: recap

Relate the probability of an event under a measure to another measure. Consider two measures $\mathbb{P}_\nu, \mathbb{P}_{\nu'}$ and an event $\mathcal{E} \in \mathcal{F}_t$, where $\mathcal{F}_t = \sigma(X_1, \ldots, X_t)$:

$$\mathbb{P}_{\nu'}(\mathcal{E}) = \mathbb{E}_{\nu'}[\mathbf{1}_{\mathcal{E}}] = \mathbb{E}_\nu\left[\mathbf{1}_{\mathcal{E}} \underbrace{\frac{\mathrm{d}\mathbb{P}_{\nu'}(X_1, \ldots, X_t)}{\mathrm{d}\mathbb{P}_\nu(X_1, \ldots, X_t)}}_{=\exp(-Z_t)}\right] = \mathbb{E}_\nu\left[\mathbf{1}_{\mathcal{E}} \exp(-Z_t)\right].$$

We can derive the the inequalities from this relationship.

Change of Measure: recap

Relate the probability of an event under a measure to another measure. Consider two measures $\mathbb{P}_\nu, \mathbb{P}_{\nu'}$ and an event $\mathcal{E} \in \mathcal{F}_t$, where $\mathcal{F}_t = \sigma(X_1, \dots, X_t)$:

$$\mathbb{P}_{\nu'}(\mathcal{E}) = \mathbb{E}_{\nu'}[\mathbf{1}_\mathcal{E}] = \mathbb{E}_\nu \left[\mathbf{1}_\mathcal{E} \underbrace{\frac{\mathrm{d}\mathbb{P}_{\nu'}(X_1, \dots, X_t)}{\mathrm{d}\mathbb{P}_\nu(X_1, \dots, X_t)}}_{=\exp(-Z_t)} \right] = \mathbb{E}_\nu \left[\mathbf{1}_\mathcal{E} \exp(-Z_t) \right].$$

We can derive the the inequalities from this relationship.

Change of Measure: recap

Relate the probability of an event under a measure to another measure. Consider two measures $\mathbb{P}_\nu, \mathbb{P}_{\nu'}$ and an event $\mathcal{E} \in \mathcal{F}_t$, where $\mathcal{F}_t = \sigma(X_1, \dots, X_t)$:

$$\mathbb{P}_{\nu'}(\mathcal{E}) = \mathbb{E}_{\nu'}[\mathbf{1}_\mathcal{E}] = \mathbb{E}_\nu \left[\mathbf{1}_\mathcal{E} \underbrace{\frac{\mathrm{d}\mathbb{P}_{\nu'}(X_1, \dots, X_t)}{\mathrm{d}\mathbb{P}_\nu(X_1, \dots, X_t)}}_{= \exp(-Z_t)} \right] = \mathbb{E}_\nu \left[\mathbf{1}_\mathcal{E} \exp(-Z_t) \right].$$

We can derive the the inequalities from this relationship.

Change of Measure: recap - 1st low-level form

$$\mathbb{P}_{\nu'}(\mathcal{E}) = \mathbb{E}_{\nu}\left[\mathbf{1}_{\mathcal{E}} \exp(-Z_t)\right].$$

First low-level form.

$$\begin{aligned}\mathbb{P}_{\nu'}(\mathcal{E}) &= \mathbb{E}_{\nu}\left[\mathbf{1}_{\mathcal{E}} \exp(-Z_t)\right], \\ &\geq \mathbb{E}_{\nu}\left[\mathbf{1}_{\mathcal{E}} \exp(-Z_t)\mathbf{1}_{\{Z_t < x\}}\right], \\ &\geq e^{-x}\mathbb{E}_{\nu}\left[\mathbf{1}_{\mathcal{E}}\mathbf{1}_{\{Z_t < x\}}\right], \\ &= e^{-x}\mathbb{P}_{\nu}\left(\mathcal{E} \cap \{Z_t < x\}\right).\end{aligned}$$

Thus

First low-level form

$$\mathbb{P}_{\nu}\left(\mathcal{E} \cap \{Z_t < x\}\right) \leq e^{x}\mathbb{P}_{\nu'}(\mathcal{E}).$$

$$\mathbb{P}_{\nu'}(\mathcal{E}) = \mathbb{E}_{\nu}\left[\mathbf{1}_{\mathcal{E}} \exp(-Z_t)\right].$$

First low-level form.

$$\begin{aligned}\mathbb{P}_{\nu'}(\mathcal{E}) &= \mathbb{E}_{\nu}\left[\mathbf{1}_{\mathcal{E}} \exp(-Z_t)\right], \\ &\geq \mathbb{E}_{\nu}\left[\mathbf{1}_{\mathcal{E}} \exp(-Z_t)\mathbf{1}_{\{Z_t < x\}}\right], \\ &\geq e^{-x}\mathbb{E}_{\nu}\left[\mathbf{1}_{\mathcal{E}}\mathbf{1}_{\{Z_t < x\}}\right], \\ &= e^{-x}\mathbb{P}_{\nu}\left(\mathcal{E} \cap \{Z_t < x\}\right).\end{aligned}$$

Thus

First low-level form

$$\mathbb{P}_{\nu}\left(\mathcal{E} \cap \{Z_t < x\}\right) \leq e^{x}\mathbb{P}_{\nu'}(\mathcal{E}).$$

$$\mathbb{P}_{\nu'}(\mathcal{E}) = \mathbb{E}_{\nu}\left[\mathbf{1}_{\mathcal{E}} \exp(-Z_t)\right].$$

First low-level form.

$$\begin{aligned}
\mathbb{P}_{\nu'}(\mathcal{E}) &= \mathbb{E}_{\nu}\left[\mathbf{1}_{\mathcal{E}} \exp(-Z_t)\right], \\
&\geq \mathbb{E}_{\nu}\left[\mathbf{1}_{\mathcal{E}} \exp(-Z_t)\mathbf{1}_{\{Z_t < x\}}\right], \\
&\geq e^{-x}\mathbb{E}_{\nu}\left[\mathbf{1}_{\mathcal{E}}\mathbf{1}_{\{Z_t < x\}}\right], \\
&= e^{-x}\mathbb{P}_{\nu}\left(\mathcal{E} \cap \{Z_t < x\}\right).
\end{aligned}$$

Thus

First low-level form

$$\mathbb{P}_{\nu}\left(\mathcal{E} \cap \{Z_t < x\}\right) \leq e^{x}\mathbb{P}_{\nu'}(\mathcal{E}).$$

Change of Measure: recap - 1st low-level form

$$\mathbb{P}_{\nu'}(\mathcal{E}) = \mathbb{E}_{\nu}\left[\mathbf{1}_{\mathcal{E}} \exp(-Z_t)\right].$$

First low-level form.

$$\begin{aligned}
\mathbb{P}_{\nu'}(\mathcal{E}) &= \mathbb{E}_{\nu}\left[\mathbf{1}_{\mathcal{E}} \exp(-Z_t)\right], \\
&\geq \mathbb{E}_{\nu}\left[\mathbf{1}_{\mathcal{E}} \exp(-Z_t) \mathbf{1}_{\{Z_t < x\}}\right], \\
&\geq e^{-x} \mathbb{E}_{\nu}\left[\mathbf{1}_{\mathcal{E}} \mathbf{1}_{\{Z_t < x\}}\right], \\
&= e^{-x} \mathbb{P}_{\nu}\left(\mathcal{E} \cap \{Z_t < x\}\right).
\end{aligned}$$

Thus

First low-level form

$$\mathbb{P}_{\nu}\left(\mathcal{E} \cap \{Z_t < x\}\right) \leq e^{x} \mathbb{P}_{\nu'}(\mathcal{E}).$$

Change of Measure: recap - 2nd low-level form

$$\mathbb{P}_{\nu'}(\mathcal{E}) = \mathbb{E}_{\nu}\left[\mathbf{1}_{\mathcal{E}} \exp(-Z_t)\right].$$

Second low-level form. From the first one we have

$$\mathbb{P}_{\nu'}(\mathcal{E}) \geq e^{-x}\mathbb{P}_{\nu}\left(\mathcal{E} \cap \{Z_t < x\}\right).$$

Use the fact that $\max(0, \mathbb{P}(A) + \mathbb{P}(B) - 1) \leq \mathbb{P}(A \cap B) \leq \min(\mathbb{P}(A), \mathbb{P}(B))$:

$$\begin{aligned}\mathbb{P}_{\nu'}(\mathcal{E}) &\geq e^{-x}\left[\mathbb{P}_{\nu}\left(\mathcal{E}\right) + \mathbb{P}_{\nu}\left(Z_t < x\right) - 1\right], \\ &\geq e^{-x}\left[\mathbb{P}_{\nu}\left(\mathcal{E}\right) - \mathbb{P}_{\nu}\left(Z_t \geq x\right)\right].\end{aligned}$$

Thus

Second low-level form

$$\mathbb{P}_{\nu}\left(\mathcal{E}\right) \leq \mathbb{P}_{\nu}(Z_t \geq x) + e^{x}\mathbb{P}_{\nu'}(\mathcal{E}).$$

Change of Measure: recap - 2nd low-level form

$$\mathbb{P}_{\nu'}(\mathcal{E}) = \mathbb{E}_{\nu}\left[\mathbf{1}_{\mathcal{E}} \exp(-Z_t)\right].$$

Second low-level form. From the first one we have

$$\mathbb{P}_{\nu'}(\mathcal{E}) \geq e^{-x}\mathbb{P}_{\nu}\left(\mathcal{E} \cap \{Z_t < x\}\right).$$

Use the fact that $\max(0, \mathbb{P}(A) + \mathbb{P}(B) - 1) \leq \mathbb{P}(A \cap B) \leq \min(\mathbb{P}(A), \mathbb{P}(B))$:

$$\begin{aligned}\mathbb{P}_{\nu'}(\mathcal{E}) &\geq e^{-x}\left[\mathbb{P}_{\nu}\left(\mathcal{E}\right) + \mathbb{P}_{\nu}\left(Z_t < x\right) - 1\right], \\ &\geq e^{-x}\left[\mathbb{P}_{\nu}\left(\mathcal{E}\right) - \mathbb{P}_{\nu}\left(Z_t \geq x\right)\right].\end{aligned}$$

Thus

Second low-level form

$$\mathbb{P}_{\nu}\left(\mathcal{E}\right) \leq \mathbb{P}_{\nu}(Z_t \geq x) + e^{x}\mathbb{P}_{\nu'}(\mathcal{E}).$$

Change of Measure: recap - 2nd low-level form

$$\mathbb{P}_{\nu'}(\mathcal{E}) = \mathbb{E}_{\nu}\left[\mathbf{1}_{\mathcal{E}} \exp(-Z_t)\right].$$

Second low-level form. From the first one we have

$$\mathbb{P}_{\nu'}(\mathcal{E}) \geq e^{-x}\mathbb{P}_{\nu}\left(\mathcal{E} \cap \{Z_t < x\}\right).$$

Use the fact that $\max(0, \mathbb{P}(A) + \mathbb{P}(B) - 1) \leq \mathbb{P}(A \cap B) \leq \min(\mathbb{P}(A), \mathbb{P}(B))$:

$$\begin{aligned} \mathbb{P}_{\nu'}(\mathcal{E}) &\geq e^{-x}\left[\mathbb{P}_{\nu}\left(\mathcal{E}\right) + \mathbb{P}_{\nu}\left(Z_t < x\right) - 1\right], \\ &\geq e^{-x}\left[\mathbb{P}_{\nu}\left(\mathcal{E}\right) - \mathbb{P}_{\nu}\left(Z_t \geq x\right)\right]. \end{aligned}$$

Thus

Second low-level form

$$\mathbb{P}_{\nu}\left(\mathcal{E}\right) \leq \mathbb{P}_{\nu}(Z_t \geq x) + e^{x}\mathbb{P}_{\nu'}(\mathcal{E}).$$

Change of Measure: recap - high-level form

Start from the beginning

$$\mathbb{P}_{\nu'}(\mathcal{E}) = \mathbb{E}_{\nu}\left[\mathbf{1}_{\mathcal{E}} \exp(-Z_t)\right] = \mathbb{E}_{\nu}\left[\mathbf{1}_{\mathcal{E}} \mathbb{E}_{\nu}[\exp(-Z_t)|\mathcal{E}]\right],$$
$$\geq \mathbb{E}_{\nu}\left[\mathbf{1}_{\mathcal{E}} \exp\left(-\mathbb{E}_{\nu}[Z_t|\mathcal{E}]\right)\right],$$
$$= \exp\left(-\mathbb{E}_{\nu}[Z_t|\mathcal{E}]\right] \mathbb{P}_{\nu}(\mathbf{1}_{\mathcal{E}} = 1) + 0 \cdot \mathbb{P}_{\nu}(\mathbf{1}_{\mathcal{E}} = 0),$$
$$= \exp\left(-\mathbb{E}_{\nu}[Z_t|\mathcal{E}]\right) \mathbb{P}_{\nu}(\mathcal{E}).$$

Repeat the same for $\mathcal{E}^c$. Hence

$$\ln \frac{P_{\nu}(\mathcal{E})}{P_{\nu'}(\mathcal{E})} \leq \mathbb{E}_{\nu}[Z_t|\mathcal{E}] \text{ and } \ln \frac{P_{\nu}(\mathcal{E}^c)}{P_{\nu'}(\mathcal{E}^c)} \leq \mathbb{E}_{\nu}[Z_t|\mathcal{E}^c].$$

Conclude by lower bounding the terms in $\mathbb{E}_{\nu}[L_t] = \mathbb{E}_{\nu}[L_t|\mathcal{E}]\mathbb{P}_{\nu}(\mathcal{E}) + \mathbb{E}_{\nu}[L_t|\mathcal{E}^c]\mathbb{P}_{\nu}(\mathcal{E}^c)$.

High level form

$$\mathbb{E}_{\nu}[L_t] \geq \mathrm{kl}(\mathbb{P}_{\nu}(\mathcal{E}), \mathbb{P}_{\nu'}(\mathcal{E})) \text{ with } \mathrm{kl}(x, y) = x\ln(x/y) + (1-x)\ln((1-x)/(1-y)).$$

Change of Measure: recap - high-level form

Start from the beginning

$$\begin{aligned}\mathbb{P}_{\nu'}(\mathcal{E}) &= \mathbb{E}_\nu\left[\mathbf{1}_{\mathcal{E}}\exp(-Z_t)\right] = \mathbb{E}_\nu\left[\mathbf{1}_{\mathcal{E}}\mathbb{E}_\nu[\exp(-Z_t)|\mathcal{E}]\right],\\ &\geq \mathbb{E}_\nu\left[\mathbf{1}_{\mathcal{E}}\exp\left(-\mathbb{E}_\nu[Z_t|\mathcal{E}]\right)\right],\\ &= \exp\left(-\mathbb{E}_\nu[Z_t|\mathcal{E}]\right]\mathbb{P}_\nu(\mathbf{1}_{\mathcal{E}}=1) + 0\cdot\mathbb{P}_\nu(\mathbf{1}_{\mathcal{E}}=0),\\ &= \exp\left(-\mathbb{E}_\nu[Z_t|\mathcal{E}]\right)\mathbb{P}_\nu(\mathcal{E}).\end{aligned}$$

Repeat the same for $\mathcal{E}^c$. Hence

$$\ln\frac{P_\nu(\mathcal{E})}{P_{\nu'}(\mathcal{E})} \leq \mathbb{E}_\nu[Z_t|\mathcal{E}] \text{ and } \ln\frac{P_\nu(\mathcal{E}^c)}{P_{\nu'}(\mathcal{E}^c)} \leq \mathbb{E}_\nu[Z_t|\mathcal{E}^c].$$

Conclude by lower bounding the terms in $\mathbb{E}_\nu[L_t] = \mathbb{E}_\nu[L_t|\mathcal{E}]\mathbb{P}_\nu(\mathcal{E}) + \mathbb{E}_\nu[L_t|\mathcal{E}^c]\mathbb{P}_\nu(\mathcal{E}^c)$.

High level form

$$\mathbb{E}_\nu[L_t] \geq \mathrm{kl}(\mathbb{P}_\nu(\mathcal{E}), \mathbb{P}_{\nu'}(\mathcal{E})) \text{ with } \mathrm{kl}(x,y) = x\ln(x/y) + (1-x)\ln((1-x)/(1-y)).$$

Change of Measure: recap - high-level form

Start from the beginning

$$\begin{aligned}\mathbb{P}_{\nu'}(\mathcal{E}) &= \mathbb{E}_\nu\left[\mathbf{1}_\mathcal{E}\exp(-Z_t)\right] = \mathbb{E}_\nu\left[\mathbf{1}_\mathcal{E}\mathbb{E}_\nu[\exp(-Z_t)|\mathcal{E}]\right],\\ &\geq \mathbb{E}_\nu\left[\mathbf{1}_\mathcal{E}\exp\left(-\mathbb{E}_\nu[Z_t|\mathcal{E}]\right)\right],\\ &= \exp\left(-\mathbb{E}_\nu[Z_t|\mathcal{E}]\right]\mathbb{P}_\nu(\mathbf{1}_\mathcal{E}=1) + 0\cdot\mathbb{P}_\nu(\mathbf{1}_\mathcal{E}=0),\\ &= \exp\left(-\mathbb{E}_\nu[Z_t|\mathcal{E}]\right)\mathbb{P}_\nu(\mathcal{E}).\end{aligned}$$

Repeat the same for $\mathcal{E}^c$. Hence

$$\ln\frac{P_\nu(\mathcal{E})}{P_{\nu'}(\mathcal{E})} \leq \mathbb{E}_\nu[Z_t|\mathcal{E}] \text{ and } \ln\frac{P_\nu(\mathcal{E}^c)}{P_{\nu'}(\mathcal{E}^c)} \leq \mathbb{E}_\nu[Z_t|\mathcal{E}^c].$$

Conclude by lower bounding the terms in $\mathbb{E}_\nu[L_t] = \mathbb{E}_\nu[L_t|\mathcal{E}]\mathbb{P}_\nu(\mathcal{E}) + \mathbb{E}_\nu[L_t|\mathcal{E}^c]\mathbb{P}_\nu(\mathcal{E}^c)$.

High level form

$$\mathbb{E}_\nu[L_t] \geq \mathrm{kl}(\mathbb{P}_\nu(\mathcal{E}), \mathbb{P}_{\nu'}(\mathcal{E})) \text{ with } \mathrm{kl}(x,y) = x\ln(x/y) + (1-x)\ln((1-x)/(1-y)).$$

Change of Measure: recap - high-level form

Start from the beginning

$$\begin{aligned}\mathbb{P}_{\nu'}(\mathcal{E}) &= \mathbb{E}_{\nu}\left[\mathbf{1}_{\mathcal{E}} \exp(-Z_t)\right] = \mathbb{E}_{\nu}\left[\mathbf{1}_{\mathcal{E}} \mathbb{E}_{\nu}[\exp(-Z_t)|\mathcal{E}]\right], \\ &\geq \mathbb{E}_{\nu}\left[\mathbf{1}_{\mathcal{E}} \exp\left(-\mathbb{E}_{\nu}[Z_t|\mathcal{E}]\right)\right], \\ &= \exp\left(-\mathbb{E}_{\nu}[Z_t|\mathcal{E}]\right] \mathbb{P}_{\nu}(\mathbf{1}_{\mathcal{E}} = 1) + 0 \cdot \mathbb{P}_{\nu}(\mathbf{1}_{\mathcal{E}} = 0), \\ &= \exp\left(-\mathbb{E}_{\nu}[Z_t|\mathcal{E}]\right) \mathbb{P}_{\nu}(\mathcal{E}).\end{aligned}$$

Repeat the same for $\mathcal{E}^c$. Hence

$$\ln \frac{P_{\nu}(\mathcal{E})}{P_{\nu'}(\mathcal{E})} \leq \mathbb{E}_{\nu}[Z_t|\mathcal{E}] \text{ and } \ln \frac{P_{\nu}(\mathcal{E}^c)}{P_{\nu'}(\mathcal{E}^c)} \leq \mathbb{E}_{\nu}[Z_t|\mathcal{E}^c].$$

Conclude by lower bounding the terms in $\mathbb{E}_{\nu}[L_t] = \mathbb{E}_{\nu}[L_t|\mathcal{E}]\mathbb{P}_{\nu}(\mathcal{E}) + \mathbb{E}_{\nu}[L_t|\mathcal{E}^c]\mathbb{P}_{\nu}(\mathcal{E}^c)$.

High level form

$$\mathbb{E}_{\nu}[L_t] \geq \mathrm{kl}(\mathbb{P}_{\nu}(\mathcal{E}), \mathbb{P}_{\nu'}(\mathcal{E})) \text{ with } \mathrm{kl}(x, y) = x \ln(x/y) + (1-x)\ln((1-x)/(1-y)).$$

Change of Measure: recap - high-level form

Start from the beginning

$$\begin{aligned}\mathbb{P}_{\nu'}(\mathcal{E}) &= \mathbb{E}_{\nu}\left[\mathbf{1}_{\mathcal{E}}\exp(-Z_t)\right] = \mathbb{E}_{\nu}\left[\mathbf{1}_{\mathcal{E}}\mathbb{E}_{\nu}[\exp(-Z_t)|\mathcal{E}]\right],\\ &\geq \mathbb{E}_{\nu}\left[\mathbf{1}_{\mathcal{E}}\exp\left(-\mathbb{E}_{\nu}[Z_t|\mathcal{E}]\right)\right],\\ &= \exp\left(-\mathbb{E}_{\nu}[Z_t|\mathcal{E}]\right]\mathbb{P}_{\nu}(\mathbf{1}_{\mathcal{E}}=1)+0\cdot\mathbb{P}_{\nu}(\mathbf{1}_{\mathcal{E}}=0),\\ &= \exp\left(-\mathbb{E}_{\nu}[Z_t|\mathcal{E}]\right)\mathbb{P}_{\nu}(\mathcal{E}).\end{aligned}$$

Repeat the same for $\mathcal{E}^c$. Hence

$$\ln\frac{P_{\nu}(\mathcal{E})}{P_{\nu'}(\mathcal{E})} \leq \mathbb{E}_{\nu}[Z_t|\mathcal{E}] \text{ and } \ln\frac{P_{\nu}(\mathcal{E}^c)}{P_{\nu'}(\mathcal{E}^c)} \leq \mathbb{E}_{\nu}[Z_t|\mathcal{E}^c].$$

Conclude by lower bounding the terms in $\mathbb{E}_{\nu}[L_t] = \mathbb{E}_{\nu}[L_t|\mathcal{E}]\mathbb{P}_{\nu}(\mathcal{E}) + \mathbb{E}_{\nu}[L_t|\mathcal{E}^c]\mathbb{P}_{\nu}(\mathcal{E}^c)$.

High level form

$$\mathbb{E}_{\nu}[L_t] \geq \mathrm{kl}(\mathbb{P}_{\nu}(\mathcal{E}),\mathbb{P}_{\nu'}(\mathcal{E})) \text{ with } \mathrm{kl}(x,y) = x\ln(x/y)+(1-x)\ln((1-x)/(1-y)).$$

Change of Measure: recap - high-level form

Start from the beginning

$$\begin{aligned}\mathbb{P}_{\nu'}(\mathcal{E}) &= \mathbb{E}_{\nu}\left[\mathbf{1}_{\mathcal{E}}\exp(-Z_t)\right] = \mathbb{E}_{\nu}\left[\mathbf{1}_{\mathcal{E}}\mathbb{E}_{\nu}[\exp(-Z_t)|\mathcal{E}]\right],\\ &\geq \mathbb{E}_{\nu}\left[\mathbf{1}_{\mathcal{E}}\exp\left(-\mathbb{E}_{\nu}[Z_t|\mathcal{E}]\right)\right],\\ &= \exp\left(-\mathbb{E}_{\nu}[Z_t|\mathcal{E}]\right]\mathbb{P}_{\nu}(\mathbf{1}_{\mathcal{E}}=1) + 0\cdot\mathbb{P}_{\nu}(\mathbf{1}_{\mathcal{E}}=0),\\ &= \exp\left(-\mathbb{E}_{\nu}[Z_t|\mathcal{E}]\right)\mathbb{P}_{\nu}(\mathcal{E}).\end{aligned}$$

Repeat the same for $\mathcal{E}^c$. Hence

$$\ln\frac{P_{\nu}(\mathcal{E})}{P_{\nu'}(\mathcal{E})} \leq \mathbb{E}_{\nu}[Z_t|\mathcal{E}] \text{ and } \ln\frac{P_{\nu}(\mathcal{E}^c)}{P_{\nu'}(\mathcal{E}^c)} \leq \mathbb{E}_{\nu}[Z_t|\mathcal{E}^c].$$

Conclude by lower bounding the terms in $\mathbb{E}_{\nu}[L_t] = \mathbb{E}_{\nu}[L_t|\mathcal{E}]\mathbb{P}_{\nu}(\mathcal{E}) + \mathbb{E}_{\nu}[L_t|\mathcal{E}^c]\mathbb{P}_{\nu}(\mathcal{E}^c)$.

High level form

$$\mathbb{E}_{\nu}[L_t] \geq \mathrm{kl}(\mathbb{P}_{\nu}(\mathcal{E}), \mathbb{P}_{\nu'}(\mathcal{E})) \text{ with } \mathrm{kl}(x,y) = x\ln(x/y) + (1-x)\ln((1-x)/(1-y)).$$

Detecting a change in a stream of data as quickly as possible

Quickest Change Detection

We now look at a different problem, called Quickest Change Detection.

- Suppose you observe a stream of random variables $X_1, X_2, X_3 \ldots$.
  - The conditional density function of X_n is $f_0(X_n|X_1, \ldots, X_{n-1})$ for $n < \nu$ and $f_1(X_n|X_1, \ldots, X_{n-1})$ for $n \geq \nu$.
- ν is an unknown change-time.
  - For $\nu = 1, 2, \ldots$ we let $\mathbb{P}_\nu$ denote the probability measure of the sequence when $\nu < \infty$, and otherwise we denote it by $\mathbb{P}_\infty$.

Quickest Change Detection

We now look at a different problem, called Quickest Change Detection.

- Suppose you observe a stream of random variables $X_1, X_2, X_3 \ldots$.
  - The conditional density function of X_n is $f_0(X_n|X_1, \ldots, X_{n-1})$ for $n < \nu$ and $f_1(X_n|X_1, \ldots, X_{n-1})$ for $n \geq \nu$.
- ν is an unknown change-time.
  - For $\nu = 1, 2, \ldots$ we let $\mathbb{P}_\nu$ denote the probability measure of the sequence when $\nu < \infty$, and otherwise we denote it by $\mathbb{P}_\infty$.

Quickest Change Detection (cont.)

Hypothesis Testing Problem

$$H_0 : \text{ no change} \quad \text{vs} \quad H_1 : \text{ a change happened}$$

- Ideally, we want an algorithm with a certain false alarm rate (type I error), i.e.,

$$E_\infty[\tau] \geq \frac{1}{\alpha} \text{ with } \alpha > 0.$$

- Performance of a detection algorithm: worst average detection delay (WADD). Let τ be the stopping time of the algorithm (that tells you when to stop, i.e., a change was detected), then

$$\bar{E}(\tau) = \sup_{\nu \geq 1} \operatorname{ess\,sup} \mathbb{E}_\nu \left[(\tau - \nu)^+ | X_1, \ldots, X_{\nu-1} \right].$$

- Minimum number of samples τ needed to detect a change with a given false alarm rate?

Quickest Change Detection: lower bound

In the i.i.d. case the information rate[1] is

$$(T^\star)^{-1} = I^\star := \lim_{n\to\infty} \frac{1}{n} \sum_{t=\nu}^{\nu+n} \ln \frac{F_1(X_t)}{F_0(X_t)} = \mathrm{KL}(F_1, F_0).$$

To generalize non the non-i.i.d. setting, we require the following assumption.

Assumption (Bound on hte information rate)

Let $Z_n = \ln \frac{f_1(X_n|X_1,\dots,X_{n-1})}{f_0(X_n|X_1,\dots,X_{n-1})}$. *We assume that* $\exists I^\star > 0$ *such that*

$$\lim_{n\to\infty} \sup_{\nu\geq 1} \operatorname{ess\,sup} \mathbb{P}_\nu \left(\max_{t\leq n} \sum_{k=\nu}^{\nu+t} Z_k \geq I^\star(1+\delta)n \mid X_1,\dots,X_{\nu-1} \right) = 0 \quad \forall \delta > 0. \tag{1}$$

That is, there exists some $I^\star$ to which $n^{-1}\sum_{\nu\leq k\leq n+\nu} Z_k$ converges to in probability.

[1] The characteristic time is $T^\star = (I^\star)^{-1}$.

Quickest Change Detection: lower bound

In the i.i.d. case the information rate[1] is

$$(T^\star)^{-1} = I^\star := \lim_{n\to\infty} \frac{1}{n} \sum_{t=\nu}^{\nu+n} \ln \frac{F_1(X_t)}{F_0(X_t)} = \mathrm{KL}(F_1, F_0).$$

To generalize non the non-i.i.d. setting, we require the following assumption.

Assumption (Bound on hte information rate)

Let $Z_n = \ln \frac{f_1(X_n | X_1,\ldots,X_{n-1})}{f_0(X_n | X_1,\ldots,X_{n-1})}$. *We assume that* $\exists I^\star > 0$ *such that*

$$\lim_{n\to\infty} \sup_{\nu\geq 1} \operatorname{ess\,sup} \mathbb{P}_\nu \left(\max_{t\leq n} \sum_{k=\nu}^{\nu+t} Z_k \geq I^\star(1+\delta)n \mid X_1,\ldots,X_{\nu-1} \right) = 0 \quad \forall \delta > 0. \tag{1}$$

That is, there exists some $I^\star$ to which $n^{-1} \sum_{\nu\leq k\leq n+\nu} Z_k$ converges to in probability.

[1] The characteristic time is $T^\star = (I^\star)^{-1}$.

Quickest Change Detection: lower bound (cont.)

The idea is to show the following for any $\delta \in (0,1)$:

$$(\mathbf{P_1}) \lim_{\alpha \to 0} \mathbb{P}_\nu \left(\tau - \nu \le T^\star (1-\delta) \ln(1/\alpha), \sum_{n=\nu}^{\tau} Z_n < (1-\delta^2) \ln(1/\alpha) \mid \tau \ge \nu \right) = 0,$$

and

$$(\mathbf{P_2}) \lim_{\alpha \to 0} \mathbb{P}_\nu \left(\tau - \nu \le T^\star (1-\delta) \ln(1/\alpha), \sum_{n=\nu}^{\tau} Z_n \ge (1-\delta^2) \ln(1/\alpha) \mid \tau \ge \nu \right) = 0,$$

which also implies that[2]

$$\liminf_{\alpha \to 0} \frac{\mathbb{E}_\nu[\tau - \nu | \tau \ge \nu]}{\ln(1/\alpha)} \ge \frac{1}{I^\star} = T^\star.$$

[2] This would conclude the proof since $\bar{E}(\tau) \ge \mathbb{E}_\nu[\tau - \nu | \tau \ge \nu]$.

Quickest Change Detection: lower bound (cont.)

The idea is to show the following for any $\delta \in (0,1)$:

$$(\mathbf{P_1}) \lim_{\alpha \to 0} \mathbb{P}_\nu \left(\tau - \nu \leq T^\star (1-\delta) \ln(1/\alpha), \sum_{n=\nu}^{\tau} Z_n < (1-\delta^2) \ln(1/\alpha) \mid \tau \geq \nu \right) = 0,$$

and

$$(\mathbf{P_2}) \lim_{\alpha \to 0} \mathbb{P}_\nu \left(\tau - \nu \leq T^\star (1-\delta) \ln(1/\alpha), \sum_{n=\nu}^{\tau} Z_n \geq (1-\delta^2) \ln(1/\alpha) \mid \tau \geq \nu \right) = 0,$$

which also implies that[2]

$$\liminf_{\alpha \to 0} \frac{\mathbb{E}_\nu[\tau - \nu | \tau \geq \nu]}{\ln(1/\alpha)} \geq \frac{1}{I^\star} = T^\star.$$

[2] This would conclude the proof since $\bar{E}(\tau) \geq \mathbb{E}_\nu[\tau - \nu | \tau \geq \nu]$.

Quickest Change Detection: lower bound (cont.)

$$(P_2):\ \mathbb{P}_\nu\left(\tau-\nu\le I^\star(1-\delta)\ln(1/\alpha),\sum_{n=\nu}^{\tau} Z_n \ge (1-\delta^2)\ln(1/\alpha)\mid \tau\ge\nu\right)$$

Let $n_\alpha = T^\star(1-\delta)\ln(1/\alpha)$ with $\delta\in(0,1)$. Then

$$(P_2) \le \operatorname{ess\,sup}\mathbb{P}_\nu\left(\tau-\nu\le T^\star(1-\delta)\ln(1/\alpha), I^\star\sum_{n=\nu}^{\tau} Z_n \ge I^\star(1-\delta)(1+\delta)\ln(1/\alpha)\mid \tau\ge\nu\right)$$

$$\le \operatorname{ess\,sup}\mathbb{P}_\nu\left(\max_{t\le n_\alpha} I^\star\sum_{n=\nu}^{\nu+t} Z_n \ge I^\star(1-\delta)(1+\delta)\ln(1/\alpha)\mid \tau\ge\nu\right)$$

$$\le \operatorname{ess\,sup}\mathbb{P}_\nu\left(\max_{t\le n_\alpha}\sum_{n=\nu}^{\nu+t} Z_n \ge I^\star(1+\delta)n_\alpha\mid \tau\ge\nu\right) \to 0 \text{ as } \alpha\to 0 \text{ by assumption.}$$

Quickest Change Detection: lower bound (cont.)

$$(P_2): \ \mathbb{P}_\nu\left(\tau - \nu \le I^\star(1-\delta)\ln(1/\alpha), \sum_{n=\nu}^{\tau} Z_n \ge (1-\delta^2)\ln(1/\alpha) \mid \tau \ge \nu\right)$$

Let $n_\alpha = T^\star(1-\delta)\ln(1/\alpha)$ with $\delta \in (0,1)$. Then

$$(P_2) \le \operatorname{ess\,sup} \mathbb{P}_\nu\left(\tau - \nu \le T^\star(1-\delta)\ln(1/\alpha), I^\star \sum_{n=\nu}^{\tau} Z_n \ge I^\star(1-\delta)(1+\delta)\ln(1/\alpha) \mid \tau \ge \nu\right)$$

$$\le \operatorname{ess\,sup} \mathbb{P}_\nu\left(\max_{t \le n_\alpha} I^\star \sum_{n=\nu}^{\nu+t} Z_n \ge I^\star(1-\delta)(1+\delta)\ln(1/\alpha) \mid \tau \ge \nu\right)$$

$$\le \operatorname{ess\,sup} \mathbb{P}_\nu\left(\max_{t \le n_\alpha} \sum_{n=\nu}^{\nu+t} Z_n \ge I^\star(1+\delta)n_\alpha \mid \tau \ge \nu\right) \to 0 \text{ as } \alpha \to 0 \text{ by assumption.}$$

Quickest Change Detection: lower bound (cont.)

$$(P_2):\ \mathbb{P}_\nu\left(\tau-\nu\le I^\star(1-\delta)\ln(1/\alpha), \sum_{n=\nu}^{\tau} Z_n \ge (1-\delta^2)\ln(1/\alpha) \mid \tau\ge\nu\right)$$

Let $n_\alpha = T^\star(1-\delta)\ln(1/\alpha)$ with $\delta\in(0,1)$. Then

$$\begin{aligned}(P_2) &\le \operatorname{ess\,sup} \mathbb{P}_\nu\left(\tau-\nu\le T^\star(1-\delta)\ln(1/\alpha), I^\star\sum_{n=\nu}^{\tau} Z_n \ge I^\star(1-\delta)(1+\delta)\ln(1/\alpha) \mid \tau\ge\nu\right)\\ &\le \operatorname{ess\,sup} \mathbb{P}_\nu\left(\max_{t\le n_\alpha} I^\star\sum_{n=\nu}^{\nu+t} Z_n \ge I^\star(1-\delta)(1+\delta)\ln(1/\alpha) \mid \tau\ge\nu\right)\\ &\le \operatorname{ess\,sup} \mathbb{P}_\nu\left(\max_{t\le n_\alpha} \sum_{n=\nu}^{\nu+t} Z_n \ge I^\star(1+\delta)n_\alpha \mid \tau\ge\nu\right) \to 0 \text{ as } \alpha\to 0 \text{ by assumption.}\end{aligned}$$

Quickest Change Detection: lower bound (cont.)

$$(P_2):\ \mathbb{P}_\nu\left(\tau-\nu\le I^\star(1-\delta)\ln(1/\alpha), \sum_{n=\nu}^{\tau} Z_n \ge (1-\delta^2)\ln(1/\alpha) \mid \tau\ge\nu\right)$$

Let $n_\alpha = T^\star(1-\delta)\ln(1/\alpha)$ with $\delta\in(0,1)$. Then

$$\begin{aligned}(P_2) &\le \operatorname{ess\,sup}\mathbb{P}_\nu\left(\tau-\nu\le T^\star(1-\delta)\ln(1/\alpha), I^\star\sum_{n=\nu}^{\tau} Z_n \ge I^\star(1-\delta)(1+\delta)\ln(1/\alpha) \mid \tau\ge\nu\right)\\
&\le \operatorname{ess\,sup}\mathbb{P}_\nu\left(\max_{t\le n_\alpha} I^\star\sum_{n=\nu}^{\nu+t} Z_n \ge I^\star(1-\delta)(1+\delta)\ln(1/\alpha) \mid \tau\ge\nu\right)\\
&\le \operatorname{ess\,sup}\mathbb{P}_\nu\left(\max_{t\le n_\alpha} \sum_{n=\nu}^{\nu+t} Z_n \ge I^\star(1+\delta)n_\alpha \mid \tau\ge\nu\right) \to 0 \text{ as } \alpha\to 0 \text{ by assumption.}\end{aligned}$$

Quickest Change Detection: lower bound (final)

To prove $(P_1) \to 0$ as $\alpha \to 0$ we can use similar arguments as in the i.i.d. case.

Lemma (Another low-level form of the fundamental inequality)

For all $x \in \mathbb{R}, t \in \mathbb{N}$ and all event $\mathcal{E} \in \mathcal{F}_t$ we have

$$\text{(Change of measure trick)} \quad \mathbb{P}_\nu(\mathcal{E} \cap \{Z_t < x\}) \leq e^x \mathbb{P}_\infty(\mathcal{E}),$$

where $Z_t = \ln \frac{\mathrm{d}\mathbb{P}_\nu(X_1,\ldots,X_t)}{\mathrm{d}\mathbb{P}_\infty(X_1,\ldots,X_t)}$ is the log-likelihood ratio.

Let $t = n_\alpha$, and $\mathcal{E} = \{\tau - \nu \leq n_\alpha\}$. Then $\mathcal{E} \in \mathcal{F}_{n_\alpha}$. As in the i.i.d. case one can prove $\mathbb{P}_\infty(\mathcal{E} | \tau \geq \nu) \leq [\ln(1/\alpha)]^2 \alpha$. Letting $x = (1-\delta^2)\ln(1/\alpha)$

$$(P_1) = \mathbb{P}_\nu(\mathcal{E} \cap \{Z_{n_\alpha} < (1-\delta^2)\ln(1/\alpha)\} \mid \tau \geq \nu) \leq [\ln(1/\alpha)]^2 \alpha^{\delta^2} \to 0 \text{ as } \alpha \to 0.$$

Hence, the result is proven.

Quickest Change Detection: lower bound (final)

To prove $(P_1) \to 0$ as $\alpha \to 0$ we can use similar arguments as in the i.i.d. case.

Lemma (Another low-level form of the fundamental inequality)

For all $x \in \mathbb{R}, t \in \mathbb{N}$ and all event $\mathcal{E} \in \mathcal{F}_t$ we have

$$\text{(Change of measure trick)} \quad \mathbb{P}_\nu(\mathcal{E} \cap \{Z_t < x\}) \leq e^x \mathbb{P}_\infty(\mathcal{E}),$$

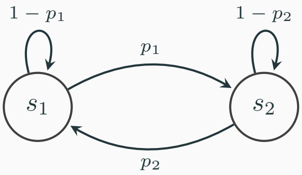
where $Z_t = \ln \frac{\mathrm{d}\mathbb{P}_\nu(X_1,\ldots,X_t)}{\mathrm{d}\mathbb{P}_\infty(X_1,\ldots,X_t)}$ is the log-likelihood ratio.

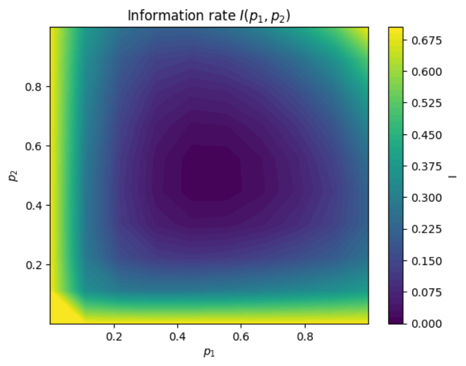
Let $t = n_\alpha$, and $\mathcal{E} = \{\tau - \nu \leq n_\alpha\}$. Then $\mathcal{E} \in \mathcal{F}_{n_\alpha}$. As in the i.i.d. case one can prove $\mathbb{P}_\infty(\mathcal{E} | \tau \geq \nu) \leq [\ln(1/\alpha)]^2 \alpha$. Letting $x = (1 - \delta^2) \ln(1/\alpha)$

$$(P_1) = \mathbb{P}_\nu(\mathcal{E} \cap \{Z_{n_\alpha} < (1-\delta^2)\ln(1/\alpha)\} \mid \tau \geq \nu) \leq [\ln(1/\alpha)]^2 \alpha^{\delta^2} \to 0 \text{ as } \alpha \to 0.$$

Hence, the result is proven.

Example with an MDP





Example with a Markov chain with 2 states. f_0 has $p_0 = p_1 = 0.5$. The quantity I[3] is $I = \mathbb{E}_{s\sim\mu}[\mathrm{KL}(P_1(s), P_2(s))]$, where μ is the stationary distribution under f_1.

[3] As $\alpha \to 0$ one can verify that the average log-likelihood ratio under $\mathbb{P}_\nu$ tends to this quantity.

Best Policy Identification: Tabular Markov Decision Processes

Introduction

- Consider an MDP $M = (S, A, P, r, \gamma)$ [4].
  - S is the state space (finite);
  - A is the action space (finite)
  - $P : S \times A \to \Delta(S)$ is the transition function.
  - $r : S \times A \to [0, 1]$ is the reward function.
  - $\gamma \in (0, 1)$ is the discount factor.
- A policy $\pi : s \to \Delta(A)$ maps states to distributions over actions.
- The value of a policy is $V^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}[Q^\pi(s, a)]$, where

$$Q^\pi(s, a) = \mathbb{E}^\pi[\sum_{t\geq 1} \gamma^{t-1} r(s_t, a_t)|s_0 = s, a_0 = a].$$

- We assume there exists a unique optimal policy $\pi^\star(s) = \arg\max_\pi V^\pi(s), \forall s \in S$ (which is deterministic...).

[4] Setting studied in [AMP21, AMGP21]

Best Policy Identification with Fixed Confidence: introduction

Estimate $\pi^\star$ as quickly as possible with confidence $\delta \in (0,1)$.

- Assume the reward function to be deterministic and known.
- As usual, define τ to be the stopping time of the algorithm.
- Let $\hat{\pi}_\tau$ be the optimal arm estimated by the algorithm at the stopping time.
- We say that an algorithm is δ-PC (Probably Correct) if $\mathbb{P}_M(\tau < \infty, \hat{\pi}_\tau = \pi^\star) \geq 1 - \delta$ for all possible models M satisfying the uniqueness of the best arm.

Best Policy Identification with Fixed Confidence: introduction

Estimate $\pi^\star$ as quickly as possible with confidence $\delta \in (0,1)$.

- Assume the reward function to be deterministic and known.
- As usual, define τ to be the stopping time of the algorithm.
- Let $\hat{\pi}_\tau$ be the optimal arm estimated by the algorithm at the stopping time.
- We say that an algorithm is δ-PC (Probably Correct) if $\mathbb{P}_M(\tau < \infty, \hat{\pi}_\tau = \pi^\star) \geq 1-\delta$ for all possible models M satisfying the uniqueness of the best arm.

Best Policy Identification with Fixed Confidence: introduction

Estimate $\pi^\star$ as quickly as possible with confidence $\delta \in (0,1)$.

- Assume the reward function to be deterministic and known.
- As usual, define τ to be the stopping time of the algorithm.
- Let $\hat{\pi}_\tau$ be the optimal arm estimated by the algorithm at the stopping time.
- We say that an algorithm is δ-PC (Probably Correct) if $\mathbb{P}_M(\tau < \infty, \hat{\pi}_\tau = \pi^\star) \geq 1-\delta$ for all possible models M satisfying the uniqueness of the best arm.

Best Policy Identification with Fixed Confidence: introduction

Estimate $\pi^\star$ as quickly as possible with confidence $\delta \in (0,1)$.

- Assume the reward function to be deterministic and known.
- As usual, define τ to be the stopping time of the algorithm.
- Let $\hat{\pi}_\tau$ be the optimal arm estimated by the algorithm at the stopping time.
- We say that an algorithm is δ-PC (Probably Correct) if $\mathbb{P}_M(\tau < \infty, \hat{\pi}_\tau = \pi^\star) \geq 1-\delta$ for all possible models M satisfying the uniqueness of the best arm.

Best Policy Identification with Fixed Confidence: introduction

Estimate $\pi^\star$ as quickly as possible with confidence $\delta \in (0,1)$.

- Assume the reward function to be deterministic and known.
- As usual, define τ to be the stopping time of the algorithm.
- Let $\hat{\pi}_\tau$ be the optimal arm estimated by the algorithm at the stopping time.
- We say that an algorithm is δ-PC (Probably Correct) if $\mathbb{P}_M(\tau < \infty, \hat{\pi}_\tau = \pi^\star) \geq 1 - \delta$ for all possible models M satisfying the uniqueness of the best arm.

Best Policy Identification with Fixed Confidence: lower bound

The δ-PC event is $\{\hat{\pi}_\tau \neq \pi^\star\}$. We define the set of confusing models according to this event!

$$\mathrm{Alt}(M) := \{M' : \pi^\star(M') \neq \pi^\star(M),\ M' \text{ has a unique optimal policy}\},$$

where $\pi^\star(M')$ is the optimal policy in M' (sim. $\pi^\star(M)$).

Why we define the set according to the δ-PC event? Because we want to check if at the stopping time the true MDP M is confusing for the MDP M_τ that we estimated.

Best Policy Identification with Fixed Confidence: lower bound

The δ-PC event is $\{\hat{\pi}_\tau \neq \pi^\star\}$. We define the set of confusing models according to this event!

$$\mathrm{Alt}(M) \coloneqq \{M' : \pi^\star(M') \neq \pi^\star(M),\ M' \text{ has a unique optimal policy}\},$$

where $\pi^\star(M')$ is the optimal policy in M' (sim. $\pi^\star(M)$).

Why we define the set according to the δ-PC event? Because we want to check if at the stopping time the true MDP M is confusing for the MDP M_τ that we estimated.

Best Policy Identification with Fixed Confidence: lower bound

The δ-PC event is $\{\hat{\pi}_\tau \neq \pi^\star\}$. We define the set of confusing models according to this event!

$$\mathrm{Alt}(M) := \{M' : \pi^\star(M') \neq \pi^\star(M),\ M' \text{ has a unique optimal policy}\},$$

where $\pi^\star(M')$ is the optimal policy in M' (sim. $\pi^\star(M)$).

Why we define the set according to the δ-PC event? Because we want to check if at the stopping time the true MDP M is confusing for the MDP M_τ that we estimated.

Best Policy Identification with Fixed Confidence: lower bound (cont.)

Consider then the log-likelihood ratio $Z_t = \ln \frac{d\mathbb{P}_M(S_1,A_1,R_1,S_1',\dots,S_t,A_t,R_t,S_t')}{d\mathbb{P}_{M'}(S_1,A_1,R_1,S_1',\dots,S_t,A_t,R_t,S_t')}$ between M and $M' \in \mathrm{Alt}(M)$[5]. Then:

$$\mathbb{E}_M[Z_\tau] = \mathbb{E}_M\left[\sum_{n=1}^{\tau}\sum_{s,a} \mathbf{1}_{\{S_n=s,A_n=a\}} \ln \frac{P(S_n'|s,a)}{P'(S_n'|s,a)}\right].$$

Let $Z_\tau(s,a) = \sum_{n=1}^{\tau} \mathbf{1}_{\{S_n=s,A_n=a\}} \ln \frac{P(S_n'|s,a)}{P'(S_n'|s,a)}$ and $N_t(s,a)$ be the time number of times (s,a) has been selected up to time t. Then

$$\mathbb{E}_M[Z_\tau(s,a)] = \mathbb{E}_M\left[\sum_{n=1}^{N_\tau(s,a)} \underbrace{\ln \frac{P(Y_n|s,a)}{P'(Y_n|s,a)}}_{W_n}\right] = \mathbb{E}_M\left[\sum_{n=1}^{\infty} \mathbf{1}_{\{N_\tau(s,a)\geq n\}} W_n\right].$$

[5] We indicate by S_n' the state observed after taking action A_n in state S_n

Best Policy Identification with Fixed Confidence: lower bound (cont.)

Consider then the log-likelihood ratio $Z_t = \ln \frac{\mathrm{d}\mathbb{P}_M(S_1,A_1,R_1,S'_1,\dots,S_t,A_t,R_t,S'_t)}{\mathrm{d}\mathbb{P}_{M'}(S_1,A_1,R_1,S'_1,\dots,S_t,A_t,R_t,S'_t)}$ between M and $M' \in \mathrm{Alt}(M)$[5]. Then:

$$\mathbb{E}_M[Z_\tau] = \mathbb{E}_M\left[\sum_{n=1}^{\tau}\sum_{s,a}\mathbf{1}_{\{S_n=s,A_n=a\}}\ln\frac{P(S'_n|s,a)}{P'(S'_n|s,a)}\right].$$

Let $Z_\tau(s,a) = \sum_{n=1}^{\tau}\mathbf{1}_{\{S_n=s,A_n=a\}}\ln\frac{P(S'_n|s,a)}{P'(S'_n|s,a)}$ and $N_t(s,a)$ be the time number of times (s,a) has been selected up to time t. Then

$$\mathbb{E}_M[Z_\tau(s,a)] = \mathbb{E}_M\left[\sum_{n=1}^{N_\tau(s,a)}\underbrace{\ln\frac{P(Y_n|s,a)}{P'(Y_n|s,a)}}_{W_n}\right] = \mathbb{E}_M\left[\sum_{n=1}^{\infty}\mathbf{1}_{\{N_\tau(s,a)\geq n\}}W_n\right].$$

[5] We indicate by S'_n the state observed after taking action A_n in state S_n

Best Policy Identification with Fixed Confidence: lower bound (cont.)

$$\mathbb{E}_M[Z_\tau(s,a)] = \mathbb{E}_M\left[\sum_{n=1}^{\infty} \mathbf{1}_{\{N_\tau(s,a)\geq n\}} W_n\right].$$

Note that the event $\{N_\tau(s,a) \geq n\} = \{N_\tau(s,a) \leq n-1\}^c \in \mathcal{F}_{n-1}$ (the filtration of the data up to and including round $n-1$). Since W_n is independent of $\mathcal{F}_{n-1}$, then we have

$$\begin{aligned}\mathbb{E}_M[Z_\tau(s,a)] &= \mathbb{E}_M\left[\sum_{n=1}^{\infty} \mathbf{1}_{\{N_\tau(s,a)\geq n\}}\right] \mathrm{KL}(P(s,a), P'(s,a)),\\ &= \sum_{n=1}^{\infty} \mathbb{P}_M(N_\tau(s,a) \geq n)\mathrm{KL}(P(s,a), P'(s,a)),\\ &= \mathbb{E}_M[N_\tau(s,a)]\mathrm{KL}(P(s,a), P'(s,a)),\end{aligned}$$

Best Policy Identification with Fixed Confidence: lower bound (cont.)

$$\mathbb{E}_M[Z_\tau(s,a)] = \mathbb{E}_M\left[\sum_{n=1}^{\infty} \mathbf{1}_{\{N_\tau(s,a)\geq n\}} W_n\right].$$

Note that the event $\{N_\tau(s,a) \geq n\} = \{N_\tau(s,a) \leq n-1\}^c \in \mathcal{F}_{n-1}$ (the filtration of the data up to and including round $n-1$). Since W_n is independent of $\mathcal{F}_{n-1}$, then we have

$$\begin{aligned}
\mathbb{E}_M[Z_\tau(s,a)] &= \mathbb{E}_M\left[\sum_{n=1}^{\infty} \mathbf{1}_{\{N_\tau(s,a)\geq n\}}\right] \mathrm{KL}(P(s,a), P'(s,a)), \\
&= \sum_{n=1}^{\infty} \mathbb{P}_M(N_\tau(s,a) \geq n) \mathrm{KL}(P(s,a), P'(s,a)), \\
&= \mathbb{E}_M[N_\tau(s,a)] \mathrm{KL}(P(s,a), P'(s,a)),
\end{aligned}$$

Best Policy Identification with Fixed Confidence: lower bound (cont.)

$$\mathbb{E}_M[Z_\tau(s,a)] = \mathbb{E}_M\left[\sum_{n=1}^{\infty} \mathbf{1}_{\{N_\tau(s,a)\geq n\}} W_n\right].$$

Note that the event $\{N_\tau(s,a) \geq n\} = \{N_\tau(s,a) \leq n-1\}^c \in \mathcal{F}_{n-1}$ (the filtration of the data up to and including round $n-1$). Since W_n is independent of $\mathcal{F}_{n-1}$, then we have

$$\begin{aligned}\mathbb{E}_M[Z_\tau(s,a)] &= \mathbb{E}_M\left[\sum_{n=1}^{\infty} \mathbf{1}_{\{N_\tau(s,a)\geq n\}}\right] \mathrm{KL}(P(s,a), P'(s,a)),\\ &= \sum_{n=1}^{\infty} \mathbb{P}_M(N_\tau(s,a) \geq n)\mathrm{KL}(P(s,a), P'(s,a)),\\ &= \mathbb{E}_M[N_\tau(s,a)]\mathrm{KL}(P(s,a), P'(s,a)),\end{aligned}$$

Best Policy Identification with Fixed Confidence: lower bound (cont.)

$$\mathbb{E}_M[Z_\tau(s,a)] = \mathbb{E}_M\left[\sum_{n=1}^{\infty} \mathbf{1}_{\{N_\tau(s,a)\geq n\}} W_n\right].$$

Note that the event $\{N_\tau(s,a) \geq n\} = \{N_\tau(s,a) \leq n-1\}^c \in \mathcal{F}_{n-1}$ (the filtration of the data up to and including round $n-1$). Since W_n is independent of $\mathcal{F}_{n-1}$, then we have

$$\begin{aligned}
\mathbb{E}_M[Z_\tau(s,a)] &= \mathbb{E}_M\left[\sum_{n=1}^{\infty} \mathbf{1}_{\{N_\tau(s,a)\geq n\}}\right] \mathrm{KL}(P(s,a), P'(s,a)), \\
&= \sum_{n=1}^{\infty} \mathbb{P}_M(N_\tau(s,a) \geq n) \mathrm{KL}(P(s,a), P'(s,a)), \\
&= \mathbb{E}_M[N_\tau(s,a)] \mathrm{KL}(P(s,a), P'(s,a)),
\end{aligned}$$

Best Policy Identification with Fixed Confidence: lower bound (cont.)

$$\mathbb{E}_M[Z_\tau] = \sum_{s,a} \mathbb{E}_M[N_\tau(s,a)] \mathrm{KL}(P(s,a), P'(s,a)).$$

Lemma (Fundamental inequality [GMS19])

For any $\mathcal{F}_\tau$-measurable r.v. $Y \in [0,1]$ we have $\mathbb{E}_{M_1}[Z_\tau(M_1, M_0)] \geq \mathrm{kl}\left(\mathbb{E}_{M_1}[Y], \mathbb{E}_{M_0}[Y]\right).$

We apply it and choose $Y = \mathbf{1}_\mathcal{E}, \mathcal{E} = \{\hat{\pi}_\tau = \pi^\star(M)\}$:

$$\mathbb{E}_M[Z_\tau] = \sum_{s,a} \mathbb{E}_M[N_\tau(s,a)] \mathrm{KL}(P(s,a), P'(s,a)) \geq \mathrm{kl}(1-\delta, \delta).$$

since $\mathbb{P}_M(\mathcal{E}) \geq 1-\delta$ and $\mathbb{P}_{M'}(\mathcal{E}) \leq \delta$ from the fact that $\mathcal{E} \subset \{\hat{\pi}_\tau \neq \pi^\star(M')\}$ under $\mathbb{P}_{M'}$.

Best Policy Identification with Fixed Confidence: lower bound (cont.)

$$\mathbb{E}_M[Z_\tau] = \sum_{s,a} \mathbb{E}_M[N_\tau(s,a)]\mathrm{KL}(P(s,a), P'(s,a)).$$

Lemma (Fundamental inequality [GMS19])

For any $\mathcal{F}_\tau$-measurable r.v. $Y \in [0,1]$ we have $\mathbb{E}_{M_1}[Z_\tau(M_1, M_0)] \geq \mathrm{kl}\left(\mathbb{E}_{M_1}[Y], \mathbb{E}_{M_0}[Y]\right)$.

We apply it and choose $Y = \mathbf{1}_\mathcal{E}, \mathcal{E} = \{\hat{\pi}_\tau = \pi^\star(M)\}$:

$$\mathbb{E}_M[Z_\tau] = \sum_{s,a} \mathbb{E}_M[N_\tau(s,a)]\mathrm{KL}(P(s,a), P'(s,a)) \geq \mathrm{kl}(1-\delta, \delta).$$

since $\mathbb{P}_M(\mathcal{E}) \geq 1-\delta$ and $\mathbb{P}_{M'}(\mathcal{E}) \leq \delta$ from the fact that $\mathcal{E} \subset \{\hat{\pi}_\tau \neq \pi^\star(M')\}$ under $\mathbb{P}_{M'}$.

Best Policy Identification with Fixed Confidence: lower bound (cont.)

We can take the infimum over the set of confusing models:

$$\inf_{M' \in \mathrm{Alt}(M)} \sum_{s,a} \mathbb{E}_M[N_\tau(s,a)] \mathrm{KL}(P(s,a), P'(s,a)) \geq \mathrm{kl}(1-\delta, \delta),$$

which yields the most confusing model.

Divide and multiply the left hand-side by $\mathbb{E}_M[\tau]$ and let $\omega_{s,a} := \mathbb{E}_M[N_\tau(s,a)]/\mathbb{E}_M[\tau]$:

$$\mathbb{E}_M[\tau] \inf_{M' \in \mathrm{Alt}(M)} \sum_{s,a} \omega_{s,a} \mathrm{KL}(P(s,a), P'(s,a)) \geq \mathrm{kl}(1-\delta, \delta).$$

Therefore, we conclude by optimizing over $\omega_{s,a} \in \Delta(S \times A)$ (the simplex states and actions):

$$\mathbb{E}_M[\tau] \sup_{\omega \in \Delta(S \times A)} \inf_{M' \in \mathrm{Alt}(M)} \sum_{s,a} \omega_{s,a} \mathrm{KL}(P(s,a), P'(s,a)) \geq \mathrm{kl}(1-\delta, \delta).$$

Best Policy Identification with Fixed Confidence: lower bound (cont.)

We can take the infimum over the set of confusing models:

$$\inf_{M' \in \mathrm{Alt}(M)} \sum_{s,a} \mathbb{E}_M[N_\tau(s,a)] \mathrm{KL}(P(s,a), P'(s,a)) \geq \mathrm{kl}(1-\delta, \delta),$$

which yields the most confusing model.

Divide and multiply the left hand-side by $\mathbb{E}_M[\tau]$ and let $\omega_{s,a} := \mathbb{E}_M[N_\tau(s,a)]/\mathbb{E}_M[\tau]$:

$$\mathbb{E}_M[\tau] \inf_{M' \in \mathrm{Alt}(M)} \sum_{s,a} \omega_{s,a} \mathrm{KL}(P(s,a), P'(s,a)) \geq \mathrm{kl}(1-\delta, \delta).$$

Therefore, we conclude by optimizing over $\omega_{s,a} \in \Delta(S \times A)$ (the simplex states and actions):

$$\mathbb{E}_M[\tau] \sup_{\omega \in \Delta(S \times A)} \inf_{M' \in \mathrm{Alt}(M)} \sum_{s,a} \omega_{s,a} \mathrm{KL}(P(s,a), P'(s,a)) \geq \mathrm{kl}(1-\delta, \delta).$$

Best Policy Identification with Fixed Confidence: lower bound (cont.)

We can take the infimum over the set of confusing models:

$$\inf_{M' \in \mathrm{Alt}(M)} \sum_{s,a} \mathbb{E}_M[N_\tau(s,a)] \mathrm{KL}(P(s,a), P'(s,a)) \geq \mathrm{kl}(1-\delta, \delta),$$

which yields the most confusing model.

Divide and multiply the left hand-side by $\mathbb{E}_M[\tau]$ and let $\omega_{s,a} := \mathbb{E}_M[N_\tau(s,a)]/\mathbb{E}_M[\tau]$:

$$\mathbb{E}_M[\tau] \inf_{M' \in \mathrm{Alt}(M)} \sum_{s,a} \omega_{s,a} \mathrm{KL}(P(s,a), P'(s,a)) \geq \mathrm{kl}(1-\delta, \delta).$$

Therefore, we conclude by optimizing over $\omega_{s,a} \in \Delta(S \times A)$ (the simplex states and actions):

$$\mathbb{E}_M[\tau] \sup_{\omega \in \Delta(S \times A)} \inf_{M' \in \mathrm{Alt}(M)} \sum_{s,a} \omega_{s,a} \mathrm{KL}(P(s,a), P'(s,a)) \geq \mathrm{kl}(1-\delta, \delta).$$

Best Policy Identification with Fixed Confidence: lower bound (final)

$$\sup_{\omega \in \Delta(S \times A)} \inf_{M' \in \mathrm{Alt}(M)} \sum_{s,a} \omega_{s,a} \mathrm{KL}(P(s,a), P'(s,a))$$

is that it? We are missing the navigation constraints! (forward model).

For ergodic models, as $\delta \to 0$, we have that ω tends to the stationary distribution over states and actions.Hence we can take the limit and find that [6]

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_M[\tau]}{\ln(1/\delta)} \geq T^\star,$$

where

$$(T^\star)^{-1} := \sup_{\omega \in \Omega(M)} \inf_{M' \in \mathrm{Alt}(M)} \sum_{s,a} w_{s,a} \mathrm{KL}(P(s,a), P'(s,a))$$

with $\Omega(M) = \{\omega \in \Delta(S \times A) : \sum_a \omega_{s,a} = \sum_{s',a'} P(s|s',a')\omega_{s',a'}\}$ (Kolmogorov equations).

[6] Find a non-asymptotic approach in the appendix.

Best Policy Identification with Fixed Confidence: lower bound (final)

$$\sup_{\omega \in \Delta(S \times A)} \inf_{M' \in \mathrm{Alt}(M)} \sum_{s,a} \omega_{s,a} \mathrm{KL}(P(s,a), P'(s,a))$$

is that it? We are missing the navigation constraints! (forward model).

For ergodic models, as $\delta \to 0$, we have that ω tends to the stationary distribution over states and actions.Hence we can take the limit and find that [6]

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_M[\tau]}{\ln(1/\delta)} \geq T^\star,$$

where

$$(T^\star)^{-1} := \sup_{\omega \in \Omega(M)} \inf_{M' \in \mathrm{Alt}(M)} \sum_{s,a} w_{s,a} \mathrm{KL}(P(s,a), P'(s,a))$$

with $\Omega(M) = \{\omega \in \Delta(S \times A) : \sum_a \omega_{s,a} = \sum_{s',a'} P(s|s',a')\omega_{s',a'}\}$ (Kolmogorov equations).

[6] Find a non-asymptotic approach in the appendix.

Best Policy Identification with Fixed Confidence: lower bound (final)

$$\sup_{\omega\in\Delta(S\times A)} \inf_{M'\in\mathrm{Alt}(M)} \sum_{s,a} \omega_{s,a} \mathrm{KL}(P(s,a), P'(s,a))$$

is that it? We are missing the navigation constraints! (forward model).

For ergodic models, as $\delta \to 0$, we have that ω tends to the stationary distribution over states and actions. Hence we can take the limit and find that [6]

$$\liminf_{\delta\to 0} \frac{\mathbb{E}_M[\tau]}{\ln(1/\delta)} \geq T^\star,$$

where

$$(T^\star)^{-1} := \sup_{\omega\in\Omega(M)} \inf_{M'\in\mathrm{Alt}(M)} \sum_{s,a} w_{s,a} \mathrm{KL}(P(s,a), P'(s,a))$$

with $\Omega(M) = \{\omega \in \Delta(S\times A) : \sum_a \omega_{s,a} = \sum_{s',a'} P(s|s',a')\omega_{s',a'}\}$ (Kolmogorov equations).

[6] Find a non-asymptotic approach in the appendix.

Best Policy Identification with Fixed Confidence: lower bound (final)

$$\sup_{\omega \in \Delta(S \times A)} \inf_{M' \in \mathrm{Alt}(M)} \sum_{s,a} \omega_{s,a} \mathrm{KL}(P(s,a), P'(s,a))$$

is that it? We are missing the navigation constraints! (forward model).

For ergodic models, as $\delta \to 0$, we have that ω tends to the stationary distribution over states and actions.Hence we can take the limit and find that [6]

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_M[\tau]}{\ln(1/\delta)} \geq T^\star,$$

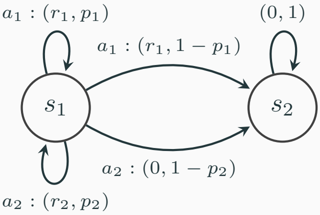
where

$$(T^\star)^{-1} := \sup_{\omega \in \Omega(M)} \inf_{M' \in \mathrm{Alt}(M)} \sum_{s,a} w_{s,a} \mathrm{KL}(P(s,a), P'(s,a))$$

with $\Omega(M) = \{\omega \in \Delta(S \times A) : \sum_a \omega_{s,a} = \sum_{s',a'} P(s|s',a')\omega_{s',a'}\}$ (Kolmogorov equations).

[6] Find a non-asymptotic approach in the appendix.

Set of confusing model is non-convex!



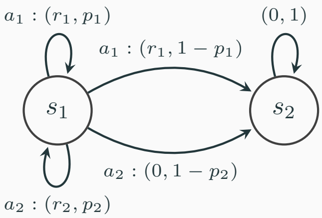
- Consider the MDP in figure, where the starting state is s_1. In each edge we indicate the action and the corresponding reward and transition probability (no action = all actions).
- The optimal Q-values in s_1 are

$$Q^\star(s_1, a_1) = r_1 + \gamma p_1 V^\star(s_1) \text{ and } Q^\star(s_1, a_2) = p_2(r_2 + \gamma V^\star(s_1)).$$

Therefore

$$V^\star(s_1) = \max\left(\frac{r_1}{1 - \gamma p_1}, \frac{p_2 r_2}{1 - \gamma p_2}\right)$$

Set of confusing model is non-convex!



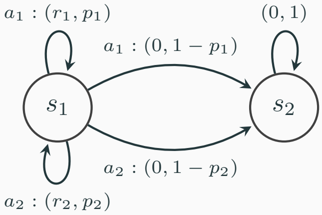
- Consider the MDP in figure, where the starting state is s_1. In each edge we indicate the action and the corresponding reward and transition probability (no action = all actions).
- The optimal Q-values in s_1 are

$$Q^\star(s_1, a_1) = r_1 + \gamma p_1 V^\star(s_1) \text{ and } Q^\star(s_1, a_2) = p_2(r_2 + \gamma V^\star(s_1)).$$

Therefore

$$V^\star(s_1) = \max\left(\frac{r_1}{1 - \gamma p_1}, \frac{p_2 r_2}{1 - \gamma p_2}\right)$$

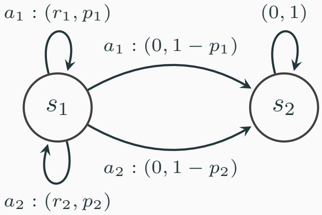
Set of confusing model is non-convex!



1. ϕ_1: fix $r_1 = 0.7, p_1 = 0.9, r_2 = 0.3, p_2 = 1$ then a_1 is optimal and $Q^\star(s_1, a_1) \approx 3.68$.
2. ϕ_2: fix $r_1 = 0.7, p_1 = 0.1, r_2 = 0.3, p_2 = 0.77$ then a_1 is optimal and $Q^\star(s_1, a_1) \approx 0.77$.
3. ϕ_{avg}: take the average between these two models. Then $p_1 = 0.5, p_2 = 0.885$. a_2 is optimal and $Q^\star(s_1, a_2) \approx 1.30$

If the first two models ϕ_1, ϕ_2 belong to Alt, then their average ϕ_{avg} does not!

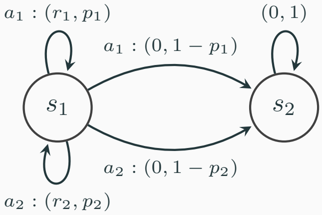
Set of confusing model is non-convex!



1. ϕ_1: fix $r_1 = 0.7, p_1 = 0.9, r_2 = 0.3, p_2 = 1$ then a_1 is optimal and $Q^\star(s_1, a_1) \approx 3.68$.
2. ϕ_2: fix $r_1 = 0.7, p_1 = 0.1, r_2 = 0.3, p_2 = 0.77$ then a_1 is optimal and $Q^\star(s_1, a_1) \approx 0.77$.
3. ϕ_{avg}: take the average between these two models. Then $p_1 = 0.5, p_2 = 0.885$. a_2 is optimal and $Q^\star(s_1, a_2) \approx 1.30$

If the first two models ϕ_1, ϕ_2 belong to Alt, then their average ϕ_{avg} does not!

Set of confusing model is non-convex!



1. ϕ_1: fix $r_1 = 0.7, p_1 = 0.9, r_2 = 0.3, p_2 = 1$ then a_1 is optimal and $Q^\star(s_1, a_1) \approx 3.68$.
2. ϕ_2: fix $r_1 = 0.7, p_1 = 0.1, r_2 = 0.3, p_2 = 0.77$ then a_1 is optimal and $Q^\star(s_1, a_1) \approx 0.77$.
3. ϕ_{avg}: take the average between these two models. Then $p_1 = 0.5, p_2 = 0.885$. a_2 is optimal and $Q^\star(s_1, a_2) \approx 1.30$

If the first two models ϕ_1, ϕ_2 belong to Alt, then their average ϕ_{avg} does not!

Set of confusing model is non-convex!

Regarding the non-convexity:

- If you check the original example from [AMP21] it is incorrect.
- We used the same reward in ϕ_1, ϕ_2 because we assumed to know the reward function!.
- Non-convexity seems to arise due to the probability values appearing both at the numerator and denominator $V^\star(s_1) = \max\left(\frac{r_1}{1-\gamma p_1}, \frac{p_2 r_2}{1-\gamma p_2}\right)$.
- However, in simple MDPs with known rewards, where $(I-\gamma P^{\pi^\star})^{-1}$ has a nice structure, maybe it is possible to have convexity...
- We have similar comments if we know the transition function but not the rewards distributions.

Set of confusing model is non-convex!

Regarding the non-convexity:

- If you check the original example from [AMP21] it is incorrect.
- We used the same reward in ϕ_1, ϕ_2 because we assumed to know the reward function!.
- Non-convexity seems to arise due to the probability values appearing both at the numerator and denominator $V^\star(s_1) = \max\left(\frac{r_1}{1-\gamma p_1}, \frac{p_2 r_2}{1-\gamma p_2}\right)$.
- However, in simple MDPs with known rewards, where $(I - \gamma P^{\pi^\star})^{-1}$ has a nice structure, maybe it is possible to have convexity...
- We have similar comments if we know the transition function but not the rewards distributions.

Set of confusing model is non-convex!

Regarding the non-convexity:

- If you check the original example from [AMP21] it is incorrect.
- We used the same reward in ϕ_1, ϕ_2 because we assumed to know the reward function!.
- Non-convexity seems to arise due to the probability values appearing both at the numerator and denominator $V^\star(s_1) = \max\left(\frac{r_1}{1-\gamma p_1}, \frac{p_2 r_2}{1-\gamma p_2}\right)$.
- However, in simple MDPs with known rewards, where $(I-\gamma P^{\pi^\star})^{-1}$ has a nice structure, maybe it is possible to have convexity...
- We have similar comments if we know the transition function but not the rewards distributions.

Set of confusing model is non-convex!

Regarding the non-convexity:

- If you check the original example from [AMP21] it is incorrect.
- We used the same reward in ϕ_1, ϕ_2 because we assumed to know the reward function!.
- Non-convexity seems to arise due to the probability values appearing both at the numerator and denominator $V^\star(s_1) = \max\left(\frac{r_1}{1-\gamma p_1}, \frac{p_2 r_2}{1-\gamma p_2}\right)$.
- However, in simple MDPs with known rewards, where $(I - \gamma P^{\pi^\star})^{-1}$ has a nice structure, maybe it is possible to have convexity...
- We have similar comments if we know the transition function but not the rewards distributions.

Set of confusing model is non-convex!

Regarding the non-convexity:

- If you check the original example from [AMP21] it is incorrect.
- We used the same reward in ϕ_1, ϕ_2 because we assumed to know the reward function!.
- Non-convexity seems to arise due to the probability values appearing both at the numerator and denominator $V^\star(s_1) = \max\left(\frac{r_1}{1-\gamma p_1}, \frac{p_2 r_2}{1-\gamma p_2}\right)$.
- However, in simple MDPs with known rewards, where $(I - \gamma P^{\pi^\star})^{-1}$ has a nice structure, maybe it is possible to have convexity...
- We have similar comments if we know the transition function but not the rewards distributions.

Can we convexify the lower bound?

Convexification

Can we convexify the lower bound?

We know that

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_M[\tau]}{\ln(1/\delta)} \geq T^\star.$$

Define $T^{-1}(\omega) = \inf_{M' \in \mathrm{Alt}(M)} \mathbb{E}_{(s,a)\sim\omega}[\mathrm{KL}(P(s,a), P'(s,a))]$.

Convexification

Can we find $U(\omega)$ s.t. for every ω we have that U is convex in ω and $T(\omega) \leq U(\omega)$?

Can we convexify the lower bound?

Convexification

Can we convexify the lower bound?

We know that

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_M[\tau]}{\ln(1/\delta)} \geq T^\star.$$

Define $T^{-1}(\omega) = \inf_{M' \in \mathrm{Alt}(M)} \mathbb{E}_{(s,a)\sim\omega}[\mathrm{KL}(P(s,a), P'(s,a))]$.

Convexification

Can we find $U(\omega)$ s.t. for every ω we have that U is convex in ω and $T(\omega) \leq U(\omega)$?

Can we convexify the lower bound?

Convexification

Can we convexify the lower bound?

We know that

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_M[\tau]}{\ln(1/\delta)} \geq T^\star.$$

Define $T^{-1}(\omega) = \inf_{M' \in \mathrm{Alt}(M)} \mathbb{E}_{(s,a)\sim\omega}[\mathrm{KL}(P(s,a), P'(s,a))]$.

Convexification

Can we find $U(\omega)$ s.t. for every ω we have that U is convex in ω and $T(\omega) \leq U(\omega)$?

Can we convexify the lower bound?

Convexification

Can we convexify the lower bound?

We know that

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_M[\tau]}{\ln(1/\delta)} \geq T^\star.$$

Define $T^{-1}(\omega) = \inf_{M' \in \mathrm{Alt}(M)} \mathbb{E}_{(s,a)\sim\omega}[\mathrm{KL}(P(s,a), P'(s,a))]$.

Convexification

Can we find $U(\omega)$ s.t. for every ω we have that U is convex in ω and $T(\omega) \leq U(\omega)$?

Ideas

$$T^{-1}(\omega) = \inf_{M' \in \mathrm{Alt}(M)} \mathbb{E}_{(s,a)\sim\omega}[\mathrm{KL}(P(s,a), P'(s,a))].$$

We want to find $U(\omega)$ s.t. $T(\omega) \leq U(\omega)$.

1. Is it possible to lower bound the sum of KL divergences so that the constraint is always satisfied?
2. Then, can we rewrite the constraints in a way that is related to the KL terms?
3. We know that the KL is roughly variance over gaps squared $\rightarrow$ try to write the constraints in terms of the sub-optimality gaps? [7].

[7] The sub-optimality gap is defined as $\Delta(s,a) = V^\star(s) - Q^\star(s,a)$.

Ideas

$$T^{-1}(\omega) = \inf_{M' \in \mathrm{Alt}(M)} \mathbb{E}_{(s,a)\sim\omega}[\mathrm{KL}(P(s,a), P'(s,a))].$$

We want to find $U(\omega)$ s.t. $T(\omega) \leq U(\omega)$.

1. Is it possible to lower bound the sum of KL divergences so that the constraint is always satisfied?
2. Then, can we rewrite the constraints in a way that is related to the KL terms?
3. We know that the KL is roughly variance over gaps squared $\rightarrow$ try to write the constraints in terms of the sub-optimality gaps? [7].

[7] The sub-optimality gap is defined as $\Delta(s,a) = V^\star(s) - Q^\star(s,a)$.

Ideas

$$T^{-1}(\omega) = \inf_{M' \in \mathrm{Alt}(M)} \mathbb{E}_{(s,a)\sim\omega}[\mathrm{KL}(P(s,a), P'(s,a))].$$

We want to find $U(\omega)$ s.t. $T(\omega) \leq U(\omega)$.

1. Is it possible to lower bound the sum of KL divergences so that the constraint is always satisfied?
2. Then, can we rewrite the constraints in a way that is related to the KL terms?
3. We know that the KL is roughly variance over gaps squared $\rightarrow$ try to write the constraints in terms of the sub-optimality gaps? [7].

[7] The sub-optimality gap is defined as $\Delta(s,a) = V^\star(s) - Q^\star(s,a)$.

Rewriting the set of confusing models [1/2]

Lemma

We have that

$$\mathrm{Alt}(M) = \cup_{s,a\neq\pi^\star(s)}\mathrm{Alt}_{s,a}(M) \text{ where } \mathrm{Alt}_{s,a}(M) = \{M' : Q^{\pi^\star}_{M'}(s,a) > V^{\pi^\star}_{M'}(s)\}.$$

where $\pi^\star$ is the optimal policy in M and $V^{\pi^\star}_{M'}$ is the evaluation of $\pi^\star$ in M'.

We begin by proving that $\mathrm{Alt}(M) \subset \cup_{s,a\neq\pi^\star(s)}\mathrm{Alt}_{s,a}(M)$ (which is more important, why?).

- By contradiction, assume $\exists M' \in \mathrm{Alt}(M)$ s.t. $\forall s, a \neq \pi^\star(s)$ we have $M' \notin \mathrm{Alt}_{s,a}(M)$.
- Therefore $Q^{\pi^\star}_{M'}(s,a) \leq V^{\pi^\star}_{M'}(s)$ for every $s, a \neq \pi^\star(s)$.
- Moreover $Q^{\pi^\star}_{M'}(s,\pi^\star(s)) = V^{\pi^\star}_{M'}(s)$ for every s.
- Hence $Q^{\pi^\star}_{M'}(s,a) \leq V^{\pi^\star}_{M'}(s)$ for every (s,a).
- By the policy improvement theorem there does not exists any action that improves the policy, hence $\pi^\star$ is optimal in $M' \Rightarrow$ contradiction!

Rewriting the set of confusing models [1/2]

Lemma

We have that

$$\mathrm{Alt}(M) = \cup_{s,a\neq\pi^\star(s)}\mathrm{Alt}_{s,a}(M) \text{ where } \mathrm{Alt}_{s,a}(M) = \{M' : Q^{\pi^\star}_{M'}(s,a) > V^{\pi^\star}_{M'}(s)\}.$$

where $\pi^\star$ is the optimal policy in M and $V^{\pi^\star}_{M'}$ is the evaluation of $\pi^\star$ in M'.

We begin by proving that $\mathrm{Alt}(M) \subset \cup_{s,a\neq\pi^\star(s)}\mathrm{Alt}_{s,a}(M)$ (which is more important, why?).

- By contradiction, assume $\exists M' \in \mathrm{Alt}(M)$ s.t. $\forall s, a \neq \pi^\star(s)$ we have $M' \notin \mathrm{Alt}_{s,a}(M)$.
- Therefore $Q^{\pi^\star}_{M'}(s,a) \leq V^{\pi^\star}_{M'}(s)$ for every $s, a \neq \pi^\star(s)$.
- Moreover $Q^{\pi^\star}_{M'}(s,\pi^\star(s)) = V^{\pi^\star}_{M'}(s)$ for every s.
- Hence $Q^{\pi^\star}_{M'}(s,a) \leq V^{\pi^\star}_{M'}(s)$ for every (s,a).
- By the policy improvement theorem there does not exists any action that improves the policy, hence $\pi^\star$ is optimal in M' $\Rightarrow$ contradiction!

Rewriting the set of confusing models [1/2]

Lemma

We have that

$$\mathrm{Alt}(M) = \cup_{s,a\neq\pi^\star(s)}\mathrm{Alt}_{s,a}(M) \text{ where } \mathrm{Alt}_{s,a}(M) = \{M' : Q^{\pi^\star}_{M'}(s,a) > V^{\pi^\star}_{M'}(s)\}.$$

where $\pi^\star$ is the optimal policy in M and $V^{\pi^\star}_{M'}$ is the evaluation of $\pi^\star$ in M'.

We begin by proving that $\mathrm{Alt}(M) \subset \cup_{s,a\neq\pi^\star(s)}\mathrm{Alt}_{s,a}(M)$ (which is more important, why?).

- By contradiction, assume $\exists M' \in \mathrm{Alt}(M)$ s.t. $\forall s, a \neq \pi^\star(s)$ we have $M' \notin \mathrm{Alt}_{s,a}(M)$.
- Therefore $Q^{\pi^\star}_{M'}(s,a) \leq V^{\pi^\star}_{M'}(s)$ for every $s, a \neq \pi^\star(s)$.
- Moreover $Q^{\pi^\star}_{M'}(s,\pi^\star(s)) = V^{\pi^\star}_{M'}(s)$ for every s.
- Hence $Q^{\pi^\star}_{M'}(s,a) \leq V^{\pi^\star}_{M'}(s)$ for every (s,a).
- By the policy improvement theorem there does not exists any action that improves the policy, hence $\pi^\star$ is optimal in $M' \Rightarrow$ contradiction!

Rewriting the set of confusing models [1/2]

Lemma

We have that

$$\mathrm{Alt}(M) = \cup_{s,a\neq\pi^\star(s)} \mathrm{Alt}_{s,a}(M) \text{ where } \mathrm{Alt}_{s,a}(M) = \{M' : Q^{\pi^\star}_{M'}(s,a) > V^{\pi^\star}_{M'}(s)\}.$$

where $\pi^\star$ is the optimal policy in M and $V^{\pi^\star}_{M'}$ is the evaluation of $\pi^\star$ in M'.

We begin by proving that $\mathrm{Alt}(M) \subset \cup_{s,a\neq\pi^\star(s)} \mathrm{Alt}_{s,a}(M)$ (which is more important, why?).

- By contradiction, assume $\exists M' \in \mathrm{Alt}(M)$ s.t. $\forall s, a \neq \pi^\star(s)$ we have $M' \notin \mathrm{Alt}_{s,a}(M)$.
- Therefore $Q^{\pi^\star}_{M'}(s,a) \leq V^{\pi^\star}_{M'}(s)$ for every $s, a \neq \pi^\star(s)$.
- Moreover $Q^{\pi^\star}_{M'}(s,\pi^\star(s)) = V^{\pi^\star}_{M'}(s)$ for every s.
- Hence $Q^{\pi^\star}_{M'}(s,a) \leq V^{\pi^\star}_{M'}(s)$ for every (s,a).
- By the policy improvement theorem there does not exists any action that improves the policy, hence $\pi^\star$ is optimal in M' $\Rightarrow$ contradiction!

Rewriting the set of confusing models [1/2]

Lemma

We have that

$$\mathrm{Alt}(M) = \cup_{s,a \neq \pi^\star(s)} \mathrm{Alt}_{s,a}(M) \text{ where } \mathrm{Alt}_{s,a}(M) = \{M' : Q^{\pi^\star}_{M'}(s,a) > V^{\pi^\star}_{M'}(s)\}.$$

where $\pi^\star$ is the optimal policy in M and $V^{\pi^\star}_{M'}$ is the evaluation of $\pi^\star$ in M'.

We begin by proving that $\mathrm{Alt}(M) \subset \cup_{s,a \neq \pi^\star(s)} \mathrm{Alt}_{s,a}(M)$ (which is more important, why?).

- By contradiction, assume $\exists M' \in \mathrm{Alt}(M)$ s.t. $\forall s, a \neq \pi^\star(s)$ we have $M' \notin \mathrm{Alt}_{s,a}(M)$.
- Therefore $Q^{\pi^\star}_{M'}(s,a) \leq V^{\pi^\star}_{M'}(s)$ for every $s, a \neq \pi^\star(s)$.
- Moreover $Q^{\pi^\star}_{M'}(s,\pi^\star(s)) = V^{\pi^\star}_{M'}(s)$ for every s.
- Hence $Q^{\pi^\star}_{M'}(s,a) \leq V^{\pi^\star}_{M'}(s)$ for every (s,a).
- By the policy improvement theorem there does not exists any action that improves the policy, hence $\pi^\star$ is optimal in $M' \Rightarrow$ contradiction!

Rewriting the set of confusing models [1/2]

Lemma

We have that

$$\mathrm{Alt}(M) = \cup_{s,a\neq\pi^\star(s)} \mathrm{Alt}_{s,a}(M) \textit{ where } \mathrm{Alt}_{s,a}(M) = \{M' : Q^{\pi^\star}_{M'}(s,a) > V^{\pi^\star}_{M'}(s)\}.$$

where $\pi^\star$ is the optimal policy in M and $V^{\pi^\star}_{M'}$ is the evaluation of $\pi^\star$ in M'.

We begin by proving that $\mathrm{Alt}(M) \subset \cup_{s,a\neq\pi^\star(s)} \mathrm{Alt}_{s,a}(M)$ (which is more important, why?).

- By contradiction, assume $\exists M' \in \mathrm{Alt}(M)$ s.t. $\forall s, a \neq \pi^\star(s)$ we have $M' \notin \mathrm{Alt}_{s,a}(M)$.
- Therefore $Q^{\pi^\star}_{M'}(s,a) \leq V^{\pi^\star}_{M'}(s)$ for every $s, a \neq \pi^\star(s)$.
- Moreover $Q^{\pi^\star}_{M'}(s,\pi^\star(s)) = V^{\pi^\star}_{M'}(s)$ for every s.
- Hence $Q^{\pi^\star}_{M'}(s,a) \leq V^{\pi^\star}_{M'}(s)$ for every (s,a).
- By the policy improvement theorem there does not exists any action that improves the policy, hence $\pi^\star$ is optimal in $M' \Rightarrow$ contradiction!

Rewriting the set of confusing models [2/2]

Lemma

We have that

$$\mathrm{Alt}(M) = \cup_{s,a\neq\pi^\star(s)}\mathrm{Alt}_{s,a}(M) \text{ where } \mathrm{Alt}_{s,a}(M) = \{M' : Q^{\pi^\star}_{M'}(s,a) > V^{\pi^\star}_{M'}(s)\}.$$

where $\pi^\star$ is the optimal policy in M and $V^{\pi^\star}_{M'}$ is the evaluation of $\pi^\star$ in M'.

We now prove $\mathrm{Alt}(M) \supset \cup_{s,a\neq\pi^\star(s)}\mathrm{Alt}_{s,a}(M)$.

- Consider a generic pair $s_0, a_0 \neq \pi^\star(s_0)$. By contradiction, assume $\exists M' \in \mathrm{Alt}_{s_0,a_0}(M)$ s.t.$M' \notin \mathrm{Alt}(M)$.
- Define the policy

$$\pi'(s) = \begin{cases} a_0 & s = s_0, \\ \pi^\star(s) & \text{otherwise.} \end{cases}$$

- Then, we have that $Q^{\pi^\star}_{M'}(s_0, \pi'(s_0)) > V^{\pi^\star}_{M'}(s_0)$. However, if $M' \notin \mathrm{Alt}(M)$, then $\pi^\star$ is optimal in M', which is not possible again by the policy improvement theorem.

Rewriting the set of confusing models [2/2]

Lemma

We have that

$$\mathrm{Alt}(M) = \cup_{s,a\neq\pi^\star(s)}\mathrm{Alt}_{s,a}(M) \text{ where } \mathrm{Alt}_{s,a}(M) = \{M' : Q^{\pi^\star}_{M'}(s,a) > V^{\pi^\star}_{M'}(s)\}.$$

where $\pi^\star$ is the optimal policy in M and $V^{\pi^\star}_{M'}$ is the evaluation of $\pi^\star$ in M'.

We now prove $\mathrm{Alt}(M) \supset \cup_{s,a\neq\pi^\star(s)}\mathrm{Alt}_{s,a}(M)$.

- Consider a generic pair $s_0, a_0 \neq \pi^\star(s_0)$. By contradiction, assume $\exists M' \in \mathrm{Alt}_{s_0,a_0}(M)$ s.t.$M' \notin \mathrm{Alt}(M)$.
- Define the policy

$$\pi'(s) = \begin{cases} a_0 & s = s_0, \\ \pi^\star(s) & \text{otherwise.} \end{cases}$$

- Then, we have that $Q^{\pi^\star}_{M'}(s_0,\pi'(s_0)) > V^{\pi^\star}_{M'}(s_0)$. However, if $M' \notin \mathrm{Alt}(M)$, then $\pi^\star$ is optimal in M', which is not possible again by the policy improvement theorem.

Rewriting the set of confusing models [2/2]

Lemma

We have that

$$\mathrm{Alt}(M) = \cup_{s,a\neq\pi^\star(s)} \mathrm{Alt}_{s,a}(M) \text{ where } \mathrm{Alt}_{s,a}(M) = \{M' : Q^{\pi^\star}_{M'}(s,a) > V^{\pi^\star}_{M'}(s)\}.$$

where $\pi^\star$ is the optimal policy in M and $V^{\pi^\star}_{M'}$ is the evaluation of $\pi^\star$ in M'.

We now prove $\mathrm{Alt}(M) \supset \cup_{s,a\neq\pi^\star(s)} \mathrm{Alt}_{s,a}(M)$.

- Consider a generic pair $s_0, a_0 \neq \pi^\star(s_0)$. By contradiction, assume $\exists M' \in \mathrm{Alt}_{s_0,a_0}(M)$ s.t. $M' \notin \mathrm{Alt}(M)$.
- Define the policy

$$\pi'(s) = \begin{cases} a_0 & s = s_0, \\ \pi^\star(s) & \text{otherwise.} \end{cases}$$

- Then, we have that $Q^{\pi^\star}_{M'}(s_0, \pi'(s_0)) > V^{\pi^\star}_{M'}(s_0)$. However, if $M' \notin \mathrm{Alt}(M)$, then $\pi^\star$ is optimal in M', which is not possible again by the policy improvement theorem.

Rewriting the set of confusing models [2/2]

Lemma

We have that

$$\mathrm{Alt}(M) = \cup_{s,a\neq\pi^\star(s)}\mathrm{Alt}_{s,a}(M) \text{ where } \mathrm{Alt}_{s,a}(M) = \{M' : Q^{\pi^\star}_{M'}(s,a) > V^{\pi^\star}_{M'}(s)\}.$$

where $\pi^\star$ is the optimal policy in M and $V^{\pi^\star}_{M'}$ is the evaluation of $\pi^\star$ in M'.

We now prove $\mathrm{Alt}(M) \supset \cup_{s,a\neq\pi^\star(s)}\mathrm{Alt}_{s,a}(M)$.

- Consider a generic pair $s_0, a_0 \neq \pi^\star(s_0)$. By contradiction, assume $\exists M' \in \mathrm{Alt}_{s_0,a_0}(M)$ s.t.$M' \notin \mathrm{Alt}(M)$.
- Define the policy

$$\pi'(s) = \begin{cases} a_0 & s = s_0, \\ \pi^\star(s) & \text{otherwise.} \end{cases}$$

- Then, we have that $Q^{\pi^\star}_{M'}(s_0, \pi'(s_0)) > V^{\pi^\star}_{M'}(s_0)$. However, if $M' \notin \mathrm{Alt}(M)$, then $\pi^\star$ is optimal in M', which is not possible again by the policy improvement theorem.

Relating the sub-optimality gaps to the KL terms

Using this decomposition we get

$$\begin{aligned}
T^{-1}(\omega) &= \inf_{M' \in \mathrm{Alt}(M)} \mathbb{E}_{(s,a)\sim\omega}[\mathrm{KL}(P(s,a), P'(s,a))], \\
&= \min_{s,a \neq \pi^\star(s)} \inf_{M' \in \mathrm{Alt}_{s,a}(M)} \mathbb{E}_{(s,a)\sim\omega}[\mathrm{KL}(P(s,a), P'(s,a))], \\
&= \min_{s,a \neq \pi^\star(s)} \inf_{M' \in \mathrm{Alt}_{s,a}(M)} \omega(s,a)\mathrm{KL}(P(s,a), P'(s,a)) \\
&\qquad + \sum_{s'} \omega_{s',\pi^\star(s')} \mathrm{KL}(P(s',\pi^\star(s')), P'(s',\pi^\star(s'))), \\
&\geq \min_{s,a \neq \pi^\star(s)} \inf_{M' \in \mathrm{Alt}_{s,a}(M)} \omega(s,a)\mathrm{KL}(P(s,a), P'(s,a)) \\
&\qquad + (\min_{s'} \omega_{s',\pi^\star(s')}) \max_{s'} \mathrm{KL}(P(s',\pi^\star(s')), P'(s',\pi^\star(s'))),
\end{aligned}$$

where we used the fact that the constraints only involve the pairs $\{(s,a), (s',\pi^\star(s'))_{s'}\}$.

Relating the sub-optimality gaps to the KL terms

Using this decomposition we get

$$
\begin{aligned}
T^{-1}(\omega) &= \inf_{M' \in \mathrm{Alt}(M)} \mathbb{E}_{(s,a)\sim\omega}[\mathrm{KL}(P(s,a), P'(s,a))], \\
&= \min_{s,a\neq\pi^\star(s)} \inf_{M'\in\mathrm{Alt}_{s,a}(M)} \mathbb{E}_{(s,a)\sim\omega}[\mathrm{KL}(P(s,a), P'(s,a))], \\
&= \min_{s,a\neq\pi^\star(s)} \inf_{M'\in\mathrm{Alt}_{s,a}(M)} \omega(s,a)\mathrm{KL}(P(s,a), P'(s,a)) \\
&\qquad + \sum_{s'} \omega_{s',\pi^\star(s')}\mathrm{KL}(P(s',\pi^\star(s')), P'(s',\pi^\star(s'))), \\
&\geq \min_{s,a\neq\pi^\star(s)} \inf_{M'\in\mathrm{Alt}_{s,a}(M)} \omega(s,a)\mathrm{KL}(P(s,a), P'(s,a)) \\
&\qquad + (\min_{s'} \omega_{s',\pi^\star(s')}) \max_{s'} \mathrm{KL}(P(s',\pi^\star(s')), P'(s',\pi^\star(s'))),
\end{aligned}
$$

where we used the fact that the constraints only involve the pairs $\{(s,a), (s',\pi^\star(s'))_{s'}\}$.

Relating the sub-optimality gaps to the KL terms

Using this decomposition we get

$$
\begin{aligned}
T^{-1}(\omega) &= \inf_{M' \in \mathrm{Alt}(M)} \mathbb{E}_{(s,a)\sim\omega}[\mathrm{KL}(P(s,a), P'(s,a))], \\
&= \min_{s, a \neq \pi^\star(s)} \inf_{M' \in \mathrm{Alt}_{s,a}(M)} \mathbb{E}_{(s,a)\sim\omega}[\mathrm{KL}(P(s,a), P'(s,a))], \\
&= \min_{s, a \neq \pi^\star(s)} \inf_{M' \in \mathrm{Alt}_{s,a}(M)} \omega(s,a)\mathrm{KL}(P(s,a), P'(s,a)) \\
&\qquad + \sum_{s'} \omega_{s', \pi^\star(s')} \mathrm{KL}(P(s', \pi^\star(s'), P'(s', \pi^\star(s'))), \\
&\geq \min_{s, a \neq \pi^\star(s)} \inf_{M' \in \mathrm{Alt}_{s,a}(M)} \omega(s,a)\mathrm{KL}(P(s,a), P'(s,a)) \\
&\qquad + (\min_{s'} \omega_{s', \pi^\star(s')}) \max_{s'} \mathrm{KL}(P(s', \pi^\star(s'), P'(s', \pi^\star(s'))),
\end{aligned}
$$

where we used the fact that the constraints only involve the pairs $\{(s,a), (s', \pi^\star(s'))_{s'}\}$.

Relating the sub-optimality gaps to the KL terms

So we have that

$$\mathrm{Alt}(M) = \cup_{s,a\neq\pi^\star(s)} \mathrm{Alt}_{s,a}(M) \text{ where } \mathrm{Alt}_{s,a}(M) = \{M' : Q_{M'}^{\pi^\star}(s,a) > V_{M'}^{\pi^\star}(s)\}.$$

How can we relate the KL terms to this constraint and to $\Delta(s,a)$? We know that $\Delta_{s,a} + Q^\star(s,a) = V^\star(s)$. Then combine the inequality with this equality to get

$$\Delta(s,a) < V^\star(s) - V_{M'}^{\pi^\star}(s) + Q_{M'}^{\pi^\star}(s,a) - Q^\star(s,a) \pm \mathbb{E}_{s'\sim P'(s,a)}[V^\star(s')].$$

from which follows that (we write in vector form)

$$\begin{aligned}\Delta(s,a) &< \Delta V(s) + \gamma P'(s,a)^\top \Delta V + \Delta P(s,a)^\top V^\star,\\ &< (\gamma P'(s,a) - \mathbf{1}_s)^\top \Delta V + \Delta P(s,a)^\top V^\star.\end{aligned}$$

where $\Delta V = V_{M'}^{\pi^\star} - V^\star, \Delta P(s,a) = P'(s,a) - P(s,a)$, which are all vectors of size $|S|$.

Relating the sub-optimality gaps to the KL terms

So we have that

$$\mathrm{Alt}(M) = \cup_{s,a\neq\pi^\star(s)}\mathrm{Alt}_{s,a}(M) \text{ where } \mathrm{Alt}_{s,a}(M) = \{M' : Q_{M'}^{\pi^\star}(s,a) > V_{M'}^{\pi^\star}(s)\}.$$

How can we relate the KL terms to this constraint and to $\Delta(s,a)$? We know that $\Delta_{s,a} + Q^\star(s,a) = V^\star(s)$. Then combine the inequality with this equality to get

$$\Delta(s,a) < V^\star(s) - V_{M'}^{\pi^\star}(s) + Q_{M'}^{\pi^\star}(s,a) - Q^\star(s,a) \pm \mathbb{E}_{s'\sim P'(s,a)}[V^\star(s')].$$

from which follows that (we write in vector form)

$$\begin{aligned}\Delta(s,a) &< \Delta V(s) + \gamma P'(s,a)^\top \Delta V + \Delta P(s,a)^\top V^\star,\\ &< (\gamma P'(s,a) - \mathbf{1}_s)^\top \Delta V + \Delta P(s,a)^\top V^\star.\end{aligned}$$

where $\Delta V = V_{M'}^{\pi^\star} - V^\star, \Delta P(s,a) = P'(s,a) - P(s,a)$, which are all vectors of size $|S|$.

Relating the sub-optimality gaps to the KL terms

So we have that

$$\mathrm{Alt}(M) = \cup_{s,a\neq\pi^\star(s)}\mathrm{Alt}_{s,a}(M) \text{ where } \mathrm{Alt}_{s,a}(M) = \{M' : Q_{M'}^{\pi^\star}(s,a) > V_{M'}^{\pi^\star}(s)\}.$$

How can we relate the KL terms to this constraint and to $\Delta(s,a)$? We know that $\Delta_{s,a} + Q^\star(s,a) = V^\star(s)$. Then combine the inequality with this equality to get

$$\Delta(s,a) < V^\star(s) - V_{M'}^{\pi^\star}(s) + Q_{M'}^{\pi^\star}(s,a) - Q^\star(s,a) \pm \mathbb{E}_{s'\sim P'(s,a)}[V^\star(s')].$$

from which follows that (we write in vector form)

$$\begin{aligned}\Delta(s,a) &< \Delta V(s) + \gamma P'(s,a)^\top \Delta V + \Delta P(s,a)^\top V^\star,\\ &< (\gamma P'(s,a) - \mathbf{1}_s)^\top \Delta V + \Delta P(s,a)^\top V^\star.\end{aligned}$$

where $\Delta V = V_{M'}^{\pi^\star} - V^\star, \Delta P(s,a) = P'(s,a) - P(s,a)$, which are all vectors of size $|S|$.

Relating the sub-optimality gaps to the KL terms

$$\Delta(s,a) < (\gamma P'(s,a) - \mathbf{1}_s)^\top \Delta V + \Delta P(s,a)^\top V^\star.$$

We upper bound ΔV using

$$\begin{aligned}
|\Delta V(s)| &= \gamma |\mathbb{E}_{s' \sim P'(s,\pi^\star(s))}[V_{M'}^{\pi^\star}(s')] - \mathbb{E}_{s' \sim P(s,\pi^\star(s))}[V^\star(s')]|, \\
&\leq \gamma(|P'(s,\pi^\star(s))^\top \Delta V| + |\Delta P(s,\pi^\star(s))^\top V^\star|), \\
&\leq \gamma(\|\Delta V\|_\infty + |\Delta P(s,\pi^\star(s))^\top V^\star|).
\end{aligned}$$

Therefore $\|\Delta V\|_\infty \leq \frac{\gamma |\Delta P(s,\pi^\star(s))^\top V^\star|}{1-\gamma}$ and

$$\Delta(s,a) < \frac{\gamma |\Delta P(s,\pi^\star(s))^\top V^\star|}{1-\gamma} + \Delta P(s,a)^\top V^\star.$$

We have rewritten the inequality in terms of the inner product $\Delta P^\top V^\star$. Can we upper bound this using the KL between P and P'?

Relating the sub-optimality gaps to the KL terms

$$\Delta(s,a) < (\gamma P'(s,a) - \mathbf{1}_s)^\top \Delta V + \Delta P(s,a)^\top V^\star.$$

We upper bound ΔV using

$$\begin{aligned} |\Delta V(s)| &= \gamma |\mathbb{E}_{s' \sim P'(s,\pi^\star(s))}[V_{M'}^{\pi^\star}(s')] - \mathbb{E}_{s' \sim P(s,\pi^\star(s))}[V^\star(s')]|, \\ &\leq \gamma(|P'(s,\pi^\star(s))^\top \Delta V| + |\Delta P(s,\pi^\star(s))^\top V^\star|), \\ &\leq \gamma(\|\Delta V\|_\infty + |\Delta P(s,\pi^\star(s))^\top V^\star|). \end{aligned}$$

Therefore $\|\Delta V\|_\infty \leq \frac{\gamma |\Delta P(s,\pi^\star(s))^\top V^\star|}{1-\gamma}$ and

$$\Delta(s,a) < \frac{\gamma |\Delta P(s,\pi^\star(s))^\top V^\star|}{1-\gamma} + \Delta P(s,a)^\top V^\star.$$

We have rewritten the inequality in terms of the inner product $\Delta P^\top V^\star$. Can we upper bound this using the KL between P and P'?

Relating the sub-optimality gaps to the KL terms

$$\Delta(s, a) < \frac{\gamma |\Delta P(s, \pi^\star(s))^\top V^\star|}{1 - \gamma} + \Delta P(s, a)^\top V^\star.$$

Can we upper bound the inner product using the KL between P and P'? Define the following quantities

- Variance of V^π in (s, a): $\mathrm{Var}_{s,a}(V^\pi) := \mathbb{E}_{s' \sim P(s,a)} \left[(V^\pi(s') - \mathbb{E}_{s'' \sim P(s,a)}[V^\pi(s'')])^2 \right]$.
- Maximum deviation of V^π in (s, a): $\mathrm{MD}_{s,a}(V^\pi) := \left\| V^\pi(s') - \mathbb{E}_{s'' \sim P(s,a)}[V^\pi(s'')] \right\|_\infty$.

Relating the sub-optimality gaps to the KL terms

Lemma

Let $(s,a) \in S \times A$. *For any policy* π *we have that*

$$|(V^\pi)^\top \Delta P(s,a)| \leq 4\mathrm{KL}(P(s,a), P'(s,a)) \Big[2\mathrm{Var}_{s,a}(V^\pi) + \sqrt{2\mathrm{KL}(P(s,a), P'(s,a))}\mathrm{MD}_{s,a}(V^\pi)^2\Big].$$

where $V^\pi \in \mathbb{R}^{|S|}$ *is the vector of values of the policy* π *and*

$$\Delta P(s,a) = \Big[P'(s_1|s,a) - P(s_1|s,a) \quad \ldots \quad P'(s_{|S|}|s,a) - P(s_{|S|}|s,a)\Big]^\top.$$

Let $\mu^\pi = \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^\pi(s')]$ and note that $(V^\pi)^\top \Delta P(s,a) = (V^\pi - \mu^\pi)^\top \Delta P(s,a)$.

$$|(V^\pi - \mu^\pi)^\top \Delta P(s,a)| \leq \Big|[(\sqrt{P'(s,a)} - \sqrt{P(s,a)}) \circ (\sqrt{P'(s,a)} + \sqrt{P(s,a)})]^\top (V^\pi - \mu^\pi)\Big|$$

where $\sqrt{x}$ is element-wise, and similarly $\circ$ is the element-wise product.

Relating the sub-optimality gaps to the KL terms

Lemma

Let $(s,a) \in S \times A$. *For any policy* π *we have that*

$$|(V^\pi)^\top \Delta P(s,a)| \le 4\mathrm{KL}(P(s,a), P'(s,a))\Big[2\mathrm{Var}_{s,a}(V^\pi) + \sqrt{2\mathrm{KL}(P(s,a), P'(s,a))}\mathrm{MD}_{s,a}(V^\pi)^2\Big].$$

where $V^\pi \in \mathbb{R}^{|S|}$ *is the vector of values of the policy* π *and*

$$\Delta P(s,a) = \Big[P'(s_1|s,a) - P(s_1|s,a) \quad \dots \quad P'(s_{|S|}|s,a) - P(s_{|S|}|s,a)\Big]^\top.$$

Let $\mu^\pi = \mathbb{E}_{s' \sim P(\cdot|s,a)}[V^\pi(s')]$ and note that $(V^\pi)^\top \Delta P(s,a) = (V^\pi - \mu^\pi)^\top \Delta P(s,a)$.

$$|(V^\pi - \mu^\pi)^\top \Delta P(s,a)| \le \Big|[(\sqrt{P'(s,a)} - \sqrt{P(s,a)}) \circ (\sqrt{P'(s,a)} + \sqrt{P(s,a)})]^\top (V^\pi - \mu^\pi)\Big|$$

where $\sqrt{x}$ is element-wise, and similarly $\circ$ is the element-wise product.

Relating the sub-optimality gaps to the KL terms

$$\begin{aligned}
|(V^\pi)^\top \Delta P(s,a)|^2 &\le \left|[(\sqrt{P'(s,a)} - \sqrt{P(s,a)}) \circ (\sqrt{P'(s,a)} + \sqrt{P(s,a)})]^\top (V^\pi - \mu^\pi))\right|^2, \\
&= \left|(\sqrt{P'(s,a)} - \sqrt{P(s,a)})^\top [(\sqrt{P'(s,a)} + \sqrt{P(s,a)}) \circ (V^\pi - \mu^\pi))]\right|^2, \\
&\overset{(a)}{\le} \left\|\sqrt{P'(s,a)} - \sqrt{P(s,a)}\right\|_2^2 \left\|(\sqrt{P'(s,a)} + \sqrt{P(s,a)}) \circ (V^\pi - \mu^\pi))]\right\|_2^2, \\
&\overset{(b)}{\le} 4H^2(P(s,a), P'(s,a)) \left[|P'(s,a) + P(s,a)|^\top (V^\pi - \mu^\pi)^{\circ 2})\right], \\
&\overset{(c)}{\le} 4\mathrm{KL}(P(s,a), P'(s,a)) \left[|P'(s,a) + 2P(s,a) - P(s,a)|^\top (V^\pi - \mu^\pi)^{\circ 2})\right], \\
&\le 4\mathrm{KL}(P(s,a), P'(s,a)) \left[2\mathrm{Var}_{s,a}(V^\pi) + \|P'(s,a) - P(s,a)\|_1 \mathrm{MD}_{s,a}(V^\pi)^2\right], \\
&\le 4\mathrm{KL}(P(s,a), P'(s,a)) \left[2\mathrm{Var}_{s,a}(V^\pi) + \sqrt{2\mathrm{KL}(P(s,a), P'(s,a))}\mathrm{MD}_{s,a}(V^\pi)^2\right].
\end{aligned}$$

(a) Cauchy-Schwarz ineq.; (b) definition of Hellinger's distance (add a factor 2) and used $(a+b)^2 \le 2(a^2+b^2)$; (c) $H(P,Q) \le \sqrt{\mathrm{KL}(P,Q)}$.

Relating the sub-optimality gaps to the KL terms

$$
\begin{aligned}
|(V^\pi)^\top \Delta P(s,a)|^2 &\le \left|[(\sqrt{P'(s,a)} - \sqrt{P(s,a)}) \circ (\sqrt{P'(s,a)} + \sqrt{P(s,a)})]^\top (V^\pi - \mu^\pi)\right|^2, \\
&= \left|(\sqrt{P'(s,a)} - \sqrt{P(s,a)})^\top [(\sqrt{P'(s,a)} + \sqrt{P(s,a)}) \circ (V^\pi - \mu^\pi)]\right|^2, \\
&\overset{(a)}{\le} \left\|\sqrt{P'(s,a)} - \sqrt{P(s,a)}\right\|_2^2 \left\|(\sqrt{P'(s,a)} + \sqrt{P(s,a)}) \circ (V^\pi - \mu^\pi))]\right\|_2^2, \\
&\overset{(b)}{\le} 4H^2(P(s,a), P'(s,a)) \left[|P'(s,a) + P(s,a)|^\top (V^\pi - \mu^\pi)^{\circ 2})\right], \\
&\overset{(c)}{\le} 4\mathrm{KL}(P(s,a), P'(s,a)) \left[|P'(s,a) + 2P(s,a) - P(s,a)|^\top (V^\pi - \mu^\pi)^{\circ 2})\right], \\
&\le 4\mathrm{KL}(P(s,a), P'(s,a)) \left[2\mathrm{Var}_{s,a}(V^\pi) + \|P'(s,a) - P(s,a)\|_1 \mathrm{MD}_{s,a}(V^\pi)^2\right], \\
&\le 4\mathrm{KL}(P(s,a), P'(s,a)) \left[2\mathrm{Var}_{s,a}(V^\pi) + \sqrt{2\mathrm{KL}(P(s,a), P'(s,a))}\mathrm{MD}_{s,a}(V^\pi)^2\right].
\end{aligned}
$$

(a) Cauchy-Schwarz ineq.; (b) definition of Hellinger's distance (add a factor 2) and used $(a+b)^2 \le 2(a^2+b^2)$; (c) $H(P,Q) \le \sqrt{\mathrm{KL}(P,Q)}$.

Relating the sub-optimality gaps to the KL terms

$$
\begin{aligned}
|(V^\pi)^\top \Delta P(s,a)|^2 &\le \left|[(\sqrt{P'(s,a)} - \sqrt{P(s,a)}) \circ (\sqrt{P'(s,a)} + \sqrt{P(s,a)})]^\top (V^\pi - \mu^\pi)\right|^2, \\
&= \left|(\sqrt{P'(s,a)} - \sqrt{P(s,a)})^\top [(\sqrt{P'(s,a)} + \sqrt{P(s,a)}) \circ (V^\pi - \mu^\pi))]\right|^2, \\
&\overset{(a)}{\le} \left\|\sqrt{P'(s,a)} - \sqrt{P(s,a)}\right\|_2^2 \left\|(\sqrt{P'(s,a)} + \sqrt{P(s,a)}) \circ (V^\pi - \mu^\pi))]\right\|_2^2, \\
&\overset{(b)}{\le} 4H^2(P(s,a), P'(s,a)) \left[|P'(s,a) + P(s,a)|^\top (V^\pi - \mu^\pi)^{\circ 2})\right], \\
&\overset{(c)}{\le} 4\mathrm{KL}(P(s,a), P'(s,a)) \left[|P'(s,a) + 2P(s,a) - P(s,a)|^\top (V^\pi - \mu^\pi)^{\circ 2})\right], \\
&\le 4\mathrm{KL}(P(s,a), P'(s,a)) \left[2\mathrm{Var}_{s,a}(V^\pi) + \|P'(s,a) - P(s,a)\|_1 \mathrm{MD}_{s,a}(V^\pi)^2\right], \\
&\le 4\mathrm{KL}(P(s,a), P'(s,a)) \left[2\mathrm{Var}_{s,a}(V^\pi) + \sqrt{2\mathrm{KL}(P(s,a), P'(s,a))}\mathrm{MD}_{s,a}(V^\pi)^2\right].
\end{aligned}
$$

(a) Cauchy-Schwarz ineq.; (b) definition of Hellinger's distance (add a factor 2) and used $(a+b)^2 \le 2(a^2+b^2)$; (c) $H(P,Q) \le \sqrt{\mathrm{KL}(P,Q)}$.

Relating the sub-optimality gaps to the KL terms

$$
\begin{aligned}
|(V^\pi)^\top \Delta P(s,a)|^2 &\le \left| [(\sqrt{P'(s,a)} - \sqrt{P(s,a)}) \circ (\sqrt{P'(s,a)} + \sqrt{P(s,a)})]^\top (V^\pi - \mu^\pi)) \right|^2, \\
&= \left| (\sqrt{P'(s,a)} - \sqrt{P(s,a)})^\top [(\sqrt{P'(s,a)} + \sqrt{P(s,a)}) \circ (V^\pi - \mu^\pi))] \right|^2, \\
&\overset{(a)}{\le} \left\| \sqrt{P'(s,a)} - \sqrt{P(s,a)} \right\|_2^2 \left\| (\sqrt{P'(s,a)} + \sqrt{P(s,a)}) \circ (V^\pi - \mu^\pi))] \right\|_2^2, \\
&\overset{(b)}{\le} 4H^2(P(s,a), P'(s,a)) \left[|P'(s,a) + P(s,a)|^\top (V^\pi - \mu^\pi)^{\circ 2}) \right], \\
&\overset{(c)}{\le} 4\mathrm{KL}(P(s,a), P'(s,a)) \left[|P'(s,a) + 2P(s,a) - P(s,a)|^\top (V^\pi - \mu^\pi)^{\circ 2}) \right], \\
&\le 4\mathrm{KL}(P(s,a), P'(s,a)) \left[2\mathrm{Var}_{s,a}(V^\pi) + \|P'(s,a) - P(s,a)\|_1 \mathrm{MD}_{s,a}(V^\pi)^2 \right], \\
&\le 4\mathrm{KL}(P(s,a), P'(s,a)) \left[2\mathrm{Var}_{s,a}(V^\pi) + \sqrt{2\mathrm{KL}(P(s,a), P'(s,a))} \mathrm{MD}_{s,a}(V^\pi)^2 \right].
\end{aligned}
$$

(a) Cauchy-Schwarz ineq.; (b) definition of Hellinger's distance (add a factor 2) and used $(a+b)^2 \le 2(a^2+b^2)$; (c) $H(P,Q) \le \sqrt{\mathrm{KL}(P,Q)}$.

Relating the sub-optimality gaps to the KL terms

$$
\begin{aligned}
|(V^\pi)^\top \Delta P(s,a)|^2 &\le \left|[(\sqrt{P'(s,a)} - \sqrt{P(s,a)}) \circ (\sqrt{P'(s,a)} + \sqrt{P(s,a)})]^\top (V^\pi - \mu^\pi))\right|^2, \\
&= \left|(\sqrt{P'(s,a)} - \sqrt{P(s,a)})^\top [(\sqrt{P'(s,a)} + \sqrt{P(s,a)}) \circ (V^\pi - \mu^\pi))]\right|^2, \\
&\overset{(a)}{\le} \left\|\sqrt{P'(s,a)} - \sqrt{P(s,a)}\right\|_2^2 \left\|(\sqrt{P'(s,a)} + \sqrt{P(s,a)}) \circ (V^\pi - \mu^\pi))]\right\|_2^2, \\
&\overset{(b)}{\le} 4H^2(P(s,a), P'(s,a)) \left[|P'(s,a) + P(s,a)|^\top (V^\pi - \mu^\pi)^{\circ 2})\right], \\
&\overset{(c)}{\le} 4\mathrm{KL}(P(s,a), P'(s,a)) \left[|P'(s,a) + 2P(s,a) - P(s,a)|^\top (V^\pi - \mu^\pi)^{\circ 2})\right], \\
&\le 4\mathrm{KL}(P(s,a), P'(s,a)) \left[2\mathrm{Var}_{s,a}(V^\pi) + \|P'(s,a) - P(s,a)\|_1 \mathrm{MD}_{s,a}(V^\pi)^2\right], \\
&\le 4\mathrm{KL}(P(s,a), P'(s,a)) \left[2\mathrm{Var}_{s,a}(V^\pi) + \sqrt{2\mathrm{KL}(P(s,a), P'(s,a))}\mathrm{MD}_{s,a}(V^\pi)^2\right].
\end{aligned}
$$

(a) Cauchy-Schwarz ineq.; (b) definition of Hellinger's distance (add a factor 2) and used $(a+b)^2 \le 2(a^2+b^2)$; (c) $H(P,Q) \le \sqrt{\mathrm{KL}(P,Q)}$.

Relating the sub-optimality gaps to the KL terms

$$
\begin{aligned}
|(V^\pi)^\top \Delta P(s,a)|^2 &\leq \left|[(\sqrt{P'(s,a)} - \sqrt{P(s,a)}) \circ (\sqrt{P'(s,a)} + \sqrt{P(s,a)})]^\top (V^\pi - \mu^\pi)\right|^2, \\
&= \left|(\sqrt{P'(s,a)} - \sqrt{P(s,a)})^\top [(\sqrt{P'(s,a)} + \sqrt{P(s,a)}) \circ (V^\pi - \mu^\pi))]\right|^2, \\
&\overset{(a)}{\leq} \left\|\sqrt{P'(s,a)} - \sqrt{P(s,a)}\right\|_2^2 \left\|(\sqrt{P'(s,a)} + \sqrt{P(s,a)}) \circ (V^\pi - \mu^\pi))]\right\|_2^2, \\
&\overset{(b)}{\leq} 4H^2(P(s,a), P'(s,a)) \left[|P'(s,a) + P(s,a)|^\top (V^\pi - \mu^\pi)^{\circ 2})\right], \\
&\overset{(c)}{\leq} 4\mathrm{KL}(P(s,a), P'(s,a)) \left[|P'(s,a) + 2P(s,a) - P(s,a)|^\top (V^\pi - \mu^\pi)^{\circ 2})\right], \\
&\leq 4\mathrm{KL}(P(s,a), P'(s,a)) \left[2\mathrm{Var}_{s,a}(V^\pi) + \|P'(s,a) - P(s,a)\|_1 \mathrm{MD}_{s,a}(V^\pi)^2\right], \\
&\leq 4\mathrm{KL}(P(s,a), P'(s,a)) \left[2\mathrm{Var}_{s,a}(V^\pi) + \sqrt{2\mathrm{KL}(P(s,a), P'(s,a))}\mathrm{MD}_{s,a}(V^\pi)^2\right].
\end{aligned}
$$

(a) Cauchy-Schwarz ineq.; (b) definition of Hellinger's distance (add a factor 2) and used $(a+b)^2 \leq 2(a^2+b^2)$; (c) $H(P,Q) \leq \sqrt{\mathrm{KL}(P,Q)}$.

Relating the sub-optimality gaps to the KL terms

$$
\begin{aligned}
|(V^\pi)^\top \Delta P(s,a)|^2 &\le \left|[(\sqrt{P'(s,a)} - \sqrt{P(s,a)}) \circ (\sqrt{P'(s,a)} + \sqrt{P(s,a)})]^\top (V^\pi - \mu^\pi))\right|^2, \\
&= \left|(\sqrt{P'(s,a)} - \sqrt{P(s,a)})^\top [(\sqrt{P'(s,a)} + \sqrt{P(s,a)}) \circ (V^\pi - \mu^\pi))]\right|^2, \\
&\overset{(a)}{\le} \left\|\sqrt{P'(s,a)} - \sqrt{P(s,a)}\right\|_2^2 \left\|(\sqrt{P'(s,a)} + \sqrt{P(s,a)}) \circ (V^\pi - \mu^\pi))]\right\|_2^2, \\
&\overset{(b)}{\le} 4H^2(P(s,a), P'(s,a)) \left[|P'(s,a) + P(s,a)|^\top (V^\pi - \mu^\pi)^{\circ 2})\right], \\
&\overset{(c)}{\le} 4\mathrm{KL}(P(s,a), P'(s,a)) \left[|P'(s,a) + 2P(s,a) - P(s,a)|^\top (V^\pi - \mu^\pi)^{\circ 2})\right], \\
&\le 4\mathrm{KL}(P(s,a), P'(s,a)) \left[2\mathrm{Var}_{s,a}(V^\pi) + \|P'(s,a) - P(s,a)\|_1 \mathrm{MD}_{s,a}(V^\pi)^2\right], \\
&\le 4\mathrm{KL}(P(s,a), P'(s,a)) \left[2\mathrm{Var}_{s,a}(V^\pi) + \sqrt{2\mathrm{KL}(P(s,a), P'(s,a))}\mathrm{MD}_{s,a}(V^\pi)^2\right].
\end{aligned}
$$

(a) Cauchy-Schwarz ineq.; (b) definition of Hellinger's distance (add a factor 2) and used $(a+b)^2 \le 2(a^2+b^2)$; (c) $H(P,Q) \le \sqrt{\mathrm{KL}(P,Q)}$.

Relating the sub-optimality gaps to the KL terms

$$\Delta(s,a) < \frac{\gamma|\Delta P(s,\pi^\star(s))^\top V^\star|}{1-\gamma} + \Delta P(s,a)^\top V^\star.$$

We also want to relate each term on the r.h.s. to a fraction of $\Delta(s,a)$ to be able to bound the individual KL terms using the gaps.

Introduce $\alpha_1, \alpha_2 \geq 0$ s.t. $\alpha_1 + \alpha_2 > 1$ and let

$$\alpha_1 \Delta(s,a) = \frac{\gamma|\Delta P(s,\pi^\star(s))^\top V^\star|}{1-\gamma}, \quad (2)$$

$$\alpha_2 \Delta(s,a) = \Delta P(s,a)^\top V^\star. \quad (3)$$

Relating the sub-optimality gaps to the KL terms

$$\Delta(s,a) < \frac{\gamma|\Delta P(s,\pi^\star(s))^\top V^\star|}{1-\gamma} + \Delta P(s,a)^\top V^\star.$$

We also want to relate each term on the r.h.s. to a fraction of $\Delta(s,a)$ to be able to bound the individual KL terms using the gaps.

Introduce $\alpha_1, \alpha_2 \geq 0$ s.t. $\alpha_1 + \alpha_2 > 1$ and let

$$\alpha_1 \Delta(s,a) = \frac{\gamma|\Delta P(s,\pi^\star(s))^\top V^\star|}{1-\gamma}, \quad (2)$$

$$\alpha_2 \Delta(s,a) = \Delta P(s,a)^\top V^\star. \quad (3)$$

Relating the sub-optimality gaps to the KL terms

Using the lemma, for $\alpha_2\Delta(s,a)$ we find

$$\underbrace{(\alpha_2\Delta(s,a))^2}_{=|\Delta P(s,a)^\top V^\star|^2} \leq 4\mathrm{KL}(P(s,a),P'(s,a))\left[2\mathrm{Var}_{s,a}(V^\pi)+\sqrt{2\mathrm{KL}(P(s,a),P'(s,a))}\mathrm{MD}_{s,a}(V^\pi)^2\right].$$

Use $a+b\leq 2\max(a,b)$. Then

$$\frac{(\alpha_2\Delta(s,a))^2}{16\mathrm{Var}_{s,a}(V^\pi)}\leq \mathrm{KL}(P(s,a),P'(s,a)) \text{ or } \frac{(\alpha_2\Delta(s,a))^{4/3}}{2^{7/3}\mathrm{MD}_{s,a}(V^\pi)^{4/3}}\leq \mathrm{KL}(P(s,a),P'(s,a)).$$

Hence

$$\min\left(\frac{(\alpha_2\Delta(s,a))^2}{16\mathrm{Var}_{s,a}(V^\pi)},\frac{(\alpha_2\Delta(s,a))^{4/3}}{2^{7/3}\mathrm{MD}_{s,a}(V^\pi)^{4/3}}\right)\leq \mathrm{KL}(P(s,a),P'(s,a)).$$

Relating the sub-optimality gaps to the KL terms

Using the lemma, for $\alpha_2\Delta(s,a)$ we find

$$\underbrace{(\alpha_2\Delta(s,a))^2}_{=|\Delta P(s,a)^\top V^\star|^2} \le 4\mathrm{KL}(P(s,a),P'(s,a))\left[2\mathrm{Var}_{s,a}(V^\pi)+\sqrt{2\mathrm{KL}(P(s,a),P'(s,a))}\mathrm{MD}_{s,a}(V^\pi)^2\right].$$

Use $a+b\le 2\max(a,b)$. Then

$$\frac{(\alpha_2\Delta(s,a))^2}{16\mathrm{Var}_{s,a}(V^\pi)} \le \mathrm{KL}(P(s,a),P'(s,a)) \text{ or } \frac{(\alpha_2\Delta(s,a))^{4/3}}{2^{7/3}\mathrm{MD}_{s,a}(V^\pi)^{4/3}} \le \mathrm{KL}(P(s,a),P'(s,a)).$$

Hence

$$\min\left(\frac{(\alpha_2\Delta(s,a))^2}{16\mathrm{Var}_{s,a}(V^\pi)},\frac{(\alpha_2\Delta(s,a))^{4/3}}{2^{7/3}\mathrm{MD}_{s,a}(V^\pi)^{4/3}}\right)\le \mathrm{KL}(P(s,a),P'(s,a)).$$

Relating the sub-optimality gaps to the KL terms

Using the lemma, for $\alpha_2\Delta(s,a)$ we find

$$\underbrace{(\alpha_2\Delta(s,a))^2}_{=|\Delta P(s,a)^\top V^\star|^2} \le 4\mathrm{KL}(P(s,a),P'(s,a))\left[2\mathrm{Var}_{s,a}(V^\pi)+\sqrt{2\mathrm{KL}(P(s,a),P'(s,a))}\mathrm{MD}_{s,a}(V^\pi)^2\right].$$

Use $a+b\le 2\max(a,b)$. Then

$$\frac{(\alpha_2\Delta(s,a))^2}{16\mathrm{Var}_{s,a}(V^\pi)} \le \mathrm{KL}(P(s,a),P'(s,a)) \text{ or } \frac{(\alpha_2\Delta(s,a))^{4/3}}{2^{7/3}\mathrm{MD}_{s,a}(V^\pi)^{4/3}} \le \mathrm{KL}(P(s,a),P'(s,a)).$$

Hence

$$\min\left(\frac{(\alpha_2\Delta(s,a))^2}{16\mathrm{Var}_{s,a}(V^\pi)},\frac{(\alpha_2\Delta(s,a))^{4/3}}{2^{7/3}\mathrm{MD}_{s,a}(V^\pi)^{4/3}}\right)\le \mathrm{KL}(P(s,a),P'(s,a)).$$

Relating the sub-optimality gaps to the KL terms

Similarly, for $\alpha_1 \Delta(s,a) = \frac{\gamma |\Delta P(s,\pi^\star(s))^\top V^\star|}{1-\gamma}$ we get

$$\min\left(\frac{(\alpha_1 \Delta_{min}(1-\gamma))^2}{16 \max_s \mathrm{Var}_{s,\pi^\star(s)}(V^\pi)}, \frac{(\alpha_1 \Delta_{min}(1-\gamma))^{4/3}}{2^{7/3} \max_s \mathrm{MD}_{s,\pi^\star(s)}(V^\pi)^{4/3}}\right) \leq \max_s \mathrm{KL}(P(s,\pi^\star(s)), P'(s,\pi^\star(s))).$$

where $\Delta_{min} = \min_{s,a\neq\pi^\star(s)} \Delta(s,a)$.

Relating the sub-optimality gaps to the KL terms

Let $B_2(s,a,\alpha_2) = \min\left(\frac{(\alpha_2\Delta(s,a))^2}{16\mathrm{Var}_{s,a}(V^\pi)}, \frac{(\alpha_2\Delta(s,a))^{4/3}}{2^{7/3}\mathrm{MD}_{s,a}(V^\pi)^{4/3}}\right)$ and
$B_1(\alpha_1) = \min\left(\frac{(\alpha_1\Delta_{min}(1-\gamma))^2}{16\max_s \mathrm{Var}_{s,\pi^\star(s)}(V^\pi)}, \frac{(\alpha_1\Delta_{min}(1-\gamma))^{4/3}}{2^{7/3}\max_s \mathrm{MD}_{s,\pi^\star(s)}(V^\pi)^{4/3}}\right)$.

Applying what we have learnt we get

$$
\begin{aligned}
T^{-1}(\omega) \geq & \min_{s,a\neq\pi^\star(s)} \inf_{M'\in\mathrm{Alt}_{s,a}(M)} \omega(s,a)\mathrm{KL}(P(s,a),P'(s,a)) \\
& + (\min_{s'} \omega_{s',\pi^\star(s')}) \max_{s'} \mathrm{KL}(P(s',\pi^\star(s')),P'(s',\pi^\star(s'))), \\
\geq & \min_{s,a\neq\pi^\star(s)} \inf_{\alpha_1+\alpha_2>1} \omega(s,a)B_2(s,a,\alpha_2) + (\min_{s'} \omega_{s',\pi^\star(s')})B_1(\alpha_1).
\end{aligned}
$$

Note that for any α satisfying $\sum_i \alpha_i > 1$ we also have that $\alpha_i/\sum_i \alpha_i$ satisfies the previous KL inequalities.

Relating the sub-optimality gaps to the KL terms

Let $B_2(s,a,\alpha_2) = \min\left(\frac{(\alpha_2\Delta(s,a))^2}{16\mathrm{Var}_{s,a}(V^\pi)}, \frac{(\alpha_2\Delta(s,a))^{4/3}}{2^{7/3}\mathrm{MD}_{s,a}(V^\pi)^{4/3}}\right)$ and $B_1(\alpha_1) = \min\left(\frac{(\alpha_1\Delta_{min}(1-\gamma))^2}{16\max_s \mathrm{Var}_{s,\pi^\star(s)}(V^\pi)}, \frac{(\alpha_1\Delta_{min}(1-\gamma))^{4/3}}{2^{7/3}\max_s \mathrm{MD}_{s,\pi^\star(s)}(V^\pi)^{4/3}}\right)$.

Applying what we have learnt we get

$$\begin{aligned}
T^{-1}(\omega) \geq & \min_{s,a\neq\pi^\star(s)} \inf_{M'\in\mathrm{Alt}_{s,a}(M)} \omega(s,a)\mathrm{KL}(P(s,a),P'(s,a)) \\
& + (\min_{s'} \omega_{s',\pi^\star(s')}) \max_{s'} \mathrm{KL}(P(s',\pi^\star(s')),P'(s',\pi^\star(s'))), \\
\geq & \min_{s,a\neq\pi^\star(s)} \inf_{\alpha_1+\alpha_2>1} \omega(s,a)B_2(s,a,\alpha_2) + (\min_{s'} \omega_{s',\pi^\star(s')})B_1(\alpha_1).
\end{aligned}$$

Note that for any α satisfying $\sum_i \alpha_i > 1$ we also have that $\alpha_i/\sum_i \alpha_i$ satisfies the previous KL inequalities.

Relating the sub-optimality gaps to the KL terms

For α_i in the simplex, we also have $\alpha_i^2 \leq \alpha_i^{4/3}$. Thus

$$T^{-1}(\omega) \geq \min_{s,a\neq\pi^\star(s)} \inf_{\alpha_i\in\Delta(2)} \omega(s,a)\alpha^2 B_2(s,a) + \alpha_1^2(\min_{s'} \omega_{s',\pi^\star(s')})B_1.$$

where $B_2(s,a) = \min\left(\frac{\Delta(s,a)^2}{16\mathrm{Var}_{s,a}(V^\pi)}, \frac{\Delta(s,a)^{4/3}}{2^{7/3}\mathrm{MD}_{s,a}(V^\pi)^{4/3}}\right)$ and $B_1 = \min\left(\frac{(\Delta_{min}(1-\gamma))^2}{16\max_s \mathrm{Var}_{s,\pi^\star(s)}(V^\pi)}, \frac{(\Delta_{min}(1-\gamma))^{4/3}}{2^{7/3}\max_s \mathrm{MD}_{s,\pi^\star(s)}(V^\pi)^{4/3}}\right)$. Optimizing over α yields

$$T^{-1}(\omega) \geq \min_{s,a\neq\pi^\star(s)} \left(\frac{1}{\omega(s,a)B_2(s,a)} + \frac{1}{\min_{s'} \omega_{s',\pi^\star(s')})B_1}\right)^{-1}.$$

Relating the sub-optimality gaps to the KL terms

For α_i in the simplex, we also have $\alpha_i^2 \leq \alpha_i^{4/3}$. Thus

$$T^{-1}(\omega) \geq \min_{s,a\neq\pi^\star(s)} \inf_{\alpha_i\in\Delta(2)} \omega(s,a)\alpha^2 B_2(s,a) + \alpha_1^2(\min_{s'} \omega_{s',\pi^\star(s')})B_1.$$

where $B_2(s,a) = \min\left(\frac{\Delta(s,a)^2}{16\mathrm{Var}_{s,a}(V^\pi)}, \frac{\Delta(s,a)^{4/3}}{2^{7/3}\mathrm{MD}_{s,a}(V^\pi)^{4/3}}\right)$ and $B_1 = \min\left(\frac{(\Delta_{min}(1-\gamma))^2}{16\max_s \mathrm{Var}_{s,\pi^\star(s)}(V^\pi)}, \frac{(\Delta_{min}(1-\gamma))^{4/3}}{2^{7/3}\max_s \mathrm{MD}_{s,\pi^\star(s)}(V^\pi)^{4/3}}\right)$. Optimizing over α yields

$$T^{-1}(\omega) \geq \min_{s,a\neq\pi^\star(s)} \left(\frac{1}{\omega(s,a)B_2(s,a)} + \frac{1}{\min_{s'}\omega_{s',\pi^\star(s')})B_1}\right)^{-1}.$$

Relating the sub-optimality gaps to the KL terms (final)

$$T^{-1}(\omega) \geq \min_{s,a\neq\pi^\star(s)} \left(\frac{1}{\omega(s,a)B_2(s,a)} + \frac{1}{\min_{s'} \omega_{s',\pi^\star(s')} B_1} \right)^{-1}.$$

Then

$$T(\omega) \leq \max_{s,a\neq\pi^\star(s)} \frac{H_{s,a}}{\omega(s,a)\Delta(s,a)^2} + \frac{H^\star}{\min_{s'} \omega_{s',\pi^\star(s')}} =: U(\omega).$$

with

$$H_{s,a} = \max\left(\frac{16\mathrm{Var}_{s,a}(V^\pi)}{\Delta(s,a)^2}, \frac{2^{7/3}\mathrm{MD}_{s,a}(V^\pi)^{4/3}}{\Delta(s,a)^{4/3}} \right),$$

$$H^\star = \max\left(\frac{16 \max_s \mathrm{Var}_{s,\pi^\star(s)}(V^\pi)}{(1-\gamma)^2\Delta_{min}^2}, \frac{2^{7/3} \max_s \mathrm{MD}_{s,\pi^\star(s)}(V^\pi)^{4/3}}{((\Delta_{min}(1-\gamma))^{4/3}} \right).$$

Relating the sub-optimality gaps to the KL terms (final)

$$T^{-1}(\omega) \geq \min_{s,a \neq \pi^\star(s)} \left(\frac{1}{\omega(s,a) B_2(s,a)} + \frac{1}{\min_{s'} \omega_{s',\pi^\star(s')} B_1} \right)^{-1}.$$

Then

$$T(\omega) \leq \max_{s,a \neq \pi^\star(s)} \frac{H_{s,a}}{\omega(s,a)\Delta(s,a)^2} + \frac{H^\star}{\min_{s'} \omega_{s',\pi^\star(s')}} =: U(\omega).$$

with

$$H_{s,a} = \max\left(\frac{16 \mathrm{Var}_{s,a}(V^\pi)}{\Delta(s,a)^2}, \frac{2^{7/3} \mathrm{MD}_{s,a}(V^\pi)^{4/3}}{\Delta(s,a)^{4/3}} \right),$$

$$H^\star = \max\left(\frac{16 \max_s \mathrm{Var}_{s,\pi^\star(s)}(V^\pi)}{(1-\gamma)^2 \Delta_{min}^2}, \frac{2^{7/3} \max_s \mathrm{MD}_{s,\pi^\star(s)}(V^\pi)^{4/3}}{((\Delta_{min}(1-\gamma))^{4/3}} \right).$$

Conclusions

$$T(\omega) \leq \max_{s,a\neq\pi^\star(s)} \frac{H_{s,a}}{\omega(s,a)\Delta(s,a)^2} + \frac{H^\star}{\min_{s'} \omega_{s',\pi^\star(s')}} =: U(\omega).$$

with

$$H_{s,a} = \max\left(\frac{16\mathrm{Var}_{s,a}(V^\pi)}{\Delta(s,a)^2}, \frac{2^{7/3}\mathrm{MD}_{s,a}(V^\pi)^{4/3}}{\Delta(s,a)^{4/3}}\right),$$

$$H^\star = \max\left(\frac{16\max_s \mathrm{Var}_{s,\pi^\star(s)}(V^\pi)}{(1-\gamma)^2\Delta_{min}^2}, \frac{2^{7/3}\max_s \mathrm{MD}_{s,\pi^\star(s)}(V^\pi)^{4/3}}{((\Delta_{min}(1-\gamma))^{4/3}}\right).$$

- If we plug in a uniform distribution $\omega(s,a) = 1/(|S||A|)$ the bound scales roughly as $O\left(\frac{|S||A|}{\Delta_{min}^2(1-\gamma)^4}\right)$. The factor on γ be improved to $1/(1-\gamma)^3$ (see [AMP21]).
- Many open questions:
  - Possible to find a tighter bound? Simpler proof?
  - Possible to characterize the gap $U(\omega) - T(\omega)$?
  - Are there some cases where the set of confusing models is convex, and we can compute $T^\star$ exactly?

Best Policy Identification: Linear Markov Decision Processes

Introduction

Consider a linear MDP $M = (S, A, P, r, \gamma)$ s.t. to each pair (s, a) is associated a feature vector $\phi(s, a) \in \mathbb{R}^d$, satisfying $\|\phi(s, a)\| \leq 1$ [8].

- S is the state space (finite);
- A is the action space (finite)
- $P(s'|s, a) = \phi(s, a)^\top \mu(s')$ and $r(s, a) = \phi(s, a)^\top \theta$ for some $\mu : S \to \mathbb{R}^d$ and $\theta \in \mathbb{R}^d$.
- $\gamma \in (0, 1)$ is the discount factor.

[8] Setting studied in [TJP23]

Lower bound

The steps are (almost) the same as before. In [TJP23] they find that

$$\begin{aligned}\sum_{s,a} \omega_{s,a} \mathrm{KL}(P(s,a), P'(s,a)) &\geq (1-\gamma)^2 \sum_{s,a} \omega_{s,a} |\phi^\top \left(\theta - \theta' + \gamma(\mu - \mu')^\top V^\star\right)|^2, \\ &= (1-\gamma)^2 \|\theta - \theta' + \gamma(\mu - \mu')^\top V^\star\|^2_{\Lambda(\omega)},\end{aligned}$$

where we are considering an alternative model M' with (ϕ', μ', θ'), and

$$\|x\|^2_{\Lambda(\omega)} = \|\Lambda(\omega)^{\frac{1}{2}} x\|_2^2, \text{ with } \Lambda(\omega) = \sum_{s,a} \omega_{s,a} \phi(s,a)\phi(s,a)^\top.$$

Lower bound

$$\sum_{s,a} \omega_{s,a} \mathrm{KL}(P(s,a), P'(s,a)) \geq (1-\gamma)^2 \|\theta - \theta' + \gamma(\mu - \mu')^\top V^\star\|^2_{\Lambda(\omega)}.$$

In [TJP23] they show that

$$\Delta_{min} \leq \frac{2}{1-\gamma} \max_{s,a} |\phi^\top \left(\theta - \theta' + \gamma(\mu - \mu')^\top V^\star\right)|$$

combine it with the lemma

$$\inf_{x \in \mathbb{R}^d : |\phi^\top x| \geq \Delta} \|x\|^2_\Lambda = \frac{\Delta^2}{\|\phi\|^2_{\Lambda^{-1}}}.$$

to obtain

$$\|\theta - \theta' + \gamma(\mu - \mu')^\top V^\star\|^2_{\Lambda(\omega)} \geq \frac{(1-\gamma)^2 \Delta^2_{min}}{4 \max_{s,a} \|\phi(s,a)\|^2_{\Lambda(\omega)^{-1}}}.$$

Lower bound

Therefore

$$(T(\omega))^{-1} = \inf_{M' \in \mathrm{Alt}(M)} \sum_{s,a} \omega_{s,a} \mathrm{KL}(P(s,a), P'(s,a)) \geq (1-\gamma)^2 \|\theta - \theta' + \gamma(\mu - \mu')^\top V^\star\|^2_{\Lambda(\omega)},$$

$$\geq \frac{(1-\gamma)^4 \Delta^2_{min}}{4 \max_{s,a} \|\phi(s,a)\|^2_{\Lambda(\omega)^{-1}}}.$$

Hence, the optimal allocation is given by

$$\omega^\star = \arg\inf_{\omega \in \Omega(M)} \max_{s,a} \|\phi(s,a)\|^2_{\Lambda(\omega)^{-1}}$$

Conclusions

Conclusions

Still many problems left...

- What is the tightest convexification we can find?
- How can we extend the results to partially observable models?
- Can we simplify the proofs?
- The bounds do not take into account the parametric uncertainty during learning.
- What is the gap between the convexified bound and the true lower bound?
- How to extend to function approximators? Use ϵ-net type discretization of the state-action space $S \times A$?

Thank you for your attention!

References i

Aymen Al Marjani, Aurélien Garivier, and Alexandre Proutiere, *Navigating to the best policy in markov decision processes*, Advances in Neural Information Processing Systems **34** (2021), 25852–25864.

Aymen Al Marjani and Alexandre Proutiere, *Adaptive sampling for best policy identification in markov decision processes*, International Conference on Machine Learning, PMLR, 2021, pp. 7459–7468.

Cheng-Der Fuh, *Sprt and cusum in hidden markov models*, The Annals of Statistics **31** (2003), no. 3, 942–977.

Aurélien Garivier, Pierre Ménard, and Gilles Stoltz, *Explore first, exploit next: The true shape of regret in bandit problems*, Mathematics of Operations Research **44** (2019), no. 2, 377–399.

Jérôme Taupin, Yassir Jedra, and Alexandre Proutiere, *Best policy identification in discounted linear mdps*, Sixteenth European Workshop on Reinforcement Learning, 2023.

Appendix

Non-asymptotic lower bound

To find a non-asymptotic lower bound with navigation constraints note that

$$\underbrace{N_\tau(s)}_{=\sum_a N_\tau(s,a)} = \mathbf{1}_{\{s_1=s\}} + \sum_{s',a'} \sum_{n=1}^{N_{\tau-1}(s',a')} \mathbf{1}_{\{W'_n=s\}}.$$

Therefore, using Wald's lemma again as in the lower bound proof

$$\mathbb{E}_M[N_\tau(s)] = \mathbb{P}_M(s_1 = s) + \sum_{s',a'} \mathbb{E}_M[N_{\tau-1}(s',a')]\mathbb{P}(s|s',a').$$

Using $\mathbb{E}_M[N_{\tau-1}(s,a)] \le \mathbb{E}[N_\tau(s,a)]$ we can write the lower bound as

$$\begin{aligned} \mathbb{E}_M[\tau] \ge \min_{n\in\mathbb{R}^{S\times A}} \quad & \sum_{s,a} n_{s,a} \\ \text{s.t.} \quad & \sum_{s,a} n_{s,a}\mathrm{KL}(P(s,a), P'(s,a)) \ge \mathrm{kl}(\delta, 1-\delta) \quad \forall M' \in \mathrm{Alt}(M), \\ & \sum_a n_{s,a} - \sum_{s',a'} n_{s',a'} P(s|s',a') \le 1. \end{aligned}$$

Non-asymptotic lower bound

To find a non-asymptotic lower bound with navigation constraints note that

$$\underbrace{N_\tau(s)}_{=\sum_a N_\tau(s,a)} = \mathbf{1}_{\{s_1=s\}} + \sum_{s',a'} \sum_{n=1}^{N_{\tau-1}(s',a')} \mathbf{1}_{\{W'_n=s\}}.$$

Therefore, using Wald's lemma again as in the lower bound proof

$$\mathbb{E}_M[N_\tau(s)] = \mathbb{P}_M(s_1 = s) + \sum_{s',a'} \mathbb{E}_M[N_{\tau-1}(s',a')]\mathbb{P}(s|s',a').$$

Using $\mathbb{E}_M[N_{\tau-1}(s,a)] \le \mathbb{E}[N_\tau(s,a)]$ we can write the lower bound as

$$\begin{aligned} \mathbb{E}_M[\tau] \ge \min_{n\in\mathbb{R}^{S\times A}} \quad & \sum_{s,a} n_{s,a} \\ \text{s.t.} \quad & \sum_{s,a} n_{s,a}\mathrm{KL}(P(s,a),P'(s,a)) \ge \mathrm{kl}(\delta,1-\delta) \quad \forall M' \in \mathrm{Alt}(M), \\ & \sum_a n_{s,a} - \sum_{s',a'} n_{s',a'}P(s|s',a') \le 1. \end{aligned}$$

Non-asymptotic lower bound

To find a non-asymptotic lower bound with navigation constraints note that

$$\underbrace{N_\tau(s)}_{=\sum_a N_\tau(s,a)} = \mathbf{1}_{\{s_1=s\}} + \sum_{s',a'} \sum_{n=1}^{N_{\tau-1}(s',a')} \mathbf{1}_{\{W'_n=s\}}.$$

Therefore, using Wald's lemma again as in the lower bound proof

$$\mathbb{E}_M[N_\tau(s)] = \mathbb{P}_M(s_1 = s) + \sum_{s',a'} \mathbb{E}_M[N_{\tau-1}(s',a')]\mathbb{P}(s|s',a').$$

Using $\mathbb{E}_M[N_{\tau-1}(s,a)] \leq \mathbb{E}[N_\tau(s,a)]$ we can write the lower bound as

$$\begin{aligned} \mathbb{E}_M[\tau] \geq \min_{n\in\mathbb{R}^{S\times A}} \quad & \sum_{s,a} n_{s,a} \\ \text{s.t.} \quad & \sum_{s,a} n_{s,a}\mathrm{KL}(P(s,a), P'(s,a)) \geq \mathrm{kl}(\delta, 1-\delta) \quad \forall M' \in \mathrm{Alt}(M), \\ & \sum_{a} n_{s,a} - \sum_{s',a'} n_{s',a'} P(s|s',a') \leq 1. \end{aligned}$$