Risk minimization from adaptively collected data: guarantees for policy learning

Antoine Chambaz[1], with

A. Bibaut[2], N. Kallus[2,3], M. Dimakopoulou[2], M. J. van der Laan[4]

[1] MAP5 (UMR CNRS 8145), Université Paris Cité

[2] Netflix

[3] Cornell University

[4] University of California, Berkeley

August 30th, 2022

Journées MAS 2022

projet **BOLD** (V. Perchet)

"Au delà de l'apprentissage séquentiel pour de meilleures prises de décisions"

Context: adaptive experiment

- $X_1, \ldots, X_t, \ldots$ iid, observed sequentially

  each $X_t \in \mathcal{X}$, a context of action

Context: adaptive experiment

- $X_1, \dots, X_t, \dots$ iid, observed sequentially

  each $X_t \in \mathcal{X}$, a context of action

- $(Y_1(a))_{a \in \mathcal{A}}, \dots, (Y_t(a))_{a \in \mathcal{A}}, \dots$ iid,hidden

  each $Y_t(a)$ the reward (in context X_t) of action a, one among $|\mathcal{A}| < \infty$ possible actions

Context: adaptive experiment

- $X_1, \ldots, X_t, \ldots$ iid, observed sequentially

  each $X_t \in \mathcal{X}$, a context of action

- $(Y_1(a))_{a \in \mathcal{A}}, \ldots, (Y_t(a))_{a \in \mathcal{A}}, \ldots$ iid,hidden

  each $Y_t(a)$ the reward (in context X_t) of action a, one among $|\mathcal{A}| < \infty$ possible actions

- at every time $t \geq 1$, $A_t \sim g_t(\cdot|X_t)$, observed,

  yields the observed reward $Y_t = Y_t(A_t)$, where

  - the law g_t is built based on $(X_1, A_1, Y_1), \ldots, (X_{t-1}, A_{t-1}, Y_{t-1})$ and known to us
  - $A_t \perp Y_t(a)|X_t$ for all $a \in \mathcal{A}$ – "randomization"

Context: adaptive experiment

- $X_1, \ldots, X_t, \ldots$ iid, observed sequentially

  each $X_t \in \mathcal{X}$, a context of action

- $(Y_1(a))_{a \in \mathcal{A}}, \ldots, (Y_t(a))_{a \in \mathcal{A}}, \ldots$ iid,hidden

  each $Y_t(a)$ the reward (in context X_t) of action a, one among $|\mathcal{A}| < \infty$ possible actions

- at every time $t \geq 1$, $A_t \sim g_t(\cdot|X_t)$, observed,

  yields the observed reward $Y_t = Y_t(A_t)$, where

  - the law g_t is built based on $(X_1, A_1, Y_1), \ldots, (X_{t-1}, A_{t-1}, Y_{t-1})$ and known to us
  - $A_t \perp Y_t(a)|X_t$ for all $a \in \mathcal{A}$ – “randomization”

Useful for

Context: adaptive experiment

- $X_1, \ldots, X_t, \ldots$ iid, observed sequentially

  each $X_t \in \mathcal{X}$, a context of action

- $(Y_1(a))_{a \in \mathcal{A}}, \ldots, (Y_t(a))_{a \in \mathcal{A}}, \ldots$ iid,hidden

  each $Y_t(a)$ the reward (in context X_t) of action a, one among $|\mathcal{A}| < \infty$ possible actions

- at every time $t \geq 1$, $A_t \sim g_t(\cdot|X_t)$, observed,

  yields the observed reward $Y_t = Y_t(A_t)$, where

  - the law g_t is built based on $(X_1, A_1, Y_1), \ldots, (X_{t-1}, A_{t-1}, Y_{t-1})$ and known to us
  - $A_t \perp Y_t(a)|X_t$ for all $a \in \mathcal{A}$ – "randomization"

Useful for regret control

Context: adaptive experiment

- $X_1, \ldots, X_t, \ldots$ iid, observed sequentially

  each $X_t \in \mathcal{X}$, a context of action

- $(Y_1(a))_{a \in \mathcal{A}}, \ldots, (Y_t(a))_{a \in \mathcal{A}}, \ldots$ iid,hidden

  each $Y_t(a)$ the reward (in context X_t) of action a, one among $|\mathcal{A}| < \infty$ possible actions

- at every time $t \geq 1$, $A_t \sim g_t(\cdot|X_t)$, observed,

  yields the observed reward $Y_t = Y_t(A_t)$, where
  - the law g_t is built based on $(X_1, A_1, Y_1), \ldots, (X_{t-1}, A_{t-1}, Y_{t-1})$ and known to us
  - $A_t \perp Y_t(a)|X_t$ for all $a \in \mathcal{A}$ – “randomization”

Useful for regret control, best arm identification

Context: adaptive experiment

- $X_1, \ldots, X_t, \ldots$ iid, observed sequentially

  each $X_t \in \mathcal{X}$, a context of action

- $(Y_1(a))_{a \in \mathcal{A}}, \ldots, (Y_t(a))_{a \in \mathcal{A}}, \ldots$ iid,hidden

  each $Y_t(a)$ the reward (in context X_t) of action a, one among $|\mathcal{A}| < \infty$ possible actions

- at every time $t \geq 1$, $A_t \sim g_t(\cdot|X_t)$, observed,

  yields the observed reward $Y_t = Y_t(A_t)$, where

  - the law g_t is built based on $(X_1, A_1, Y_1), \ldots, (X_{t-1}, A_{t-1}, Y_{t-1})$ and known to us
  - $A_t \perp Y_t(a)|X_t$ for all $a \in \mathcal{A}$ – “randomization”

Useful for regret control, best arm identification, policy learning

Context: adaptive experiment

- $X_1, \ldots, X_t, \ldots$ iid, observed sequentially
  each $X_t \in \mathcal{X}$, a context of action
- $(Y_1(a))_{a \in \mathcal{A}}, \ldots, (Y_t(a))_{a \in \mathcal{A}}, \ldots$ iid,hidden
  each $Y_t(a)$ the reward (in context X_t) of action a, one among $|\mathcal{A}| < \infty$ possible actions
- at every time $t \geq 1$, $A_t \sim g_t(\cdot|X_t)$, observed,
  yields the observed reward $Y_t = Y_t(A_t)$, where
  - the law g_t is built based on $(X_1, A_1, Y_1), \ldots, (X_{t-1}, A_{t-1}, Y_{t-1})$ and known to us
  - $A_t \perp Y_t(a)|X_t$ for all $a \in \mathcal{A}$ – "randomization"

Useful for regret control, best arm identification, policy learning, etc.

Risk minimization for policy learning

Given a class of stochastic policies,

$$\mathcal{F} = \left\{ f : \mathcal{X} \times \mathcal{A} \to \mathbb{R}_+ : \sum_{a \in \mathcal{A}} f(a|x) = 1 \text{ for all } x \in \mathcal{X} \right\}$$

and a risk R identifying a best policy $f^\star$,

$$R(f) = E\left[\sum_{a \in \mathcal{A}} g^\star(a|X) \times f(a|X) \times (-Y(a)) \right], \quad \text{all } f \in \mathcal{F}$$

$$f^\star \in \arg\min_{f \in \mathcal{F}} R(f)$$

how to learn $f^\star$?

- Here, $g^\star$ serves as a reference action mechanism
- If $g^\star(a|x) = |\mathcal{A}|^{-1}$ for all $(a, x) \in \mathcal{A} \times \mathcal{X}$, then $R(f)$ is minus the value of policy f

How to learn $f^\star$?

Recall that

$$R(f) = E\left[\sum_{a\in\mathcal{A}} g^\star(a|X) \times (-Y(a)) \times f(a|X)\right], \quad \text{all } f \in \mathcal{F}$$

$$f^\star \in \arg\min_{f\in\mathcal{F}} R(f)$$

Introduce

- the loss function $f \mapsto \ell(f)$ such that $\ell(f)(x, a, y) = -y \times f(a|x)$
- the importance-sampling-weighted empirical risk

$$\widehat{R}_T(f) = \frac{1}{T}\sum_{t=1}^{T} \frac{g^\star(A_t|X_t)}{g_t(A_t|X_t)} \times \ell(f)(X_t, A_t, Y_t), \quad \text{all } f \in \mathcal{F}$$

  - key-fact: $E\left[\frac{g^\star(A_t|X_t)}{g_t(A_t|X_t)} \times \ell(f)(X_t, A_t, Y_t)\right] = R(f)$ (proof)

Define the estimator

$$\widehat{f}_T \in \arg\min_{f\in\mathcal{F}} \widehat{R}_T(f)$$

How to study $\widehat{f}_T$?

We wish to control the excess risk of $\widehat{f}_T$,

$$0 \leq R(\widehat{f}_T) - R(f^\star) \leq ?$$

Main challenge: controlling the martingale sequence difference

$$f \mapsto \frac{1}{T} \sum_{t=1}^{T} \frac{g^\star(A_t|X_t)}{g_t(A_t|X_t)} \xi_t(f)$$

with $\xi_t(f)$ $\sigma(X_1, A_1, Y_1, \ldots, X_{t-1}, A_{t-1}, Y_{t-1}) = \mathcal{S}_t$-measurable such that $E[\xi_t(f)|\mathcal{S}_t] = 0$

- Need to introduce a notion of sequential bracketing entropy

  Our proof adapts that of van Handel (2011)

- Utmost care in handling the deterministic sequence $(\gamma_t)_{t\geq 1}$ defined such that

$$\left\| \frac{g^\star}{g_t} \right\|_\infty \leq \gamma_t, \quad t \geq 1$$

Theorem

Suppose that

- $|Y_t| \leq M$ for all $t \geq 1$
- $\ell(\mathcal{F})$ has finite $\|\cdot\|_\infty$- and $\|\cdot\|_{2,g^\star}$-diameters
- there exists $p > 0$, $p \neq 2$, such that $\log N_{[\cdot]}(\varepsilon M, \ell(\mathcal{F}), \|\cdot\|_{2,g^\star}) \lesssim \varepsilon^{-p}$ for all $\varepsilon > 0$ (A)

Define

$$\gamma_T^{\text{avg}} = \frac{1}{T}\sum_{t=1}^{T} \gamma_t, \quad \gamma_T^{\max} = \max_{t \leq T} \gamma_t$$

Then, for all $\delta \in]0, \frac{1}{2}[$, with probability at least $(1-\delta)$,

$$0 \leq R(\widehat{f}_T) - R(f^\star) \lesssim M\left\{\left(\frac{\gamma_T^{\text{avg}}}{T}\right)^{1/p} \mathbf{1}\{p > 2\} + \sqrt{\frac{\gamma_T^{\text{avg}}}{T}}\sqrt{\log\left(\frac{1}{\delta}\right)} + \frac{\gamma_T^{\max}}{T}\log\left(\frac{1}{\delta}\right)\right\}$$

- If $p = 2$, same as case $p > 2$ with polylog terms
- (A) quantifies the complexity of $\mathcal{F}$
  - $p < 2$, Donsker class
  - $p > 2$, possibly non-Donsker class (bigger than Donsker)

Corollary

In the same context, suppose that $p < 2$ ($\ell(\mathcal{F})$ is a Donsker class) and that the adaptive experiment implements an ε-greedy exploration with $\varepsilon = t^{-\beta}$, $\beta \in]0, 1[$: for all $t \geq t_0$,

$$g_t(a|x) \in \left\{ \frac{t^{-\beta}}{|\mathcal{A}| - 1}, 1 - t^{-\beta} \right\}, \quad \text{all } (a, x) \in \mathcal{A} \times \mathcal{X}$$

Then

- $\gamma_T^{\text{avg}} = O(T^{\beta})$ and $\gamma_T^{\text{max}} = O(T^{\beta})$
- $0 \leq E\left[R(\widehat{f}_T) - R(f^\star)\right] = O\left(T^{-\frac{1}{2}+\frac{\beta}{2}}\right)$

Corollary

In the same context, suppose that $p < 2$ ($\ell(\mathcal{F})$ is a Donsker class) and that the adaptive experiment implements an ε-greedy exploration with $\varepsilon = t^{-\beta}$, $\beta \in]0, 1[$: for all $t \geq t_0$,

$$g_t(a|x) \in \left\{ \frac{t^{-\beta}}{|\mathcal{A}| - 1}, 1 - t^{-\beta} \right\}, \quad \text{all } (a, x) \in \mathcal{A} \times \mathcal{X}$$

Then

- $\gamma_T^{\text{avg}} = O(T^\beta)$ and $\gamma_T^{\max} = O(T^\beta)$
- $0 \leq E\left[R(\widehat{f}_T) - R(f^\star)\right] = O\left(T^{-\frac{1}{2}+\frac{\beta}{2}}\right)$
  - this matches the lower bound obtained by Zhan et al (2021)

Corollary

In the same context, suppose that $p < 2$ ($\ell(\mathcal{F})$ is a Donsker class) and that the adaptive experiment implements an ε-greedy exploration with $\varepsilon = t^{-\beta}$, $\beta \in]0, 1[$: for all $t \geq t_0$,

$$g_t(a|x) \in \left\{ \frac{t^{-\beta}}{|\mathcal{A}| - 1}, 1 - t^{-\beta} \right\}, \quad \text{all } (a, x) \in \mathcal{A} \times \mathcal{X}$$

Then

- $\gamma_T^{\text{avg}} = O(T^\beta)$ and $\gamma_T^{\max} = O(T^\beta)$
- $0 \leq E\left[R(\widehat{f}_T) - R(f^\star)\right] = O\left(T^{-\frac{1}{2}+\frac{\beta}{2}}\right)$
  - this matches the lower bound obtained by Zhan et al (2021)
  - difficult to compare our upper bound to theirs (they use the "Natarayan dimension" to control the complexity of $\ell(\mathcal{F})$)

Corollary

In the same context, suppose that $p < 2$ ($\ell(\mathcal{F})$ is a Donsker class) and that the adaptive experiment implements an ε-greedy exploration with $\varepsilon = t^{-\beta}$, $\beta \in]0, 1[$: for all $t \geq t_0$,

$$g_t(a|x) \in \left\{ \frac{t^{-\beta}}{|\mathcal{A}| - 1}, 1 - t^{-\beta} \right\}, \quad \text{all } (a, x) \in \mathcal{A} \times \mathcal{X}$$

Then

- $\gamma_T^{\text{avg}} = O(T^{\beta})$ and $\gamma_T^{\max} = O(T^{\beta})$
- $0 \leq E\left[R(\widehat{f}_T) - R(f^\star)\right] = O\left(T^{-\frac{1}{2}+\frac{\beta}{2}}\right)$
  - this matches the lower bound obtained by Zhan et al (2021)
  - difficult to compare our upper bound to theirs (they use the "Natarayan dimension" to control the complexity of $\ell(\mathcal{F})$)
  - but when $\mathcal{F}$ is parametrized by a finite-dimensional parameter set, we close the gap and they do not

Discussion

There is more in our article, Bibaut et al (2021):

- faster rates under a margin condition, assuming $\mathcal{F}$ contains the absolute best policy
- same kind of results in regression and classification settings

And there remains many open questions, for instance:

- how to deal with changing classes $(\mathcal{F}_t)_{t \geq 1}$?
- can using a doubly-robust estimator of $R(f)$ instead of $\widehat{R}_T(f)$ yield better finite sample performance?

Discussion

There is more in our article, Bibaut et al (2021):

- faster rates under a margin condition, assuming $\mathcal{F}$ contains the absolute best policy
- same kind of results in regression and classification settings

And there remains many open questions, for instance:

- how to deal with changing classes $(\mathcal{F}_t)_{t \geq 1}$?
- can using a doubly-robust estimator of $R(f)$ instead of $\widehat{R}_T(f)$ yield better finite sample performance?

Merci

Short bibliography

- A. Bibaut, N. Kallus, M. Dimakopoulou, A. Chambaz, M. J. van der Laan, *Risk minimization from adaptively collected data: guarantees for supervised and policy learning*, NeurIPS 2021, 34:19261–19273, 2021
- R. van Handel, *On the minimal penalty for markov order estimation*, Probability theory and related fields, 150(3-4):709–738, 2011
- R. Zhan, Z. Ren, S. Athey, Z. Zhou, *Policy learning with adaptively collected data*, arXiv preprint arXiv:2105.02344, 2021

Faster rates

Define

$$\mu(a, X) = E(Y(a)|X), \quad \text{all } a \in \mathcal{A}$$

$$\mu^\star(X) = \max_{a \in \mathcal{A}} \mu(a, X)$$

and $a^\star(X)$ such that $\mu(a^\star(X), X) = \mu^\star(X)$

Suppose that

- $R(f^\star) = -E[\mu^\star(X)]$ (the class $\mathcal{F}$ is well-specified)
- margin assumption: there exists $\nu > 0$ such that, for all $s > 0$,

$$P\left(0 < \mu^\star(X) - \max_{a \neq a^\star(X)} \mu(a, X) \leq s\right) \lesssim s^\nu$$

Consider for simplicity the same adaptive experiment as in the corollary, with $t^{-\beta}$-greedy exploration, and the case that p is very small. Then, for all $\delta \in]0, \frac{1}{2}[$, with probability at least $(1 - \delta)$,

$$0 \leq E\left[R(\widehat{f}_T) - R(f^\star)\right] = O\left(T^{-\left(\frac{1}{2} + \frac{\beta}{2}\right) \times \frac{2+2\nu}{2+\nu}}\right)$$

Introduce $L(f)(A_t, X_t) = E[\ell(f)(X_t, A_t, Y_t)|A_t, X_t]$. It holds that

$$E\left[\frac{g^\star(A_t|X_t)}{g_t(A_t|X_t)} \times \ell(f)(X_t, A_t, Y_t)\right]$$

Introduce $L(f)(A_t, X_t) = E[\ell(f)(X_t, A_t, Y_t)|A_t, X_t]$. It holds that

$$
\begin{aligned}
&E\left[\tfrac{g^\star(A_t|X_t)}{g_t(A_t|X_t)} \times \ell(f)(X_t, A_t, Y_t)\right] \\
&= E\left[\tfrac{g^\star(A_t|X_t)}{g_t(A_t|X_t)} \times L(f)(X_t, A_t)\right]
\end{aligned}
$$

back

Introduce $L(f)(A_t, X_t) = E[\ell(f)(X_t, A_t, Y_t)|A_t, X_t]$. It holds that

$$
\begin{aligned}
& E\left[\frac{g^\star(A_t|X_t)}{g_t(A_t|X_t)} \times \ell(f)(X_t, A_t, Y_t)\right] \\
&= E\left[\frac{g^\star(A_t|X_t)}{g_t(A_t|X_t)} \times L(f)(X_t, A_t)\right] \\
&= E\left[\left(\sum_{a \in \mathcal{A}} \mathbf{1}\{A_t = a\}\right) \times \frac{g^\star(A_t|X_t)}{g_t(A_t|X_t)} \times L(f)(X_t, A_t)\right]
\end{aligned}
$$

back

Introduce $L(f)(A_t, X_t) = E[\ell(f)(X_t, A_t, Y_t)|A_t, X_t]$. It holds that

$$
\begin{aligned}
& E\left[\frac{g^\star(A_t|X_t)}{g_t(A_t|X_t)} \times \ell(f)(X_t, A_t, Y_t)\right] \\
&= E\left[\frac{g^\star(A_t|X_t)}{g_t(A_t|X_t)} \times L(f)(X_t, A_t)\right] \\
&= E\left[\left(\sum_{a\in\mathcal{A}} \mathbf{1}\{A_t = a\}\right) \times \frac{g^\star(A_t|X_t)}{g_t(A_t|X_t)} \times L(f)(X_t, A_t)\right] \\
&= \sum_{a\in\mathcal{A}} E\left[\mathbf{1}\{A_t = a\} \times \frac{g^\star(a|X_t)}{g_t(a|X_t)} \times L(f)(X_t, a)\right]
\end{aligned}
$$

back

Introduce $L(f)(A_t, X_t) = E[\ell(f)(X_t, A_t, Y_t)|A_t, X_t]$. It holds that

$$
\begin{aligned}
& E\left[\tfrac{g^\star(A_t|X_t)}{g_t(A_t|X_t)} \times \ell(f)(X_t, A_t, Y_t)\right] \\
&= E\left[\tfrac{g^\star(A_t|X_t)}{g_t(A_t|X_t)} \times L(f)(X_t, A_t)\right] \\
&= E\left[\left(\sum_{a \in \mathcal{A}} \mathbf{1}\{A_t = a\}\right) \times \tfrac{g^\star(A_t|X_t)}{g_t(A_t|X_t)} \times L(f)(X_t, A_t)\right] \\
&= \sum_{a \in \mathcal{A}} E\left[\mathbf{1}\{A_t = a\} \times \tfrac{g^\star(a|X_t)}{g_t(a|X_t)} \times L(f)(X_t, a)\right] \\
&= \sum_{a \in \mathcal{A}} E\left[P\{A_t = a|X_t\} \times \tfrac{g^\star(a|X_t)}{g_t(a|X_t)} \times L(f)(X_t, a)\right]
\end{aligned}
$$

Introduce $L(f)(A_t, X_t) = E[\ell(f)(X_t, A_t, Y_t)|A_t, X_t]$. It holds that

$$
\begin{aligned}
&E\left[\frac{g^\star(A_t|X_t)}{g_t(A_t|X_t)} \times \ell(f)(X_t, A_t, Y_t)\right] \\
&= E\left[\frac{g^\star(A_t|X_t)}{g_t(A_t|X_t)} \times L(f)(X_t, A_t)\right] \\
&= E\left[\left(\sum_{a\in\mathcal{A}} \mathbf{1}\{A_t = a\}\right) \times \frac{g^\star(A_t|X_t)}{g_t(A_t|X_t)} \times L(f)(X_t, A_t)\right] \\
&= \sum_{a\in\mathcal{A}} E\left[\mathbf{1}\{A_t = a\} \times \frac{g^\star(a|X_t)}{g_t(a|X_t)} \times L(f)(X_t, a)\right] \\
&= \sum_{a\in\mathcal{A}} E\left[P\{A_t = a|X_t\} \times \frac{g^\star(a|X_t)}{g_t(a|X_t)} \times L(f)(X_t, a)\right] \\
&= \sum_{a\in\mathcal{A}} E\left[g_t(a|X_t) \times \frac{g^\star(a|X_t)}{g_t(a|X_t)} \times L(f)(X_t, a)\right]
\end{aligned}
$$

back

Introduce $L(f)(A_t, X_t) = E[\ell(f)(X_t, A_t, Y_t)|A_t, X_t]$. It holds that

$$
\begin{aligned}
& E\left[\frac{g^\star(A_t|X_t)}{g_t(A_t|X_t)} \times \ell(f)(X_t, A_t, Y_t)\right] \\
&= E\left[\frac{g^\star(A_t|X_t)}{g_t(A_t|X_t)} \times L(f)(X_t, A_t)\right] \\
&= E\left[\left(\sum_{a \in \mathcal{A}} \mathbf{1}\{A_t = a\}\right) \times \frac{g^\star(A_t|X_t)}{g_t(A_t|X_t)} \times L(f)(X_t, A_t)\right] \\
&= \sum_{a \in \mathcal{A}} E\left[\mathbf{1}\{A_t = a\} \times \frac{g^\star(a|X_t)}{g_t(a|X_t)} \times L(f)(X_t, a)\right] \\
&= \sum_{a \in \mathcal{A}} E\left[P\{A_t = a|X_t\} \times \frac{g^\star(a|X_t)}{g_t(a|X_t)} \times L(f)(X_t, a)\right] \\
&= \sum_{a \in \mathcal{A}} E\left[g_t(a|X_t) \times \frac{g^\star(a|X_t)}{g_t(a|X_t)} \times L(f)(X_t, a)\right] \\
&= E\left[\sum_{a \in \mathcal{A}} g^\star(A_t|X_t) \times L(f)(X_t, A_t)\right]
\end{aligned}
$$

Introduce $L(f)(A_t, X_t) = E[\ell(f)(X_t, A_t, Y_t)|A_t, X_t]$. It holds that

$$
\begin{aligned}
& E\left[\tfrac{g^\star(A_t|X_t)}{g_t(A_t|X_t)} \times \ell(f)(X_t, A_t, Y_t)\right] \\
&= E\left[\tfrac{g^\star(A_t|X_t)}{g_t(A_t|X_t)} \times L(f)(X_t, A_t)\right] \\
&= E\left[\left(\sum_{a\in\mathcal{A}} \mathbf{1}\{A_t = a\}\right) \times \tfrac{g^\star(A_t|X_t)}{g_t(A_t|X_t)} \times L(f)(X_t, A_t)\right] \\
&= \sum_{a\in\mathcal{A}} E\left[\mathbf{1}\{A_t = a\} \times \tfrac{g^\star(a|X_t)}{g_t(a|X_t)} \times L(f)(X_t, a)\right] \\
&= \sum_{a\in\mathcal{A}} E\left[P\{A_t = a|X_t\} \times \tfrac{g^\star(a|X_t)}{g_t(a|X_t)} \times L(f)(X_t, a)\right] \\
&= \sum_{a\in\mathcal{A}} E\left[g_t(a|X_t) \times \tfrac{g^\star(a|X_t)}{g_t(a|X_t)} \times L(f)(X_t, a)\right] \\
&= E\left[\sum_{a\in\mathcal{A}} g^\star(A_t|X_t) \times L(f)(X_t, A_t)\right] \\
&= E\left[\sum_{a\in\mathcal{A}} g^\star(A_t|X_t) \times \ell(f)(X_t, A_t, Y_t)\right] = R(f).
\end{aligned}
$$

back