HAL
open science

Targeting a simple statistical bandit problem

Antoine Chambaz, Wenjing Zheng

To cite this version:

Antoine Chambaz, Wenjing Zheng. Targeting a simple statistical bandit problem. 2016. hal-01359222

HAL Id: hal-01359222

https://hal.science/hal-01359222

Preprint submitted on 2 Sep 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Targeting a simple statistical bandit problem

Antoine Chambaz[1,2,3] and Wenjing Zheng[2]

[1] Modal'X, Université Paris Nanterre
[2] Division of Biostatistics, School of Public Health, UC Berkeley
[3] MAP5 (UMR CMRS 8145), Université Paris Descartes

1 Introduction

Statistical challenge. An infinite sequence of independent and identically distributed (iid) random variables $(W_n, Y_n(0), Y_n(1))_{n\geq 1}$ drawn from a common law Q_0 is to be sequentially and partially disclosed during the course of a controlled experiment. The first component, W_n, describes the nth context in which we will have to carry out one action out of two, denoted $a = 0$ and $a = 1$. The second and third components, $Y_n(0)$ and $Y_n(1)$, are the rewards that actions $a = 0$ and $a = 1$ would grant. The set $\mathcal{W}$ of contexts may be high-dimensional. The rewards take their values in $]0, 1[$.

The controlled experiment will unfold as follows. Sequentially, we will be informed of the new context W_n. We will then carry out a randomized action $A_n \in \{0, 1\}$ with probability either $g_n(1|W_n)$ or $g_n(0|W_n) \equiv 1 - g_n(1|W_n)$ to go for either action $a = 1$ or action $a = 0$, where $g_n(\cdot|W_n)$ will be determined by us based on observations $O_1, \ldots, O_{n-1}$ accrued so far during the course of the experiment. We will then be granted reward $Y_n \equiv Y_n(A_n)$ corresponding to the action undertaken, the alternative reward being kept undisclosed, hence the nth observation $O_n \equiv (W_n, A_n, Y_n)$. This setting is one of the simplest bandits settings in the machine learning literature, hence the expression "simple bandit problem" in the title of this manuscript.

Our objective justifies why the expression actually reads "simple *statistical* bandit problem". Indeed, it consists in inferring the optimal rule

$$r_0(W) \equiv \arg\max_{a=0,1} E_{Q_0}\left(Y(a)|W\right)$$

(by convention, $r_0(W) = 1$ if equality occurs) with r_n and the mean reward under r_0,

$$\psi_0 \equiv E_{Q_0}\big(Y(r_0(W))\big),$$

trying to get a narrow confidence interval (CI) for ψ_0 and a sense of how well we sequentially determined our actions through the estimation of the following regret:

$$\mathcal{R}_n \equiv \frac{1}{n}\sum_{i=1}^{n} Y_i(r_n(W_i)) - Y_i.$$

Regret is one the most central notion in the bandits literature. Seen here as a data-adaptive parameter, $\mathcal{R}_n$ compares the actual average of the rewards granted at step n, $n^{-1}\sum_{i=1}^{n} Y_i$, with the counterfactual average of the rewards we would have been be granted at step n if we had constantly used r_n from the start of the experiment to decide which action to carry out at the n successive steps, $n^{-1}\sum_{i=1}^{n} Y_i(r_n(W_i))$. We emphasize that the former average is known to us but the latter is not, since it may occur that $A_i \neq r_n(W_i)$ for some $1 \leq i \leq n$, in which case $Y_i(r_n(W_i))$ is the reward that was kept secret from us at step i. If all actions A_i coincide with $r_n(W_i)$ $(1 \leq i \leq n)$, a very unlikely event, then $\mathcal{R}_n = 0$. In general, $n\mathcal{R}_n$ equals

$$\sum_{\substack{1 \leq i \leq n \\ A_i \neq r_n(W_i)}} Y_i(1 - A_i) - Y_i(A_i),$$

this alternative expression showing that $n\mathcal{R}_n$ is the counterfactual sum of the differences between the two possible rewards at each step i where the randomized action A_i differs from the optimal action $r_n(W_i)$ according to the estimate of the optimal rule at step n. Since the optimal action is that which has the larger conditional mean given the context, as opposed to that action which grants the larger reward, it is not guaranteed that $\mathcal{R}_n$ is non-negative.

Inference of data-adaptive parameters are at the core of the present manuscript. We will derive CIs for ψ_0 and for $\mathcal{R}_n$, the first data-adaptive parameter we introduced, from a targeted minimum loss estimator (TMLE, which also stands for targeted minimum loss estima*tion*) of the second data-adaptive parameter

$$\psi_{r_n,0} \equiv E_{Q_0}\big(Y(r_n(W))\big),$$

the mean reward under r_n, thus justifying entirely the title of the manuscript. There is much more to $\psi_{r_n,0}$ than being a convenient proxy for the inference of ψ_0. In fact, we may argue that $\psi_{r_n,0}$ is more interesting than ψ_0 itself because it is the mean reward under rule r_n that we know and can use concretely. The same reasoning motivates our choice of regret $\mathcal{R}_n$ instead of its counterpart with r_0 substituted for r_n.

Quick review of literature. Chakraborty and Moodie (2013) present an excellent unified overview on the estimation of optimal rules. Their focus is on dynamic rules, which actually prescribe successive actions at successive time points based on time-dependent contexts. The estimation of the optimal rule from iid observations has been studied extensively, with a recent interest in the use of machine learning algorithms to reach this goal (Qian and Murphy, 2011; Zhao et al., 2012, 2015; Zhang et al., 2012a,b; Rubin and van der Laan, 2012; Luedtke and van der Laan, 2016). The estimation of the mean reward under the optimal rule is more challenging. Zhao et al. (2012, 2015) use their theoretical risk bounds evaluating the statistical performance of the estimator of the optimal rule as measures of statistical performance of the resulting estimators of the mean reward under the optimal rule. However, this approach does not yield CIs.

Constructing CIs for the mean reward under the optimal rule is known to be more difficult when there exists a stratum of context where no action dominates the other (if action means treatment, no treatment is neither beneficial nor harmful) (Robins, 2004). In this so called "exceptional" case, the definition of the optimal rule has to be disambiguated. Assuming non-exceptionality, Zhang et al. (2012a) derive CIs for the mean reward under the (sub-) optimal rule defined as the optimal rule over a parametric class of candidate rules. Luedtke and van der Laan (2015a) derive CIs for the actual mean reward under the optimal rule. In the more general case where exceptionality can occur, different approaches have been considered (Chakraborty et al., 2014; Goldberg et al., 2014; Laber et al., 2014b; Luedtke and van der Laan, 2015b). Here, we focus on the non-exceptional case under a companion margin assumption (Mammen and Tsybakov, 1999).

We already unveiled that our pivotal TMLE is actually conceived as an estimator of the mean reward under the current estimate of the optimal rule. Worthy of interest on its own, this data-adaptive statistical parameter (or similar ones) has also been considered in (Chakraborty et al., 2014; Laber et al., 2014a,b; Luedtke and van der Laan, 2015a,b).

Our main result is a central limit theorem (CLT), which enables the construction of various CIs. The analysis (for the proofs that we omit here, see the full-blown Chambaz et al., 2016) builds upon previous studies on the construction and statistical analysis of targeted, covariate-adjusted, response-adaptive trials also based on TMLE (Chambaz and van der Laan, 2014; Zheng et al., 2015; Chambaz et al., 2015). The asymptotic variance in the CLT takes the form of the variance of an efficient influence curve at a limiting distribution, allowing to discuss the efficiency of inference. One of the cornerstones of the theoretical study is a new maximal inequality for martingales with respect to (wrt) the uniform entropy integral. Proved by decoupling (de la Peña and Giné, 1999), symmetrization and chaining, it allows us to control several empirical processes indexed by random functions.

Organization. The manuscript is organized as follows. Section 2 presents our sampling strategy and how we implement TMLE. Section 3 describes the convergence of the data-adaptive sampling strategy, states the CLT satisfied by the TMLE, and Section 4 discusses the construction of CIs based on it. Section 5 illustrates the manuscript with the results of a simulation study. Section 6 concludes the manuscript (on a twist).

2 Sampling strategy and targeted minimum loss estimation

Let us introduce some notation. We let $\bar{Q}_{0,Y}$ and $\bar{q}_{0,Y}$ respectively denote the true conditional expectation $\bar{Q}_{0,Y}(a,W) \equiv E_{Q_0}(Y(a)|W)$ (for $a = 0,1$) and related "blip function" $\bar{q}_{0,Y}(W) \equiv \bar{Q}_{0,Y}(1,W) - \bar{Q}_{0,Y}(0,W)$. More generally, every (measurable) function $\bar{Q}_Y$ from $\{0,1\} \times \mathcal{W}$ to $]0,1[$ is associated with its blip function $\bar{q}_Y(W) \equiv \bar{Q}_Y(1,W) - \bar{Q}_Y(0,W)$. Thus,

$$r_0(W) = \underset{a=0,1}{\arg\max}\, \bar{Q}_{0,Y}(a,W) = \mathbf{1}\{\bar{q}_{0,Y}(W) \geq 0\} \equiv R(\bar{Q}_{0,Y})(W) \tag{0.1}$$

(recall that, by convention, $r_0(W) = 1$ if equality occurs), ψ_0 equals $E_{Q_0}\big(\bar{Q}_{0,Y}(r_0(W),W)\big)$ and $\psi_{r_n,0}$ equals $E_{Q_0}\big(\bar{Q}_{0,Y}(r_n(W),W)\big)$.

The adaptive sampling strategy and TMLE rely on a working model $\bar{\mathcal{Q}}_Y$ and loss function L_Y for $\bar{Q}_{0,Y}$ that we determine prior to starting the controlled experiment. Requirements on the complexity of $\bar{\mathcal{Q}}_Y$ will be given in Section 3. They also rely on a non-decreasing, Lipschitz function G from $[-1,1]$ to $[0,1]$ such that $G(0) = 1/2$ and, for some fixed and small real numbers $p, \xi > 0$, $|x| > \xi$ implies $G(x) = p$ if $x < 0$ and $G(x) = (1-p)$ if $x > 0$

2.1 Sampling strategy

The first n_0 randomized actions $A_1, \ldots, A_{n_0}$ are drawn from the Bernoulli distribution with parameter $1/2$. In other words, we set $g_i = g^b$ for $i = 1, \ldots, n_0$ where $g^b(1|W) = 1 - g^b(0|W) \equiv 1/2$, thus giving equiprobable chance to each action to be carried out as long as deemed necessary to start estimating $\bar{Q}_{0,Y}$ from the accrued observations. Suppose now that $O_1, \ldots, O_{n-1}$ have been observed. Explaining how the next observation is obtained will complete the description of the sampling strategy.

We estimate $\bar{Q}_{0,Y}$ with

$$\bar{Q}_{n,Y} \in \underset{\bar{Q}_Y \in \bar{\mathcal{Q}}_Y}{\arg\min} \frac{1}{n-1} \sum_{i=1}^{n-1} L_Y(\bar{Q}_Y)(O_i) \frac{g^b(A_i|W_i)}{g_i(A_i|W_i)}. \tag{0.2}$$

The weights $g^b(A_i|W_i)/g_i(A_i|W_i)$ $(i = 1, \ldots, n)$ compensate for the fact that our observations are not identically distributed. We associate the above estimator with its blip function $\bar{q}_{n,Y}$ and rule $r_n \equiv R(\bar{Q}_{n,Y})(W) \equiv \mathbf{1}\{\bar{q}_{n,Y}(W) \geq 0\}$. They are substitution estimators of $\bar{q}_{0,Y}$ and r_0. We now define

$$g_{n+1}(1|W) = 1 - g_{n+1}(0|W) \equiv G(\bar{q}_{n,Y})(W),$$

and thus are in a position to sample O_{n+1}: we request the disclosure of W_{n+1}, draw A_{n+1} from the Bernoulli distribution with parameter $g_{n+1}(1|W_{n+1})$, carry out action A_{n+1}, are granted reward $Y_{n+1} = Y_{n+1}(A_{n+1})$ and form $O_{n+1} \equiv (W_{n+1}, A_{n+1}, Y_{n+1})$.

The randomized action A_{n+1} rarely differs from the deterministic action $r_n(W_{n+1})$ in the sense that

$$|g_{n+1}(1|W_{n+1}) - r_n(W_{n+1})|\, \mathbf{1}\{|\bar{q}_{n,Y}(W_{n+1})| > \xi\} = p: \tag{0.3}$$

if $|\bar{q}_{n,Y}(W_{n+1})|$ is sufficiently away from 0, meaning that we confidently believe that one action is superior to the other, then A_{n+1} equals $r_n(W_{n+1})$ with (large) probability $(1-p)$. On the contrary, if $|\bar{q}_{n,Y}(W_{n+1})|$ is small, meaning that it is unclear whether an action is superior to the other or not, then the probability that A_{n+1} be equal $r_n(W_{n+1})$ lies between $(1-t)$ and $1/2$, and is continuously closer to $1/2$ as $|\bar{q}_{n,Y}(W_{n+1})|$ gets closer to 0.

2.2 TMLE

The initial substitution estimator of $\psi_{r_n,0}$,

$$\psi_n^0 \equiv \frac{1}{n}\sum_{i=1}^n \bar{Q}_{n,Y}(r_n(W_i), W_i),$$

may fail to be $\sqrt{n}$-consistent and must therefore be enhanced. Fortunately, we can rely on TMLE. Indeed, just like any mapping $\Psi_\rho : P_{Q,g} \mapsto E_Q\left(Y(\rho(W))\right)$ with a fixed rule ρ from $\mathcal{W}$ to $\{0,1\}$, the data-adaptive Ψ_{r_n} is pathwise differentiable from the nonparametric set of all possible data-generating distributions $P_{Q,g}$ of $O \equiv (W, A, Y)$ with g bounded away from 0 to $[0,1]$ (Luedtke and van der Laan, 2015a,b). Its efficient influence curve at $P_{Q,g}$ is $\Delta_{r_n}(Q,g)$ where, for every rule $\rho : \mathcal{W} \to \{0,1\}$, $\Delta_\rho(Q,g)$ is characterized by

$$\Delta_\rho(Q,g)(O) = (Y - \bar{Q}_Y(\rho(W), W))\frac{\mathbf{1}\{A = \rho(W)\}}{g(A|W)} + \bar{Q}_Y(\rho(W), W) - \Psi_\rho(P_{Q,g}). \tag{0.4}$$

We let ℓ denote the quasi negative-log-likelihood loss function, which is characterized by

$$-\ell(\bar{Q}_Y)(O) \equiv Y\log(\bar{Q}_Y(A,W)) + (1-Y)\log(1-\bar{Q}_Y(A,W)),$$

and introduce the one-dimensional regression model through $\bar{Q}_{n,Y}$ given by

$$\text{logit}\left(\bar{Q}_{n,Y}(\epsilon)(A,W)\right) \equiv \text{logit}\left(\bar{Q}_{n,Y}(A,W)\right) + \epsilon\frac{\mathbf{1}\{A = r_n(W)\}}{g_n(A|W)}$$

for all $\epsilon \in \mathbb{R}$. It is tailored to the estimation of $\psi_{r_n,0} = \Psi_{r_n}(P_{Q_0,g_n})$ in the sense that $\frac{\partial}{\partial\epsilon}\ell(\bar{Q}_{n,Y}(\epsilon))(O)|_{\epsilon=0}$ equals $(Y - \bar{Q}_{n,Y}(A,W))\mathbf{1}\{A = r_n(W)\}/g_n(A|W)$, the component of $\Delta_{r_n}(Q_n, g_n)$ which is orthogonal to the set of P_{Q_n,g_n}-square-integrable and centered functions of W. Here, Q_n denotes any distribution of $(W, Y(0), Y(1))$ such that $E_{Q_n}(Y(a)|W) = \bar{Q}_{n,Y}(a,W)$ for each $a = 0, 1$, Q_n-almost surely.

The optimal fluctuation parameter is

$$\epsilon_n \in \arg\min_{\epsilon\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^n \ell(\bar{Q}_{n,Y}(\varepsilon))(O_i)\frac{g_n(A_i|W_i)}{g_i(A_i|W_i)}.$$

Setting $\bar{Q}^*_{n,Y} \equiv \bar{Q}_{n,Y}(\epsilon_n)$, the TMLE of $\psi_{r_n,0}$ finally writes

$$\psi_n^* \equiv \frac{1}{n}\sum_{i=1}^n \bar{Q}^*_{n,Y}(r_n(W_i), W_i).$$

3 Convergence of sampling strategy and asymptotic normality of TMLE

We must choose the working model $\bar{\mathcal{Q}}_Y$ and loss function L_Y for $\bar{Q}_{0,Y}$ in such a way that $\bar{\mathcal{Q}}_Y$ and the subsequent working models $L_Y(\bar{\mathcal{Q}}_Y) \equiv \{L(\bar{Q}_Y) : Q_Y \in \bar{\mathcal{Q}}_Y\}$ and $R(\bar{\mathcal{Q}}_Y) \equiv \{R(\bar{Q}_Y) : Q_Y \in \bar{\mathcal{Q}}_Y\}$ be reasonably large/complex relative to a measure of complexity central to the theory of empirical processes (van der Vaart and Wellner, 1996). Specifically, we must choose them so that $\bar{\mathcal{Q}}_Y$, $L(\bar{\mathcal{Q}}_Y)$, $R(\bar{\mathcal{Q}}_Y)$ be separable (countable would be sufficient) and that each admit a finite uniform entropy integral wrt an envelope function (van der Vaart and Wellner, 1996, Sections 2.5.1 and 2.6).

Introduce the norm $\|\cdot\|_{Q_0}$ characterized by $\|f\|^2_{Q_0} \equiv E_{P_{Q_0,g^b}}(f^2(O))$. We will assume that $\bar{\mathcal{Q}}_Y$ satisfies the following assumption:

A1. There exists $\bar{Q}_{1,Y} \in \bar{\mathcal{Q}}_Y$ such that $\bar{Q}_Y \mapsto E_{P_{Q_0,g^b}}(L_Y(\bar{Q}_Y)(O))$ from $\bar{\mathcal{Q}}_Y$ to $\mathbb{R}$ is minimized at $\bar{Q}_{1,Y}$. Moreover, $\bar{Q}_{1,Y}$ is well-separated in the sense that, for all $\delta > 0$,

$$E_{P_{Q_0,g^b}}\left(L_Y(\bar{Q}_{1,Y})(O)\right) < \inf\left\{E_{P_{Q_0,g^b}}\left(L_Y(\bar{Q}_Y)(O)\right) : \bar{Q}_Y \in \bar{\mathcal{Q}}_Y, \|\bar{Q}_Y - \bar{Q}_{1,Y}\|_{Q_0} \geq \delta\right\}.$$

Finally, $\bar{q}_{1,Y} = \bar{q}_{0,Y}$.

The most stringent condition is the equality of the blip functions.

Our second assumption concerns the fluctuation/targeting step in the construction of the TMLE. Let g_0 be given by

$$g_0(1|W) = 1 - g_0(0|W) \equiv G(\bar{q}_{0,Y}(W)). \tag{0.5}$$

Just like g_n is an approximation to r_n, see (0.3) and its comment, g_0 is an approximation to the optimal rule r_0. We will soon see that g_0 is the limit of g_n. For every rule $\rho : \mathcal{W} \to \{0,1\}$, consider the one-dimensional regression model through $\bar{Q}_{1,Y}$ characterized by

$$\text{logit}\left(\bar{Q}_{1,Y,\rho}(\epsilon)(A,W)\right) \equiv \text{logit}\left(\bar{Q}_{1,Y}(A,W) + \epsilon\frac{\mathbf{1}\{A = \rho(W)\}}{g_0(A|W)}\right) \tag{0.6}$$

for all $\epsilon \in \mathbb{R}$. We will assume that:

A2. For every rule $\rho : \mathcal{W} \to \{0,1\}$, there exists a unique $\epsilon_0(\rho) \in \mathbb{R}$ which minimizes the real-valued mapping $\epsilon \mapsto E_{P_{Q_0,g_0}}\left(\ell(\bar{Q}_{1,Y,\rho}(\epsilon))(O)\right)$ over $\mathbb{R}$.

The third and last assumption concerns Q_0:

A3. The conditional distributions of $Y(0)$ and $Y(1)$ given W under Q_0 is not degenerated. Moreover, there exist $\gamma_1, \gamma_2 > 0$ such that, for all $t \geq 0$,

$$P_{Q_0}\left(0 \leq |\bar{q}_{0,Y}(W)| \leq t\right) \leq \gamma_1 t^{\gamma_2}. \tag{0.7}$$

Taking $t = 0$ in (0.7) yields $\bar{q}_{0,Y}(W) = 0$ with probability zero under Q_0. In words, the optimal action $r_0(W)$ is defined without ambiguity Q_0-almost surely. In the terminology of (Robins, 2004), Q_0 is non-exceptional. More generally, (0.7) for $t > 0$ is a known as a margin assumption. Inspired from the seminal article (Mammen and Tsybakov, 1999), **A3** formalizes a tractable concentration of $\bar{q}_{0,Y}(W)$ around 0, where our inference task is the most challenging.

We may now state our results. According to the first proposition, the sampling strategy nicely converges as n tends to infinity:

Proposition 0.1. *Under* **A1**, **A2** *and* **A3**, *it holds that* $\|\bar{Q}_{n,Y} - \bar{Q}_{1,Y}\|_{Q_0}$, $\|\bar{q}_{n,Y} - \bar{q}_{0,Y}\|_{Q_0}$, $\|r_n - r_0\|_{Q_0}$, $\|g_n - g_0\|_{Q_0}$ *and the non-negative data-adaptive parameter* $\psi_0 - \psi_{r_n,0}$ *all converge in probability to zero as* n *tends to infinity.*

The second proposition establishes the asymptotic normality of $\sqrt{n}(\psi_n^* - \psi_{r_n,0})$. Let us introduce $\bar{Q}_{1,Y}^* \equiv \bar{Q}_{1,Y,r_0}(\epsilon_0(r_0))$ (see (0.6) and **A2**), D_1^* given by

$$D_1^*(O) \equiv (Y - \bar{Q}_{1,Y}^*(A,W))\frac{\mathbf{1}\{A = r_0(W)\}}{g_0(A|W)} + \bar{Q}_{1,Y}^*(r_0(W),W) - \psi_0, \tag{0.8}$$

and $\sigma_1^2 \equiv E_{P_{Q_0,g_0}}\left(D_1^*(O)^2\right)$. Analogously, recalling the definition of $\bar{Q}_{n,Y}^* \equiv \bar{Q}_{n,Y}(\epsilon_n)$, let us define

$$D_{ni}^*(O_i) \equiv (Y_i - \bar{Q}_{n,Y}^*(A_i,W_i))\frac{\mathbf{1}\{A_i = r_n(W_i)\}}{g_i(A_i|W_i)} + \bar{Q}_{n,Y}^*(r_n(W_i),W_i) - \psi_n^* \quad (\text{each } 1 \leq i \leq n)$$

then $\sigma_n^2 \equiv n^{-1}\sum_{i=1}^n D_{ni}^*(O_i)^2$.

Proposition 0.2. *Under* **A1**, **A2** *and* **A3**, ψ_n^* *consistently estimates* $\psi_{r_n,0}$ *hence* ψ_0 *as well by Proposition 0.1. Moreover,* σ_n^2 *consistently estimates* σ_1^2*, which is positive, and* $\sqrt{n/\sigma_n^2}(\psi_n^* - \psi_{r_n,0})$ *converges in law to the standard normal distribution as* n *tends to infinity.*

Obviously, the larger is γ_2 from **A3**, the less concentrated is $\bar{q}_{0,Y}(W)$ around zero under Q_0, the less difficult is our inference task. If we assume that $\gamma_2 \geq 1$ and that the rate of convergence of $\bar{q}_{n,Y}$ to $\bar{q}_{0,Y}$ is sufficiently fast, then a first corollary to Proposition 0.2 shows that $\sqrt{n}(\psi_n^* - \psi_0)$ is also asymptotically normal. Introduce $\gamma_3 \equiv \frac{1}{4} + \frac{1}{2(1+\gamma_2)}$.

Corollary 0.1. *Under* **A1**, **A2** *and* **A3**, *if* $\gamma_2 \geq 1$ *hence* $\gamma_3 \in (\frac{1}{4}, \frac{1}{2}]$ *and if* $n^{\gamma_3}\|\bar{q}_{n,Y} - \bar{q}_{0,Y}\|_{Q_0}$ *converges in probability to zero as* n *tends to infinity, then the data-adaptive parameter* $\sqrt{n}(\psi_{r_n,0} - \psi_0)$ *converges in probability to zero as* n *tends to infinity. Therefore,* $\sqrt{n/\sigma_n^2}(\psi_n^* - \psi_0)$ *converges in law to the standard normal distribution as* n *tends to infinity.*

The proofs of Propositions 0.1, 0.2 and Corollary 0.1 rely on arguments typical of empirical processes theory and the analysis of TMLEs (Chambaz et al., 2016). The underlying martingale structure of the empirical process proves again a nice extension to an iid structure.

Let Q_1^* be any distribution of $(W, Y(0), Y(1))$ such that W has the same distribution under Q_0 and Q_1^* and $E_{Q_1^*}(Y(a)|W) = \bar{Q}_{1,Y}^*(a, W)$ for each $a = 0, 1$, Q_0-almost surely. The influence function D_1^* in (0.8) equals $\Delta_{r_0}(Q_1^*, g_0)$, the efficient influence curve of Ψ_{r_0} at $P_{Q_1^*,g_0}$ (0.4). Consequently, $\sigma_1^2 = E_{P_{Q_0,g_0}}\left(\Delta_{r_0}(Q_1^*, g_0)(O)^2\right)$.

If $\bar{Q}_{1,Y} = \bar{Q}_{0,Y}$ (a stronger condition than equality $\bar{q}_{1,Y} = \bar{q}_{0,Y}$ in **A1**), then $\bar{Q}_{1,Y}^* = \bar{Q}_{0,Y}$ (because $\epsilon_0(r_0)$ from **A2** equals zero) hence $\sigma_1^2 = E_{P_{Q_0,g_0}}\left(\Delta_{r_0}(Q_0, g_0)(O)^2\right)$: the asymptotic variance of $\sqrt{n}(\psi_n^* - \psi_{r_n,0})$ coincides with the generalized Cramér-Rao lower bound for the asymptotic variance of any regular and asymptotically linear estimator of $\Psi_{r_0}(P_{Q_0,g_0}) = \psi_0$ when sampling independently from P_{Q_0,g_0} (Luedtke and van der Laan, 2015b). Otherwise, the discrepancy between σ_1^2 and $E_{P_{Q_0,g_0}}\left(\Delta_{r_0}(Q_0, g_0)(O)^2\right)$ will vary subtly depending on that between $\bar{Q}_{1,Y}$ and $\bar{Q}_{0,Y}$, hence in particular on our working model $\bar{\mathcal{Q}}_Y$.

4 Confidence intervals

Set a confidence level $\alpha \in]0, 1/2[$ and let $\xi_{1-\alpha/2}$ be the corresponding $(1-\alpha/2)$-quantile of the standard normal distribution. By Proposition 0.2 and Corollary 0.1, the TMLE can be used to construct CIs for the data-adaptive parameter $\psi_{r_n,0}$ or ψ_0 itself, as stated in this second corollary to Proposition 0.2:

Corollary 0.2. *Under the assumptions of Proposition 0.2,*

$$\left[\psi_n^* \pm \xi_{1-\alpha/2}\frac{\sigma_n}{\sqrt{n}}\right] \tag{0.9}$$

contains $\psi_{r_n,0}$ *with probability tending to* $(1-\alpha)$ *as* n *tends to infinity. Moreover, under the stronger assumptions of Corollary 0.1, the above CI also contains* ψ_0 *with probability tending to* $(1-\alpha)$ *as* n *tends to infinity.*

Deriving a CI for $\mathcal{R}_n$ is not as immediate because of its counterfactual nature. We need to introduce a new assumption:

A4. There exist an infinite sequence $(U_n)_{n\geq 1}$ of iid random variables independent from $(W_n)_{n\geq 1}$ and taking values in $\mathcal{U}$ and a deterministic (measurable) function $\bar{\mathbb{Q}}_{0,Y}$ mapping $\{0,1\} \times \mathcal{U} \times \mathcal{W}$ to $]0,1[$ such that $Y_n(a) = \bar{\mathbb{Q}}_{0,Y}(a, U_n, W_n)$ for all $n \geq 1$ and both $a = 0, 1$.

With **A4**, we frame the present discussion in the context of non-parametric structural equations models (Pearl, 2000). The notation $\bar{\mathbb{Q}}_{0,Y}$ is justified by the equalities

$$\bar{Q}_{0,Y}(a, W_n) = E_{Q_0}(Y_n(a)|W_n) = E_{Q_0}(\bar{\mathbb{Q}}_{0,Y}(a, U_n, W_n)|W_n)$$

showing that, for each $n \geq 1$ and $a = 0, 1$, the conditional mean of $Y_n(a)$ given W_n is obtained by averaging out U_n from $\bar{\mathbb{Q}}_{0,Y}(a, U_n, W_n)$ conditionally on W_n.

Introduce

$$s_1^2 \equiv E_{P_{Q_0,g_0}}\left(\left(D_1^*(O) + \psi_0 - \bar{Q}_{0,Y}(r_0(W), W)\right)^2\right),$$
$$s_n^2 \equiv \frac{1}{n}\sum_{i=1}^{n}\left(D_{ni}^*(O_i) + \psi_n^0 - \bar{Q}_{n,Y}(r_n(W_i), W_i)\right)^2.$$

The latter is an empirical counterpart to and estimator of the former. We may now state the last result of this manuscript, which exhibits a conservative CI for $\mathcal{R}_n$:

Proposition 0.3. *Under* **A1**, **A2**, **A3** *and* **A4**, s_n^2 *consistently estimates* s_1^2, *which is positive. Moreover,*

$$\left[\psi_n^* - \frac{1}{n}\sum_{i=1}^{n} Y_i \pm \xi_{1-\alpha/2}\frac{s_n}{\sqrt{n}}\right] \tag{0.10}$$

contains $\mathcal{R}_n$ *with probability converging to* $(1-\alpha') \geq (1-\alpha)$ *as* n *tends to infinity.*

The proof of Proposition 0.3 unfolds as follows. Pretending, contrary to facts, that U_n is also observed at each step though not used to define the TMLE, which is thus the same as before, we adapt the proof of Proposition 0.2 to obtain a similar CLT. The normalization factor involved now depends on $U_1, \ldots, U_n$ as well. We straightforwardly derive from it a CI for $\mathcal{R}_n$ whose width λ_n depends on $U_1, \ldots, U_n$ too. Fortunately, we can prove that the width of the CI in (0.10) is always larger than λ_n. Since it is free of $U_1, \ldots, U_n$, this yields the desired result. This clever scheme of proof draws its inspiration from (Balzer et al., 2015).

5 Simulation study

We now illustrate Sections 2, 3 and 4 with a simulation study. Section 5.1 presents its settings and Section 5.2 its results.

5.1 Settings

Under Q_0, the baseline covariate W decomposes as $W \equiv (U, V) \in [0,1] \times \{1,2,3\}$, where U and V are independent random variables respectively drawn from the uniform distributions on $[0,1]$ and $\{1,2,3\}$. Moreover, $Y(0)$ and $Y(1)$ are conditionally drawn given W from Beta distributions with a constant variance set to 0.1 and means $\bar{Q}_{0,Y}(0, W)$ and $\bar{Q}_{0,Y}(1, W)$ satisfying

$$\bar{q}_{0,Y}(W) = \bar{Q}_{0,Y}(1, W) - \bar{Q}_{0,Y}(0, W) \equiv \tfrac{9}{8}\left(U^2 - \tfrac{5}{2}U + \tfrac{2}{3}\right) + \tfrac{3\sqrt{V}}{4\sqrt{3}}\mathbf{1}\{U \geq \tfrac{1}{4}\lfloor\tfrac{V+3}{3}\rfloor\}$$

and

$$\bar{Q}_{0,Y}(1, W) + \bar{Q}_{0,Y}(0, W) \equiv \tfrac{4}{5} + \tfrac{1}{3\sqrt{V}}\left(\cos\left(\tfrac{\pi}{2}\tfrac{4U}{V}\right)\mathbf{1}\{4U \leq V\} + \sin\left(\tfrac{\pi}{2}\tfrac{4U-V}{4-V}\right)\mathbf{1}\{4U > V\} - \tfrac{1}{2}\right).$$

The conditional means $\bar{Q}_{0,Y}(0, \cdot)$, $\bar{Q}_{0,Y}(1, \cdot)$ and associated blip function $\bar{q}_{0,Y}$ are represented in Figure 0.2 (left plots). We compute the numerical values of the following parameters: $\psi_0 \approx 0.5570$ (mean reward under optimal rule r_0); $\mathrm{Var}_{P_{Q_0,g^b}} \Delta(Q_0, g^b)(O) \approx 0.1812^2$ (variance under P_{Q_0,g^b} of the efficient influence curve of Ψ at P_{Q_0,g^b}, *i.e.*, under Q_0 with equiprobability of carrying out action $a = 1$ or $a = 0$);

$\mathrm{Var}_{P_{Q_0,g_0}} \Delta(Q_0, g_0)(O) \approx 0.1548^2$ (variance under P_{Q_0,g_0} of the efficient influence curve of Ψ at P_{Q_0,g_0}, *i.e.*, under Q_0 and the approximation g_0 to r_0); and $\mathrm{Var}_{P_{Q_0,r_0}} \Delta(Q_0, r_0)(O) \approx 0.1512^2$ (variance under P_{Q_0,r_0} of the efficient influence curve of Ψ at P_{Q_0,r_0}, *i.e.*, under Q_0 and r_0).

We set $p = 10\%$, $\xi = 1\%$ and choose G characterized over $[-1, 1]$ by

$$G(x) \equiv p\mathbf{1}\{x \leq -\xi\} + \left(-\tfrac{1/2-p}{2\xi^3}x^3 + \tfrac{1/2-p}{2\xi/3}x + \tfrac{1}{2}\right)\mathbf{1}\{-\xi \leq x \leq \xi\} + (1-p)\mathbf{1}\{x \geq \xi\}.$$

Reducing p to 5% did not change the results significantly (not shown). Working model $\bar{\mathcal{Q}}_Y$ consists of functions $\bar{Q}_{Y,\beta}$ mapping $\{0,1\} \times \mathcal{W}$ to $[0,1]$ such that, for each $a = 0, 1$ and $v \in \{1, 2, 3\}$, $\mathrm{logit}\, \bar{Q}_{Y,\beta}(a, (U, v))$ is a linear combination of $1, U, U^2, \ldots, U^5$ and $\mathbf{1}\{\frac{j-1}{10} \leq U < \frac{j}{10}\}$ $(1 \leq j \leq 10)$. The resulting global parameter β belongs to $\mathbb{R}^{96}$. Neither $\bar{Q}_{0,Y}$ nor $\bar{q}_{0,Y}$ belongs to $\bar{\mathcal{Q}}_Y$ or $\{\bar{q}_{Y,\beta} : \bar{Q}_{Y,\beta} \in \bar{\mathcal{Q}}_Y\}$. However, $\mathrm{expit}(\bar{q}_{Y,0})$ does belong to the latter working model.

The targeting steps are performed when sample size is a multiple of 25, at least 200 and no more than 1000, when the experiment is stopped. Working model $\bar{\mathcal{Q}}_Y$ is fitted wrt quasi log-likelihood loss function ℓ using the `cv.glmnet` function from package `glmnet` (Friedman et al., 2010), with weights given in (0.2) and the option `"lambda.min"`. This means imposing (data-adaptive) upper-bounds on the ℓ^1- and ℓ^2-norms of parameter β (via penalization), hence the search for a sparse optimal parameter.

We repeat $N = 1000$ times, independently, the strategy described in Section 2. Each time a targeting step is performed, we construct the CIs of Corollary 0.2 and Proposition 0.3, with a nominal coverage set to $(1 - \alpha) = 95\%$ for each of them.

5.2 Results

Figures 0.1 and 0.2 illustrate a typical realization. Figure 0.2 represents $\bar{Q}_{0,Y}$, $\bar{q}_{0,Y}$ and their estimators $\bar{Q}_{n,Y}$, $\bar{q}_{n,Y}$ at final sample size $n = 1000$. The top plot of Figure 0.1 shows the 95%-CI I_n in (0.9) at every sample size n where a CI is derived. By Corollary 0.2, the probability of the event "$\psi_{r_n,0} \in I_n$" is more likely to be close to 95% than the probability of the event "$\psi_0 \in I_n$" in the sense that the latter property requires that the rate of convergence of $\bar{q}_{n,Y}$ to $\bar{q}_{0,Y}$ be sufficiently fast. Nevertheless, we observe on this realization that each I_n contains both its corresponding data-adaptive parameter $\psi_{r_n,0}$ (pink cross) and ψ_0 (blue line). Moreover, the difference between the length of I_n and that of the vertical segment joining the two curves of the same nuance of darker gray at sample size n gets smaller as n grows. This indicates that the variance of ψ_n^* gets closer to the optimal variance $\mathrm{Var}_{P_{Q_0,r_0}} \Delta(Q_0, r_0)(O)$ as n grows.

The bottom plot of Figure 0.1 shows the actual value of $\mathcal{R}_n$ (green cross) and 95%-CI in (0.10) at every sample size n where a CI is derived. We observe on this realization that the regrets are all positive, a fact that was not granted. Moreover, each CI contains its corresponding data-adaptive parameter $\mathcal{R}_n$.

We can evaluate if our 95%-CIs achieve their nominal 95%-coverage. To do so, we carry out binomials tests. By construction, the empirical number of CIs which cover $\psi_{r_n,0}$ is a random variable drawn from a Binomial distribution with parameters (N, π). We choose to test the null "$\pi \geq 95\%$" against its one-sided alternative "$\pi < 95\%$". A large p-value is interpreted as the absence of empirical evidence supporting that the CI does not achieve its nominal coverage. We do the same for ψ_0 and $\mathcal{R}_n$, *mutatis mutandis*.

Instead of reporting $3 \times 33 = 99$ empirical proportions of coverage and related p-values, we simply plot the logarithms of the p-values of the tests evaluating the coverage of $\psi_{r_n,0}$ and ψ_0, see Figure 0.3. Overall, the orange curve dominates the green one, indicating that empirical coverage tends to be higher for $\psi_{r_n,0}$ (it ranges between 0.917 and 0.955 with an average of 0.940) than for ψ_0 (it ranges between 0.919 and 0.946 with an average of 0.937). This does not come as a surprise, as argued in the first paragraph of this section. Moreover, a majority of the p-values are larger than 5% (top grey horizontal line), and even more of them are larger than the Bonferroni-corrected threshold of 5/33%. Furthermore, the smallest p-values correspond to sample sizes $n = 200$ and $n = 225$, where inference is based on little information. As for the coverage of $\mathcal{R}_n$, it is far above the nominal 95%-coverage, ranging between 0.951 and 0.990 with an average of 0.997. This does not come as a surprise either since the CIs for $\mathcal{R}_n$ are conservative by construction.

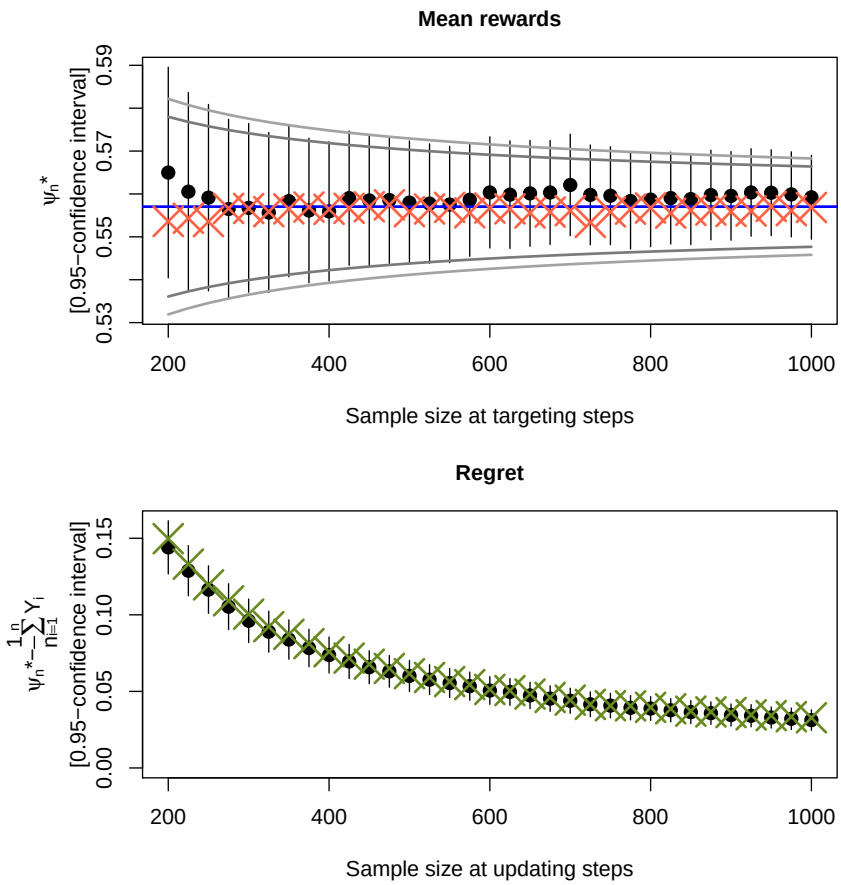

Fig. 0.1. Illustrating the data-adaptive inference of the optimal rule, its mean reward and the related regret (see also Figure 0.2). *Top plot.* The blue horizontal line represents the value of the mean reward under the optimal rule, ψ_0. The gray curves represent the mapping $n \mapsto \psi_0 \pm \xi_{97.5\%}\sigma_k/\sqrt{n}$ $(k = 1, 2)$, where $\sigma_1 \approx 0.1512$ is the square root of $\mathrm{Var}_{P_{Q_0,r_0}} \Delta(Q_0, r_0)(O)$ (darker gray) and $\sigma_2 \approx 0.1812$ is the square root of $\mathrm{Var}_{P_{Q_0,g^b}} \Delta(Q_0, g^b)(O)$ (lighter gray). Thus, at a given sample size n, the length of the vertical segment joining the two darker gray curves equals the length of a CI based on a regular, asymptotically efficient estimator of ψ_0. The pink crosses represent the successive values of the data-adaptive parameters $\psi_{r_n,0}$. The black dots represent the successive values of ψ_n^*, and the vertical segments centered at them represent the successive 95%-CIs for $\psi_{r_n,0}$ and, under additional assumptions, for ψ_0 as well. *Bottom plot.* The green crosses green represent the successive values of regret $\mathcal{R}_n$. The black dots represent the successive values of $\psi_n^* - n^{-1}\sum_{i=1}^n Y_i$, and the vertical segments represent the successive 95%-CIs for $\mathcal{R}_n$.

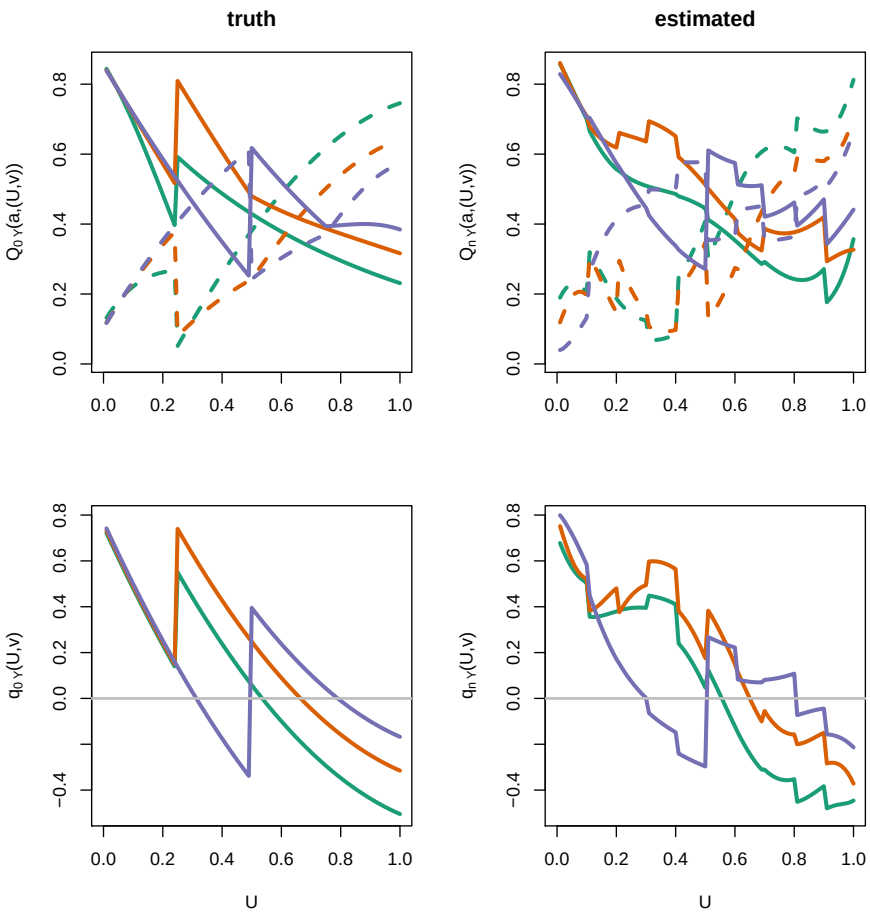

Fig. 0.2. Illustrating the data-adaptive inference of the optimal rule, its mean reward and the related regret through the representation of the conditional mean $\bar{Q}_{0,Y}$, blip function $\bar{q}_{0,Y}$ and their estimators (see also Figure 0.1). *Top left plot.* The solid curves represent $U \mapsto \bar{Q}_{0,Y}(1,(U,v))$ for $v = 1$ (in dark green, lowest value in 1), $v = 2$ (in dark orange, middle value in 1) and $v = 3$ (in dark blue, largest value in 1). The dashed curves represent $U \mapsto \bar{Q}_{0,Y}(0,(U,v))$ for $v = 1$ (in dark green, largest value in 1), $v = 2$ (in dark orange, middle value in 1) and $v = 3$ (in dark blue, smallest value in 1). *Bottom left plot.* The curves represent $U \mapsto \bar{q}_{0,Y}(U,v)$ for $v = 1$ (in dark green, smallest value in 1), $v = 2$ (in dark orange, middle value in 1) and $v = 3$ (in dark blue, largest value in 1). *Right plots.* Counterparts to the left plots, where $\bar{Q}_{0,Y}$ and $\bar{q}_{0,Y}$ are replaced with $\bar{Q}_{n,Y}$ and $\bar{q}_{n,Y}$ for $n = 1000$, the final sample size.

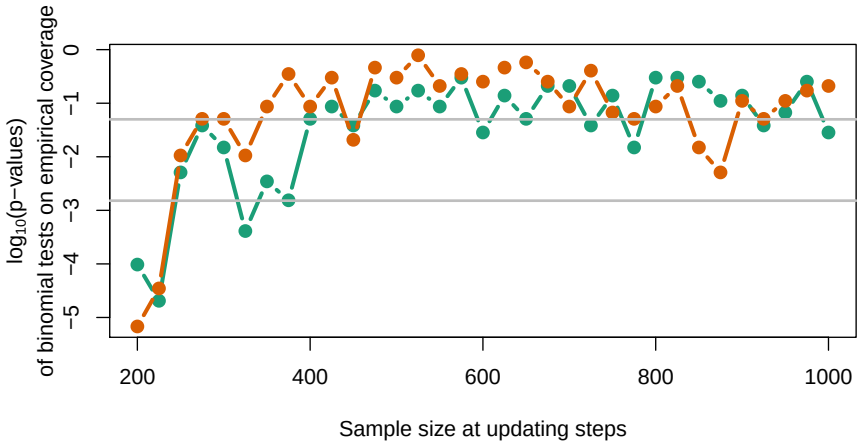

Fig. 0.3. Empirical evaluation of the coverage of the CIs. The curves represent the logarithms of p-values of binomial tests of adequate coverage (null) *vs.* inadequate coverage (alternative). A large p-value is interpreted as the absence of empirical evidence supporting that the related CI does not achieve its nominal coverage of 95%. The dark green curve corresponds with CIs for $\psi_{r_n,0}$, and the dark orange with CIs for ψ_0. The gray curves show the threshold of 5% (top) and the Bonferonni-corrected threshold of 5/33% (bottom).

6 Conclusion (on a twist)

We acknowledged that assuming the equality $\bar{q}_{1,Y} = \bar{q}_{0,Y}$ in **A1** is a stringent condition. It happens that the equality is mandatory only in the context of Corollary 0.1, which provides sufficient conditions for the TMLE to estimate ψ_0, the mean reward under r_0. Yet we argued that we are more interested in the data-adaptive parameter $\psi_{r_n,0}$, the mean reward under r_n, than in ψ_0. What can be said then without assuming $\bar{q}_{1,Y} = \bar{q}_{0,Y}$?

Let **A1*** be assumption **A1** deprived of its condition $\bar{q}_{1,Y} = \bar{q}_{0,Y}$. In light of (0.1) and (0.5), let rule r_1 and its approximation g_1 be given by $r_1(W) \equiv \mathbf{1}\{\bar{q}_{1,Y}(W) \geq 0\}$ and $g_1(1|W) = 1 - g_1(0|W) \equiv G(\bar{q}_{1,Y}(W))$. Introduce

$$\psi_1 \equiv E_{Q_0}\left(Y(r_1(W))\right),$$

the mean reward under rule r_1. Now, let **A2*** be assumption **A2** with $\epsilon \mapsto E_{P_{Q_0,g_1}}\left(\ell(\bar{Q}'_{1,Y,\rho}(\epsilon))(O)\right)$ substituted for $\epsilon \mapsto E_{P_{Q_0,g_0}}\left(\ell(\bar{Q}_{1,Y,\rho}(\epsilon))(O)\right)$, where $\bar{Q}'_{1,Y,\rho}(\epsilon)$ is defined as in (0.6) using g_1 in lieu of g_0. Introduce $\bar{Q}'^{*}_{1,Y,r_1} \equiv \bar{Q}'_{1,Y,r_1}(\epsilon_0(r_1))$ and, in light of (0.8), $D_1'^{*}$ given by

$$D_1'^{*}(O) \equiv (Y - \bar{Q}'^{*}_{1,Y}(A,W))\frac{\mathbf{1}\{A = r_1(W)\}}{g_1(A|W)} + \bar{Q}'^{*}_{1,Y}(r_1(W),W) - \psi_1,$$

then $\Sigma_1^2 \equiv E_{P_{Q_0,g_1}}(D_1'^{*}(O)^2)$. Finally, consider the following counterpart to **A3**:

A3*. The conditional distributions of $Y(0)$ and $Y(1)$ given W under Q_0 is not degenerated. Moreover, there exist $\gamma_1, \gamma_2 > 0$ such that, for all $t \geq 0$,

$$P_{Q_0}\left(0 \leq |\bar{q}_{1,Y}(W)| \leq t\right) \leq \gamma_1 t^{\gamma_2}. \tag{0.11}$$

In addition, the ratio $|\bar{q}_{0,Y}/\bar{q}_{1,Y}|$ can be defined and its (essential) supremum is finite.

The margin condition in **A3*** now concerns the limit blip function $\bar{q}_{1,Y}$. The true blip function $\bar{q}_{0,Y}$ needs not take positive values Q_0-almost surely anymore. As for the constraint on the ratio $|\bar{q}_{0,Y}/\bar{q}_{1,Y}|$ (which is obviously met when $\bar{q}_{1,Y} = \bar{q}_{0,Y}$), we could simply enforce it by choosing $\bar{\mathcal{Q}}_Y$ in such a way that $|\bar{q}_Y| \geq \delta > 0$ for all $\bar{Q}_Y \in \bar{\mathcal{Q}}_Y$. We may now state the final result of this manuscript.

Proposition 0.4. *Under* **A1***, **A2*** *and* **A3***, *it holds that* $\|\bar{Q}_{n,Y} - \bar{Q}_{1,Y}\|_{Q_0}$, $\|\bar{q}_{n,Y} - \bar{q}_{1,Y}\|_{Q_0}$, $\|r_n - r_1\|_{Q_0}$, $\|g_n - g_1\|_{Q_0}$ *and the data-adaptive parameter* $\psi_1 - \psi_{r_n,0}$ *all converge in probability to zero as* n *tends to infinity. Furthermore,* ψ_n^* *consistently estimates* $\psi_{r_n,0}$ *hence* ψ_1 *as well. It does so in such a way that* $\sqrt{n/\sigma_n^2}(\psi_n^* - \psi_{r_n,0})$ *converges in law to the standard normal distribution as* n *tends to infinity, where* σ_n^2 *consistently estimates the positive* Σ_1^2.

Therefore, under the assumptions of Proposition 0.4, the CI defined in (0.9) still contains $\psi_{r_n,0}$ with probability tending to $(1-\alpha)$ as n tends to infinity. The most important result of the manuscript is thus preserved without assuming that the limit blip function and the true one coincide.

References

L. B. Balzer, M. L. Petersen, and M. J. van der Laan. Targeted estimation and inference for the sample average treatment effect. Technical Report 334, U.C. Berkeley Division of Biostatistics Working Paper Series, 2015. URL http://biostats.bepress.com/ucbbiostat/paper334.

B. Chakraborty and E. E. M. Moodie. *Statistical methods for dynamic treatment regimes*. Statistics for Biology and Health. Springer, New York, 2013. doi: 10.1007/978-1-4614-7428-9. URL http://dx.doi.org/10.1007/978-1-4614-7428-9. Reinforcement learning, causal inference, and personalized medicine.

B. Chakraborty, E. B. Laber, and Y-Q. Zhao. Inference about the expected performance of a data-driven dynamic treatment regime. *Clin. Trials*, 11(4):408–417, 2014.

A. Chambaz and M. J. van der Laan. Inference in targeted group-sequential covariate-adjusted randomized clinical trials. *Scand. J. Stat.*, 41(1):104–140, 2014. doi: 10.1111/sjos.12013. URL http://dx.doi.org/10.1111/sjos.12013.

A. Chambaz, M. J. van der Laan, and W. Zheng. Targeted covariate-adjusted response-adaptive lasso-based randomized controlled trials. In A. Sverdlov, editor, *Modern Adaptive Randomized Clinical Trials: Statistical, Operational, and Regulatory Aspects*, pages 345–368. CRC Press, 2015.

A. Chambaz, W. Zheng, and M. J. van der Laan. Data-adaptive inference of the optimal treatment rule and its mean reward. The masked bandit. Technical report, 2016. URL https://hal.archives-ouvertes.fr/hal-01301297.

V. H. de la Peña and E. Giné. *Decoupling*. Probability and its Applications (New York). Springer-Verlag, New York, 1999. doi: 10.1007/978-1-4612-0537-1. URL http://dx.doi.org/10.1007/978-1-4612-0537-1. From dependence to independence, Randomly stopped processes. U-statistics and processes. Martingales and beyond.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL http://www.jstatsoft.org/v33/i01/.

Y. Goldberg, R. Song, D. Zeng, and M. R. Kosorok. Comment on "Dynamic treatment regimes: Technical challenges and applications". *Electron. J. Stat.*, 8:1290–1300, 2014.

E. B. Laber, D. J. Lizotte, M. Qian, W. E. Pelham, and S. A. Murphy. Dynamic treatment regimes: Technical challenges and applications. *Electron. J. Stat.*, 8(1):1225–1272, 2014a.

E. B. Laber, D. J. Lizotte, M. Qian, W. E. Pelham, and S. A. Murphy. Rejoinder of "Dynamic treatment regimes: Technical challenges and applications". *Electron. J. Stat.*, 8(1):1312–1321, 2014b.

A. R. Luedtke and M. J. van der Laan. Targeted learning of the mean outcome under an optimal dynamic treatment rule. *Journal of Causal Inference*, 3(1):61–95, 2015a.

A. R. Luedtke and M. J. van der Laan. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Ann. Statist.*, 2015b. To appear.

A. R. Luedtke and M. J. van der Laan. Super-learning of an optimal dynamic treatment rule. *International Journal of Biostatistics*, 2016. To appear.

E. Mammen and A. B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27(6):1808–1829, 1999. doi: 10.1214/aos/1017939240. URL http://dx.doi.org/10.1214/aos/1017939240.

J. Pearl. *Causality: Models, Reasoning and Inference*, volume 29. Cambridge University Press, Cambridge, 2000.

M. Qian and S. A. Murphy. Performance guarantees for individualized treatment rules. *Ann. Statist.*, 39(2): 1180–1210, 2011. doi: 10.1214/10-AOS864. URL http://dx.doi.org/10.1214/10-AOS864.

J. M. Robins. Optimal structural nested models for optimal sequential decisions. In D. Y. Lin and P. Heagerty, editors, *Proc. Second Seattle Symp. Biostat.*, pages 189–326, 2004.

D. B. Rubin and M. J. van der Laan. Statistical issues and limitations in personalized medicine research with clinical trials. *Int. J. Biostat.*, 8(1), 2012. Article 1.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence*. Springer, 1996.

B. Zhang, A. Tsiatis, M. Davidian, M. Zhang, and E. Laber. A robust method for estimating optimal treatment regimes. *Biometrics*, 68:1010–1018, 2012a.

B. Zhang, A. Tsiatis, M. Davidian, M. Zhang, and E. Laber. Estimating optimal treatment regimes from a classification perspective. *Stat*, 68(1):103–114, 2012b.

Y. Zhao, D. Zeng, A. J. Rush, and M. R. Kosorok. Estimating individualized treatment rules using outcome weighted learning. *J. Amer. Statist. Assoc.*, 107(499):1106–1118, 2012. doi: 10.1080/01621459.2012.695674. URL http://dx.doi.org/10.1080/01621459.2012.695674.

Y. Zhao, D. Zeng, E. B. Laber, and M. R. Kosorok. New statistical learning methods for estimating optimal dynamic treatment regimes. *J. Amer. Statist. Assoc.*, 110(510):583–598, 2015. doi: 10.1080/01621459.2014.937488. URL http://dx.doi.org/10.1080/01621459.2014.937488.

W. Zheng, A. Chambaz, and M. J. van der Laan. Drawing valid targeted inference when covariate-adjusted response-adaptive rct meets data-adaptive loss-based estimation, with an application to the LASSO. Technical Report 339, U.C. Berkeley Division of Biostatistics Working Paper Series, 2015. URL http://biostats.bepress.com/ucbbiostat/paper339.