1 Introduction

1.1 Objective

We consider an experiment where a player repeatedly chooses and carries out one among two actions to receive a random reward. Each time, the reward depends on both the action undertaken and a random context preceding it, which is given to the player before she makes her choice. The law of the context and conditional law of the reward given the context and action are fixed throughout the experiment. The player's objective is to obtain as large a cumulated sum of rewards as possible.

In this framework, a policy is a rule that maps any context to an action. The value of a policy is the expectation of the reward in the experiment where the action carried out is the action recommended by the policy. Given a class Π of candidate policies, the regret of a policy $\pi \in \Pi$ is the difference between the largest value achievable within Π and the value of π.

Learning the optimal policy within a class of candidate policies is meaningful whenever the goal is to make recommendations. This is, for instance, the case in personalized medicine, also known as precision medicine. There, the context would typically consist of the description of a patient, the actions would correspond to two strategies of treatment, and the policies are rather called individualized treatment rules.

The objective of this article is not to establish optimal regret bounds for optimal policy estimators. It is, rather, to show that rates faster than $n^{-1/2}$ can be demonstrated under much more general conditions than have previously been discussed in the policy learning literature.

1.2 A Brief Literature Review

There has been a surge of interest in developing flexible methods for estimating optimal policies in recent years. Here we give a deeply abbreviated overview, and refer the reader