# Tools and Techniques for Machine Learning
# Homework 3

**Instructions**: Your answers to the questions below, including plots and mathematical work, should be submitted as a single PDF file. It's preferred that you write your answers using software that typesets mathematics (e.g. LaTeX, LyX, or Jupyter), though if you need to you may scan handwritten work. For submission, you can also export your Jupyter notebook and merge that PDF with your PDF for the written solutions into one file. **Don't forget to complete the Jupyter notebook as well, for the programming part of this assignment**.

## 1 Derivation of importance-weighted reward imputation

Suppose we have a contextual bandit where context $X \in \mathcal{X}$ has probability density function $p(x)$ and reward vector $R \in \mathbb{R}^k$ has conditional distribution given by $P_{R|X}$. We want to use the direct method to evaluate the performance of a static policy $\pi$. That is, we want to use

$$\begin{aligned}\hat{V}_{\mathrm{dm}}(\pi) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{a=1}^{k}\hat{r}(X_i,a)\pi(a \mid X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{A_i\sim\pi(\cdot|X_i)}\left[\hat{r}(X_i,A_i)\right],\end{aligned}$$

where $\hat{r}(x,a)$ is some estimate for $\mathbb{E}\left[R(A) \mid X=x, A=a\right] = \mathbb{E}\left[R(a) \mid X=x\right]$ and

$$\left(X_1, A_1, R_1(A_1)\right), \ldots, \left(X_n, A_n, R_n(A_n)\right)$$

is the logged bandit feedback from static policy $\pi_0$ on the same contextual bandit distribution. The basic approach to fitting $\hat{r}$ from some hypothesis space $\mathcal{H}$ is least squares:

$$\hat{r} = \operatorname*{arg\,min}_{r\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\left(r(X_i,A_i) - R_i(A_i)\right)^2.$$

1. With this approach, what is the covariate distribution in training? Explain why we have a covariate shift between the train and target distribution.

2. Give an importance-weighted objective function $J(r)$ for finding $\hat{r}$, and use the change of measure theorem to show that $\mathbb{E}\left[J(r)\right] = \mathbb{E}\left[r(X,A) - R(A)\right]^2$, where $X \sim p(x)$, $R \mid X \sim P_{R|X}$ and $A \mid X \sim \pi(a \mid x)$. In other words, the objective function is an unbiased estimate of the expected square loss (i.e. the risk) of $r$ w.r.t. the target distribution.