1. **Causal Inference 1:** Suppose we are in a randomized trial, i.e. we have $n$ individuals sampled from some distribution and of these $n$ individuals $n/2$, (assume $n$ is even for simplicity) are randomly assigned to receive treatment and the remaining $n/2$ receive control. Assume that the outcomes are binary, and for each unit denote the observed outcomes as $Y_{i\text{obs}}$. Denote the units receiving treatment by $T$ and those receiving control by $C$. We want to estimate the average treatment effect:

$$
\tau = \mathbb{E}[Y(1) - Y(0)],
$$

and we use the natural estimator:

$$
\hat{\tau}_1 = \frac{2}{n} \sum_{i \in T} Y_{i\text{obs}} - \frac{2}{n} \sum_{i \in C} Y_{i\text{obs}}.
$$

Use Hoeffding’s inequality to give a bound on $|\hat{\tau}_1 - \tau|$ that holds with probability at least $1 - \delta$. Provide a reason as to why Hoeffding is a reasonable tool here.

2. **Causal Inference 2:** Now we suppose that we are in an observational set-up, but we have no unmeasured confounding, i.e. for some covariates $X$ we have that $W \perp \perp (Y(1), Y(0))|X$. We assume that in this setting we observe $n$ i.i.d. samples of the form $(X_i, W_i, Y_{i\text{obs}})$.

To begin with let us suppose that the propensity scores $\mathbb{P}(W = 1|X)$ are known, i.e. we know the probability of receiving treatment for each unit in the population. This is not an unreasonable assumption in settings where say a doctor prescribes treatment say after conducting some tests (i.e. measuring some covariates).

To simplify things we will assume that there are constants $\pi_{\text{min}} > 0$ and $\pi_{\text{max}} < 1$ such that,

$$
\pi_{\text{min}} \leq \mathbb{P}(W = 1|X) \leq \pi_{\text{max}} \quad \forall \ x.
$$

A natural estimator then is the Horvitz-Thompson estimator:

$$
\hat{\tau}_2 = \frac{1}{n} \sum_{i=1}^{n} \left[ \frac{Y_{i\text{obs}} W_i}{\pi(X_i)} - \frac{Y_{i\text{obs}} (1 - W_i)}{1 - \pi(X_i)} \right].
$$

Use Hoeffding’s inequality to give a bound on $|\hat{\tau}_2 - \tau|$ that holds with probability at least $1 - \delta$. You need to reason about why we needed the $\pi_{\text{min}}, \pi_{\text{max}}$ assumption, as well as why Hoeffding is a reasonable tool here.
3. Causal Inference 3: Now we turn to the general case when the propensity scores are not known. For the remaining questions, to simplify things we are going to assume that we are given the following things:

(a) An estimator \( \hat{\mu}_1 \) of the conditional expectation \( \mathbb{E}[Y_{obs}|W = 1, X] \), which satisfies:

\[
\sup_x |\mu_1(x) - \hat{\mu}_1(x)| \leq C \left( \frac{\log n}{n} \right)^{\beta/(2\beta + d)},
\]

where \( d \) is the dimensionality of the covariates, \( C \) is some positive constant and \( \beta \) is the smoothness of the function \( \mu_1 \). This is similar to the bounds we have discussed for non-parametric regression, but requiring that the estimate is close everywhere (as opposed to just on average) costs a log-factor.

(b) An analogous estimator \( \hat{\mu}_0 \) of the conditional expectation \( \mathbb{E}[Y_{obs}|W = 0, X] \), which satisfies:

\[
\sup_x |\mu_0(x) - \hat{\mu}_0(x)| \leq C \left( \frac{\log n}{n} \right)^{\beta/(2\beta + d)}.
\]

For simplicity we assume that \( \mu_1 \) and \( \mu_0 \) share the same smoothness parameter. You can also assume that your estimates are bounded, i.e.

\[
\max\{\hat{\mu}_1(x), \hat{\mu}_0(x)\} \leq \mu_{\max} \quad \forall \ x.
\]

(c) An estimator \( \hat{\pi} \) of the propensity score \( \mathbb{P}(W = 1|X) \), which satisfies:

\[
\sup_x |\pi(x) - \hat{\pi}(x)| \leq C \left( \frac{\log n}{n} \right)^{\gamma/(2\gamma + d)},
\]

where \( \gamma \) denotes the smoothness of the propensity score. You can also assume that the estimator is bounded from above and below, i.e.

\[
0 < \hat{\pi}_{\min} \leq \hat{\pi}(x) \leq \hat{\pi}_{\max} < 1 \quad \forall \ x.
\]

In practice you would obtain these estimators by regression, and the bounds would hold with high-probability but we will assume they hold deterministically to simplify things. You can also assume that these estimators were constructed on a separate sample, i.e. you can treat them as fixed functions that do not depend on the sample you are examining.

First, we consider the regression-based estimator we proposed in class:

\[
\hat{\tau}_3 = \frac{1}{n} \sum_{i=1}^{n} (\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)).
\]
Provide a bound on $|\hat{\tau}_3 - \tau|$ that holds with probability at least $1 - \delta$, again using Hoeffding’s inequality and the assumptions given above. Observe that the rate of convergence is now slower than in either a trial or when the propensity scores are known.

**Hint:** Try to consider the intermediate quantity:

$$\bar{\tau} = \frac{1}{n} \sum_{i=1}^{n} (\mu_1(X_i) - \mu_0(X_i)).$$

4. **Causal Inference 4:** Assume the same set-up (and everything else as before). Now, we consider the Horvitz-Thompson estimator with estimated propensity scores, i.e.

$$\hat{\tau}_4 = \frac{1}{n} \sum_{i=1}^{n} \left[ \frac{Y_{i,\text{obs}} W_i}{\hat{\pi}(X_i)} - \frac{Y_{i,\text{obs}} (1 - W_i)}{1 - \hat{\pi}(X_i)} \right].$$

Provide a bound on $|\hat{\tau}_4 - \tau|$ that holds with probability at least $1 - \delta$, again using Hoeffding’s inequality and the assumptions given above. Compare the rate of convergence to that in the previous question.

5. **Causal Inference 5:** In the previous two questions we have seen roughly that if the propensity score is easy to estimate, i.e. $\gamma > \beta$ then we should use the Horvitz-Thompson type estimator and if not we should use the regression estimator. Unfortunately, in practice we do not know which function is easier to estimate. However, there is a somewhat miraculous estimator that combines the benefits of these two estimators, and improves on both of them. This is called the doubly robust estimator.

Here is the estimator:

$$\hat{\tau}_5 = \frac{1}{n} \sum_{i=1}^{n} \left[ \frac{\hat{\mu}_1(X_i) + W_i}{\hat{\pi}(X_i)} (Y_{i,\text{obs}} - \hat{\mu}_1(X_i)) - \frac{1 - W_i}{1 - \hat{\pi}(X_i)} (Y_{i,\text{obs}} - \hat{\mu}_0(X_i)) \right].$$

You can view this estimator roughly as correcting the regression estimator using propensity weighting on the residual (i.e. the difference between the observed and predicted regression value). In order to study this estimator, one way to proceed is to
show that it can be re-written as:

\[
\hat{\tau}_5 = \hat{\tau}_2 + \frac{1}{n} \sum_{i=1}^{n} \left[ \hat{\mu}_1(X_i) - \frac{W_i}{\pi(X_i)} \hat{\mu}_1(X_i) \right] - \frac{1}{n} \sum_{i=1}^{n} \left[ \hat{\mu}_0(X_i) - \frac{1 - W_i}{1 - \pi(X_i)} \hat{\mu}_0(X_i) \right]
\]

\[
+ \frac{1}{n} \sum_{i=1}^{n} W_i (Y^\text{obs}_i - \mu_1(X_i)) \left[ \frac{1}{\hat{\pi}(X_i)} - \frac{1}{\pi(X_i)} \right]
\]

\[
- \frac{1}{n} \sum_{i=1}^{n} (1 - W_i) (Y^\text{obs}_i - \mu_0(X_i)) \left[ \frac{1}{1 - \hat{\pi}(X_i)} - \frac{1}{1 - \pi(X_i)} \right]
\]

\[
+ \frac{1}{n} \sum_{i=1}^{n} W_i (\mu_1(X_i) - \hat{\mu}_1(X_i)) \left[ \frac{1}{\hat{\pi}(X_i)} - \frac{1}{\pi(X_i)} \right]
\]

\[
- \frac{1}{n} \sum_{i=1}^{n} (1 - W_i) (\mu_0(X_i) - \hat{\mu}_0(X_i)) \left[ \frac{1}{1 - \hat{\pi}(X_i)} - \frac{1}{1 - \pi(X_i)} \right].
\]

Provide a bound on \(|\hat{\tau}_5 - \tau|\) that holds with probability at least 1 - \(\delta\), again using Hoeffding’s inequality and the assumptions given above. In particular, you should use Hoeffding on the terms in the first, second and third lines and the assumptions for the other two. You have already dealt with \(\hat{\tau}_2\), for the other terms, remember that the functions \(\hat{\mu}_1, \hat{\mu}_0\) do not depend on the data you are analyzing (i.e. they were fit on a separate sample). Compare the rate of convergence to those of the previous two questions.

Observe the important double robustness property of the estimator you just constructed: the estimator is consistent if either your estimate for the regression functions or of the propensity score is consistent (i.e. either \(\beta > 0\) or \(\gamma > 0\)). Furthermore, when both estimators are consistent (i.e. both \(\beta > 0\) and \(\gamma > 0\)) your estimator dominates both \(\hat{\tau}_3\) and \(\hat{\tau}_4\).

While it might seem to you that the doubly robust estimator came out of nowhere, it in fact comes from a general method of trying to improve plug-in estimators (like the regression or Horvitz-Thompson estimator) using what are known are influence functions. The course in causal inference will likely be a good introduction to this idea.