Causal Estimation with Sample Reweighting under Covariate Shift

(And also some notes about distribution shifts and density ratios)

Daniel Csillag

2024-03-21

1. Introduction and Setup

We are working under the Potential Outcomes framework. That means we have the following setup:

$$\left(X_i, T_i, Y_i^{\text{obs}}, Y_i^{(1)}, Y_i^{(0)}\right)_{i=1}^{n} \sim \mathcal{P}, \tag{1}$$

where X_i are the covaraites/variables/features we are considering, $T_i \in \{0, 1\}$ is whether the i-th individual has been exposed to the treatment, Y_i^{obs} is the observed outcome for that individual, and $Y^{(1)}, Y^{(0)}$ are the potential outcomes. We will also operate under the assumptions of SUTVA and (strong) ignorability:

Assumption 1 (Consistency) . *It holds that*

$$Y^{obs} = \begin{cases} Y^{(1)} \ if \ T = 1 \\ Y^{(0)} \ if \ T = 0 \end{cases}.$$

Assumption 2 (No interference) . *There is no interference among our samples, i.e., they are i.i.d.. Under this assumptions, we can refer to random variables $X, T, Y^{obs}, Y^{(1)}, Y^{(0)}$, of which the indexed versions presented in Equation 1 are realizations thereof – something we will make great use of throughout this document.*

Assumption 3 (SUTVA) . *Assumption 1 (consistency) and Assumption 2 (no interference) both hold.*

Assumption 4 (Strong Ignorability) . *It holds that $\left(Y^{(1)}, Y^{(0)}\right) \perp\!\!\!\perp T \mid X$.*

It's worth noting that a number of the results we discuss here also hold only under weak ignorability:

Assumption 5 (Weak Ignorability). *For all $a \in \{0, 1\}$, it holds that $Y^{(a)} \perp\!\!\!\perp T \mid X$.*

Nevertheless, for presentation purposes we will stick to strong ignorability (which is also the more usual assumption in practice). It should be reasonably evident when weak ignorability suffices.

Finally, later on in this document we will introduce an additional *positivity* assumption:

Assumption 6 (Positivity) . *For every possible x and for all treatments a,*

$$0 < \mathbb{P}[T = a \mid X = x] < 1. \qquad (note\ the\ strict\ inequalities!)$$

Intuitively, this asserts that no possible set of covariates force make it so that one particular treatment is assigned almost surely.

1.1. Equivalence between ignorability and covariate shift

1.1.1. Distribution shifts

Distribution shifts are a "classical" topic in the Machine Learning literature. The usual motivation is that there is typically a mismatch between the distribution of the training data (i.e., the distribution of the data for which you were able to acquire target labels for) and the distribution of the data that you

will actually receive when deploying your model. Moreover, this "real-world" distribution may (and often does) change over time, adding another layer of complexity.

Now, unfortunately, simple stating that the test distribution(s) differs from the training one with no further information is not enough to argue about how well your model (or your inference) will perform once there is this shift in distribution (from the training distribution to the test distribution). To amend this, we consider some few particular cases of distribution shifts, which are tractable.

Let the training and test distributions be $P^{\text{train}}_{X,Y}$ and $P^{\text{test}}_{X,Y}$, respectively (both distributions over random variables X [covariates/variables/features] and Y [outcomes/labels/targets]). It is worth noting that regardless of any additional assumptions, we can factor them as products/compositions of a conditional distribution with a marginal one:

$$\begin{aligned} P^{\text{train}}_{X,Y} &= P^{\text{train}}_{Y|X} \cdot P^{\text{train}}_{X} & \qquad P^{\text{test}}_{X,Y} &= P^{\text{test}}_{Y|X} \cdot P^{\text{test}}_{X} \\ &= P^{\text{train}}_{X|Y} \cdot P^{\text{train}}_{Y} & \qquad &= P^{\text{test}}_{X|Y} \cdot P^{\text{test}}_{Y} \end{aligned}$$

With this factorization in mind we can present the more usual forms of distribution shift. They are:

1. **Covariate Shift:** $P^{\text{train}}_{Y|X} = P^{\text{test}}_{Y|X}$ (the marginals over X may vary arbitrarily). For the purposes of this document, this is the most important.

   Intuitively, this assumption says that the relation between the inputs X and the outputs Y has not changed. There are many situations where you can assume this to be the case. A nice example is if you are making a model that receives as inputs X sattelite images of the Amazon rainforest and produces segmentations of forestation in these images Y. The relation between the inputs and outputs is expected to remain constant – our view of what is or not a forest in an image shouldn't change[1].

2. **Concept Shift:** $P^{\text{train}}_{X} = P^{\text{test}}_{X}$.

3. **Label Shift:** $P^{\text{train}}_{X|Y} = P^{\text{test}}_{X|Y}$ (the marginals over Y may vary arbitrarily).

   This differs the most from the Covariate Shift assumption because the space of labels ($\mathcal{Y}$) is oftentimes much simpler than the space of covariates ($\mathcal{X}$). In particular, it is often discrete (e.g., in classification problems)! This allows for alternate arguments that are generally not applicable to the Covariate Shift, and moreover allows for much more tractable analyses than what can be done under general Concept Shift, while still being quite applicable.

Once we assume one of these forms of distribution shift, we can now do inference and analyses over the test/training distributions.

1.1.2. The equivalence

As it turns out, assuming ignorability (Assumption 4) (as well as SUTVA) is equivalent to assuming that the distributions $P_{Y^{(a)},X}$ and $P_{Y^{\text{obs}},X|T=a}$ differ by a covariate shift.

Proposition 1. *Assume SUTVA (Assumption 3, i.e., consistency [Assumption 1] and no interference [Assumption 2]). Then $Y^{(a)} \perp\!\!\!\perp T \mid X$ if and only if $P_{Y^{(a)}|X} = P_{Y^{obs}|X,T=a}$.*

Proof. ($\Longrightarrow$) If ignorability and consistency hold, then

[1] Unless there are particularly tricky images for which there is some improved technical understanding that changes the judgements we make. But if that happens then a domain expert or the team implementing the system is bound to know, and it may be easier to just retrain the model.

$$\begin{aligned} P_{Y^{\text{obs}}|X,T=a} &= P_{Y^{(a)}|X,T=a} && \text{(by consistency)} \\ &= P_{Y^{(a)}|X}. && \text{(by ignorability)} \end{aligned}$$

($\Longleftarrow$) We have that

$$\begin{aligned} P_{Y^{(a)}|X} &= P_{Y^{\text{obs}}|X,T=a} && \text{(by hypothesis)} \\ &= P_{Y^{(a)}|X,T=a}. && \text{(by consistency)} \end{aligned}$$

$P_{Y^{(a)}|X} = P_{Y^{(a)}|X,T=a}$ means that, conditional on X, $Y^{(a)}$ and X are independent: $Y^{(a)} \perp\!\!\!\perp T \mid X$. ■

1.2. Sample reweighting

The idea of sample reweighting is common in statistics. It is a core tenet of techniques such as importance sampling and importance reweighting. Indeed, the sample reweighting ideas here presented are occasionally called "sample reweighting à la importance sampling" or similar.

The main idea revolves around the following lemma:

Lemma 1 . *Consider two distributions P and Q (with densities $\mathrm{d}P$ and $\mathrm{d}Q$) over the same space $\mathcal{U}$, and any $\phi : \mathcal{U} \to \mathbb{R}$. Then*

$$\mathbb{E}_{W_P \sim P}[\phi(W_P)] = \mathbb{E}_{W_Q \sim Q}\left[\frac{\mathrm{d}P(W_Q)}{\mathrm{d}Q(W_Q)} \cdot \phi(W_Q)\right].$$

Proof (informal!). Write out the expectations as integrals:

$$\begin{aligned} \mathbb{E}_{W_Q \sim Q}\left[\frac{\mathrm{d}P(W_Q)}{\mathrm{d}Q(W_Q)} \cdot \phi(W_Q)\right] &= \int \frac{\mathrm{d}P(w)}{\mathrm{d}Q(w)} \cdot \phi(w) \cdot \mathrm{d}Q(w) \;\; \mathrm{d}w \\ &= \int \phi(w) \cdot \mathrm{d}P(w) \;\; \mathrm{d}w = \mathbb{E}_{W_P \sim P}[\phi(W_P)]. \end{aligned}$$

■

Remark 1. *This lemma (and proof) can be fully formalized, in all its generality, by using some measure theory. Irrespective of having the precise densities $\mathrm{d}P$ and $\mathrm{d}Q$, we just have that $\mathrm{d}P/\,\mathrm{d}Q$ is the Radon-Nikodym derivative of the measure P in relation to the measure Q. We can then use the Radon-Nikodym theorem to prove this fact.*

This means that we can use data from a distribution we know (Q) reweighted by the density ratio $\mathrm{d}P/\,\mathrm{d}Q$ in order to infer things about hte distribution we don't know (P). But of course, this assumes that we know the true density ratio $\mathrm{d}P/\,\mathrm{d}Q$, which is easier said than done.

1.3. Estimation of Density Ratios

A naive procedure to estimate density ratios would be to first individually estimate both densities (e.g., by Kernel Density Estimation) and then divide the two estimated densities. However, we can do *much* better. As it turns out, density ratio estimation is much more tractable than density estimation!

Entire books have been written on this topic; in particular, (Sugiyama et al., 2012) is a great reference. In this section, we will only present one particularly neat and practical way of doing this estimation, based on probabilistic classification. But we emphasize that there are other more sophisticated (and possibly more accurate) methods available.

Consider we have two distributions, P and Q, and we have samples $W_1^{(P)}, \ldots, W_{n_P}^{(P)}$ and $W_1^{(Q)}, \ldots, W_{n_Q}^{(Q)}$ from these distributions.

To do our estimation, let's consider the following game: first, someone flips a coin. If it's heads, then they give us a sample from P. If it's tails, they give us a sample from Q. We only receive the sample, without knowing whether it came from P or from Q. Our goal is to guess that.

Let's write by $D \in \{P, Q\}$ (a random variable) the distribution from which the other player took the sample, and write the sample as $W^{(D)}$. Note how the following reduces to the density ratio between P and Q:

$$\frac{\mathbb{P}\left[W^{(D)} = w \mid D = P\right]}{\mathbb{P}\left[W^{(D)} = w \mid D = Q\right]} = \frac{\mathrm{d}P(w)}{\mathrm{d}Q(w)}.$$

Now, by Bayes' rule:

$$\begin{aligned}\frac{\mathbb{P}\left[W^{(D)} = w \mid D = P\right]}{\mathbb{P}\left[W^{(D)} = w \mid D = Q\right]} &= \frac{\mathbb{P}\left[D = P \mid W^{(D)} = w\right] \cdot \mathbb{P}\left[W^{(D)} = w\right] / \mathbb{P}[D = P]}{\mathbb{P}\left[D = Q \mid W^{(D)} = w\right] \cdot \mathbb{P}\left[W^{(D)} = w\right] / \mathbb{P}[D = Q]} \\ &= \frac{\mathbb{P}[D = Q]}{\mathbb{P}[D = P]} \cdot \frac{\mathbb{P}\left[D = P \mid W^{(D)} = w\right]}{\mathbb{P}\left[D = Q \mid W^{(D)} = w\right]}.\end{aligned}$$

Moreover, writing $p(w) = \mathbb{P}\left[D = P \mid W^{(D)} = w\right]$,

$$\frac{\mathbb{P}[D = Q]}{\mathbb{P}[D = P]} \cdot \frac{\mathbb{P}\left[D = P \mid W^{(D)} = w\right]}{\mathbb{P}\left[D = Q \mid W^{(D)} = w\right]} = \frac{\mathbb{P}[D = Q]}{\mathbb{P}[D = P]} \cdot \frac{p(w)}{1 - p(w)}.$$

The nice thing is that we can estimate p with "plain" probabilistic classification! All we need to do is to train p on a dataset formed by both our samples from P and the ones from Q (concatenated together), and labelled 1 if the sample was from P and 0 if it was from Q. But it's important to note that our predictions should be well-calibrated – which is especially important if we are using Machine Learning or Bayesian models to do this classification. To do so, one can refer to methods such Platt Scaling, Isotonic Regression and Histogram Binning, as well as more recent methods such as Venn-ABERS predictors.

And more, it will hold that

$$\frac{\mathbb{P}[D = Q]}{\mathbb{P}[D = P]} \approx \frac{n_Q / (n_Q + n_P)}{n_P / (n_Q + n_P)} = \frac{n_Q}{n_P}.$$

Alternatively, we can approximate

$$\mathbb{E}_{W \sim Q}\left[\frac{p(W)}{1 - p(W)}\right] \approx \mathbb{E}_{W \sim Q}\left[C \cdot \frac{\mathrm{d}P(W)}{\mathrm{d}Q(W)}\right] = \mathbb{E}_{W \sim P}[C] = C \approx \frac{\mathbb{P}[D = Q]}{\mathbb{P}[D = P]}.$$

It should be noted, though, that oftentimes when density ratios are involved we only need to know it up to a proportional constant. An example of this is Section 3. In this case, for convenience (and more accurate estimation), we would ignore the multiplication by $\mathbb{P}[D = Q] / \mathbb{P}[D = P]$, as that is merely constant (since it does not involve w).

2. Means of potential outcomes via sample reweighting

Consider the problem of inferring $\mathbb{E}\left[Y^{(a)}\right]$. Our goal here is to use Lemma 1 along with our assumptions to do this.

As pointed out in Section 1.1.2, we can consider the two following distributions, which differ by a covariate shift (under SUTVA and ignorability):

$$\underbrace{P_{Y^{(a)},X}}_{\text{desired, unobservable}} \qquad \underbrace{P_{Y^{\text{obs}},X \mid T=a}}_{\text{observable!}} = P_{Y^{(a)},X \mid T=a}.$$

Let's use Lemma 1. We get that

$$\mathbb{E}\left[Y^{(a)}\right] = \mathbb{E}\left[\frac{\mathrm{d}P_{Y^{(a)},X}}{\mathrm{d}P_{Y^{\text{obs}},X \mid T=a}} \cdot Y^{\text{obs}} \mid T = 1\right]$$

So all that remains is for us to better figure out what is this density ratio and how to estimate it. Well,

$$\begin{aligned}
\frac{\mathrm{d}P_{Y^{(a)},X}}{\mathrm{d}P_{Y^{\text{obs}},X \mid T=a}} &= \frac{\mathrm{d}P_{Y^{(a)},X}}{\mathrm{d}P_{Y^{(a)},X \mid T=a}} && \text{(by consistency)} \\
&= \frac{\mathrm{d}P_{Y^{(a)} \mid X} \cdot \mathrm{d}P_X}{\mathrm{d}P_{Y^{(a)} \mid X,T=a} \cdot \mathrm{d}P_{X \mid T=a}} && \text{(factoring distributions)} \\
&= \frac{\mathrm{d}P_{Y^{(a)} \mid X} \cdot \mathrm{d}P_X}{\mathrm{d}P_{Y^{(a)} \mid X} \cdot \mathrm{d}P_{X \mid T=a}} = \frac{\mathrm{d}P_X}{\mathrm{d}P_{X \mid T=a}} && \text{(by ignorability)} \\
&= \frac{\mathrm{d}P_X}{\mathrm{d}P_{T=a \mid X} \cdot \mathrm{d}P_X / \mathrm{d}P_{T=a}} = \frac{\mathrm{d}P_{T=a}}{\mathrm{d}P_{T=a \mid X}} = \frac{\mathbb{P}[T=a]}{\mathbb{P}[T=a \mid X]}. && \text{(by Bayes' rule)}
\end{aligned}$$

Therefore:

Theorem 1. *Under SUTVA and ignorability,*

$$\mathbb{E}\left[Y^{(a)}\right] = \mathbb{E}\left[\frac{\mathbb{P}[T=a]}{\mathbb{P}[T=a|X]} \cdot Y^{obs} \mid T = 1\right]. \tag{2}$$

The quantity $\mathbb{P}[T=a|X]$ is actually quite handy in general, and is called the *propensity score* (discussed in e.g., Chapter 11 of (Ding, 2023)), and usually denoted as

$$e(x) := \mathbb{P}[T=a \mid X=x].$$

But note that Equation 2 uses the *true* propensity score, which is unknown in practice. Nevertheless, it can be proven that approximate versions of this equality hold when using a well-estimated propensity score (though this is nontrivial to show).

It's also *very* important to note that Equation 2 (and the derivation above) makes the positivity assumption (Assumption 6), so that the inverse of the propensity score is well-defined (no division by zero). But even more than that: though the math holds, things can get problematic if the propensity score gets even "too close" to zero, as in such cases the weight for some samples will be enormous. This would not be a problem if we had infinite samples, as the weights would end up compensating each other. But in reality we have only finite samples, and these blow-ups can be quite unstable.

Remark 2. *Indeed, essentially any work using these techniques for estimation in finite samples assumes not only positivity but that we have known bounds on the propensity scores that are not too large:*

$$0 < e_{min} \leq e(x) \leq e_{max} < 1 \qquad \forall x.$$

If $[e_{min}, e_{max}]$ is large (i.e., close to $[0,1]$), then the intervals produced by these finite-sample methods becomes extremely large and uninformative.

3. Risk minimization via sample reweighting

Another scenario is that we want to learn models $f^{(1)}, f^{(0)}$ that predict the potential outcomes:

$$f^{(1)}(X) \approx \mathbb{E}\left[Y^{(1)} \mid X\right] \qquad f^{(0)}(X) \approx \mathbb{E}\left[Y^{(0)} \mid X\right].$$

As we have seen previously, we can accomplish this under SUTVA and ignorability by simply regressing on the covariates conditional on the respective treatments:

$$f^{(a)} = \arg\min_f \frac{1}{n} \sum_{\substack{i=1 \\ T_i = a}}^{n} \left(f(X_i) - Y_i^{\text{obs}}\right)^2 \approx \arg\min_f \mathbb{E}\left[\left(f(X) - Y^{\text{obs}}\right)^2 \mid T = a\right].$$

We could argue, however, that what we'd *really* want to solve is

$$f^{(a)}_{\text{ideal}} = \arg\min_f \mathbb{E}\left[\left(f(X) - Y^{(a)}\right)^2\right],$$

i.e., minimizing the mean squared error of the potential outcome *over the complete distribution*, not just the part where $T = a$. While the actual global minimum will be equal (under SUTVA&ignorability), it can be argued that a small MSE conditional on $T = a$ does not necessarily imply a small MSE marginally.

To this end, we can again use sample reweighting to approximate the unconditional distribution from the distribution conditional on T:

$$\begin{aligned}
\arg\min_f \mathbb{E}\left[\left(f(X) - Y^{(a)}\right)^2\right] &= \arg\min_f \mathbb{E}\left[\frac{\mathrm{d}P_{Y^{(a)},X}}{\mathrm{d}P_{Y^{(a)},X|T=a}} \cdot \left(f(X) - Y^{(a)}\right)^2 \mid T = a\right] \\
&= \arg\min_f \mathbb{E}\left[\frac{\mathbb{P}[T=a]}{\mathbb{P}[T=a|X]} \cdot \left(f(X) - Y^{(a)}\right)^2 \mid T = a\right] \\
&= \arg\min_f \mathbb{P}[T=a]\ \mathbb{E}\left[\frac{1}{\mathbb{P}[T=a|X]} \cdot \left(f(X) - Y^{(a)}\right)^2 \mid T = a\right] \\
&= \arg\min_f \mathbb{E}\left[\frac{1}{\mathbb{P}[T=a|X]} \cdot \left(f(X) - Y^{(a)}\right)^2 \mid T = a\right].
\end{aligned}$$

Again, $\mathbb{P}[T = a|X]$ is the propensity score. So this amounts to reweighting our samples by the inverse of the propensity score. This can be quite intuitive: if a sample is unlikely to have $T = a$, then $\mathbb{P}[T = a|X]$ will be small and thus its inverse will be large, leading us to give more weight to this sample. Conversely, if a sample is likely to have $T = a$, then $\mathbb{P}[T = a|X]$ will be large and thus its inerse will be small, yielding a smaller weight and effectively giving more space for the more important (read, unlikely to be observed) samples to be taken into account.

Remark 3. *Everything we've done in this section refers to the mean squared error. As it happens, we can generalize everything to a general loss $\ell(f(X), Y)$, including the mean absolute error, pinball loss, log loss, and more.*

Bibliography

Ding, P. (2023). *A First Course in Causal Inference*.

Sugiyama, M., Suzuki, T., & Kanamori, T. (2012). *Density ratio estimation in machine learning*. Cambridge University Press.