Identifying the unidentifiable: combining knowledge from destructive measurements

Jackson Loper

April 2, 2018

Abstract

When many different measurement tools are applied, how can we best combine information from the different different tools? If we can apply multiple measurement tools to the same specimen, we can at least begin to understand how the tools they are related. We call this “joint measurement.” What can we do if joint measurement is unavailable? In this paper we investigate a simple assumption that can make it possible do obtain joint distributions even when joint measurement isn’t possible.

The modern setting is rife with experimental methodologies, and it can be very frustrating to understand how the output of these methods relate to one another. This is particularly difficult if joint measurement is unavailable, i.e. it is impossible or impractical to observe a single specimen with multiple methodologies. For example, after processing a neuron with a given single-cell rna sequencing technique, we cannot generally go back and process that same neuron with a different technique. This makes it seemingly impossible to calibrate the techniques against each other. By contrast, it is easy to calibrate thermometers against each other, because we can simply measure the same water with two different thermometers and see how the measurements relate. Without joint measurement, this isn’t possible.

The absence joint measurement gives rise to a number of problems:

- Different tools may teach us different things, and we need to combine tools to learn everything we can. Let A, B denote two different quantities of interest about cells (e.g. morphology and gene expression). Let us say one tool enables us to measure A and another tool to measure B. We would like to be able to say something about what morphologies are associated with what gene expressions.
- Two labs may used different methods, but may wish to pool results. Each lab looks at the other lab’s data and asks themselves "what would their data have looked like if they had used our method?" When they find they don’t know the answer, it greatly complicates collaboration.
- One may attempt to cluster the data. It is simple to cluster the data from one technique, and cluster the data from another technique – but how can we know how those clusters correspond to one another?

In this work we show that it is indeed sometimes possible to get useful and statistically rigorous bounds for these kinds of questions, even when joint measurements are unavailable. In particular, we consider the case that we can obtain samples from a variety of subpopulations. If we can assume that the *relationship* between the measurements is the same in each subpopulation, we can use these samples to learn something about that relationship – even though we never see both measurements for any single member of the population.

We apply our method to single cell RNA data. We show how different clusterings from different single cell RNA modalities can be connected. Our method yields results which are

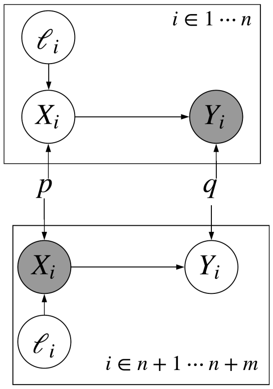

Figure 1: We assume that we have two datasets, both governed by the same process. In the top dataset, the results of technique II, denoted by Y_i, are hidden from us (we indicate this by making the circle gray). In the bottom dataset, the results of technique I, denoted by X_i, are hidden from us. In each dataset, we know what subpopulation of specimens we are considering, and this is designated by ℓ_i. Thus, for any given specimen i we can observe either ℓ_i, X_i or ℓ_i, Y_i but never ℓ_i, X_i, Y_i. Using observations of X_i we can certainly learn the parameters p which govern the relationship between X_i and the subpopulation, ℓ_i. However, if we believe that the relationship between X and Y is the same regardless of the population value ℓ, we can also learn something about the parameters q which govern the relationship between X and Y.

generally consistent with the scientist's understanding of the data and clusters, but also reveal potential gaps which could be important for further study.

This work stands on the shoulders of a long history of relating probabilistic assumptions to probabilistic inequalities. Much of this literature comes from research into causality. For example, in [1] Bonet uses polytopes not unlike the ones seen here to explore whether a variable can be used as an instrument. The famous Clauser-Horne-Shimony-Holt inequality was designed to help answer causality questions in quantum physics, but it also sheds light on what distributions are consistent with certain assumptions [2]. Indeed the physics literature has contributed many key inequalities (cf. [3], [4], and the references therein). Perhaps the closest work to this one would be [5], which used two marginal distributions to get bounds on a property of the joint distribution (namely the distribution of the sum). We build on this approach, both by using subpopulations to considerably refine our estimates and by considering the entire space of possible joint distributions instead of simply a particular aspect of the joint.

1 Mathematical formulation

Mathematically, we formulate our model as follows:

- Let us say we have $n+m$ individual members of a population. These could be cells, humans, or whatever the smallest sample unit may be for a given problem.
- For each individual member, we have some readily observable features by which we can group the members, such as the size of a cell, where an individual lives, or some other basic information. We designate these features as ℓ_i for the ith individual.
- For the first n members, we have made an observation using the first technique. This gives us the values $X_1 \cdots X_n$.

- The last m members were observed using the second technique. This gives us the values $Y_{n+1} \cdots Y_{n+m}$.

For simplicity, we here assume that ℓ_i, X_i, Y_i take values in finite sets, $\Omega_\ell, \Omega_X, \Omega_Y$. Extensions to more general cases should be straightforward, but we must leave them to future work.

We further posit a set of *counterfactual* variables – variables which could not ever be obtained in practice, but which help us organize our thinking. These are sometimes referred to as "potential outcomes" (cf. [6]).

- Let $Y_1 \cdots Y_n$ denote the observations we *would have gotten* if we had applied the second technique to the first n members.
- Let $X_{n+1} \cdots X_n$ denote the observations we would have gotten if we had applied the first technique to the last m members.

Our main assumption is that

$$\mathbb{P}(X_i = x, Y_i = y | L_i = \ell) = p^*(x_i | \ell_i) q^*(y_i | x_i) \tag{1}$$

for some distributions p^*, q^*. Moreover, we assume that the vectors $\{(p^*(x_1|\ell) \cdots p^*(x_n|\ell))\}_\ell$ are linearly independent. If this seems unlikely, it may make sense to collapse similar values of ℓ together. This assumption can be articulated by the plate diagram found in Figure 1.

The validity of these assumptions for a given situation should be closely contemplated. There are several key questions to answer when deciding whether this assumption is applicable. Does the process by which samples are gathered gathered depend only upon ℓ? In particular, is it not at all statistically related to which measurement technique was applied, except through ℓ? Are the particular measurement biases of both techniques the same for every value of ℓ? Is the statistical relationship between the quantities being measured by the two techniques the same for every value of ℓ?[1] If the answer to all of these questions is yes, the assumption that (X_i, Y_i) is drawn from $p(x_i|\ell_i)p(y_i|x_i)$ may apply.

We can now articulate our goal. Informally, our goal is to see what we can understand about the relationship between the observations X_i from one measurement modality and the other measurement modality Y_i, i.e. the distribution q^*. Unfortunately, given the limitations of the data, this may not be possible. Indeed, it is easy to see that q is dramatically unidentifiable in many cases.

For example, consider the case that ℓ lies in the set $1, 2$, X lies in the set $1, \cdots 100$ and Y lies in the set $1, 2$. As long as m, n are large enough, we can effectively estimate p^* and $h^*(Y|\ell) = \sum_x p^*(x|\ell) q(Y|x)$. However, this does not tell us very much about what q might be. Indeed, q lies in a 100-dimensional space, and the restriction that q must satisfy $h^*(Y|\ell) = \sum_x p^*(x|\ell) q(Y|x)$ for each ℓ, Y actually only introduces 2 new constraints. Thus there is a 98-dimensional space that q may lie in, and we simply cannot know where in that space q may lie. However, in practice, we find that the *inequality constraints*, namely that $q(y|x) > 0$ for every x, y, can actually force this apparently vast space to be tightly centered around a single point.

This suggests a somewhat less amibtious goal: to understand the *set of values* of q which are consistent with the data. Specifically, we will seek to produce an estimator for the set

[1] Note that this is automatically true if both techniques are measuring the same thing. More generally, if two techniques have something that they both measure, we can restrict attention to that one common phenomenon.

$$\Theta^* \triangleq \left\{ q: \sum_y p^*(x|\ell)q(y|x) = \sum_y p^*(x|\ell)q^*(y|x) \ \forall \ell, y \right\}$$

2 Algorithm

We begin by estimating

- p^* using the empirical distribution of the observations $\{X_i\}$, $\hat{p}(x|\ell)$
- $h^*(y|\ell) \triangleq \sum_x p^*(x|\ell)q^*(y|\ell)$ using the empirical distribution of $\{Y_i\}$, $\hat{h}(y|\ell)$

These empirical distributions may not be quite consistent with the original assumption that the distribution of $X, Y|\ell$ may be written as $p^*(x|\ell)q^*(y|x)$. In particular, it may be that there is *no* value of q such that $\sum_x \hat{p}(x|\ell)q(y|x) = \hat{h}(y|\ell)$. There may also be many such values. To obtain a well-defined estimator, we take a value of q which is decent, and use a slight regularization to ensure it is uniquely defined. Let $N_{X,\ell} = \#\{i \leq n: \ \ell_i = \ell\}, N_{Y,\ell} = \#\{i > n: \ \ell_i = \ell\}$. We take

$$\hat{q} = \arg\max_q \sum_\ell N_{Y,\ell} \sum_y \hat{h}(y|\ell) \log\left(\sum_x \hat{p}(x|\ell)q(y|x)\right) + \kappa \sum_{xy} \log q(y|x)$$

Note in the second term we apply an entropic penalty to q. This is both necessary to ensure that $\hat{q}$ is uniquely defined in terms of $\hat{q}, \hat{h}$, as well as a useful regularization. This regularization pushes us slightly towards uniform distributions, which makes sense if we believe that $q(y|x) > 0$ for each x, y. Other regularizations could also be used and would potentially enjoy similar theoretical guarantees to the ones we will show below.

This method gives us an estimate for $\hat{q}$, but we emphasize this estimate is very likely to be *inconsistent* for the true value of q^*. As we described in the previous section, this is simply a limitation of the data we have available – attempting to deduce q^* may be too much to ask for. However, subject to suitable assumptions, the estimator

$$\hat{\Theta} \triangleq \left\{ q: \sum_y p(x|\ell)q(y|x) = \sum_y p(x|\ell)\hat{q}(y|x) \ \forall \ell, y \right\}$$

is indeed consistent in the sense that we can be asymptotically assured that the true q^* lies very close to some point in $\hat{\Theta}$.

Theorem. *Fix any $\kappa > 0$. Let $N_{X,\ell}, N_{Y,\ell} \to \infty$ in such a way that $N_{Y,\ell'} / \sum_\ell N_{Y,\ell} \geq \rho > 0$ for each ℓ'. Let us assume that $q^*(y|x) > 0$ for every x, y and the rows of p^* are linearly independent. Then $\inf_{q \in \hat{\Theta}} \sup_{x,y} |q^*(y|x) - q(y|x)| \to 0$ in probability.*

Proof. We defer the proof to Appendix B. □

We close by remarking that the assumption $q^*(y|x) > 0$ is very likely to be unnecessary. However, it substantially simplifies the proof. We therefore leave a more complete theorem for future work.

3 Empirical results

Our motivation for this problem arose from looking at Allen Institute cell-type assignment of cells, performed using two different modalities. The only thing connecting the two modalities was that each modality had a variety of samples drawn from a variety of sub-populations, and the same sub-populations were used for both modalities. Thus our input was two tables: (technique I cluster × sub-population), and (technique II cluster × sub-population) (Figure 2). Our output was a polytope $\hat{\Theta}$ of possible ways that technique I clusters could line up against the technique II clusters (Figure 3).

Empirically, we see that this polytope appears to be very small, and tightly centered around our estimate for $\hat{q}$. Thus $\hat{q}$ may indeed serve as a useful indication of how the different technique's clusters line up. There are certain aspects of q, however, which are more uncertain than others. We were able to measure this by sampling uniformly from the polytope $\hat{\Theta}$ and seeing which entries varied the most (cf. Appendix A). Detecting these entries is useful because it suggests what further experiments would be helpful to more fully clarify the relationship between the two modalities.

Another way to look at $\hat{\Theta}$ is to try to understand its scope of the polytope along various directions. In this effort, we selected a direction uniformly at random and found the two extremal vertices in this direction. We then computed the distance between those vertices, as well as the distance between those vertices projected to the direction. Doing this procedure many times and making a scatterplot yields Figure 4. This figure suggests that the polytope may be only about 8 units in diameter, although we caution that general results on polytope diameter estimation are not encouraging, cf. [7]. However, considering that the magnitude of this diameter would be distributed over 1624 entries of the matrix, each of which lies in $[0, 1]$, it would seem that the polytope may indeed be constrained to be very close to the output we showed in Figure 3.

4 Conclusions

When joint measurement is impossible, it can be difficult to calibrate two methods against each other or understand how they may be related. Here we show that a simple Markov assumption can make it possible to actually learn quite a lot. Although the exact relationship may not be identifiable, a polytope of possible relationships can be identified, and this polytope may in fact be quite small indeed. By exploring this polytope, we can understand what we know – and what we don't know – about the relationship between measurements.

The Markov assumption is of course not the only one that we could have used, and may not be valid in every case. For example, it has been speculated that some cell types tend to die more often in one experimental modality than another, and these cells are not part of the data. This violates our assumptions. However, assuming this death rate can be roughly measured, it can be adjusted for, yielding a different but equally meaningful assumption about the data. Moreover, if this method yields bizarre results, it may give useful clues as to exactly how cell death may happen differently in the two modalities.

Once we accept that what we're interested in may not be fully identifiable, any of a wide variety of assumptions can help us obtain practical bounds. Although we may not be able to learn exactly what we want, we can learn a set of possibilities. By probing this set carefully with uniform samplers and extremal tests, we can learn what the data actually has to say and what experiments we need to do to learn more.

References

[1] Blai Bonet. Instrumentality tests revisited. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 48–55. Morgan Kaufmann Publishers Inc., 2001.

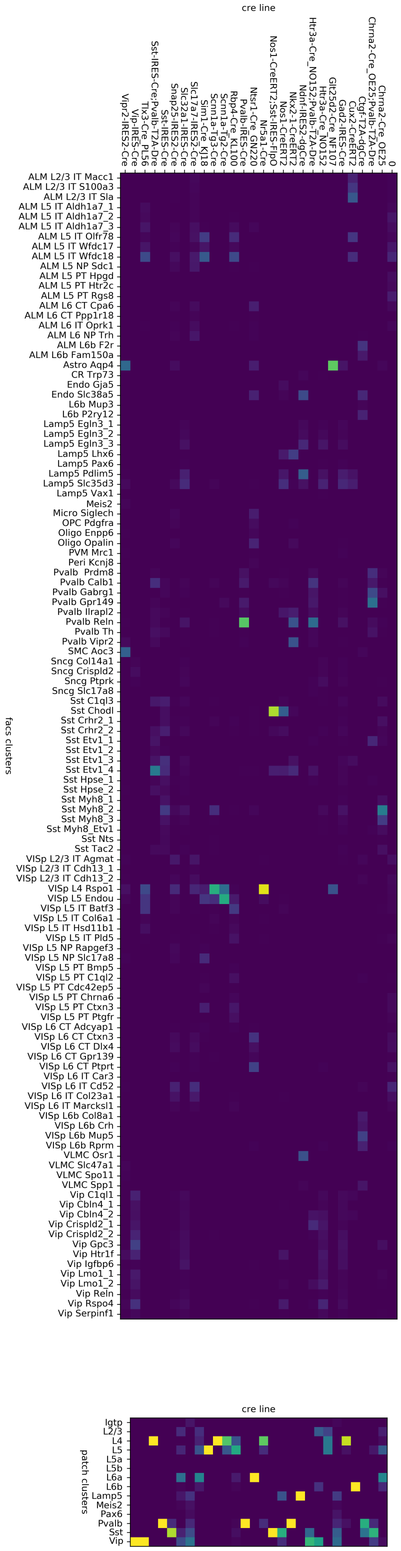

Figure 2: Input: two tables. The Allen Institute had access to various cre-line-based cell selection techniques. Each technique pulls out a different group of cells. Once the cells were selected, they were then either subjected to technique I ('facs') or technique II ('patch'). Technique I gives a very complete analysis of the gene expression of the cell. Technique II gives a less complete analysis, but yields additional electrophysiological data that may be of interest. The results of technique I was used to assign it to one of 116 categories, and the results of technique II was used to assign it to one of 14 categories. The tables above show the distribution of category assignment for each selection technique. The goal is to use these tables to be able to say something about how the 14 categories from Technique II might line up with the 116 categories from technique I.

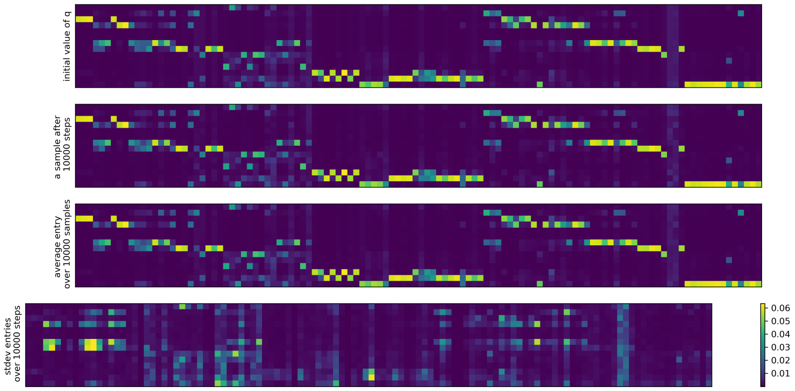

Figure 3: In the same setup as Figure 2. Using our method we produce a set of possible qs which are consistent with the data tables given. Using our method we first produce a single table, $\hat{q}$, that suggests how the outputs of the two techniques might line up. We are able to do this even though we were never able to observe a cell processed by both techniques. We emphasize, however, that this table is not guaranteed to be a consistent estimator for the true value of q^*. However, we can guarantee that asymptotically the true q^* will lie very close to a certain set $\hat{\Theta}$, containing $\hat{q}$. Fortunately, this set appears to be very small indeed, and tightly centered around $\hat{q}$. To see this, we used a Dikin sampler to sample uniformly from $\hat{\Theta}$. We saw that almost every sample was very close to $\hat{q}$. The second plot shows an example of such a sample. The third plot is an average value of each entry of q over many samples. The final plot indicates the standard deviation of each entry over the samples. Where this is high we have some substantial ambiguity, and so it may be important to perform further experiments to narrow down the correspondence.

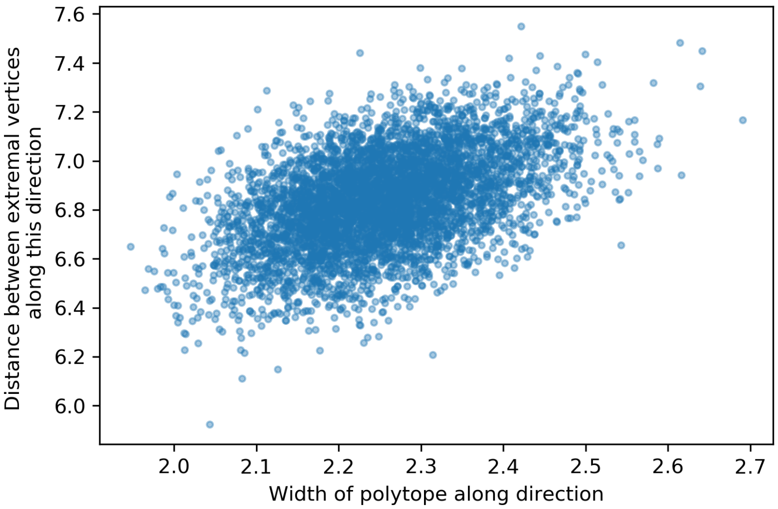

Figure 4: How wide is $\hat{\Theta}$? Each dot above corresponds to a randomly selected direction. The horizontal position of the dot indicates the width of the polytope along that direction, and the vertical position indicates the distance between the two extremal vertices along that direction.

[2] John F Clauser, Michael A Horne, Abner Shimony, and Richard A Holt. Proposed experiment to test local hidden-variable theories. *Physical review letters*, 23(15):880, 1969.

[3] Rafael Chaves, Lukas Luft, Thiago O Maciel, David Gross, Dominik Janzing, and Bernhard Schölkopf. Inferring latent structures via information inequalities. *arXiv preprint arXiv:1407.2256*, 2014.

[4] Aditya Kela, Kai von Prillwitz, Johan Aberg, Rafael Chaves, and David Gross. Semidefinite tests for latent causal structures. *arXiv preprint arXiv:1701.00652*, 2017.

[5] GD Makarov. Estimates for the distribution function of a sum of two random variables when the marginal distributions are fixed. *Theory of Probability & its Applications*, 26(4):803–806, 1982.

[6] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[7] Andreas Brieden, Peter Gritzmann, Ravi Kannan, Victor Klee, László Lovász, and Miklós Simonovits. Approximation of diameters: Randomization doesn't help. In *Foundations of Computer Science, 1998. Proceedings. 39th Annual Symposium on*, pages 244–251. IEEE, 1998.

[8] Ravindran Kannan and Hariharan Narayanan. Random walks on polytopes and an affine interior point method for linear programming. *Mathematics of Operations Research*, 37(1):1–20, 2012.

A Dikin sampler

Consider a convex polytope $T = \{x : Ax \leq b\}$. We have implemented a method for sampling from this polytope, based on the paper [8]. This method makes use of the Dikin ellipsoids, $E(x)$. For any x, these are defined by

- Computing the distance from x to each facet of the polytope, i.e. $d_i = b_i - \sum_j A_{ij} X_j$.
- Constructing $\tilde{A}$ as $\tilde{A}_{ij} = A_{ij}/d_i$.
- Define $E(x) = \{y : |\tilde{A}(X - y)| \leq 1\}$.

We can use these ellpsoids to efficiently sample the polytope T. At each step, we have some point $X \in T$, and we would to use this point to obtain a new sample Y, such that by iterating this process we asymptotically obtain samples which are uniform in T. Here is how we use

X to get Y:

```
Algorithm 1: Dikin sampler step
Data: A point $X \in T$
Result: A point $Y \in T$

Sample a proposal $\tilde{Y}$, uniformly from $E(X)$;
if $X \in E(\tilde{Y})$ then
    Sample $U \sim \text{Uniform}[0,1]$;
    if $U \leq \text{Vol}(E(X))/\text{Vol}(E(\tilde{Y})) \leq 1$ then
        Let $Y \leftarrow \tilde{Y}$;
    else
        Let $Y \leftarrow X$;
    end
else
    Let $Y \leftarrow X$;
end
```

It is easy to show that the stationary distribution of the Markov chain found by iterating these Dikin sampler steps is indeed uniform on T. To ensure an numerically robust method in the face of high-dimensional and nearly degenerate matrices, we take the following approach to robustly sampling from the ellipsoid:

```
Algorithm 2: Ellipsoid sampler
Data: An $n \times m$ matrix $\tilde{A}$
Result: A point $X$ sampled uniformly from $\{x : |Ax| \leq 1\}$

Sample $Z$ as an $n$-dimensional normal variables vector;
Let $X$ denote the solution to the least squares problem $\min_x |\tilde{A}x - Z|$;
Normalize $X$ by $X \leftarrow X/|\tilde{A}X|$;
Sample $U \sim \text{Uniform}[0,1]$;
Scale $X$ by $X \leftarrow X \times U^{1/m}$;
```

B Proof of the theorem

For the benefit of the reader, we here repeat the statement of our theorem in more explicit terms.

- Let $|\cdot|_\infty$ denote the uniform norm (i.e. the maximum absolute value) and $|\cdot|$ denote the Euclidean norm (i.e. the square root of the sum of the squares). In the case of matrices, this Euclidean norm goes by the name of the Frobenius norm. Recall that in this norm matrices satisfy a Cauchy-Schwarz like equality, $|pq| \leq |p|\,|q|$. Also recall that $|a|_\infty \leq ENa \leq \sqrt{n}\,|a|_\infty$ where n is the number of entries in a.
- Let $T_{a,b}$ denote the transition matrix polytope, i.e. the set of $a \times b$ matrices whose rows sum to 1 and whose entries are all positive.
- Let $|\Omega_\ell|, |\Omega_X|, |\Omega_Y| \in \mathbb{N}$.
- Let $p^* \in T_{|\Omega_\ell|,|\Omega_X|}$.
- Let $q^* \in T_{|\Omega_X|,|\Omega_Y|}$.
- We require the matrix q^* has strictly positive entries, $q^*_{xy} \geq c > 0$.

- We require that the rows of p^* are linearly independent.
- Let $\hat{p}$ denote an empirical transition matrix drawn by obtaining $N_{X,\ell}$ samples for each row of p^*, i.e. we have samples $(\ell_1,X_1)\cdots(\ell_1,X_n)$ such that $\mathbb{P}(X_i = x) = p^*_{\ell_i,x}$, $N_{X,\ell} = \sum_{i=1}^n \mathbb{I}_{\ell_i=\ell}$, and $\hat{p}_{\ell x} = \sum_{i=1}^n \mathbb{I}_{X_i=x,\ell_i=\ell}/N_{X,\ell}$.
- Let $\hat{h}$ denote an empirical transition matrix drawn by obtainin $N_{Y,\ell}$ samples for each row of h^*.

Now fix any $\kappa > 0$. Let

$$\hat{q} = \arg\max_q \left(\sum_\ell N_{Y,\ell} \sum_y \hat{h}(y|\ell) \log \left(\sum_x \hat{p}(x|\ell) q(y|x) \right) + \kappa \sum_{xy} \log q(y|x) \right) \tag{2}$$

and $\hat{\Theta} = \{q : \hat{p}\hat{q} = \hat{p}q\} \cap T_{|\Omega_X|,|\Omega_Y|}$.

Theorem. *If $N_{X,\ell}, N_{Y,\ell} \to \infty$ in such a way that $N_{Y,\ell'}/\sum_\ell N_{Y,\ell} \geq \rho > 0$ for each ℓ', then $\inf_{q\in\hat{\Theta}} |q^* - q|_\infty \to 0$ in probability.*

Proof. It is well-known that $\hat{p} \to p^*$ in probability (in both the uniform or the Euclidean norm, which are of course equivalent in this case). It is easy to see that the same goes for $\hat{p}\hat{q} \to h^*$ (see Lemma 1). Thus, intuitively, the difficulty is this: by allowing ourselves to ensure $|\hat{p} - p^*|_\infty, |\hat{p} - p^*|, |\hat{p}\hat{q} - h^*|_\infty, |\hat{p}\hat{q} - h^*|$ sufficiently small, can we find some $\tilde{q} \in \hat{\Theta}$ so that $|\tilde{q} - q^*|_\infty$ is arbitrarily small? It turns out we can.

Recall that $c > 0$ is the smallest value of q^*_{xy}. Fix any $\epsilon < c, p^*, q^*$. Let the right inverse of a matrix be defined by $a^\dagger \triangleq a^T(aa^T)^{-1}$. Note that since p^* has linearly independent rows, this is well-defined and continuous in a small neighborhood around p^*. Let $M = \left|(p^*)^\dagger\right|$. Find δ small enough so that if $|p - p^*|_\infty < \delta$ then $\left|p^\dagger\right| < 2M$. Taking a further smaller δ if necessary, ensure that if $|p^* - p|_\infty < \delta$ then $|p^* - p|$ is less than $\epsilon/4M\sqrt{|\Omega_X||\Omega_Y|}$. Now fix any $\hat{p}, \hat{q}$ with $|\hat{p} - p^*|_\infty < \delta$ and $|\hat{p}\hat{q} - p^*q^*| < \epsilon/4M$. Take

$$\tilde{q} = q^* + \hat{p}^\dagger \hat{p}(\hat{q} - q^*)$$

Then we make the following observations:

- Let us compute $|\tilde{q} - q^*|$. We have

$$\begin{aligned} |\tilde{q} - q^*| &= \left|\hat{p}^\dagger \hat{p}(\hat{q} - q^*)\right| \leq 2M\,|\hat{p}\hat{q} - \hat{p}q^*| \\ &\leq 2M\,|\hat{p}\hat{q} - p^*q^*| + 2M\,|(p^* - \hat{p})q^*| \\ &\leq 2M\frac{\epsilon}{4M} + \frac{2M\epsilon}{4M\sqrt{|\Omega_X||\Omega_Y|}}\sqrt{|\Omega_X||\Omega_Y|}\,|q^*|_\infty \leq \epsilon \end{aligned}$$

- $\hat{p}\tilde{q} = \hat{p}q^* + \hat{p}\hat{q} - \hat{p}q^* = \hat{p}\hat{q}$
- The rows of $\tilde{q}$ sum to 1. This is easy to see, because the rows of q^* sum to 1 and the rows of $\hat{q}$ sum to 1, and so $\tilde{q}\mathbf{1} = q^*\mathbf{1} + \hat{p}^\dagger\hat{p}(\hat{q} - q^*)\mathbf{1} = \mathbf{1} + 0$ as desired.
- The entries of $\tilde{q}$ are positive. Indeed, the the smallest value of q^* is c, and we have already argued that $|\tilde{q} - q^*|_\infty \leq \epsilon$. Thus the smallest value of $\tilde{q}$ is at least $c - \epsilon$, and we have required $\epsilon < c$.

Thus $|\tilde{q} - q^*|_\infty < \epsilon$ and $\tilde{q} \in \hat{\Theta}$.

In conclusion, we see that by taking $\hat{p}$ sufficiently close to p^* and $\hat{p}\hat{q}$ sufficiently close to p^*q^*, we can ensure that the set $\hat{\Theta}$ contains a close which is arbitrarily close to the true q^*. Since $\hat{p}$ and $\hat{p}\hat{q}$ are themselves consistent estimators, this completes the proof. □

Lemma 1. *If $N_{X,\ell}, N_{Y,\ell} \to \infty$ in such a way that $N_{Y,\ell'}/\sum_\ell N_{Y,\ell} \geq \rho > 0$ for each ℓ', then $|p^*q^* - \hat{p}\hat{q}|_\infty, |p^*q^* - \hat{p}\hat{q}| \to 0$ in probability.*

Proof. Our first task is to make a short study of the continuity of KL divergences on categorical distributions when the probabilities are bounded away from zero. Recall that we have insisted $q^*_{xy} \geq c > 0$ for every x, y – and this also means $(pq^*)_{\ell y}$ for every ℓy, since each row of p is itself a probability distribution. Moreover, observe that the KL-divergence on $|\Omega_Y|$-dimensional distributions, $D(\hat{r}||\tilde{r}) \triangleq \sum_y \hat{r}_y \log \hat{r}_y/\tilde{r}_y$, is *uniformly* continuous on the space of such distributions whose minimum probability is greater than any fixed positive constant. It follows that the map $h, p, q \mapsto D(h_\ell||(pq)_\ell)$ is also uniformly continuous on a space where h and q are strictly greater than some fixed positive constant.

With this in hand, the remainder of the proof follows naturally, using the well-known results that $\hat{p} \to p^*$ and $\hat{h} \to p^*q^*$ in probability.

Fix any ϵ, π. Let δ the modulus of continuity in the norm $|\cdot|_\infty$ at level $\epsilon\rho$ for $h, p, q \mapsto D(h_\ell||(pq)_\ell)$ when $h, q > c/2$. Select N large enough so that $\frac{1}{N_{Y,\ell}}\kappa|\Omega_X||\Omega_Y|\log\frac{1}{c} < \epsilon$ for each ℓ and so that with probability at least π we have that $\hat{h}, \hat{p}$ so that $\left|\hat{h} - p^*q^*\right|_\infty, |\hat{p} - p^*|_\infty \leq \delta, \left|\hat{h} - p^*q^*\right|_\infty < c/2$. Then

$$\begin{aligned} |D\left(\hat{h}_\ell||(\hat{p}\hat{q})_\ell\right) - D\left(h^*_\ell||(\hat{p}\hat{q})_\ell\right)| &\leq \rho\epsilon \\ D\left(\hat{h}_\ell||(\hat{p}q^*)_\ell\right) = |D\left(\hat{h}_\ell||(\hat{p}q^*)_\ell\right) - D\left((p^*q^*)_\ell||(p^*q^*)_\ell\right)| &\leq \rho\epsilon \end{aligned}$$

Now, since $\hat{q}$ is defined as the maximizer of a certain quantity (Equation 2), we may be sure that it is greater than the same quantity evaluated at $q = q^*$. That is,

$$\begin{aligned} 0 &\leq \sum_\ell N_{Y,\ell} \sum_y \hat{h}(y|\ell) \log \frac{\sum_x \hat{p}(x|\ell)\hat{q}(y|x)}{\sum_x \hat{p}(x|\ell)q^*(y|x)} + \kappa \sum_{xy} \log \frac{\hat{q}(y|x)}{q^*(y|x)} \\ &= \sum_\ell N_{Y,\ell} \left(D\left(\hat{h}_\ell||(\hat{p}q^*)_\ell\right) - D\left(\hat{h}_\ell||(\hat{p}\hat{q})_\ell\right)\right) + \kappa \sum_{xy} \log \frac{\hat{q}(y|x)}{q^*(y|x))} \end{aligned}$$

Applying our continuity results, it follows that

$$\sum_\ell N_{Y,\ell}\left(D\left(\hat{h}^*_\ell||(\hat{p}\hat{q})_\ell\right)\right) \leq 2\left(\sum_\ell N_{Y,\ell}\right)\epsilon + \kappa|\Omega_X||\Omega_Y|\log\frac{1}{c}$$

Note that the left-hand summands are all positive. So, in particular, it follows that for each ℓ', applying the uniformity condition ρ, we have

$$D\left(\hat{h}^*_{\ell'}||(\hat{p}\hat{q})_{\ell'}\right) \leq 2\epsilon + \frac{1}{N_{Y,\ell'}}\kappa|\Omega_X||\Omega_Y|\log\frac{1}{c} \leq 3\epsilon$$

That is, we have shown convergence of probability in the KL sense: for any ϵ, π we can find N high enough so that $D\left(\hat{h}^*_{\ell'}||(\hat{p}\hat{q})_{\ell'}\right) < \epsilon$ with at least probability π. This, in turn yields convergence in probability in the Euclidean or uniform metrics by Pinsker's inequality. □