Chapter 2

Couplings

These notes describe some couplings of random variables, on $\mathbb{N}$ or $\mathbb{R}^d$, as opposed to couplings of stochastic processes that will be seen in following chapters. There will be an emphasis on simulations and the relation to Monte Carlo methods, compared to references on coupling such as Lindvall [2002], Thorisson [2000].

2.1 Couplings, joint distributions and copulas

2.1.1 Related concepts but different contexts

Consider two random variables, X following a distribution p and Y following a distribution q, both on the same space $\mathbb{X}$ which might be a subset of $\mathbb{R}^d$ or of $\mathbb{N}^d$. A coupling of these two random variables refers to a joint distribution on $\mathbb{X} \times \mathbb{X}$ such that its first marginal is p and its second is q. We will refer indifferently to couplings of (X, Y) and to couplings of (p, q). For example, a coupling of (X, Y) is specified if we state that X is independent of Y. Then their joint distribution is of the form $p \otimes q$ i.e. their density reads $(x, y) \mapsto p(x)q(y)$ for all $x, y \in \mathbb{X}$. This can be called the independent coupling.

As an example, suppose that X and Y are Bernoulli variables with parameters p and q respectively; Bernoulli variables are simply coin flips, taking values in $\{0, 1\}$. The joint distributions on X and Y are described by matrices, say Γ, as in Table 2.1. If we impose $\mathbb{P}(X = 1) = p = 1 - \mathbb{P}(X = 0)$ and $\mathbb{P}(Y = 1) = q = 1 - \mathbb{P}(Y = 0)$, we put restrictions on the row sums and column sums of the matrix Γ. There is still some freedom in the specification of Γ. In Table 2.2 the left table corresponds to an independent coupling, in the case $p = 0.5$, $q = 0.6$. By adding the matrix in the middle, we are guaranteed to preserve the row and column sums. We obtain another coupling represented by the table on the right. Note that, given a matrix Γ of size $d \times d$, with d^2 entries, the cost to sample a pair $(X, Y) \sim \Gamma$ is, in general, equal to the cost of sampling from a Categorical distribution with $K = d^2$ categories. With the inverse transform method, one needs K operations to obtain cumulative probabilities in an arbitrary order. We can draw a uniform variable U in $[0, 1]$ and locate the category k corresponding to the inverse cdf transform of U in $\log K$ operations by binary search; overall the cost is of the order of K, see Malefaki and Iliopoulos [2007] for references. However, for certain couplings the sampling cost can be lower. For instance sampling X and Y independently would cost only of the order of d operations, and not d^2 as for generic couplings.

	$j = 0$	$j = 1$	
$i = 0$	Γ_{00}	Γ_{01}	$\to 1 - p$
$i = 1$	Γ_{10}	Γ_{11}	$\to p$
	$\downarrow 1 - q$	$\downarrow q$	

Table 2.1: Matrix Γ representing a coupling of two Bernoulli variables with parameters p and q.

.2	.3	$\to .5$			+1	-1		.3	.2	$\to .5$
.2	.3	$\to .5$	+	0.1$\times$	-1	+1	=	.1	.4	$\to .5$
$\downarrow .4$	$\downarrow .6$							$\downarrow .4$	$\downarrow .6$	

Table 2.2: Examples of couplings of two Bernoulli random variables represented as matrices.

An initial reaction to the above definition is that any joint distribution on a space $\mathbb{X} \times \mathbb{X}$ is a coupling of two marginal distributions, and thus it might seem unnecessary to introduce a name "coupling" to refer to what is just a multivariate probability distribution. It might become clearer later, after going through enough examples, that there is something specific with coupling techniques. There, the pair of variables is introduced with a specific goal such as proving an inequality or a convergence result, or developing a new algorithm, and the relation between X and Y is defined in a very specific way, while the marginal distributions p and q are given to us by the context. Thus we can think of a coupling as the definition of a special relationship between X and Y under the "marginal constraints" that $X \sim p$ and $Y \sim q$.

There is another term that comes back often in statistics and that has to do with the description of joint distributions, the term "copula". The current Wikipedia page defines this as follows: "In probability theory and statistics, a copula is a multivariate cumulative distribution function for which the marginal probability distribution of each variable is uniform." It seems that copulas are just another word for couplings of two uniform distributions on $[0, 1]$. There are indeed some overlap between the literature on couplings and that on copulas, but in practice the context is usually very different. The literature on copulas primarily focuses on the task of modeling data, and the dependencies between some observed variables or transformations thereof, using parametric distributions and with the usual associated inference questions. For instance we might observe a sample of pairs, $(X_n, Y_n)_{n\in[N]}$, and we might look for an adequate joint distribution to fit these samples. We might want to decompose the modeling task into 1) the modeling of the marginal distributions of X and Y, and 2) the modeling of the dependencies between X and Y, via a proposed coupling of uniform variables $F_X(X)$ and $F_Y(Y)$, where F_X and F_Y refer to the cumulative distribution functions of the marginals. Then we would be in the world of copulas (I think... I've actually never been there). The starting point is Sklar's theorem [Sklar, 1959] which states that we can model any joint distribution using such decomposition. In the literature on copulas one might find various parametric classes of joint distributions on $[0, 1]^2$ with uniform marginals, named for instance Gaussian copulas or Archimedean copulas. A book on copulas is Nelsen [2007].

2.1.2 Some couplings of Bernoulli variables

Let us go back to the example of coupling Bernoulli distributions, with examples in Table 2.2. We can see that, in this example, all couplings can be obtained by adding a scalar multiple of the matrix with -1's and $+1$'s in the middle of Table 2.2. Indeed if we add any value to an entry Γ_{ij}, we must subtract that value on the other cell of the same row and of the same column, in order for the marginal constraints to be satisfied, and in turn the same value must then be added to the cell located diagonally from Γ_{ij}. Furthermore the added value must be restricted so that the resulting matrix Γ has entries that are all in the interval $[0, 1]$; indeed they must represent probabilities $\mathbb{P}(X = i, Y = j) \in [0, 1]$. We have the two

following examples, corresponding to multiplicative factors -0.2 and 0.2,

$$\Gamma^{(1)} = \begin{pmatrix} 0.0 & 0.5 \\ 0.4 & 0.1 \end{pmatrix} \quad \text{and} \quad \Gamma^{(2)} = \begin{pmatrix} 0.4 & 0.1 \\ 0 & 0.5 \end{pmatrix}.$$

On the left, the matrix Γ is such that the event $\{X = Y\}$ would occur with probability 0.1, equal to the sum of the diagonal entries. We can sample from this coupling by drawing $U \sim \text{Uniform}(0, 1)$ and defining $X = \mathbb{1}(U \leq 0.5)$ and $Y = \mathbb{1}(U > 0.4)$. On the right, the coupling corresponds to drawing $U \sim \text{Uniform}(0, 1)$ and defining $X = \mathbb{1}(U \leq 0.5)$ and $Y = \mathbb{1}(U \leq 0.6)$. We can convince ourselves that we cannot have a larger sum of diagonal entries than 0.9, and also we cannot have less than 0.1, without breaking the marginal constraints.

This example illustrates two important points about couplings: 1) there are infinitely many of them, and 2) some of them are special, for instance they achieve the minimum or maximum probabilities for the event $\{X = Y\}$, as in $\Gamma^{(1)}$ and $\Gamma^{(2)}$ above respectively. We next focus on couplings that are called "maximal", which are precisely the ones corresponding to the largest probability of $\{X = Y\}$, or the minimum probability of $\{X \neq Y\}$.

Figure 2.1 shows different couplings of the same two marginal distributions, two of which are "maximal".

2.2 Maximal couplings and total variation distance

2.2.1 Various representations of the total variation distance

The next result is often called the coupling lemma or coupling inequality [Chapter 4 of Thorisson, 2000]. First, for any pair of variables X, Y defined on a general state space $\mathbb{X}$ with measurable sets $\mathcal{X}$, we have

$$\forall A \in \mathcal{X} \quad \mathbb{P}(X \in A) - \mathbb{P}(Y \in A) \leq \mathbb{P}(X \neq Y), \tag{2.1}$$

$$\text{or equivalently,} \quad \sup_{A \in \mathcal{X}} \mathbb{P}(X \in A) - \mathbb{P}(Y \in A) \leq \mathbb{P}(X \neq Y).$$

Notably the left-hand side is an object that only involves the marginal distributions of X and Y, whereas the right hand side concerns a relation between X and Y.

Proof. We can start by writing

$$\mathbb{P}(X \in A) - \mathbb{P}(Y \in A) = \mathbb{P}(X \in A \text{ and } X \neq Y) - \mathbb{P}(Y \in A \text{ and } X \neq Y),$$

and then

$$\mathbb{P}(X \in A \text{ and } X \neq Y) - \mathbb{P}(Y \in A \text{ and } X \neq Y) \leq \mathbb{P}(X \in A \text{ and } X \neq Y),$$

which holds because $\mathbb{P}(Y \in A \text{ and } X \neq Y) \geq 0$, and finally we can upper bound the right-hand side by $\mathbb{P}(X \neq Y)$ because $A \subset \mathbb{X}$. □

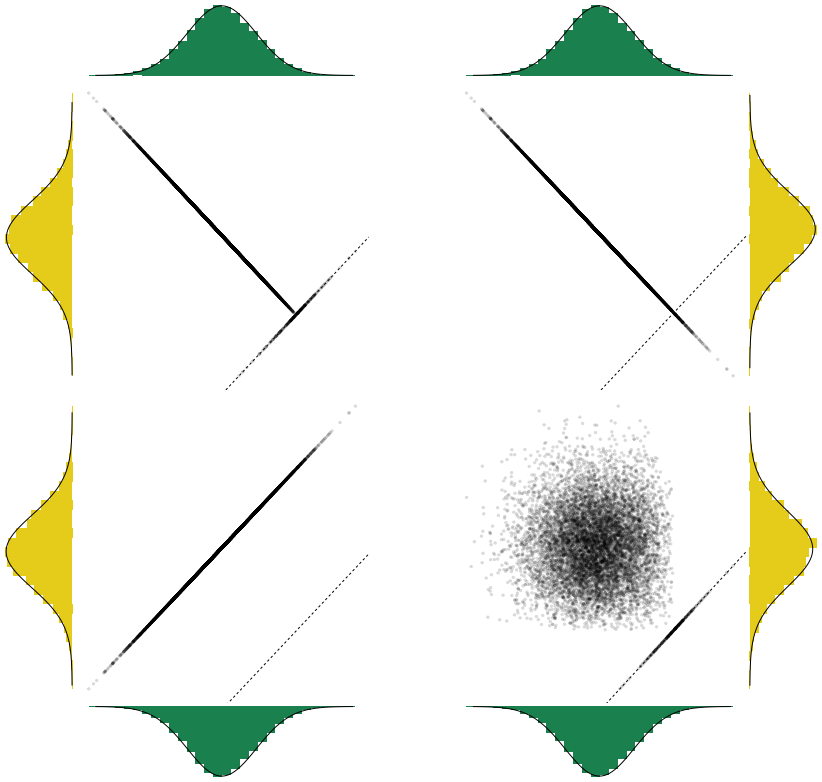

Figure 2.1: Tetra-coupling of scatter plots with marginal histograms.

The quantity $\sup_{A \in \mathcal{X}} \{\mathbb{P}(X \in A) - \mathbb{P}(Y \in A)\}$ is also known as the "total variation distance" or TV between the distributions of X and Y. It is the maximal difference that we can observe between the probabilities assigned by the distributions of X and Y to any measurable subset A. We will see that the inequalities in (2.1) are in fact sharp in the sense that there really is a pair (X, Y) such that $\sup_{A \in \mathcal{X}} \mathbb{P}(X \in A) - \mathbb{P}(Y \in A)$ is equal to the probability of $\{X \neq Y\}$. There are in fact more than one such coupling, and any coupling with this property is called a "maximal coupling".

Before showing the existence of such maximal couplings, we will express the TV in various ways. Figure 2.2 might be good to keep in mind. We will use p, q to also denote probability density functions with respect to a common dominating measure; we recall that $\mathbb{X}$ can be discrete as $\{0, 1\}$ or continuous as $\mathbb{R}^d$. First, we can find explicitly the set that maximizes $\mathbb{P}(X \in A) - \mathbb{P}(Y \in A)$ over $A \in \mathcal{X}$. Consider the measure $\nu = p - q$, which is a signed measure with a mass of zero. By Hahn-Jordan decomposition there is a set D such that $\nu = \nu^+ - \nu^-$ where $\nu^+(\cdot) = (p - q)(\cdot \cap D)$ is a positive measure and $\nu^-(\cdot) = (q - p)(\cdot \cap D^c)$ is a positive measure, with D^c denoting the complement of D in $\mathbb{X}$. Essentially, D is

the set $\{x \in \mathbb{X} : p(x) \geq q(x)\}$.

First note that $\sup_A \nu(A) = \nu(D)$. Indeed $\nu(A) = \nu^+(A) - \nu^-(A)$, $\nu^+(A) \leq \nu^+(D)$ for all A because ν^+ is a positive measure, and $-\nu^-(A) \leq 0 = -\nu^-(D)$ because ν^- is positive and $D \cap D^c = \emptyset$. Thus,

$$\|p - q\|_{\mathrm{TV}} := \sup_{A \in \mathcal{X}} \mathbb{P}(X \in A) - \mathbb{P}(Y \in A) = \mathbb{P}(X \in D) - \mathbb{P}(Y \in D). \tag{2.2}$$

Next, consider the quantity defined as

$$\mathfrak{D}(p, q) = \sup_{|h| \leq 1} \left| \int h(x)p(x)dx - \int h(x)q(x)dx \right|. \tag{2.3}$$

The supremum is over all functions bounded by one. In X, Y notation this is $\sup_{|h|\leq 1} |\mathbb{E}_p[h(X)] - \mathbb{E}_q[h(Y)]|$. Then $\mathfrak{D}(p, q)$ is attained over all h by the function $x \mapsto \mathbb{1}(x \in D) - \mathbb{1}(x \in D^c)$, i.e. the function taking largest positive values on D and the most negative values on D^c. We obtain $\mathfrak{D}(p, q) = (p-q)(D)+(q-p)(D^c)$. Since the total mass of $p - q$ is zero, we have $(p - q)(D) = -(p - q)(D^c) = (q - p)(D^c)$ so

$$\mathfrak{D}(p, q) = 2 \times \{\mathbb{P}(X \in D) - \mathbb{P}(Y \in D)\} = 2\|p - q\|_{\mathrm{TV}}.$$

We can also find a representation of the TV that is more explicit in the density functions p and q. Recalling that $D = \{x : p(x) \geq q(x)\}$,

$$\begin{aligned} \mathbb{P}(X \in D) - \mathbb{P}(Y \in D) &= \int_{\{x:p(x)\geq q(x)\}} p(x)dx - \int_{\{x:p(x)\geq q(x)\}} q(x)dx \\ &= \int_{\{x:p(x)<q(x)\}} q(x)dx - \int_{\{x:p(x)<q(x)\}} p(x)dx, \end{aligned}$$

and then we can add the two equations above to conclude that

$$\|p - q\|_{\mathrm{TV}} = \frac{1}{2} \int |p(x) - q(x)| \, dx, \tag{2.4}$$

i.e. the TV is the "area between the two pdf curves". Furthermore, by writing $|a - b| = a + b - 2\min(a, b)$ for all $a, b \in \mathbb{R}$, we also have

$$\|p - q\|_{\mathrm{TV}} = 1 - \int \min(p(x), q(x))dx. \tag{2.5}$$

Sometimes $\min(a, b)$ is written $a \wedge b$, where $a, b \in \mathbb{R}$, and $p \wedge q$ might refer to the function defined as the point-wise minimum between p and q.

Now let us come back to (2.1) and the question of finding a coupling such that $\mathbb{P}(X \neq Y)$ achieves the lower bound $\sup_{A\in\mathcal{X}} \mathbb{P}(X \in A) - \mathbb{P}(Y \in A)$. We construct such a coupling explicitly in Algorithm 1. This algorithm is well-known, it was mentioned in Section 4.5 of Thorisson [2000], and is called a γ-coupling in Johnson [1998], to give some examples. We next justify that the algorithm provides a maximum coupling of p and q. Checking that $X \sim p$ can be done by

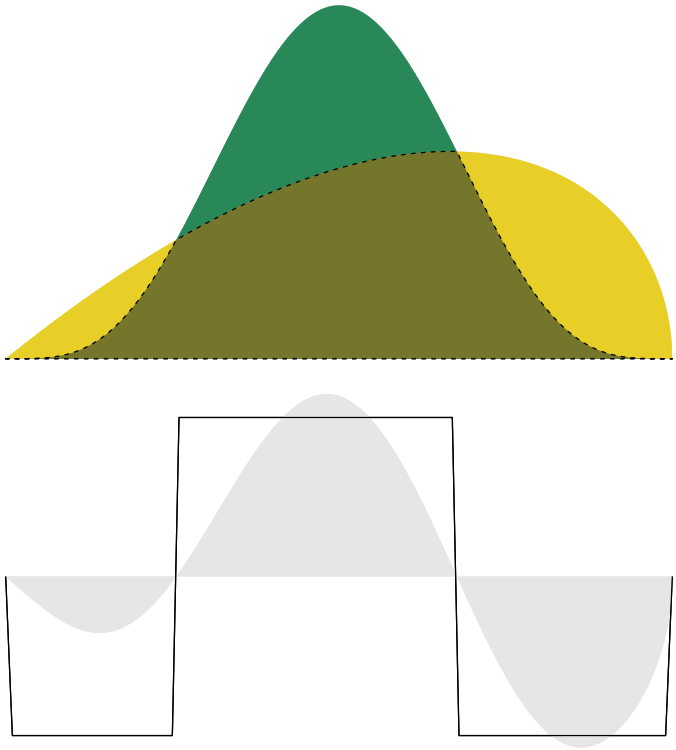

Figure 2.2: Total variation distance.

looking at the algorithm. Next we check that $Y \sim q$. For a measurable set A, we write

$$\mathbb{P}(Y \in A) = \mathbb{P}(Y \in A, W \leq q(X)/p(X)) + \mathbb{P}(Y \in A, W > q(X)/p(X)),$$

which separates step 2.(a) and 2.(b) in two cases. In the first case, $Y = X$ and $X \sim p$, so the term $\mathbb{P}(Y \in A, W \leq q(X)/p(X))$ can be written

$$\begin{aligned} &\int \mathbb{1}(y \in A)\mathbb{1}(x = y)\mathbb{1}(w \leq q(x)/p(x))\mathbb{1}(0 \leq w \leq 1)p(x)dwdx \\ &= \int_A \min(1, q(x)/p(x))p(x)dx \\ &= \int_A \min(p(x), q(x))dx. \end{aligned}$$

With the choice $A = \mathbb{X}$ we obtain $\mathbb{P}(W \leq q(X)/p(X)) = \int \min(p(x), q(x))dx$, which we'll use later on. At this point, if we want $\mathbb{P}(Y \in A)$ to be equal to

$\int \mathbb{1}(x \in A) q(x) dx$, we can see that we need

$$\mathbb{P}(Y \in A, W > q(X)/p(X)) = \int \mathbb{1}(x \in A) \left(q(x) - \min(p(x), q(x))\right) dx,$$

or equivalently,

$$\mathbb{P}(Y \in A | W > q(X)/p(X)) = \frac{\int \mathbb{1}(x \in A) \left(q(x) - \min(p(x), q(x))\right) dx}{1 - \int \min(p(x), q(x)) dx},$$

i.e. we need to check that Y sampled in step 2.(b) is distributed according to a distribution with density

$$\tilde{q} : x \mapsto \frac{q(x) - \min(p(x), q(x))}{1 - \int \min(p(x), q(x)) dx}. \tag{2.6}$$

We can see that step 2.(b) as a rejection sampler, where the target distribution is indeed $\tilde{q}$, the proposal distribution is q, and draws are accepted if $W^\star > p(Y^\star)/q(Y^\star)$ i.e. with probability $\max(0, 1 - p(Y^\star)/q(Y^\star))$. This is precisely equal to $\tilde{q}(Y^\star)/q(Y^\star)$ up to the multiplicative constant $1 - \int \min(p(x), q(x)) dx$. Thus step 2.(b) is a rejection sampler as in the previous chapter, using unnormalized target density evaluations and an appropriate upper bound "M". This concludes the justification that Y follows q marginally.

Finally, since we have seen that $\mathbb{P}(W \leq q(X)/p(X)) = \int \min(p(x), q(x)) dx$ and since $\{X = Y\}$ only occurs when $\{W \leq q(X)/p(X)\}$ occurs (can you see why?), we have $\mathbb{P}(X \neq Y) = \|p - q\|_{\mathrm{TV}}$ via (2.5).

Algorithm 1 Sampling from a maximal coupling of p and q.

1. Sample $X \sim p$
2. Sample $W \sim \mathrm{Uniform}(0, 1)$, independently of X.
   (a) If $W \leq q(X)/p(X)$, set $Y = X$,
   (b) otherwise, sample $Y^\star \sim q$ and $W^\star \sim \mathrm{Uniform}(0, 1)$ (indep. of $Y^\star$) until $W^\star > p(Y^\star)/q(Y^\star)$, and set $Y = Y^\star$.
3. Return (X, Y).

We summarize the different representations of the total variation distance

between $X \sim p$ and $Y \sim q$,

$$\begin{aligned}
\|p - q\|_{\mathrm{TV}} &= \sup_{A \in \mathcal{X}} \mathbb{P}(X \in A) - \mathbb{P}(Y \in A) \\
&= \frac{1}{2} \sup_{|f| \leq 1} |\int f(x)p(x)dx - \int f(x)q(x)dx|. \\
&= \frac{1}{2} \int |p(x) - q(x)| \, dx \\
&= 1 - \int \min(p(x), q(x))dx \\
&= \inf_{(X,Y) \in \mathrm{coupling}(p,q)} \mathbb{P}(X \neq Y).
\end{aligned}$$

Here coupling(p, q) refers to the set of joint distributions for (X, Y) with marginals $X \sim p$ and $Y \sim q$. See Figure 2.2 for some illustrations of the above equalities.

2.2.2 The mixture view on maximal coupling

The proof of the existence of maximal couplings, as e.g. in Lindvall [Chapter I, 2002], can be established by constructing the coupling in the following way. Recall the definition of $\tilde{q}$ in (2.6), and define $\tilde{p}$ similarly by swapping the letters p and q. Denote also by $\int p \wedge q$ the mass of the overlap between p and q,

$$\int p \wedge q = \int (p \wedge q)(x) \, dx, \tag{2.7}$$

and define the overlap density $c : x \mapsto (p \wedge q)(x) / \int p \wedge q$. Then write

$$\begin{aligned}
p(\cdot) &= \left(\int p \wedge q \right) \times c(\cdot) + \left(1 - \int p \wedge q \right) \times \tilde{p}(\cdot), \\
q(\cdot) &= \left(\int p \wedge q \right) \times c(\cdot) + \left(1 - \int p \wedge q \right) \times \tilde{q}(\cdot).
\end{aligned} \tag{2.8}$$

In this view, p and q are both represented as mixtures of two components. The first one is common, has weight $\int p \wedge q$ and density $x \mapsto c(x)$. The second component has weight $1 - \int p \wedge q$ and density $x \mapsto \tilde{p}(x)$ and $x \mapsto \tilde{q}(x)$ respectively for p and q. Note that $\tilde{p}$ and $\tilde{q}$ have no common mass; they were indeed obtained by removing the "common mass" from p and q. With this representation, consider Algorithm 2.

The pair (X, Y) generated by Algorithm 2 is such that $X \sim p$ and $Y \sim q$, and $\mathbb{P}(X = Y) = \int p \wedge q$, so the coupling is indeed maximal. If X and Y are drawn independently in step 2.(b), then we can write the joint probability $\gamma_{\max}$ of (X, Y) as

$$\gamma_{\max}(\cdot) = \left(\int p \wedge q \right) \times \psi_\sharp c(\cdot) + \left(1 - \int p \wedge q \right) \times \tilde{p} \otimes \tilde{q}(\cdot),$$

Algorithm 2 Sampling from a maximal coupling of p and q.

1. Draw $U \sim \text{Uniform}(0, 1)$.
2. If $U \leq \int p \wedge q$,
   (a) draw $X \sim c$ with $c : x \mapsto (p \wedge q)(x) / \int p \wedge q$, and set $Y = X$,
   (b) otherwise, draw $X \sim \tilde{p}$ and $Y \sim \tilde{q}$, with $\tilde{p} \propto p - p \wedge q$, $\tilde{q} \propto q - p \wedge q$.
3. Return (X, Y).

where $\psi : x \mapsto (x, x)$, $\psi_\sharp c$ is the push-forward measure of c through ψ, i.e. for all measurable set $B \in \mathcal{X} \times \mathcal{X}$, $\psi_\sharp c(B) = c(\psi^{-1}(B)) = c(\{x \in \mathbb{X} : (x, x) \in B\})$, and $\tilde{p} \otimes \tilde{q}$ is a distribution defined as a product of marginals $\tilde{p}$ and $\tilde{q}$. Algorithm 1 in fact samples from that joint distribution $\gamma_{\max}$. In discrete spaces, where p and q are represented by vectors of size K, then $\gamma_{\max}$ is a $K \times K$ matrix defined as $\text{diag}(p \wedge q) + (1 - \int p \wedge q)\, \tilde{p}\, \tilde{q}^T$.

The mixture view makes it apparent that different couplings can be maximal: any coupling of $\tilde{p}$ and $\tilde{q}$ can be used instead of $\tilde{p} \otimes \tilde{q}$ above, without modifying the probability of $\{X \neq Y\}$. Sampling from the mixture using Algorithm 2 requires access to the value of $\int p \wedge q$, and assumes the ability to sample from c, $\tilde{p}$ and $\tilde{q}$. Algorithm 1 bypasses the need to evaluate $\int p \wedge q$. On the other hand, in discrete state spaces Algorithm 2 might be preferable to Algorithm 1. Indeed, if p and q are vectors of K probabilities on a common discrete set, then Algorithm 2 can be run in the order of K operations and with a deterministic cost, while Algorithm 1 presents a while loop.

2.2.3 Sampling from a maximal coupling

The above discussions provide different ways of sampling from maximal couplings. Algorithm 1 samples from a maximal coupling of p and q provided that we can sample from p and q, and evaluate their probability density functions, including their normalizing constants. Algorithm 2 can be implemented when it is possible to sample an event of probability $\int p \wedge q$, and to sample from c, $\tilde{p}$ and $\tilde{q}$.

Let us look at the computational cost of Algorithm 1 in more detail. We define a unit of cost to be a draw from either p or q, and an evaluation of the ratio p/q. Then the cost is one unit if the algorithm terminates at step 2.(a). This occurs with probability equal to $\mathbb{P}(W < q(X)/p(X))$, where $W \sim \text{Uniform}(0,1)$ and $X \sim p$. We have already computed $\mathbb{P}(W < q(X)/p(X)) = 1 - \|p - q\|_{\text{TV}}$, in the justification of Algorithm 1. If the algorithm enters step 2.(b) the number of trials before completion follows a Geometric variable, with parameter equal to $\|p - q\|_{\text{TV}}$. Indeed, the parameter of that Geometric is $\mathbb{P}(W > p(Y)/q(Y))$

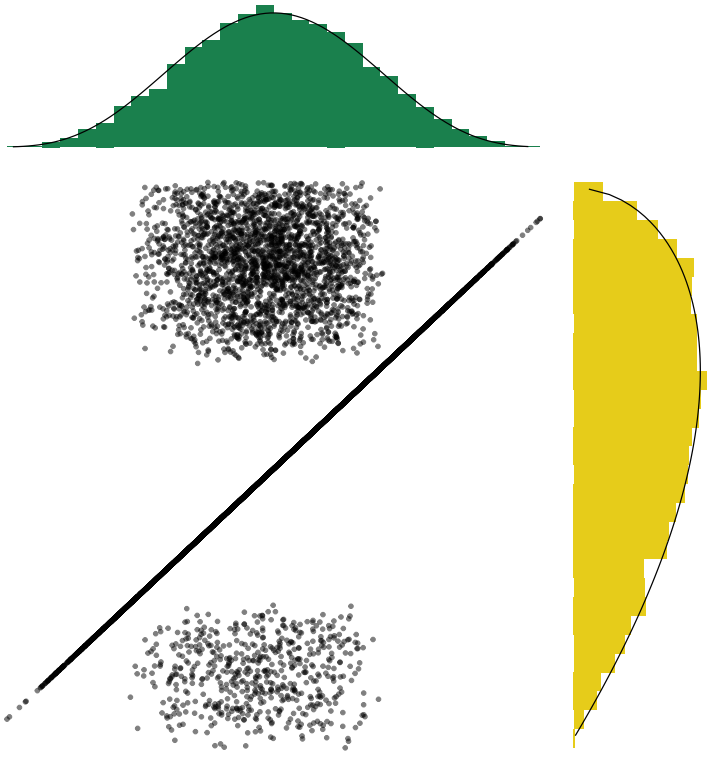

Figure 2.3: A maximal coupling with independent residuals.

where $W \sim \text{Uniform}(0,1)$ and $Y \sim q$. We can compute

$$\begin{aligned}\mathbb{P}(W > p(Y)/q(Y)) &= \int \mathbb{1}(w > p(y)/q(y))q(y)\mathbb{1}(w \in (0,1))dydw \\ &= \int_{p(y)<q(y)} (1 - p(y)/q(y))q(y)dy \\ &= \int_{p(y)<q(y)} (q-p)\,(y)dy = \mathbb{P}(Y \in D^c) - \mathbb{P}(X \in D^c).\end{aligned}$$

We recognize one of the representations of the TV. Next, a Geometric variable with parameter p has expectation $1/p$ and variance $(1-p)/p^2$. Thus the algorithm has a cost that satisfies

$$\mathbb{E}[\text{cost}] = (1 - \|p-q\|_{\text{TV}}) \times 1 + \|p-q\|_{\text{TV}} \times (1 + \|p-q\|_{\text{TV}}^{-1}) = 2.$$

The expected completion time of the algorithm is equal to 2 units, irrespective of p and q. However, the variance of the cost depends on $\|p-q\|_{\text{TV}}$ in an unpleasant way: it goes to infinity as $\|p-q\|_{\text{TV}} \to 0$. To see this, compute the variance of

the cost by conditioning on whether the algorithm enters step 2.(a) or step 2.(b). On the other hand, if $\|p - q\|_{\mathrm{TV}} = 0$ then $p = q$ and the algorithm stops at step 2.(a) with probability one, thus the variance of the cost is then equal to zero.

There are multiple alternatives to Algorithm 1, in cases where the variance of its cost becomes a source of anxiety. For the particular case $p = \mathcal{N}(\mu_1, \Sigma), q = \mathcal{N}(\mu_2, \Sigma)$, we can use a *reflection-maximal coupling* [Bou-Rabee et al., 2018, Eberle et al., 2019] which has a deterministic computational cost and is also maximal. This is given in Algorithm 3 below, where s denotes the probability density function of a d-dimensional standard Normal. In the case $\dot{Y} = \dot{X} + z$ below, we get an event $\{X = Y\}$. We can check that this coupling satisfies

Algorithm 3 A reflection-maximal coupling of $\mathcal{N}(\mu_1, \Sigma)$ and $\mathcal{N}(\mu_2, \Sigma)$.

Let $z = \Sigma^{-1/2}(\mu_1 - \mu_2)$ and $e = z/\|z\|$.
Sample $\dot{X} \sim \mathcal{N}(0_d, I_d)$, and $W \sim \mathrm{Uniform}(0, 1)$
If $s(\dot{X})W \leq s(\dot{X} + z)$
Set $\dot{Y} = \dot{X} + z$
Else set $\dot{Y} = \dot{X} - 2(e^T \dot{X})e$
Set $X = \Sigma^{1/2}\dot{X} + \mu_1, Y = \Sigma^{1/2}\dot{Y} + \mu_2$, and return (X, Y)

the marginal constraints and that it is maximal. First we can check that $X = \mu_1 + \Sigma^{1/2}\dot{X}$ is $\mathcal{N}(\mu_1, \Sigma)$. Then, for Y let us check that $\dot{Y}$ is $\mathcal{N}(0, I)$. We have

$$\mathbb{P}(Y \in A) = \int \mathbb{1}(x + z \in A) \min(1, \frac{s(x+z)}{s(x)}) s(x) dx + \int \mathbb{1}(x - 2(e^T x)e \in A) \max(0, 1 - \frac{s(x+z)}{s(x)}) s(x) dx.$$

With a change of variable $x \to x + z$ the first integral is $\int \mathbb{1}(x \in A) \min(s(x - z), s(x))dx$. With a change of variable $x \to x - 2(e^T x)e$ which corresponds to a reflection with respect to a plane orthogonal to e, we obtain $\int \mathbb{1}(x \in A) \max(0, s(x) - s(x - z))dx$, thanks to the spherical symmetry of the standard Normal (pictures might help here). Summing up the two integrals, we obtain $\int_A s(x)dx$ and thus $\dot{Y} \sim s$. Finally we can compute

$$\mathbb{P}(X \neq Y) = \mathbb{P}(\dot{Y} \neq \dot{X} + z) = 1 - \int s(x) \wedge s(x - z) dx,$$

and we can check that this is indeed the total variation distance between the two Normals $p = \mathcal{N}(\mu_1, \Sigma), q = \mathcal{N}(\mu_2, \Sigma)$, perhaps upon a few changes of variables. We can generalize Algorithm 3 to other spherically symmetrical distributions. The completion time of Algorithm 3 is deterministic. This is another illustration of the fact that there are more than one maximal coupling. In passing, how could you extend Algorithm 3 to Normals with different covariance matrices?

We can also modify Algorithm 1 as described in a comment by Mathieu Gerber and Anthony Lee [add reference], so that the variance of the cost would be upper

bounded for all p and q, but with a decrease of the probability of $\{X = Y\}$, i.e. the algorithm then samples from a sub-maximal coupling. So, imagine that we are OK with $\{X = Y\}$ occurring less often if this solves the "large variance of the cost" issue. Decreasing the probability of $\{X = Y\}$ amounts to accepting $Y = X$ with a lower probability than in Algorithm 1. Define $w : x \mapsto q(x)/p(x)$. Let's draw $X \sim p$ as in Algorithm 1, but instead of step 2.(a), we set $Y = X$ with probability $\min(C, w(X))$ with $C < 1$. With a reasoning similar to the one establishing the validity of Algorithm 1, we can see that $\{X = Y\}$ then occurs with probability $\int \min(Cp(x), q(x))dx \leq \int p \wedge q$. Furthermore, Y conditional on $\{X = Y\}$ follows a distribution with density $\min(Cp(x), q(x))/\int \min(Cp(x'), q(x'))dx'$. Thus, we can replace the acceptance probability in step 2.(b) by an acceptance probability corresponding to a rejection sampler using q as a proposal and targeting a distribution with density proportional to $y \mapsto q(y) - \min(Cp(y), q(y))$. Namely, we can repeatedly sample $Y^\star \sim q$ and $W^\star \sim \text{Uniform}(0, 1)$ until $W^\star > Cp(Y^\star)/q(Y^\star)$, at which point we set $Y = Y^\star$. Computing the variance of the cost, we could see that it is now bounded by a finite constant that depends on C but not on p and q.

2.2.4 Maximal coupling as a proof technique

Our focus here is on couplings as building blocks of Monte Carlo methods, but most of the original interest in couplings was in establishing theoretical results, such as inequalities or limits relating random variables. An example [Thorisson, 1995, Pollard, 2002, Chapter 10] is the use of couplings to formalize the Poisson approximation to the Binomial distribution. Let X be Binomial(λ, N), so X is the sum of N independent Bernoulli(λ), denoted by $X_1, \ldots, X_N$. Let Y be Poisson(λN), that is equal to the sum of N independent Poisson(λ) variables denoted by $Y_1, \ldots, Y_N$, with pmf $k \mapsto \lambda^k e^{-\lambda}/k!$. Note that we have $X_n \in \{0, 1\}$ and $Y_n \in \{0, 1, 2, \ldots\}$, so they can take common values equal to 0 or 1. Next couple X_n and Y_n maximally. The "overlap mass" is equal to $(1 - \lambda) \wedge e^{-\lambda} + \lambda \wedge \lambda e^{-\lambda}$, summing over the two possible cases for the value of X_n, and this simplifies to $1 - \lambda + \lambda e^{-\lambda}$. Thus $\mathbb{P}(X_n \neq Y_n) = \lambda(1 - e^{-\lambda})$, and

$$\mathbb{P}(X \neq Y) \leq \sum_{n \in [N]} \mathbb{P}(X_n \neq Y_n) \leq N\lambda(1 - e^{-\lambda}).$$

The first inequality comes from a repeated application of $\mathbb{P}((X_1, X_2) \neq (Y_1, Y_2)) \leq \mathbb{P}(X_1 \neq Y_1) + \mathbb{P}(X_2 \neq Y_2)$, which itself comes from $\{(X_1, X_2) \neq (Y_1, Y_2)\}$ being the union of $\{X_1 \neq Y_1\}$ and $\{X_2 \neq Y_2\}$, and the "union bound". Using the inequality $1 - e^{-\lambda} \leq \lambda$ and replacing λ by λ/N we finally obtain

$$\|\text{Binomial}(N, \lambda/N) - \text{Poisson}(\lambda)\|_{\text{TV}} \leq \lambda^2/N,$$

which is one way of formalizing a Poisson approximation of a Binomial distribution in the fixed λ and large N regime.

2.3 Common random numbers and optimal transport

Couplings appear naturally in the context of variance reduction of Monte Carlo estimators, which has inspired many articles on "common random numbers" strategies. These are closely related to the broader topic of "optimal transport" between probability distributions, which, in turn, inspires various Monte Carlo strategies and provides technical tools for the analysis of MCMC algorithms.

2.3.1 Common random numbers

Suppose that the object of interest is the expected difference $\mathbb{E}[X - Y]$ between two real-valued variables X and Y. Perhaps we can draw X, Y independently, and compute $X - Y$ with a variance $\mathbb{V}[X] + \mathbb{V}[Y]$. We could obtain an estimator with a smaller variance if X and Y were positively correlated. Indeed in general $\mathbb{V}[X - Y] = \mathbb{V}[X] + \mathbb{V}[Y] - 2\mathbb{C}\mathrm{ov}(X, Y)$ thus the variance is decreased when $\mathbb{C}\mathrm{ov}(X, Y)$ is increased. If the marginal distributions p and q are fixed, maximizing $\mathbb{C}\mathrm{ov}(X, Y)$ is equivalent to maximizing $\mathbb{C}\mathrm{or}(X, Y)$. So overall, minimizing $\mathbb{V}[X - Y]$ under fixed marginals is equivalent to maximizing $\mathbb{C}\mathrm{or}(X, Y)$, which is also equivalent to minimizing $\mathbb{E}[(X - Y)^2]$. The latter quantity is the expected squared distance between X and Y. So, how do we do that?

Suppose that X and Y are generated using inverse transform sampling: $X = F_X^-(U)$ and $Y = F_Y^-(V)$ where U, V are uniform variables in $[0, 1]$. In this setting, it appears intuitive that positive correlations between U and V leads to positive correlations between X and Y. In fact, we can set $V = U$, thus inducing perfect correlation between U and V. Is it really a good strategy to obtain large correlations between X and Y?

The proposed strategy (setting $V = U$) is an instance of what is commonly called "common random numbers" or "CRN" strategy [Whitt, 1976, Gal et al., 1984, Glasserman and Yao, 1992]. It can be traced back to the dawn of humanity. Among various results obtained on CRN strategies, note first that the answer to the above question is affirmative. Among related questions stands the maximization of $\mathbb{E}[\psi(X, Y)]$ over all couplings of X and Y where ψ is supermodular and right-continuous [Cambanis et al., 1976, Glasserman and Yao, 1992]. Next we will describe the relation between common random numbers and optimal transport (and thus Monge–Kantorovich, Wasserstein distances, etc), as this relation extends the above reasoning to d-dimensional spaces and provides connections to other topics as well.

Consider again the use of a single $U \sim \mathrm{Uniform}(0, 1)$ in the construction of both $X = F_X^-(U)$ and $Y = F_Y^-(U)$. That strategy defines a joint distribution on $\mathbb{R} \times \mathbb{R}$ by "pushing forward" the uniform distribution on $[0, 1]$ through the function $u \mapsto (F_X^-(u), F_Y^-(u))$. If F_X is a continuous function, recall that $F_X(X)$ is uniformly distributed in $[0, 1]$. We can then view the pair (X, Y) in this slightly different way: $Y = F_Y^- \circ F_X(X)$, therefore is a deterministic function of X. The function $T_{\mathrm{opt}} = F_Y^- \circ F_X$ will be called the optimal transport map between p

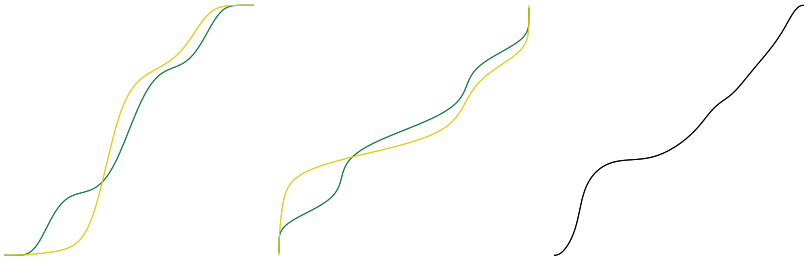

Figure 2.4: Cumulative distribution functions F_X and F_Y, their inverses, and optimal transport map $F_Y^- \circ F_X$.

and q. See Figure 2.4.

The map $T_{\text{opt}} = F_Y^- \circ F_X$ and the coupling $\gamma_{\text{opt}} = (F_X^-, F_Y^-)_\sharp \text{Uniform}(0,1)$ are solutions respectively of the Monge problem

$$\inf_{T : T_\sharp p = q} \mathbb{E}_{X \sim p} \left[c(X, T(X)) \right],$$

and the Kantorovich problem

$$\inf_{\gamma \in \text{coupling}(p,q)} \mathbb{E}_{X,Y \sim \gamma}[c(X,Y)],$$

which coincide sometimes, and in particular when $\mathbb{X} = \mathbb{R}$ and when F_X, F_Y are continuous. These are "optimal transport" problem, in that they formalize the question of finding the optimal way of moving the "mass" spread out according to a distribution p into another distribution q. For Monge, this was literally the question of optimally moving piles of soil, from a set of source locations to another set of destinations. The cost of moving some soil can be defined as the product of the amount of mass being moved, times the distance over which the mass is transported. In the above problem, p would be the distribution of sources, q the distribution of destinations, and the total mass of soil to move would be equal to one.

The "cost" $c(x,y)$ can be taken as $(x-y)^2$, in which case we look again at the problem of maximizing $\mathbb{C}\text{ov}(X,Y)$ following the above reasoning. The Monge version of the problem is about finding a map T such that $T(X)$ follows q whenever $X \sim p$ while minimizing the total "effort" $\int c(x, T(x)) p(dx)$ of moving the mass from p to q. The Kantorovich problem is a "relaxation", in that it has solutions in various cases where Monge's problem has none. Kantorovich asks for the least costly "plan", where a plan describes how mass should be moved from some infinitesimally small location x to another y, or vice versa, in order to move the total mass from a configuration p to a configuration q, or vice versa. In the Kantorovich formulation the mass located at x can be "split" and moved to different locations, whereas in Monge's problem the mass at x must be moved to the unique location $T(x)$; thus the Kantorovich version is completely symmetric in p and q, whereas Monge's is not.

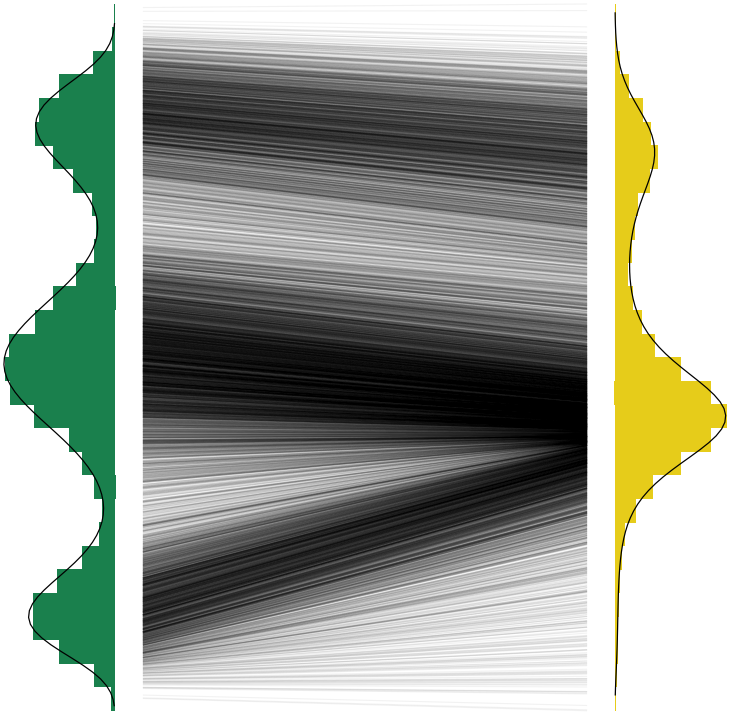

Figure 2.5: Monotone rearrangement of univariate distributions.

In the univariate case $\mathbb{X} = \mathbb{R}$, the "common random numbers" coupling is optimal for any cost $c(x, y)$ that can be written as $h(y-x)$ for a convex function h. And if h is strictly convex, that coupling is the only solution [Santambrogio, 2015, Chapter 2]. For example, the same pair $(F_X^-(U), F_Y^-(U))$ minimizes $\mathbb{E}[(X-Y)^k]$ over all couplings (X, Y) of p and q, for all even integer $k \geq 2$. See Figure 2.5 for an illustration of the "optimal rearrangement" associated with a pair of distributions.

2.3.2 Monge-Kantorovich-Wasserstein distances

On top of the connection with CRN, the Monge-Kantorovich problem appears useful in the Monte Carlo literature through its association with distances between probability distributions, often called Wasserstein distances. In a general metric space $\mathbb{X}$ with "ground distance" $(x, y) \mapsto d(x, y)$, the k-Wasserstein distance between p and q is

$$W_k(p, q) = \left(\inf_{(X,Y) \in \text{coupling}(p,q)} \mathbb{E}\left[d(X, Y)^k \right] \right)^{1/k},$$

for distributions p and q for which $\int d(x, x_0)^k p(dx) < \infty$ for at least one x_0, and likewise for q. Here we will mention a few properties of that distance that resonate with properties of the total variation distance.

First, let us go back to the one-dimensional case, with $d(x, y) = |x - y|$. There we have an explicit solution of the transport problem for any $k \geq 1$, given by $X = F_X^-(U), Y = F_Y^-(U)$ for $U \sim \text{Uniform}(0, 1)$. This can be used to obtain

$$W_k^k(p, q) = \int_{\mathbb{R}} |F_X(t) - F_Y(t)|^k dt,$$

which is an explicit expression in terms of "integrated difference between cumulative distribution functions". This contrast with the "area between the probability density function" view of the total variation distance. See Figure 2.4.

Back to the d-dimensional case, but focusing on $k = 1$, we can show that

$$W_1(p, q) = \sup_{f \in 1-\text{Lipschitz}} \int f(x)p(dx) - \int f(y)q(dy),$$

i.e. the 1-Wasserstein distance is the infimum of $\mathbb{E}[f(X) - f(Y)]$ over all measurable 1-Lipschitz functions. These are functions f such that $|f(x) - f(y)| \leq d(x, y)$ for all $x, y \in \mathbb{X}$. This is to be compared with an infimum over all measurable functions bounded by one for the total variation. The above representation of W_1 comes from "Kantorovich-Rubinstein" duality [Villani, 2008, Chapters 5-6] and W_1 is sometimes called the Kantorovich-Rubinstein distance.

The distance W_k metrizes the weak convergence on the space of probability measures for which it is defined, i.e. $p_n \to p$ weakly if and only if $W_k(p_n, p) \to 0$. This is not the case for the total variation distance; for instance the empirical measure supported by n independent draws from a continuous distribution with finite first moment converges weakly, and thus in W_k if enough moments are finite, but does not converge in total variation.

In general there is no explicit construction for the solution of an optimal transport problem, but there are numerous numerical approaches [Peyré and Cuturi, 2019]. In the case of discrete state spaces (or distributions on a continuous space but with discrete support), the optimization problem can be solved exactly using linear programming techniques. Indeed, introducing a coupling matrix Γ of size $K \times K$, with p and q being vectors of size K, the cost can be represented by a $K \times K$ matrix C with $C_{i,j}$ being the cost associated with the "locations" (i, j), the transport problem becomes

$$\inf_{\Gamma \in \text{coupling}(p,q)} \sum_{i,j} C_{i,j}\Gamma_{i,j},$$

where the minimization is over all matrices Γ with rows summing to p and columns summing to q. The objective is linear in Γ: this is a "linear program". Solvers for such programs have a worst-case computing cost of the order of K^3. Since the definition of the program includes a cost matrix C with K^2 entries, there is little hope for any general solution with a cost inferior to K^2, that is the

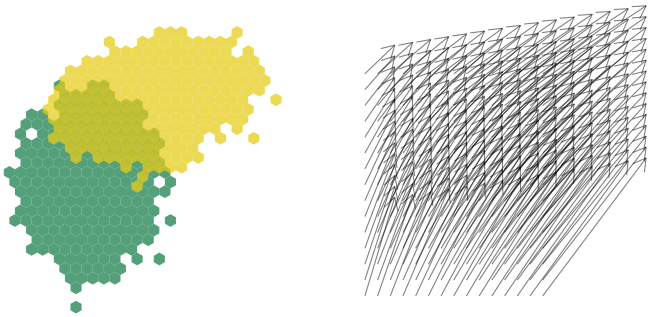

Figure 2.6: Optimal transport between two Normal distributions.

cost of reading each entry of C once. Approximate solutions, for example based on entropic regularization, have received a lot of attention lately and indeed have a cost of order K^2 [Peyré and Cuturi, 2019, Chapter 4] and various appealing properties.

2.3.3 Transport and Monte Carlo methods

Let us consider again the function $T_{\text{opt}} = F_Y^- \circ F_X$. Given a sample $X \sim p$, applying that map yields a sample $Y \sim q$. Next imagine that p is a distribution that is easy to sample from and that q is a distribution that we wish to approximate. If we have access to T_{opt} we can transform draws from p into draws from q by application of the map.

Optimal transport maps are easy to come up with if the two distributions are just translated or scaled versions of each other. Thus, there is an explicit optimal transport map between two Normal distributions, with the ground cost $c(x, y) = \|x - y\|_2^2$. For $p = \mathcal{N}(m, \Sigma)$ and $q = \mathcal{N}(\tilde{m}, \tilde{\Sigma})$, Peyré and Cuturi [2019, equation (2.40)] gives

$$T_{\text{opt}} : x \mapsto \tilde{m} + A(x - m),$$
$$\text{with} \quad A = \Sigma^{-1/2} \left(\Sigma^{1/2} \tilde{\Sigma} \Sigma^{1/2} \right)^{1/2} \Sigma^{-1/2},$$

which is a linear map. The transport cost itself has an explicit form,

$$W_2^2(\mathcal{N}(m, \Sigma), \mathcal{N}(\tilde{m}, \tilde{\Sigma})) = \|m - \tilde{m}\|_2^2 + \left\{ \text{trace} \left(\Sigma + \tilde{\Sigma} - 2 \left(\Sigma^{1/2} \tilde{\Sigma} \Sigma^{1/2} \right)^{1/2} \right) \right\}^2.$$

See Figure 2.6.

Apart from the Gaussian case, it is rare for an optimal transport map to be directly available in multivariate settings. There is always a transport map, that is not optimal but still related to optimal transport, called Knothe–Rosenblatt rearrangement [Carlier et al., 2010]. Let us describe it in the bivariate setting,

where $X = (X_1, X_2)$ and $Y = (Y_1, Y_2)$. We can optimally transport X_1 using the map $T_1 = Q_1^- \circ P_1$, the composition of the cdf of X_1 with the inverse cdf of Y_1. Given a sample X_1, we define $Y_1 = T_1(X_1)$. Then, we introduce $T_{2|1}(\cdot|X_1, Y_1) = Q_{2|1}^-(\cdot|Y_1) \circ P_{2|1}(\cdot|X_1)$, which corresponds to an optimal transport between the conditional distributions of X_2 given X_1 and that of Y_2 given Y_1. Then we define $Y_2 = T_{2|1}(X_2|X_1, Y_1)$. Since $Y_1 = T_1(X_1)$, the map $T_{2|1}$ is in fact a function of X_1, X_2 only. More generally this construction can be extended to define a function from $X_1, \ldots, X_d$ to $Y_1, \ldots, Y_d$, which is "triangular": T_1 depends only on X_1, T_2 only on X_1, X_2, etc. If p and q are absolutely continuous, it can be shown that this is in fact the unique lower triangular map transporting p to q.

In general, if we sample $X \sim p$ and define $Y = T(X)$ and if T is a diffeomorphism, the distribution of Y has density $y \mapsto p(T^{-1}(y))|\det \nabla T^{-1}(y)|$. This might give some clues on how to come up with a map T in such a way that Y follows q exactly, or approximately. If one wants to correct for the difference between the marginal distribution of Y and q, for example by importance sampling, it would be convenient to choose T in such a way that $\det \nabla T$ can be computed for a moderate cost.

Some references on the use of deterministic transformations for Monte Carlo purposes include Meng and Schilling [2002] or Chorin et al. [2010]. One way of coming up with transport maps is to parametrize a class of maps and then to optimize some criterion [El Moselhy and Marzouk, 2012, Marzouk et al., 2016]. The latter articles consider triangular maps, so that $\det \nabla T$ is easier to compute "by design". Among a class of parameterized maps $T(\cdot, \theta)$, we can distinguish the one that solves

$$\min_\theta \mathrm{KL}\left(T(\cdot, \theta)_\sharp p, q\right)$$
$$\text{such that } \det \nabla T(\cdot, \theta) > 0.$$

The objective is defined with numerical convenience in mind, in cases where $\bar{q}$ is only available through its unnormalized density function.

Another Monte Carlo method related to transport ideas is proposed in Heng et al. [2015]. There the construction of maps comes from flows [see also Reich, 2011], which are solutions of differential equations. There are various differential equations associated with transport maps. In Heng et al. [2015] the proposal is to define a flow that "moves particles using a velocity field designed to track the full conditional distributions", called a "Gibbs flow". These moves can be computed by solving one-dimensional integrals only, and the appropriate determinants can be computed as well. The generated samples are then weighted in importance sampling steps.

These uses of transport maps are also related to "normalizing flows" [Tabak and Turner, 2013]. The authors write that "normalizing the data x_j is finding a map $y(x)$ such that $y_j = y(x_j)$ have a prescribed distribution". The context in Tabak and Turner [2013] is that of density estimation for $x \sim \rho$, and μ is prescribed as a standard Gaussian. Given a map T we could estimate ρ by

$T^{-1} \circ \mu$. This type of ideas has been used for various statistical and computational purposes, see Papamakarios et al. [2019] for a review.

2.4 Prokhorov distance and couplings

There is another distance between probability distributions with a coupling interpretation: the Prokhorov distance, sometimes called "Prohorov" or Lévy-Prokhorov; see Dudley [Chapter 11, 2002], Pollard [Chapter 10 2002] or Billingsley [Section 6 2013]. It is somewhat of a "mix" between total variation and Wasserstein distances, in the sense that it relies on a choice of metric but does not assume the finiteness of moments. Consider again a metric space with distance $(x, y) \mapsto d(x, y)$. For a set A, denote by A^ε the union $\cup_{x \in A} B(x, \varepsilon)$ of balls centered at $x \in A$ with radius ε. The set A^ε is called the ε-neighborhood of the set A. The Prokhorov distance between p and q is defined as

$$\Delta(p, q) = \inf \{\varepsilon > 0 : p(A) \leq q(A^\varepsilon) + \varepsilon \quad \forall A \in \mathcal{X}\}.$$

It does not look very symmetrical but it is; or we could define it as the smallest ε such that both $p(A) \leq q(A^\varepsilon) + \varepsilon$ and $q(A) \leq p(A^\varepsilon) + \varepsilon$. One can check that $\Delta(p, q)$ is a distance between probability measures. Furthermore, if $\Delta(p_n, p) \to 0$ then $p_n \to p$ weakly. If the state space is separable and $p_n \to p$ then $\Delta(p_n, p) \to 0$. In other words, on metric separable spaces the Prokhorov distance "metrizes" weak convergence. Weak convergence of p_n to p is equivalent to $\Delta(p_n, p) \to 0$.

The following coupling theorem is due to Strassen 1965; we state the result as in Pollard 2002, Chapter 10. Let $\varepsilon, \varepsilon' > 0$. Then there exists a pair (X, Y) marginally following (p, q) such that $\mathbb{P}(d(X, Y) > \varepsilon) \leq \varepsilon'$, if and only if $p(A) \leq q(A^\varepsilon) + \varepsilon' \quad \forall A \in \mathcal{X}$. In Billingsley 2013, Section 6 this is phrased as follows, with a single "epsilon". Let $\varepsilon > 0$, then $\Delta(p, q) \leq \varepsilon$ if and only if there exists (X, Y) marginally following (p, q) such that $\mathbb{P}(d(X, Y) > \varepsilon) \leq \varepsilon$. See also Whitt 1976.

2.5 Summary and references

There are infinitely many couplings of two variables $X \sim p$ and $Y \sim q$. The variables X, Y can jointly satisfy some properties, such as: $\mathbb{P}(X = Y)$ is maximal, or $\mathbb{E}[\|X - Y\|^2]$ is minimal, or $Y = T(X)$ for some function T... and this makes certain couplings remarkable. Some couplings are intrinsically related to various notions of distance between probability distributions, such as the total variation, the Monge-Kantorovich-Wasserstein or the Prokhorov distance. Coupling is a useful proof technique, sometimes used to establish asymptotic approximations, or inequalities or properties of random variables. In later chapters we will see how couplings can be used to study the convergence of some MCMC algorithms.

Certain couplings can also be simulated, depending on available information about the distirbutions p and q. Such simulations can then be exploited in methods, and in Monte Carlo algorithms in particular. In later chapters we will

see examples of such methods, for example a perfect sampling method called "coupling from the past" [Propp and Wilson, 1996]. Before that we need to discuss the concept of couplings in the context of Markov chains.

Detailed treatments of coupling methods in probability tend to be more theoretical than applied, which is partly the motivation for writing the present notes. They include Lindvall [2002], Thorisson [2000] and den Hollander [2012]. An accessible overview is given in Thorisson [1995]. Common random number strategies are mentioned in most books on Monte Carlo, for instance Owen [Chapter 8, 2013], while it falls under the umbrella of "antithetic variables" in Robert and Casella [Chapter 3, 1999]. Textbooks on optimal transport include Santambrogio [2015], Peyré and Cuturi [2019] which are perhaps more accessible to applied mathematicians and statisticians than the reference in mathematics, Villani [2008].

Bibliography

P. Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013. 20

N. Bou-Rabee, A. Eberle, and R. Zimmer. Coupling and convergence for Hamiltonian Monte Carlo. *arXiv preprint arXiv:1805.00452*, 2018. 12

S. Cambanis, G. Simons, and W. Stout. Inequalities for Ek (X, Y) when the marginals are fixed. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 36(4):285–294, 1976. 14

G. Carlier, A. Galichon, and F. Santambrogio. From Knothe's transport to Brenier's map and a continuation method for optimal transport. *SIAM Journal on Mathematical Analysis*, 41(6):2554–2576, 2010. 18

A. Chorin, M. Morzfeld, and X. Tu. Implicit particle filters for data assimilation. *Communications in Applied Mathematics and Computational Science*, 5(2): 221–240, 2010. 19

F. den Hollander. Probability theory: The coupling method. *Lecture notes available online (http://websites. math. leidenuniv. nl/probability/lecturenotes/-CouplingLectures. pdf)*, 2012. 21

R. M. Dudley. *Real analysis and probability*. Cambridge University Press, 2002. 20

A. Eberle, A. Guillin, and R. Zimmer. Couplings and quantitative contraction rates for Langevin dynamics. *The Annals of Probability*, 47(4):1982–2010, 2019. 12

T. A. El Moselhy and Y. M. Marzouk. Bayesian inference with optimal maps. *Journal of Computational Physics*, 231(23):7815–7850, 2012. 19

S. Gal, R. Rubinstein, and A. Ziv. On the optimality and efficiency of common random numbers. *Mathematics and computers in simulation*, 26(6):502–512, 1984. 14

P. Glasserman and D. D. Yao. Some guidelines and guarantees for common random numbers. *Management Science*, 38(6):884–908, 1992. 14

J. Heng, A. Doucet, and Y. Pokern. Gibbs flow for approximate transport with applications to Bayesian computation. *arXiv preprint arXiv:1509.08787*, 2015. 19

V. E. Johnson. A coupling-regeneration scheme for diagnosing convergence in Markov chain Monte Carlo algorithms. *Journal of the American Statistical Association*, 93(441):238–248, 1998. 6

T. Lindvall. *Lectures on the coupling method*. Dover Books on Mathematics, 2002. 1, 9, 21

S. Malefaki and G. Iliopoulos. Simulating from a multinomial distribution with large number of categories. *Computational statistics & data analysis*, 51(12): 5471–5476, 2007. 2

Y. Marzouk, T. Moselhy, M. Parno, and A. Spantini. *Sampling via Measure Transport: An Introduction*, pages 1–41. Springer International Publishing, Cham, 2016. ISBN 978-3-319-11259-6. doi: 10.1007/978-3-319-11259-6_23-1. URL https://doi.org/10.1007/978-3-319-11259-6_23-1. 19

X.-L. Meng and S. Schilling. Warp bridge sampling. *Journal of Computational and Graphical Statistics*, 11(3):552–586, 2002. 19

R. B. Nelsen. *An introduction to copulas*. Springer Science & Business Media, 2007. 3

A. B. Owen. *Monte Carlo theory, methods and examples*. 2013. 21

G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *arXiv preprint arXiv:1912.02762*, 2019. 20

G. Peyré and M. Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019. 17, 18, 21

D. Pollard. *A user's guide to measure theoretic probability*, volume 8. Cambridge University Press, 2002. 13, 20

J. G. Propp and D. B. Wilson. Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures & Algorithms*, 9 (1-2):223–252, 1996. 21

S. Reich. A dynamical systems framework for intermittent data assimilation. *BIT Numerical Mathematics*, 51(1):235–249, 2011. 19

C. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 1999. 21

F. Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015. 16, 21

A. Sklar. Fonctions de répartitions à n dimensions et leurs marges. *Publ. Inst. Statistique Univ. Paris 8*, pages 229–231, 1959. 3

V. Strassen. The existence of probability measures with given marginals. *The Annals of Mathematical Statistics*, 36(2):423–439, 1965. 20

E. G. Tabak and C. V. Turner. A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164, 2013. 19

H. Thorisson. Coupling methods in probability theory. *Scandinavian journal of statistics*, pages 159–182, 1995. 13, 21

H. Thorisson. *Coupling, stationarity, and regeneration*, volume 14. Springer New York, 2000. 1, 4, 6, 21

C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008. 17, 21

W. Whitt. Bivariate distributions with given marginals. *The Annals of statistics*, 4(6):1280–1289, 1976. 14, 20