Brief lecture notes on Markov random fields

for a graduate course given in the spring of 2014[1]

Olle Häggström

These notes do not convey the full content of the course, but are merely meant as a complement to [GHM] and [W].

1 Wednesday, March 19

We'll adhere as far as possible to the notation laid down in Section 3.1 of Winkler [W], but I have some reservations regarding parts of his terminology. In particular, for a finite index set S, a finite set $\mathbf{X}_s$ of attainable values at each $s \in S$, and $\mathbf{X} = \prod_{s \in S} \mathbf{X}_s$, he defines a **random field** as a probability measure Π on $\mathbf{X}$ satisfying

$$\Pi(x) > 0 \text{ for all } x \in \mathbf{X}\,. \tag{1}$$

Spontaneously I'd call Π a random field regardless of whether or not (1) holds. In many cases (1) is a very useful assumption, but to require it for a probability measure on $\mathbf{X}$ to qualify as a random field seems to me unnatural. I'll sometimes consider examples violating (1), unabashedly calling them random fields. And whenever (1) is needed I'll emphasize it explicitly, sometimes calling it *Winkler's positivity condition* (a bit sloppily, as Winkler is far from the first to employ it).

*

Please pay attention to Definitions 3.1.1 (of neighborhood systems and cliques) and 3.1.2 (of Markov fields). Winkler's defining property of Markov fields in Definition 3.2.1 is what I would call the **local** Markov property. More generally, one can ask, for any $A \subset S$, whether

$$\Pi(X_A = x_A \,|\, X_{S\setminus A} = x_{S\setminus A}) \;=\; \Pi(X_A = x_A \,|\, X_{\partial(A)} = x_{\partial(A)}) \tag{2}$$

holds for all $x \in \mathbf{X}$, where $\partial(A) = \bigcup_{s\in A} \partial(s) \setminus A$. I propose the following terminology.

[1] See also `http://www.math.chalmers.se/~olleh/MarkovRandomFieldsVT2014.html`

- If (2) holds for all *singletons* $A = \{s\}$, then we say Π satisfies **the local Markov property**.
- If (2) holds for all *finite* $A \subset S$, then we say Π satisfies **the regional Markov property**.
- If (2) holds for *all* $A \subset S$, then we say Π satisfies **the global Markov property**.

At this point, it might seem a bit moronic to distinguish between the regional and global Markov properties, because S is assumed to be finite, so every $A \subset S$ is automatically finite, and the regional and global properties trivially coincide. But have patience, later in the course we will move on to countably infinite S, and then the distinction will be real.[2] In any case, we have, trivially, that the global Markov property implies the regional, and the regional implies the local. What about the other directions?

We will see in a later lecture that if we assume Winkler's positivity condition, then the local Markov property does imply the regional Markov property, while without the positivity assumption we'll see a counterexample to the hoped-for implication. As to the regional Markov property implying the global, we'll see in the setting of countably infinite S that there are counterexamples (even assuming natural extensions of Winkler's positivity to that setting).

*

The Ising model (Example 3.1.2) is defined as follows. Fix a finite S and a neighborhood structure ∂, and let $S_s = \{-1, +1\}$ for each $s \in S$, so that $X = \{-1, +1\}^S$. (For concreteness, we may, e.g., take S to be a square grid $\{0, 1, \ldots, n\}^2$, with edges connecting sites at Euclidan distance 1 from each other.) For fixed $\beta > 0$ (the so-called inverse temperature parameter), the **energy** $H(x)$ of a configuration $x \in \mathbf{X}$ is defined as

$$H(x) = -\beta \sum_{\langle s,t \rangle} x_s x_t \tag{3}$$

where $\langle s, t \rangle$ means that we sum over all neighboring pairs of sites in S, counting each such pair once. We then define a probability measure Π on $\mathbf{X}$

[2] Note, however, that when S is infinite, then $\mathbf{X}$ will be uncountable, so that most $x \in \mathbf{X}$ will get probability 0, and we need to take some care with conditional probabilities, e.g., by writing "If Π admits conditional probabilities such that (2) holds..." in place of "If (2) holds...".

by setting, for each $x \in \mathbf{X}$,

$$\Pi(x) = \frac{1}{Z} \exp(-H(x)) \tag{4}$$

where $Z = \sum_{y \in \mathbf{X}} e^{-H(y)}$ is a normalizing constant making the probabilities sum to 1.

Probabilities of the form (4) are called **Gibbs measure**. Other choices of energy function are possible, but with the present choice, we call Π **the Ising model on S and at inverse temperature β.**

Thousands of mathematics papers have been written on the Ising model, and even more physics papers. Yet, it may look odd at first? Why is this a natural choice of probability measure? There are many reasons, I'll offer two:

First, it's a Markov random field. To see this, fix $s \in S$ and $x \in \mathbf{X}$, and consider the odds ratio

$$\begin{aligned}
\frac{\Pi(X_s = +1 \,|\, X_r = x_r, \forall r \neq s)}{\Pi(X_s = -1 \,|\, X_r = x_r, \forall r \neq s)} &= \frac{\Pi(X_s = +1, X_r = x_r, \forall r \neq s)}{\Pi(X_s = -1, X_r = x_r, \forall r \neq s)} \\
&= \frac{\frac{1}{Z}\exp(-H(x \text{ with a } +1 \text{ at } s))}{\frac{1}{Z}\exp(-H(x \text{with a } -1 \text{ at } s))} \\
&= \frac{\exp\left(\beta\left(\sum_{\substack{\langle t,r\rangle \\ t,r\neq s}} x_t r_t + \sum_{t \in \partial(s)} x_t\right)\right)}{\exp\left(\beta\left(\sum_{\substack{\langle t,r\rangle \\ t,r\neq s}} x_t r_t - \sum_{t \in \partial(s)} x_t\right)\right)} \\
&= \exp\left(2\beta\left(\sum_{t \in \partial(s)} x_t\right)\right)
\end{aligned}$$

which only depends on x via its values on $\partial(s)$.

Second, exp(sum) = product, and product means independence (a fundamental building block in almost all probabilistic modelling) so that Gibbs measures with energy function $H(x) = \sum$ exhibit some independence (or, more precisely, conditional independence) structure. We'll see in the Hammersley–Clifford Theorem next week, that **every Markov random field** (in Winkler's sense) **can be written as a Gibbs measure with H equal to a sum over cliques**. Here are a couple of really simple rewrites into Gibbs measures, building up towards the Ising model:

Example 0. Let $\{X_s\}_{s\in S}$ be i.i.d. random variables with

$$\begin{cases} \Pi(X_s = +1) = p \\ \Pi(X_s = -1) = 1 - p\,. \end{cases}$$

For $x \in \mathbf{X} = \{-1, +1\}^S$,

$$\Pi(x) \;=\; p^{(\#+1\text{'s in } x)}(1-p)^{(\#-1\text{'s in } x)} \;=\; \exp(-H(x))$$

where

$$H(x) = -\sum_{s\in S}(\log(p)\mathbf{1}_{\{x_s=+1\}} + \log(1-p)\mathbf{1}_{\{x_s=-1\}})\,.$$

Example 1. Let $S = \{0, 1, \ldots, n\}$, and define $X = (X_0, X_1, \ldots, X_n)$ as a (symmetric, two state) Markov chain with initial value X_0 equal to -1 or $+1$ with probability $1/2$ each, and transition matrix

$$\begin{bmatrix} p & 1-p \\ 1-p & p \end{bmatrix}.$$

Then any given $x \in \mathbf{X}$ has probability

$$\begin{aligned} \Pi(x) &= \frac{1}{2}\prod_{i=1}^{n} p^{\{\mathbf{1}_{x_{i-1}=x_i}\}}(1-p)^{\{\mathbf{1}_{x_{i-1}\neq x_i}\}} \\ &= \cdots \\ &= \frac{(p(1-p))^{n/2}}{2}\exp\left(\sum_{i=1}^{n}\log\left(\sqrt{\frac{p}{1-p}}\right)x_{i-1}x_i\right) \\ &= \frac{(p(1-p))^{n/2}}{2}e^{-H(x)} \end{aligned}$$

with $H(x) = -\log\left(\sqrt{\frac{p}{1-p}}\right)\sum_{i=1}^{n} x_{i-1}x_i$, so this is the Ising model on S (and neighborhood system ∂ where i and j are neighbors whenever $|i-j| = 1$) at inverse temperature $\beta = \log\left(\sqrt{\frac{p}{1-p}}\right)$.

This last example reflects a more general fact that (under mild conditions), a Markov chain is also a Markov random field with a 1-dimensional dependence structure. If we now want to introduce similar interactions on a grid or a more general graph structure, we have the Ising model.

2 Wednesday, March 26

As before, we take S finite, $\mathbf{X}_s$ finite for each $s \in S$, $\mathbf{X} = \prod_{s \in S} \mathbf{X}_s$, and Π a strictly positive probability measure on $\mathbf{X}$. If Π can be written as

$$\Pi(x) = \frac{1}{Z} e^{-H(x)}$$

for a given function $H : \mathbf{X} \to \mathbf{R}$, then Π is said to be a **Gibbs measure for energy function** H**.** Every strictly positive Π is a Gibbs measure for some H, and by adding a constant to H we are even free to choose our own favorite value of the normalizing constant Z. Indeed, defining H by

$$H(x) = -\log(\Pi(x)) - \log(Z)$$

gives

$$\begin{aligned} \frac{1}{Z} e^{-H(x)} &= \frac{1}{Z} e^{\log(\Pi(x)) + \log(Z)} \\ &= \frac{1}{Z} \Pi(x) Z \; = \; \Pi(x) \, . \end{aligned}$$

Hence, being a Gibbs measure is in itself not a remarkable property. More interesting is if Π is a **neighbor Gibbs measure** for a given neighborhood system ∂, meaning that

$$H(x) = \sum_C U_c(x)$$

where the sum ranges over cliques $C \subset S$ for ∂, and $U_C(x)$ depends on $x \in \mathbf{X}$ only via $\{x_s\}_{s \in C}$.

Part of Proposition 3.2.1 in [W]: *If* Π *is a neighbor Gibbs measure for the neighborhood system* ∂*, then* Π *satisfies the local Markov property for the same* ∂*.*

To prove this, it suffices to show that for any $s \in S$, $x \in \mathbf{X}$ and $y_s, z_s \in \mathbf{X}_s$, the odds ratio

$$\frac{\Pi(X_s = y_s \,|\, X_r = x_r, \forall r \neq s)}{\Pi(X_s = z_s \,|\, X_r = x_r, \forall r \neq s)}$$

depends on x only via $x_{\partial(s)}$. To do this, proceed as in the proof of the local Markov property for the Ising model in my first lecture – and enjoy all the cancellation! (Or see [W], p 55–56.)

A much deeper result (in my view) is the following partial converse:

The Hammersley–Clifford Theorem (Part of Thm 3.3.2 in [W]): *If* Π *is a (strictly positive) Markov random field for* ∂*, then it is also a neighbor Gibbs measure for* ∂*.*

See [W] for the proof, which yields an explicit formula for $U_C(x)$. It involves a series of calculations, and proceeds via two other results – Lemma 3.3.1 (the Möbius Inversion Formula) and Theorem 3.3.1.

3 Friday, March 28

Staying as before in the finite setting (with both S and $\{bfX\}$ finite), recall from Lecture 1 my definitions of *local* versus *regional* Markov properties. The regional Markov property trivially implies the local, but how about the other direction? This lecture will be devoted to proving the following result.

Theorem L3:

(a) *Under Winkler's positivity condition, the local Markov property implies the regional.*

(b) *Without Winlker's positivity, there are conterexamples to show that the local markov property does not imply the regional.*

Part (b) is relatively the easier prat, so let's begin with that.

Proof of Thm L3 (b): Set $S = \{s_1, s_2, s_3, s_4, s_5\}$, $\mathbf{X}_s = \{0, 1\}$ for each $s \in S$, and defined the neighborhood system ∂ in such a way that $\langle s_1, s_2\rangle$, $\langle s_1, s_3\rangle$, $\langle s_2, s_3\rangle$, $\langle s_3, s_4\rangle$, $\langle s_3, s_5\rangle$ and $\langle s_5, s_6\rangle$ are neighbors (draw the graph - a bowtie!). Define Π as the probability measure on $\mathbf{X}$ with

$$\begin{cases} \Pi(0,0,0,0,0) = \frac{1}{4} \\ \Pi(0,0,1,0,0) = \frac{1}{4} \\ \Pi(1,1,0,1,1) = \frac{1}{4} \\ \Pi(1,1,1,1,1) = \frac{1}{4} \\ \Pi(x) = 0 \text{ for all other } x \in \mathbf{X} \end{cases}$$

The local Markov property is easy to check: $\Pi(X_{s_3} = 0 | X_{S\setminus s_3} = x_{S\setminus s_3}) = \frac{1}{2}$ regardless of $x \in \mathbf{X}$; while

$$\Pi(X_{s_1} = 0 | X_{S\setminus s_1} = x_{S\setminus s_1}) = \begin{cases} 1 & \text{if } x_{s_2} = 0 \\ 0 & \text{if } x_{s_2} = 1 \end{cases}$$

which only depends on x via $x_{\partial(s_1)}$; and similarly for X_{s_2}, X_{s_4} and X_{s_5}. So the local Markov property holds.

On the other hand, take $A = \{s_1, s_2\}$ so that $\partial(A) = \{s_3\}$, and note that

$$\Pi(X_A = (0,0) | X_{s_3,s_4,s_5} = (0,0,0)) = 1$$

while

$$\Pi(X_A = (0,0) | X_{s_3} = 0) = \frac{1}{2}$$

so that the regional Markov property fails. $\diamond$

As to Theorem L3 (a), this can be proved using the methods involving Möbius Inversion discussed in Lecture 2, and is in fact part of Winkler's Theorem 3.3.2. I'll offer a completely different proof, which proceeds via coupling of Markov chains – a method I find more illuminating and therefore preferable, although I realize this may be mostly a matter of taste.

Proof of Thm L3 (a): Fix S, $\mathbf{X}$, ∂ and a distribution Π on $\mathbf{X}$ satisfying both the local Markov property and Winkler's positivity. Also fix $A \subset S$, and $x, x' \in \mathbf{X}$ such that $x_{\partial(A)} = x'_{\partial(A)}$. We need to show that

$$\Pi(X_A = x_A | X_{S\setminus A} = x_{S\setminus A}) = \Pi(X_A = x_A | X_{S\setminus A} = x'_{S\setminus A})\,. \tag{5}$$

A small piece of extra notation will be convenient: let $\Pi_{|x_{S\setminus A}}$ denote Π conditioned on taking values x on $S \setminus A$, and define $\Pi_{|x'_{S\setminus A}}$ analogously.

We will define two $\mathbf{X}$-valued Markov chains $(X_0, X_1, \ldots)$ and $(X'_0, X'_1, \ldots)$, designed in such a way that for every k,

$$\begin{cases} X(k) \text{ has distribution } \Pi_{|x_{S\setminus A}} \\ X'(k) \text{ has distribution } \Pi_{|x'_{S\setminus A}} \end{cases}$$

To this end, we pick the initial values $X(0)$ and $X'(0)$ with these respective distributions, and let the two chains evolve according to transition mechanisms that preserve their respective distributions. Namely, at each time $k \geq 1$, select $s \in A$ at random (uniformly), and let

$$\begin{cases} X_s(k) = & \text{a new value chosen according to } \Pi \\ & \text{conditioned on agreeing with} \\ & X(k-1) \text{ on } S \setminus s \\ X_t(k) = & X_t(k-1), \forall t \in S \setminus s\,. \end{cases}$$

The $(X'_0, X'_1, \ldots)$ chain will in fact have the same transition kernel, choosing $s \in A$ at random (uniformly) and letting

$$\begin{cases} X'_s(k) = & \text{a new value chosen according to } \Pi \\ & \text{conditioned on agreeing with} \\ & X'(k-1) \text{ on } S \setminus s \\ X'_t(k) = & X'_t(k-1), \forall t \in S \setminus s\,. \end{cases}$$

(This Markov chain transition kernel is a variant of the so-called **Gibbs sampler** for Π; see Section 5.1 in [W].)

This defines the two chains separately, but we will *couple* them, i.e., run them jointly on the same probability space, and then we need to specify their interdependence. First, pick the initial values $X(0)$ and $X'(0)$ *independently*. Second, at each time k, let the two chains pick *the same* $s \in A$ to update. Third, the new values $X_s(k)$ and $X'_s(k)$ are chosen

$$\begin{cases} \text{to be } \textit{identical} & \text{if } X_{\partial(s)}(k-1) = X_{,\partial(s)}(k-1) \text{ (this is possible by} \\ & \text{the assumed local Markov property of } \Pi) \\ \textit{independently} & \text{otherwise.} \end{cases} \tag{6}$$

The great thing about this rule is that

$$\begin{array}{l} \text{as soon as the two chains coincide on } A \text{ (and thus on } A \cup \partial(A)), \\ \text{they will do so forever more.} \end{array} \tag{7}$$

And they will almost surely do so, eventually. One way to see this by noting that if on $|A|$ consequtive updates, the choices of s happen to scan through all of A, and if each time the two chains happen to pick the same value at s, then they will coincide on A at the end of the scan. The event that such a succesful turn of events happens during times $1, \ldots, k$ is easily seen to have probability at least

$$\left(\frac{\delta^2}{|A|}\right)^{|A|}, \tag{8}$$

where

$$\delta = \min_{s \in A} \min_{x \in \mathbf{X}} \Pi(X_s = x_s | X_{S \setminus s} = x_{S \setminus s})$$

which is > 0 since we assumed Winkler's positivity. The probability in (8) may be a very small, yet strictly positive, and the point is that if the event happens to fail during times $1, \ldots, k$, then it has another chance at times $k+1, \ldots, 2k$, and another at times $2k+1, \ldots, 3k$, and so on. The probability of seeing such a coalescence by time km is therefore

$$1 - \left(1 - \left(\frac{\delta^2}{|A|}\right)^{|A|}\right)^m$$

which tends to 1 as $m \to \infty$. Hence, in combination with (7), we get

$$\lim_{k \to \infty} P(X_A(k) \neq X'_A(k)) = 0.$$

It follows that for any configuration $y_a \in \mathbf{X}_A$ we have

$$\lim_{k\to\infty} |P(X_A(k) = y_A) - P(X'_A(k) = y_A)| = 0,$$

i.e. for any $\varepsilon > 0$, there is a $k < \infty$ such that

$$|P(X_A(k) = y_A) - P(X'_A(k) = y_A)| < \varepsilon\,.$$

But since the chains are stationary, this means we also have

$$|P(X_A(0) = y_A) - P(X'_A(0) = y_A)| < \varepsilon\,,$$

and since $\varepsilon > 0$ and $y_A \in \mathbf{X}_A$ were arbitrary, we get that the distributions of $X_A(0)$ and $X'_A(0)$ coincide. We thus have (5), as desired. $\diamond$

*

Exercise: Theorem 3.3.1 (b) in [W] states that if Π and Π' are strictly positive probability distributions on $\mathbf{X}$ such that for all $s \in S$ and all $x \in \mathbf{X}$ we have

$$\Pi(X_s = x_s | X_{S\setminus s} = x_{S\setminus s}) = \Pi'(X'_s = x_s | X'_{S\setminus s} = x_{S\setminus s})$$

then we also have $\Pi = \Pi'$. Reprove this result using today's Markov chain technique!

4 Friday, April 4

Two of the main motivations for studying Markov random fields come from (a) image analysis, and (b) statistical mechanics. I'll leave you with Winkler [W] to learn about (a), and I'll talk here about what I know better, which is (b), picking up most of that stuff from my paper [GHM] with Hans-Otto Georgii and Christian Maes. Notation in [GHM] clashes with that in [W], but I'll try in these lectures to stick with the [W] notation I started out with.

S used to be finite, but now we'll take it to be a countably infinite set (typically $S = \mathbf{Z}^d$) and define a neighborhood system $\partial = \{\partial(s)\}_{s\in S}$ such that each $\partial(s)$ is finite (typically, with $S = \mathbf{Z}^d$, $\partial(s)$ consists of the $2d$ sites sitting at Euclidean distance 1 from s). For each $s \in S$, let $\mathbf{X}_s$ be finite, and let $\mathbf{X} = \prod_{s\in S} \mathbf{X}_s$ (typically, $\mathbf{X}_s$ is the same for all s, in the Ising case with $\mathbf{X}_s = \{-1, +1\}$).

Definition. *The probability measure* Π *on* $\mathbf{X}$ *is said to be a* **Markov random field** *if it satisfies the* **regional Markov property**, *i.e., if* Π *admits conditionl probabilities such that for any finite* $A \subset S$ *and* Π*-almost all* $x \in \mathbf{X}$, *we have*

$$\Pi(X_A = x_A | X_t = x_t, \forall t \in S \setminus A) = \Pi(X_A = x_A | X_t = x_t, \forall t \in \partial(A)) \,.$$

(See [GHM], p 10, eq (5).)

Apologies for the inconsistency in defining MRF in terms of the *regional* Markov property, rather than the *local* Markov property is we did following Winkler in the finite case. I just find the regional Markov property a so much more natural definition. Perhaps we'd better always be explicit about which Markov property we have in mind.

We saw in the finite case that under Winkler's positivity condition, the two properties are equivalent. The same is true in the present setting of countably infinite S, although we have to be careful what we mean by the condition in this case. We cannot (as in the finite case) ask that every $x \in \mathbf{X}$ has positive Π-probability, because $\mathbf{X}$ is (in nondegenerate cases) uncountable, so that's simply impossible. Instead, we have two candidate positivity conditions that make sense:

(a) For any finite $A \subset S$ and any $x_A \in \mathbf{X}_A$, $\Pi(X_A = x_a) > 0$.

(b) Π admits conditional probabilities such that for any finite $AsubsetS$, any $x_A \in X_A$ and any $x_{S\setminus A} \in \mathbf{X}_{S\setminus A}$ we have

$$\Pi(X_A = x_a | X_{S\setminus A} = x_{S\setminus A}) > 0 \,.$$

Condition (a) may seem simpler, but (b) turns out to be even more important in statistical mechanics and percolation theory, where it is known as **the finite energy condition**. (b) implies (a), obviously, but the following example shows that the reverse implication fails.

Example: *Let* $S = \mathbf{Z}^2$ *(or whatever countably infinite set you want) and* $\mathbf{X}_s = \{0, 1\}$ *for each* $s \in S$. *Let* Π *be the probability measure corresponding to first tossing a fair coin, and then,* **if heads**, *let* $X_s = s$ *for all* $s \in S$, *while* **if tails**, *let all the* X_s *values be determined by i.i.d. fair coin tosses.*

Clearly property (a) holds, whereas (b) fails, since if we condition on having all 1*'s outside* A *(an event with positive probability), then the conditional probability of seeing any* 0 *in* A *is* 0.

Still, the weaker condition (a) turns out to suffice for the asked-for equivalence between local and regional Markov properties.

*

Now let us define the Ising model on $\mathbf{Z}^d$ (with the standard neighborhood structure) at inverse temperature $\beta \geq 0$. Recall first that for finite S, we defined it as the probability measure on $\{-1,+1\}^S$ given by

$$\Pi(x) = \frac{1}{Z}\exp(-H(x)) \text{ where } H(x) = -\beta \sum_{\langle x,y\rangle} x_s x_t \,.$$

For infinite S this won't do, because the sum $\sum_{\langle x,y\rangle} x_s x_t$ will diverge. Instead:

Definition: *A probability measure Π on $\{-1,+1\}^{\mathbf{Z}^d}$ is said to be a Gibbs measure for the Ising model on $\mathbf{Z}^d$ (with the given neighbothood structure ∂) at inverse temperature $\beta \geq 0$ if it is a Markov random field such that for all finite $A \subset \mathbf{Z}^d$, all $x_{\partial(A)} \in \{-1,+1\}^{\partial(A)}$ and all $x_A \in \{-1,+1\}^A$ we have*

$$\Pi(X_a = x_A | X_{\partial(A)} = x_{\partial(A)}) = \frac{1}{Z}\exp\left(\beta \sum_{\substack{\langle s,t\rangle \\ s,t\in A}} x_s x_t + \beta \sum_{\substack{\langle s,t\rangle \\ s\in A, t\in\partial(A)}} x_s x_t\right) \quad (9)$$

where

$$Z = \sum_{y_A \in \mathbf{X}_A} \exp\left(\beta \sum_{\substack{\langle s,t\rangle \\ s,t\in A}} y_s y_t + \beta \sum_{\substack{\langle s,t\rangle \\ s\in A, t\in\partial(A)}} y_s x_t\right)$$

is a normalizing constant.

The first thing to realize at this point is that the conditional distributions given by (9) coincide with those that we get for the Ising model on a finite S. Next, the two basic questions are

(a) Given $\beta \geq 0$, does such a Π on $\{-1,+1\}^{\mathbf{Z}^d}$ exist?

(b) If yes, then is it unique?

We'll answer these questions in the next lecture. (SPOILER ALERT: The answer to (a) is "yes", and the answer to (b) is "that depends on β".)

5 Wednesday, April 9

In response to the questions (a) and (b) at the end of the previous lecture, let's construct (fairly explicitly) a particular Gibbs measure Π^+ on $\{-1,+1\}^{\mathbf{Z}^d}$. It arises as a limit as $n \to \infty$ of probability measures Π_n^+ on $\{-1,+1\}^{\mathbf{Z}^d}$. Define the box $\Lambda_n = \{-n,\ldots,n\}^d$, and let Π_n^+ be the distribution of the $\{-1,+1\}^{\mathbf{Z}^d}$-valued random object X that arises by

(i) setting $X_s = +1$ for all $s \in \mathbf{Z}^d \setminus \Lambda_n$,

(ii) picking X_{Λ_n} according to the conditional distribution given in (9), with all $+1$'s on $\partial(\Lambda_n)$.

Π_n^+ is certainly *not* a Gibbs measure for the Ising model on $\mathbf{Z}^d$, since spins outside Λ_n are forced to take value $+$, violating (9). But inside Λ_n things behave as they should, and by sending $n \to \infty$ the misbehaving region will disappear on us.

But why would the limiting measure exist, and in what sense? The key to understanding this is **coupling** and **stochastic domination**.

Let $\preceq$ denote **coordinatewise partial order** on $\{-1,+1\}^S$ (with S finite or countably infinite), i.e., for $x, y \in \{-1,+1\}^S$ we say $x \preceq y$ if $x_s \le y_s$ for all $s \in S$.

Definition GHM 4.5, Stochastic domination: *For two probability measures* Π *and* Π' *on* $\{-1,+1\}^S$*, we say* $\Pi \preceq_{\mathcal{D}} \Pi'$ *if*

$$\Pi(f) \le \Pi'(f)$$

for every increasing (w.r.t. $\preceq$*) and bounded* $f : \{-1,+1\}^S \to \mathbf{R}$.

Theorem GHM 4.6, Strassen's Theorem: $\Pi \preceq_{\mathcal{D}} \Pi'$ *if and only if there exists a coupling of two* $\{-1,+1\}^S$*-valued random objects* X *and* X' *such that* X *has distribution* Π*,* X' *has distribution* Π'*, and* $P(X \preceq X') = 1$.

The "if" direction here is obvious. The "only if" direction is deeper, and proving it would take us too far, so we'll skip the proof.

Definition GHM 4.5 and Theorem GHM 4.6, as phrased a bit narrowly here, extend to the case where $\{-1,+1\}$ is replaced by $\mathbf{R}$. A major tool for establishing stochastic domination is the following.

Theorem GHM 4.8, Holley's Theorem: *Let* S *be finite, and* R *a finite subset of* $\mathbf{R}$*. Let* Π *and* Π' *be strictly positive probability measures on* R^S*,*

and assume that for all $s \in S$, all $y_s \in R$ and all $x, x' \in R^{S \setminus s}$ such that $x \preceq x'$ we have

$$\Pi(X_s \geq y_s | X_t = x_t, \forall t \neq s) \leq \Pi(X'_s \geq y_s | X'_t = x'_t, \forall t \neq s)\,. \qquad (10)$$

Then $\Pi \preceq_{\mathcal{D}} \Pi'$.

It is important that you understand the proof of this result (based on coupling of two R^S-valued Markov chains known as *Gibbs samplers* for Π and Π'), but I refer you to [GHM] for the proof.

Holley's Theorem has the following important consequence for the Ising model.

Lemma GHM 4.13: *Fix n (and d and β) and let $x, x' \in \{-1, +1\}^{\partial(\lambda_n)}$ be boundary conditions satisfying $x \preceq x'$. Let Π and Π' be two probability measures on $\{-1, +1\}^{\Lambda_n}$ representing the ising model conditional distributions on Λ_n with respective boundary conditions x and x'. Then $\Pi \preceq_{\mathcal{D}} \Pi'$.*

The proof is just a matter of checking that the single-site conditional probabilities under Π and Π' satisfy (10), and invoking Theorem GHM 4.8. Make sure you know how to do that!

Lemma pre-Prop GHM 4.14: *For any $n \geq 1$,*

$$\Pi_n^+ \succeq_{\mathcal{D}} \Pi_{n+1}^+\,. \qquad (11)$$

Proof: Here's a coupling of two $\{-1, +1\}^{\mathbf{Z}^d}$-valued random objects X_n^+ and X_{n+1}^+ establishing (11).

Set $X_n^+ = X_{n+1}^+ \equiv +1$ on $\mathbf{Z}^d \setminus \lambda_{n+1}$.

Set $X \equiv +1$ on $\Lambda_{n-1} \setminus \Lambda_n$.

Pick the X_{n+1}^+ configuration on $\Lambda_{n-1} \setminus \Lambda_n$ according to whatever is its correct marginal distribution.

Pick the X_n^+ and X_{n+1}^+ configurations on Λ_n in such a awy that $X_n^+ \succeq X_{n+1}^+$ on this box; such a coupling exists by Lemma GHM 4.13 using the corresponding domination on $\Lambda_{n-1} \setminus \Lambda_n$ ensured by steps 2 and 3.

This gives a coupling such that, a.s., $X_n^+ \succeq X_{n+1}^+$ on all of $\mathbf{Z}^d$. $\diamond$

So now we have pairwise couplings witnessing $\Pi_n^+ \succeq_{\mathcal{D}} \Pi_{n+1}^+$ for each n. This defines, for each n, a conditional distribution of X_{n+1}^+ given X_n^+. By applying these conditional distributions sequentially, we obtain a simultaneous coupling of all of them, with

$$X_1^+ \succeq X_2^+ \succeq X_3^+ \succeq \cdots$$

and a limiting configuration $X^+ \in \{-1, +1\}^{\mathbf{Z}^d}$ whose distribution we denote Π^+ and call **the plus measure for the Ising model on $\mathbf{Z}^d$** (at inverse temperature β). This probability measure has the following important properties.

1. **Π^+ is a Gibbs measure for the Ising model on $\mathbf{Z}^d$ with parameter β.** To see this, we just need to verify that for any finite $A \subset \mathbf{Z}^d$, the conditional distribution of X_A^+ given $X_{\mathbf{Z}^d \setminus A}^+$ is Markov with the prescribed distribution. This holds for Π_n^+ in place of Π^+ as soon as n is large enough so the Λ_n contains A. Taking limits, this property is inherited by Π^+.

2. A similar limiting Gibbs measure Π^- can also be obtained, with minuses instead of pluses outside Λ_n in the finite sages of the construction. Π^+ and Π^- look the same except with the roles of pluses and minuses interchanged.

3. **$\Pi^+ \succeq_{\mathcal{D}} \Pi$ for any Gibbs measure with the given parameter.** By the same argument as in Lemma pre-Prop GHM 4.14 we get $\Pi_n^+ \succeq_{\mathcal{D}} \Pi$ and the corresponding coupling $X_n^+ \succeq X$. The claim follows by sending $n \to \infty$.

   This gives in particular

   $$\Pi^- \preceq_{\mathcal{D}} \Pi \preceq_{\mathcal{D}} \Pi^+$$

   so that Gibbsian uniqueness is equivalent to having $\Pi^- = \Pi^+$.

4. Π^+ is translation invariant. To see this, note that we can build a similar Gibbs measure $\Pi_{shifted}^+$ with the boxes Λ_n replaced by boxes centered not at the origin but somewhere else. This gives a shift of Π^+, and by the same argument as for Π^+ we see that $\Pi_{shifted}^+$ stochastically dominates all other Gibbs measures. Hence $\Pi^+ \preceq_{\mathcal{D}} \Pi_{shifted}^+$ and $\Pi_{shifted}^+ \preceq_{\mathcal{D}} \Pi^+$, and it's not hard to see that this implies $\Pi^+ = \Pi_{shifted}^+$, and translation invariance follows.

6 Monday, April 14

The following is perhaps the most famous result for the Ising model.

Theorem: *For the Ising model on* $\mathbf{Z}^d$ *with* $d \geq 2$*, there exists a critical value* $\beta_c = \beta_c(d)$ *satisfying* $0 < \beta_c < \infty$ *such that*

$$\begin{cases} \beta < \beta_c & \Rightarrow \quad \Pi^+ = \Pi^-, \ \textit{Gibbsian uniqueness} \\ \beta > \beta_c & \Rightarrow \quad \Pi^+ \neq \Pi^-, \ \textit{Gibbsian nonuniqueness} \end{cases}$$

This excludes the case $d = 1$, because there there is no phase transition: the Ising model on $\mathbf{Z}^1$ has a unique Gibbs measure regardless of β (a fact closely related to the fact that a finite-state irreducible aperiodic Markov chain has a unique stationary distribution).

The statement of the theorem can be separated in three parts:

(i) for β sufficiently close to 0 we have uniqueness,

(ii) for β sufficiently large we have nonuniqueness, and

(iii) for β_1, β_2 such that $\beta_1 < \beta_2$, nonuniqueness at β_1 implies nonuniqueness at β_2.

All three are proved using percolation-theoretic methods in Chapters 5 and 6 of [GHM]. In this course we'll be less ambitious and restrict to $d = 2$ and parts (i) and (ii). (The proof we'll give for (i) extends in straightforward manner to $d \geq 3$, whereas the same thing for (ii) is highly demanding.)

Proposition small-β: *For the Ising model on* $\mathbf{Z}^2$ *with* $\beta < \frac{1}{8}\log(\frac{5}{3})$ *we get* $\Pi^+ = \Pi^-$.

Proof: Let us define a $\{-1,+1\}^{\mathbf{Z}^d}$-valued Markov chain $(X^+(0), X^+(1), \ldots)$ as a kind of **massively parallel Gibbs sampler for** Π^+, as follows. Start by picking $X^+(0)$ according to Π^+. Then use separate transition mechanisms for even and odd times k, as follows. Let $\mathbf{Z}^2_{even}$ denote the set of vertices in $\mathbf{Z}^2$ whose sum of coordinates is even, and define $\mathbf{Z}^2_{odd}$ analogously.

For k **even,** set $X^+_s(k) = X^+_s(k-1)$ for all $s \in \mathbf{Z}^2_{odd}$, whereas for all $s \in \mathbf{Z}^2_{even}$ independently, update its value as in the single-site Gibbs sampler.

For k **odd,** set $X^+_s(k) = X^+_s(k-1)$ for all $s \in \mathbf{Z}^2_{even}$, whereas for all $s \in \mathbf{Z}^2_{odd}$ independently, update its value as in the single-site Gibbs sampler.

Clearly, this dynamics preserves Π^+, as well as it would preserve any other Gibbs measure for the Ising model at parapeter β that we'd care to start with. So let's start another Markov chain $(X^-(0), X^-(1), \ldots)$ to run in parallel with the first, but started with $X^-(0)$ chosen according to Π^- (and independently of $X^+(0)$). We need to specify how the chains are run together. Here's how:

Define, for all $k = 1, 2, \ldots$ and all $s \in \mathbf{Z}^2$, i.i.d. uniform $[0,1]$ random variables $U(s,k)$. When a node s is updated at time k, we set

$$X_s^+(k) = \begin{cases} -1 & \text{if } U(s,k) < \dfrac{\exp\left(-\beta \sum_{t\in\partial(s)} X_t^+(k-1)\right)}{\exp\left(\beta \sum_{t\in\partial(s)} X_t^+(k-1)\right) + \exp\left(-\beta \sum_{t\in\partial(s)} X_t^+(k-1)\right)} \\ +1 & \text{otherwise} \end{cases}$$

and

$$X_s^-(k) = \begin{cases} -1 & \text{if } U(s,k) < \dfrac{\exp\left(-\beta \sum_{t\in\partial(s)} X_t^-(k-1)\right)}{\exp\left(\beta \sum_{t\in\partial(s)} X_t^-(k-1)\right) + \exp\left(-\beta \sum_{t\in\partial(s)} X_t^-(k-1)\right)} \\ +1 & \text{otherwise.} \end{cases}$$

Note now that the expression

$$\frac{\exp\left(-\beta \sum_{t\in\partial(s)} x_t\right)}{\exp\left(\beta \sum_{t\in\partial(s)} x_t(k-1)\right) + \exp\left(-\beta \sum_{t\in\partial(s)} x_t\right)} = \frac{1}{1 + \exp\left(-2\beta \sum_{t\in\partial(s)} x_t\right)}$$

is maximized when $\sum_{t\in\partial(s)} x_t = 4$ and minimized when $\sum_{t\in\partial(s)} x_t = -4$, giving values $\frac{1}{1+e^{-8}}$ and $\frac{1}{1+e^{8}}$, respectively. Denote by α, the difference between these two thresholds: $\alpha = \frac{1}{1+e^{-8}} - \frac{1}{1+e^{8}}$. For reasons that will soon be clear, we want the α to be less than $\frac{1}{4}$. Two lines of secondary-school algebraic manipulation gives that $\alpha < \frac{1}{4}$ is equivalent to the condition $\beta < \frac{1}{8}\log(\frac{5}{3})$ in the lemma.

Consider an update at a site s at time k, and denote by $A_{s,k}$ that at least one of the vertices t in $\partial(s)$ has a discrepancy at time $k-1$ between $X_t^-(k-1)$ and $X_t^+(k-1)$. For a discrepancy to happen at s after the update, necessary conditions are (a) $A_{s,k}$, and (b) that $U(s,k)$ takes a value in the length-α interval $[\frac{1}{1+e^{8}}, \frac{1}{1+e^{-8}}]$. Hence, with $D(k)$ denoting the probability that a newly updated site s at time k suffers from a discrepancy $(X_t^-(k) \neq X_t^+(k))$, we get

$$D(k) \quad = \quad P(X_s^-(k) \neq X_s^+(k))$$

$$
\begin{aligned}
&= P(A_{s,k})P(X_s^-(k) \neq X_s^+(k)|A_{s,k}) + P(\neg A_{s,k})P(X_s^-(k) \neq X_s^+(k)|\neg A_{s,k}) \\
&= P(A_{s,k})P(X_s^-(k) \neq X_s^+(k)|A_{s,k}) \\
&< 4D(k-1)\alpha.
\end{aligned}
$$

This recursive relation starts with $D(0) \leq 1$ (trivially), so we get

$$D(k) \leq (4\alpha)^k \to 0 \text{ as } k \to \infty\,.$$

For any finite region $A \subset \mathbf{Z}^d$, the expected number of discrepancies in A at time k is at most $|A|(4\alpha)^k$, which again tends to 0, so

$$\lim_{k\to\infty} P(X_A^-(k) = X_A^+(k)) = 1\,.$$

Hence Π^- and Π^+ agree on A, and since A was arbitrary we have $\Pi^- = \Pi^+$. $\diamond$

Proposition large-β: *For the Ising model on* $\mathbf{Z}^2$ *with* $\beta > \log(2\sqrt{3})$*, we get* $\Pi^+ \neq \Pi^-$.

Proof: Write $\mathbf{0}$ for the origin $(0,0) \in \mathbf{Z}^2$. If $\Pi^+ = \Pi^-$, then, by symmetry, $\Pi^+(X_{\mathbf{0}} = +1) = \frac{1}{2}$. Hence, in order to prove the proposition, it is enough to show that

$$\liminf_{n\to\infty} \Pi_n^+(X_{\mathbf{0}} = -1) < \frac{1}{2}\,. \tag{12}$$

If, under Π_n^+, we obtain $X_{\mathbf{0}} = -1$, then $\mathbf{0}$ must sit in a finite connected component of minus spins, with what I'll call a finite $+/-$ **contour** surrounding it (it's a hassle to define in words, so instead see Figure 1 in Bonati, C., The Peierls argument for higher dimensional Ising models, *Eur. J. Phys.* 2014, `http://iopscience.iop.org/0143-0807/35/3/035002/article`), because under Π_n^+ all spins outside Λ_n are 1, a.s.

Fix a finite contour C inside Λ_n surrounding $\mathbf{0}$, and a configuration $x \in \{-1,+1\}^{\mathbf{Z}^2}$ taking value -1 at $\mathbf{0}$ and $+1$ everywhere outside Λ_n, and for which C is the outermost $+/-$ contour surrounding $\mathbf{0}$. Let $\tilde{x} \in \{-1,+1\}^{\mathbf{Z}^2}$ be the configuration obtained from x by flipping all spins inside C and leaving all spins outside C unchanged. The energy difference between the two configurations arise exclusively from the pair interactions across C, and if C has length m we get

$$\frac{\Pi_n^+(X = x)}{\Pi_n^+(X = \tilde{x})} = \exp(-2\beta m)\,.$$

Hence

$$\begin{aligned}
\Pi_n^+(C \text{ is a } +/- \text{ contour for } X) &= \sum_{\substack{x\in\{-1,+1\}^{\mathbf{Z}^2} \\ C +/- \text{ contour for } x}} \Pi_n^+(X=x) \\
&= \frac{\displaystyle\sum_{\substack{x\in\{-1,+1\}^{\mathbf{Z}^2} \\ C +/- \text{ contour for } x}} \Pi_n^+(X=x)}{\displaystyle\sum_{\substack{x\in\{-1,+1\}^{\mathbf{Z}^2} \\ C +/- \text{ contour for } x}} \Pi_n^+(X=\tilde{x})} \\
&= \exp(-2\beta m)\,.
\end{aligned}$$

The number of possible lenth-m contours around $\mathbf{0}$ is at most

$$m3^{m-1}$$

where the initical m comes from the contour's leftmost crossing of the x-axis, and the 3 comes from the at most 3 choices of where to go next when traversing the contour clockwise. Hence,

$$\begin{aligned}
\Pi_n^+(X_{\mathbf{0}}=-1) &= \Pi_n^+(\exists \text{ some } +/- \text{ contour around } \mathbf{0}) \\
&\le \sum_{\text{contours } C} \Pi_n^+(C \text{ id a contour around } \mathbf{0}) \\
&= \sum_{m=4}^{\infty} \sum_{\text{length-}m \text{ contours } C} \Pi_n^+(C \text{ id a contour around } \mathbf{0}) \\
&\le \sum_{m=4}^{\infty} m3^{m-1}\exp(-2\beta m) \\
&\quad \{\text{now use the crude estimate } m \le 2^m\} \\
&\le \frac{1}{3}\sum_{m=4}^{\infty} 6^m e^{-2\beta m} = \frac{1}{3}\sum_{m=4}^{\infty}(6e^{-2\beta})^m
\end{aligned}$$

which is $< \frac{1}{2}$ provided $6e^{-2\beta} < \frac{1}{2}$, i.e., when $\beta > \log(2\sqrt{3})$, which is the condition in the proposition, so (12) is established and we are done. $\diamond$

This is the famous contour argument of Rudolph Peierls from 1936!

7 Wednesday, April 16

Recall Holley's Theorem (Theorem GHM 4.8) from lecture 5. An important consequence is the following correlation inequality.

Theorem GHM 4.11, the FKG Inequality: *For S finite and $R \subset \mathbf{R}$ finite, let Π be a strictly positive probability measure on R^S such that for all $s \in S$, all $y \in R$ and all $x, x' \in R^{S\setminus s}$ with $x \preceq x'$ we have*

$$\Pi(X_s \geq y | X_t = x_t, \forall t \neq s) \leq \Pi(X_s \geq y | X_t = x'_t, \forall t \neq s)\,. \tag{13}$$

Then, for any two increasing (with respect to $\preceq$) functions $f, g : R^S \to \mathbf{R}$ we have

$$\Pi(fg) \geq \Pi(f)\Pi(g)\,. \tag{14}$$

It's easy to check that (13) holds for the Ising model on a finite S, so the FKG inequality applies. Hence, for example, the spin values at any two sites are positively correlated.

Sketch proof of the FKG inequality: Since R^S is finite, f and g are bounded. We may assume without loss og generality that g is strictly positive, because replacing g by $g + c$ for some contant c means just adding $c\Pi(f)$ to each side of (14). We can then define the **g-weighted modification of** Π as the probability measure Π' that on R^S that to each $x \in R^S$ assigns probability $\Pi(x)g(x)$ divided by a normalizing constant Z making Π a probability measure. But then $Z = \sum_{y\in R^S} \Pi(y)g(y)$, so

$$\Pi'(x) = \frac{\Pi(x)g(x)}{\sum_{y\in R^S} \Pi(y)g(y)}\,.$$

The key step of the proof now is to establish that

$$\Pi(x) \preceq_{\mathcal{D}} \Pi'(x)\,. \tag{15}$$

To show this, we need to check that Π and Π' satisfy condition (10) in Holley's Theorem (Theorem GHM 4.8 in Lecture 5) – make sure you know how to do that (and if nothing else helps, consult the proof in [GHM])! Holley's Theorem then kicks in to ensure (15). And once we have (15), the proof is concluded by noting that

$$\begin{aligned} \Pi(f) \;&\leq\; \Pi'(f) \;=\; \sum_{x\in R^S} \Pi'(x)f(x) \\ &=\; \sum_{x\in R^S} \frac{\Pi(x)g(x)f(x)}{\sum_{y\in R^S} \Pi(y)g(y)} \;=\; \frac{\Pi(fg)}{\Pi(g)} \end{aligned}$$

and multiplying both sides with $\Pi(g)$ gives (14). ◇

The rest of this lecture will be spent on **the inhomogeneous Ising model on $\mathbf{Z}^1$**, which will serve mostly as a counterexample-generator. The neighborhood structure ∂ well be the obvious choice: $x, y \in \mathbf{Z}$ are neighbors iff $|x - y| = 1$.

The finite case first. For finite n and parameters $\beta_{-n}, \beta_{-n+1}, \ldots, \beta_{n-2}, \beta_{n-1}$, define the Ising model on $\{-n, \ldots, n\}$ with these parameters as the probability measure on $\{-1, +1\}^{\{-n,\ldots,n\}}$ that to each $x \in \{-1, +1\}^{\{-n,\ldots,n\}}$ assigns probability

$$\Pi_n(x) = \frac{1}{Z} \exp\left(\sum_{i=-n}^{n-1} \beta_i x_i x_{i+1} \right) .$$

For each i, define $y_i = x_i x_{i+1}$ (and for the corresponding random variables, similarly, $Y_i = X_i X_{i+1}$). For any x such that $x_i x_{i+1} = +1$, define another configuration

$$\tilde{x} = \begin{cases} x & \text{up to site } i \\ -x & \text{from site } i+1 \text{ onwards.} \end{cases}$$

Then

$$\frac{\Pi_n(x)}{\Pi_n(\tilde{x})} = \exp(2\beta_i) .$$

A configuration x is uniquely determined if we know x_{-n} and all the flip values $Y_{-n}, y_{-n+1}, \ldots, y_{n-1}$. So if we know X_{-n} and all flip variables *except* Y_i, then we know we're in either a given x or in $\tilde{x}$, so

$$\Pi(Y_i = 1 | Y_j = y_j \forall j \neq i) = \frac{e^{2\beta_i}}{e^{2\beta_i} + 1} = \frac{1}{1 + e^{-2\beta_i}}$$

so the Y_i-variables are *independent* taking values

$$\begin{cases} +1 & \text{w.p. } \frac{1}{1+e^{-2\beta}} \\ -1 & \text{w.p. } \frac{e^{-2\beta}}{1+e^{-2\beta}} \end{cases} \tag{16}$$

Now fix the *bi-infinite sequence*

$$\ldots, \beta_{-2}, \beta_{-1}, \beta_0, \beta_1, \beta_2, \ldots$$

and send $n \to \infty$ in the above construction. Property (16) is preserved in the limit, and one can check that this gives a Gibbs measure Π on $\{-1, +1\}^{\mathbf{Z}}$.

We haven't specified the β_i values so far, but note that if $\beta_n \to \infty$ as $|n| \to \infty$ fast enough so that

$$\sum_{-\infty}^{\infty} \frac{e^{-2\beta}}{1 + e^{-2\beta}} < \infty \tag{17}$$

(which is the same as $\sum_{-\infty}^{\infty} e^{-2\beta} < \infty$), then the expected number of spin flips is finite, so there will a.s. be only finitely many spin flips. We want this property, and choose to set $\beta_i = |i|$ for each i (this satisfies (17)), so that with Π-probability 1, the limits

$$X_{-\infty} = \lim_{i \to -\infty} X_i$$

and

$$X_{+\infty} = \lim_{i \to +\infty} X_i$$

exist (and equal $+1$ or -1). Since $\beta_0 = 0$, we have that $Y_0 = +1$ or -1 with probability $\frac{1}{2}$ each, independently of all other Y_i's. Note also that flipping Y_0 changes whether $X_{-\infty} = X_{+\infty}$ or not. This, together with the ± 1 symmetry of the model, gives $(X_{-\infty}, X_{+\infty}) = (-1,-1)$, $(-1,+1)$, $(+1,-1)$ or $(+1,+1)$, each with probability $\frac{1}{4}$.

Next we'll do something slightly unusual, namelt let Π^{mix} be the probability measure on $\{-1,+1\}^{\mathbf{Z}}$ that arises by conditioning on the event that $(X_{-\infty}, X_{+\infty})$ is either $(-1,+1)$ or $(+1,-1)$. That is a tail event with respect to the X_i variables, and conditioning on a tail event doesn't change the conditional distributions on finite sets (which are the defining properties of Gibbs measures), so Π^{mix} is a Gibbs measure for the inhomogeneous Ising model on $\mathbf{Z}$ with the given parameter. This Gibbs measure will serve as a conterexample to two properties one might otherwise naively suspect to hold in general:

FKG. We saw in connection with the FKG inequality that for the Ising model on finite S, two spin values are always positively correlated. This fails in general for Ising model Gibbs measures in the infinite setting, as exemplified by Π^{mix}. The $\pm$ symmetry gives $E[x_i] = 0$ for all i. On the other hand, since $X_{-\infty}X_{+\infty} = -1$ with probability 1, we get that $E[X_{-i}X_i]$ tends to -1 as $i \to \infty$, and hence must be strictly negative for large enough i. For such i, we thus get $E[X_{-i}X_i] < E[X_{-i}]E[X_i]$, which is the desired counterexample.

Global Markov property. Let $A = \{1,2,3\ldots\}$ and consider the conditional distribution (under Π^{mix}) of X_A given $X_{\mathbf{Z}\setminus A}$. Π^{mix} satisfying the global markov property would imply that this conditional distribution would only depend on $X_{\mathbf{Z}\setminus A}$ via $X_{\partial(A)}$, i.e., via X_0. It is easy to see that conditional on X_0, the value of $X_{+\infty}$ (which is a function of X_A) can be either $+1$ or -1, each with positive probability. But if we condition further on all of $X_{\mathbf{Z}\setminus A}$, we can read off $X_{-\infty}$, and then

the conditional probability that $X_{+\infty} = +1$ changes to either 0 or 1, so the global Markov property is violated.

8 Friday, April 25

Let's say we're interested in the Ising model on (S, ∂) with parameter β, and say $S = |1'000'000|$ (not by any means an unusually large system in practice). Suppose we want to calculate the expectation $\Pi(f)$ of some quantity f such as

(i) $f(X) = X_s$ for a given $s \in S$,

(ii) $f(X) = X_s X_t$ for given $s, t \in S$,

(iii) $f(x) = \mathbf{1}_{\{\sum_{s \in S} \geq 200'000\}}$.

Somtimes we can find clever arguments to find $\Pi(f)$, such as is the cae with (i), where the ± 1 symmetry of the model gives $\Pi(f) = 0$. Cases (ii) and (iii) are less obvious, although *in principle* trivial, because this is a finite problem, and

$$\Pi(f) = \sum_{x \in \mathbf{X}} \Pi(x) f(x) \,. \tag{18}$$

But *in practice* the obstacle to simply calculating this sum is the prohibitive number of terms in the sum: $|\mathbf{X}| = 2^{1'000'000}$.

What to do? **Assuming we had a machine** for simulsting i.i.d. $\mathbf{X}$-valued random objects with distribution Π, then we could take a sample

$$X(1), \ldots, X(n)$$

from that machine, and estimate $\Pi(f)$ with $\hat{Pi}_n(f) = \frac{1}{n} \sum_{i=1}^n f(X_i)$. This is unbiased, and since in cases (i), (ii) and (iii) above f is bounded between -1 and $+1$ we have $Var[X(i)] \leq 1$, so that

$$Var[\hat{\Pi}_n(f)] = \frac{1}{n^2} \sum_{i=1}^n \mathbf{Var}[X_i] \leq \frac{1}{n}$$

and Chebyshev's inequality[3] yields

$$P(|\hat{\Pi}_n(f) - \Pi(f)| > \epsilon) \leq \frac{1}{n\epsilon^2} \,.$$

[3] If the random variable Y has finite second moment, then $P(Y - E[Y] > \epsilon) \leq Var[Y]/\epsilon^2$

So to get this probability below a given p, just pick $n \geq \frac{\epsilon^2}{p}$.

Now, we don't have such a machine, but MCMC provides a kind of approximate such machine. The idea is to devise an irreducible aperiodic Markov chain $X(1), X(2), \ldots$ on $\mathbf{X}$ whose unique stationary distribution is Π. The convergence theorem for finite-state irreducible aperiodic markov chains gives us that if we sample at sufficiently long intervals, say m, we get a sample which is approximately i.i.d. (in a sense that can and will be specified), so a sensible estimator might be

$$\frac{1}{n}\sum_{i=1}^{n} f(X(mi))\,.$$

But if so, then for any $j \in (0, 1, \ldots, m-1)$ the estimators

$$\frac{1}{n}\sum_{i=1}^{n} f(X(mi+j))$$

seem about equally good. And when we have m such good estimators, it makes sense to reduce variance further by taking the average of then, which is tantamount to sampling the X chain at every time point (after an initial burn-in of length m). This is often done in practice.

But how long do we need to run the chain? Winkler states and proves the following Markov chain analogue of the above Chebyshev estimate.

Theorem 4.3.2: *Let $(X(0), X(1), \ldots)$ be a time-homogeneous, irreducible and aperiodic Markov chain with finite state space $\mathbf{X}$ and invariant distribution μ. Then, for any $f : \mathbf{X} \to \mathbf{R}$, we have (ergardless of how f is chosen), that the estimator $\hat{\mu}_n(f) = \frac{1}{n}\sum_{i=1}^{n} f(X(i))$ converges in probability (or in L^2 as Winkler says – since f is automatically bounded these modes of convergence are equivalent). Qunatitatively, for any $\epsilon > 0$*

$$P(|\hat{\mu}_n(f) - \mu(f))| > \epsilon) \leq \frac{13\,f\|^2}{1 - c(P))n\epsilon^2} \tag{19}$$

where

$$\|f\| = \sum_{x \in \mathbf{X}} |f(\mathbf{X})|$$

and $c(P)$ is the so-called **contraction coefficient** *of the chain's transition kernel P:*

$$c(P) = \max_{\substack{x,y \in \mathbf{X} \\ B \subset \mathbf{X}}} |P(x, B) - P(y, B)|\,.$$

The contraction coefficint $c(P)$ is an important concept, so please pay attention to Winkler's Section 4.2. Regarding the concept $\|f\|$, however, I stronly advice *against* using it in the present context. Winkler's use of $\|f\|$ is just plain lazy, and if we instead use $\max_{x\in\mathbf{X}}|f(x)|$ we'll get a variant of Theorem 4.3.2 which – unlike the present Theorem 4.3.2 – can actually be *useful*. Here's why the present Theorem 4.3.2 is so bad.

Suppose we take $\epsilon = 0.01$ and that what we know (as in cases (i), (ii) and (iii) above) is that $|f(x)| \leq 1 \forall x$ and that $|\mathbf{X}| = 2^{1'000'000}$. In order to bound the probability in (19) by 0.1, we'll have to take

$$n \geq \frac{13 \cdot 2^{2'000'000}}{0.1(1-c(P)]0.01^2} = \frac{130'000 \cdot 2^{2'000'000}}{1-c(P)}$$

so that even if $c(P) = 0$ (which is best possible), we'll have to run the chain for an amount of time that, even on the snazziest computer, makes the age of the universe seem like hardly even a blink-of-the-eye.

In general, running time bounds that (like this one) grow exponentially in $|S|$ tend to be useless. And in the frivolous science fiction scenario that we *do* have such incredible amounts of time at our disposal, we might as well use the method of directly calculation the sum in (18), rather than reverting to MCMC simulation.

*

Next, how to concretely construct the Markov chain. I'll focus on the **Gibbs sampler**. For the related and more flexible **Metropolis–Hastings algorithm** I'll refer the reader to Winkler's chapter on that.

Given the Markov random field distribution Π on $\mathbf{X}$, a Gibbs sampler is a Markov chain $(X(0), X(1), \ldots)$ constructed as follows. At each time k, select an $s \in S$ (according to some rule, deterministic or random), set

$$X_t(k) = X_t(k-1) \text{ for all } t \in S \setminus s$$

and pick a fresh value of $X_s(k)$ according to the Π-conditional distribution of X_s given a configuration on $S \setminus s$ agreeing with $X_{S\setminus s}(k-1)$.

Obviously, if $X(k-1) \sim \Pi$, then $X(k) \sim \Pi$, so Π is a stationary distribution for the chain. But is it the *only* stationery distribution, and do we have *convergence* towards it as $k \to \infty$?

This depends on the mechanism for chossing which vertex to update. For instanec, always choosing *the same* vertex to update is a stupid rule, under which the answer to both questions are "no". Two other choices, which are popular and which under Winkler's positivity condition gives answer "yes" to both questions, are to select s

- at random (i.i.d., and unifromly on S), and
- according to a systematic weep: deterministically go through all of S in the first $|S|$ updates, and then repeat.

The former is sometimes more convenient to work with, because it is time-homogeneous, so we can immediately apply the basic convergence theorem for finite-state Markov chains to deduce uniqueness of the stationary distribution and convergence to it as $k \to \infty$. The latter allows the same conclusion, but only if we redefine "time" by considering the embedded Markov chain obtained by looking at the original one only at times that are multiples of $|S|$.

Concerning quantitatively the *rate* of convergence, the contraction coefficient $c(P)$ from Theorem 4.3.2 plays an important role. If we look at a single step of the Gibbs sampler, we get $c(P) = 1$ (no contraction at all), which is useless, but by considering the embedded chain by viewing $|S|$ updates as a single step, we get $c(P) < 1$. It can be very close to 1, however, and this is one of the reasons why the rate of convergence in Winkler's Theorem 5.1.4 for the Gibbs sampler is so terribly bad.

The quantitative part of Winkler's Theorem 5.1.4 is that

$$P(|\frac{1}{n}\sum_{i=1}^{n} f(X(i)) - \Pi(f)| > \epsilon) \leq \frac{c}{n\epsilon^2} e^{\sigma\Delta}$$

where c, σ and Δ are as follows.

- c is a constant depending on the updating scheme and on the catastrophic factor $\|f\|^2$. This is enough to render Theorem 5.1.4 useless, but this part can actually be fixed by replacing Theorem 4.3.2 by a more sensible variant.
- $\Delta = \max_{s \in S} \max_{x,yin\mathbf{X}: x_{S\setminus s} = y_{S\setminus s}}$ says roughly (if Δ is small) that the conditional distribution at any s given $X_{S\setminus s}$ does not deviate much from uniform distribution. This condition works to get results, but does not capture the essence of the "fast convergence" problem. What's needed is not being close to uniform, but rather not depending too heavily on $X_{S\setminus s}$.
- $\sigma = |S|$, and since the σ factor sits in the exponent we obtain another catastrophic factor $\exp(1'000'000\Delta)$ rendering Theorem 5.1.4 useless. This time fixing Theorem 4.3.2 won't help.

Nevertheless, not all is lost, and next week I'll offer you sensible replacement for Theorems 5.1.4 and 4.3.2!

*

The needed irreducibility of the Markov chain follows under Winkler's positivity, but there are important examples where we can get away without it, such as in **the hard-core model**. Here $\mathbf{X} = \{0,1\}^S$, and a 1 at s is thought of as a particle, and a 0 at s as the absence of a particle. Particles cannot be packed too tightly, and a configuration $x \in \{0,1\}^S$ is called **legal** if there are no two neighbors s and t with $x_s = x_t = 1$. The hard-core model with parameter $\lambda > 0$ is the probability measure Π on $\{0,1\}^S$ given by

$$\Pi(x) = \frac{1}{Z}\lambda^{(\#\text{ 1's in } x)}\mathbf{1}_{\{x \text{ is legal}\}}$$

The parameter λ quantify the model's tendency to have many 1's, and sending $\lambda \to \infty$ is tantamount to trying harder and harder to find an "optimal packing", i.e., maximizing the number of 1's without making x illegal.

The Gibbs sampler for this model is irreducible in the sense that for any two configurations with positive probability (i.e., any two legal configurations) you can reach one from the other, by forst successively removing one 1 after the other, to reach the "all 0's configuration", and then adding new 1's agreeing with target configuration, one after the other.

9 Wednesday, April 30

Today we'll replace Winkler's useless Theorems 4.3.2 and 5.1.4. Straight to the point:

Theorem Replace-4.3.2: *Let $(X(o), X(1), \ldots)$ be a time-homogeneous, irreducible and aperiodic Markov chain with finite state space $\mathbf{X}$ and invariant distribution μ. Then, for any $\epsilon > 0$, any $a > 0$ and any $f : \mathbf{X} \to \mathbf{R}$ with $\max_{x\in\mathbf{X}} |f(x)| \le a$, we have, defining*

$$\hat{\mu}_n(f) = \frac{1}{n}\sum_{i=1}^{n} f(X(i))\,,$$

that

$$P(|\hat{\mu}_n - \mu| > \epsilon) \le \frac{16a^2}{(1-c(P))n\epsilon^2}\,.$$

Here the contraction coefficient is, as before,

$$c(P) = \max_{\substack{x,y\in\mathbf{X}\\ B\subset\mathbf{X}}} |P(x,B) - P(y,B)|\,.$$

We can also write $c(P)$ as

$$c(P) = \frac{1}{2} \max_{x,y \in \mathbf{X}} \|P(x,\cdot) - P(y,\cdot)\|$$

where, for any two probability distributions Π and ν on $\mathbf{X}$, $\|\Pi - \nu\|$ denotes the toal variation norm $\sum_{x \in \mathbf{X}} |\Pi(x) - \nu(x)|$. (This total variation distance ranges between 0 (identical distributions) and 2 (disjoint distributions), and differs from the total variation distance in [GHM] by a factor 2. Both definitions are fine, but it is dangerous of course to mix them up, so be careful when exploring the literature.)

A central result on coupling (see Proposition 4.4 of [GHM]) is that for any two distributions Π and ν we can couple two random objects $X \sim \Pi$ and $Y \sim \nu$ in such a way that $P(X \neq Y) = \frac{1}{2}\|\Pi - \nu\|$; this is called a **maximal coupling**, because it cannot be improved. We'll use this result, but will not dig into its proof.

Another result we'll use without digging into its proof (but do have a look at Winkler's Lemma 4.2.2 in you want to understand it) is that the n-step transition kernel P^n of a Markov chain with transition kernel P satisfies

$$c(P^n) \leq (c(P))^n \,.$$

Proof oh Theorem Replace-4.3.2: Assume for simplicity that $\mu(f) = 0$ (we'll fix that at the end). In order to apply Chebyshev, we need to estimate

$$\begin{aligned} Var\left[\frac{1}{n}\sum_{i=1}^{n} f(X(i))\right] &= E\left[(\frac{1}{n}\sum_{i=1}^{n} f(X(i)))^2\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} E[f(X_i)f(X_j)] \,. \end{aligned}$$

Imagine (for the time being) the chain starting in a fixed $x \in \mathbf{X}$, and couple $(X(0), X(1), \ldots)$ with another $\mathbf{X}$-valued Markov chain with the same trnasition kernel but starting in stationarity μ. For fixed $k \geq 1$, we can couple $X(k)$ and $X'(k)$ in such a way that

$$\begin{aligned} P(X(k) \neq X'(k)) &\leq \frac{1}{2} \max_{x,y \in \mathbf{X}} \|P^k(x,\cdot) - P^k(y,\cdot)\| \\ &= c(P^k) \leq (c(P))^k \,. \end{aligned}$$

Then

$$E[f(X(0))f(X(k))] = f(x)E[f(X(k))]$$

$$
\begin{aligned}
&= f(x)E[f(X(k)) + f(X'(k)) - f(X'(k))] \\
&= f(x)E[f(X'(k))] + f(x)E[f(X(k)) - f(X'(k))] \\
&\leq f(x) \cdot 0 + f(x) \cdot 2a(c(P))^k \\
&\leq 2a^2(c(P))^k \,.
\end{aligned}
$$

Dropping the assumption that the chain X starts in a fixed $x \in \mathbf{X}$, thus allowing random $X(0)$, we obtain $E[f(X(0))f(X(k))]$ as a weighted average of terms that are at most $2a^2(c(P))^k$, so the conclusion

$$E[f(X(0))f(X(k))] \leq a^2(c(P))^k$$

remains valid. And for similar reasons, for any $m \geq 0$

$$E[f(X(m))f(X(m+k))] \leq a^2(c(P))^k \,.$$

We get

$$
\begin{aligned}
E\left[(\frac{1}{n}\sum_{i=1}^{n} f(X(i)))^2\right] &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} E[f(X_i)f(X_j)] \\
&\leq \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} 2a^2(c(P))^{|i-j|} \\
&\leq \frac{2a^2}{n^2}\sum_{i=1}^{n}\sum_{j=-\infty}^{\infty} (c(P))^{|i-j|} \\
&\leq \frac{4a^2}{n^2}\sum_{i=1}^{n}\sum k = 0^{\infty}(c(P))^k \\
&= \frac{4a^2}{n(1-c(P)} \,.
\end{aligned}
$$

So Chebyshev gives

$$P(|\frac{1}{n}\sum_{i=1}^{n} f(X(i))| > \epsilon) \leq \frac{4a^2}{\epsilon^2 n(1-c(P)} \,. \tag{20}$$

But that was for f such that $\mu(f) = 0$. For $\mu(f) \neq 0$, we just apply the above to $g = f - \mu(f)$, and $|f(x)| \leq a$ gives $|g(x)| \leq 2a$. So in this more general case the bound in (20) becomes

$$\frac{4(2a)^2}{\epsilon^2 n(1-c(P)} = \frac{16a^2}{\epsilon^2 n(1-c(P)}$$

which is the bound claimed in the theorem. $\diamond$

Next, we want to replace Theorem 5.1.4 on convergence of a Gibbs sampler. With our new replacement of Theorem 4.3.2 there is hope to get something useful, as long as we can get a reasonable bound on $c(P^n)$. There is no hope of getting a very general result, because the single-site Gibbs sampler for the Ising model on large finite subsets of $\mathbf{Z}^d$ *is* exponentially slow in the phase transition regime $\beta > \beta_c$. But we should at least be able to obtain a result that gives fast convergence for *some* nontrivial Markov random fields (including the small-β Ising model).

We'll stick, for definiteness, to a systematic sweep Gibbs sampler, and exploit the idea of the small-β $\Pi^+ = \Pi^-$ result for the $\mathbf{Z}^2$ Ising model in Lecture 6.

There we assumed $\beta < \frac{1}{8}\log(\frac{5}{3})$, and if we inspect the proof, we see that the crucial property that follows from this is that

$$\max_{s\in S} \max_{x,y\in S\setminus s} \|\Pi(X_s \in \cdot | S_{S\setminus s} = x) - \Pi(X_s \in \cdot | S_{S\setminus s} = y)\| < \frac{1}{2}\,, \qquad (21)$$

so that by employing the maximal coupling when we uptate s at time k, we can make sure that $P(X_s(k) \neq X'_s(k)] < \frac{1}{4}$. Or more generally we want $P(X_s(k) \neq X'_s(k)] < \frac{1}{d_{max}}$ where $d_{max} = \max_{s\in S} |\partial(s)|$, so we want the right-hand side of (21) to be $\frac{2}{d_{max}}$. We will make that the central assumption of our theorem. To arrive there, let's denote the left-hand side of (21) by α, and note that after the first weep through S, employing the maximal coupling at each update guarantees that each site has probability at most $\frac{\alpha}{2}$ of exhibiting a discrepancy between its values in the X chain and the X' chain.

When a site s is chosen at a time k in the second sweep, we get

$$\begin{aligned} P(X_s(k) \neq X'_s(k)) &= P(\text{some discrepancy in } \partial(s)) \\ &\quad \cdot P(X_s(k) \neq X'_s(k) | \text{ some discrepancy in } \partial(s)) \\ &\leq \frac{\alpha d_{max}}{2}\frac{\alpha}{2}\,. \end{aligned}$$

So after the second sweep, all sites have probability at most $\frac{\alpha d_{max}}{2}\frac{\alpha}{2}$ of exhibiting discrepancy. Applying the same reasoning iteratively, we get after the n:th sweep that all sites have probability at most $\frac{\alpha}{2}(\frac{\alpha d_{max}}{2})^{n-1}$. When $\alpha < \frac{2}{d_{max}}$, this tends to 0 (fast), so we're in business. Specifically, note that $\frac{|S|\alpha}{2}(\frac{\alpha d_{max}}{2})^{n-1}$ is an upper bound for the expected number of discrepancies after the n:th sweep. Take n so large that this bound satisfies

$$\frac{|S|\alpha}{2}\left(\frac{\alpha d_{max}}{2}\right)^{n-1} \leq \frac{1}{2}$$

which via a few steps of high school-level algebraic manipulation is seen to be equivalent to

$$n \geq \frac{\log(\alpha|S|)}{\log(\frac{2}{\alpha d_{max}})} + 1 \,.$$

Then the probability that the two chains have coalesced (at every site) after n sweeps is at least $\frac{1}{2}$. Our theorem, therefore, is as follows.

Theorem Replace-5.1.4: *Let Π be a Markov random field, and define*

$$\alpha = \max_{s \in S} \max_{x,y \in S \setminus s} \|\Pi(X_s \in \cdot | S_{S\setminus s} = x) - \Pi(X_s \in \cdot | S_{S\setminus s} = y)\| \,.$$

If $\alpha < \frac{2}{d_{max}}$, then the Markov chain in which a single step represents at least $\frac{\log(\alpha|S|)}{\log(\frac{2}{\alpha d_{max}})} + 1$ full sweeps of the Gibbs sampler has a contraction coefficient $c(P) \leq \frac{1}{2}$.

This result can then readily be plugged into Theorem Replace-4.3.2 to get bounds on how long we need to run the chain in order to get good estimates of whatever expectation $\Pi(f)$ we wish to calculate. The fact that it takes just of the order $\log(|S|)$ sweeps, i.e., $|S| \log(|S|)$ single site updates, in order to reach a useful contraction coefficient (as opposed to in Winkler's Theorem 5.1.4 where it took exponentially many) means that we're in business with a sampling algorithm that may actually be useful.

10 Wednesday, May 7

Now is time to discuss statistical estimation of parameters in Markov random fields. I'll focus on a simple example: estimating β in the Ising model (this will put you in a better position to digest Winkler's more general treatment of the topic).

A configuration $X \in \{-1, +1\}^S$, with S finite, is observed. Me may assume it comes from the Ising model Gibbs measure Π for this known S with known neighborhood ∂, but we do not know the inverse temperature β. How to estimate β?

I'll focus on maximum likelihood (ML) and related techniques. Winkler mentions two highly desirable properties of estimates $\hat{\beta}_n$ (where n is some measure of the amount of data):

(i) So-called **consistency**: $\hat{\beta}_n$ should tend to β as $n \to \infty$ (in whatever mode of convergence we can get, almost sure convergence, convergence in probability, in L^2, or...)

(ii) $\hat{\beta}_n$ should be computationally feasible.

Property (i) is not quite what we want in practice, because at the end of the day we need to know for some fixed finite n that $\hat{\beta}_n$ is likely to be close to the true value β, and asymptotic statements like (i) do not answer that. But it's a good start.

What do we mean by the amount of data n? Winkler discusses two cases:

(a) n i.i.d. samples $X(1), \ldots, X(n)$ from Π on a fixed finite S.

(b) Let S be countably infinite, such as $S = \mathbf{Z}^2$ with the standard neighborhood structure, and look at the configuration on $\Lambda_n = \{-n, \ldots, n\}^2$.

He says (b) is more relevant in image analysis and focuses mainly on that; I'm very happy to go along.

The simplification we get from focusing on estimating β in the Ising model, compared to Winkler's more general setting, is that of studying just a single parameter, so that when taking derivatives of the log-likelihood function L (as we almost always need to do when studying ML estimators) we don't need to handle vector-valued gradients ∇L and Hessian matrices $\nabla^2 L$.

Before treating the Ising model, let's remind ourselves of how ML works by considering the very simplest situation of **possibly biased coin tosses**: let $X(1), X(2), \ldots$ be i.i.d. $\{0, 1\}$ valued random variables taking value 1 with an unknown probability p. The task here is to estimate p based on $X(1), \ldots X(m)$.

Given $X(1), \ldots X(m)$ such that $\sum_{i=1}^{m} X(i) = k$, the likelihood becomes

$$l(p) = p^k (1-p)^{m-k}$$

and the log likelihood

$$L(p) = \log(l(p)) = k \log(p) + (m-k) \log(1-p)\,.$$

How do we maximize $L(p)$? We solve for $\frac{dL(p)}{dp} = 0$ and check whether $\frac{d^2 L}{dp^2} < 0$.[4] We get

$$\frac{dL(p)}{dp} = \frac{k}{p} - \frac{m-k}{1-p}$$

[4] In addition, we of course need to check what happens on the boundary of the parameter space. This turns out to make no difference to the present example, and I'll just skip that.

which is 0 when $p = \frac{k}{m}$, and

$$\frac{d^2L}{dp^2} = \frac{-k}{p^2} - \frac{m-k}{(1-p)^2} < 0\,,$$

so $\hat{p}_m = \frac{k}{m}$ is our ML estimate of p. From the strong law of large numbers, we know that

$$\lim_{m\to\infty} \hat{p}_m = \lim_{m\to\infty} \frac{1}{m} \sum_{i=1}^{m} X(i) = p$$

almost surely, so the ML estimate is consistent!

Now let's try to imitate this in the Ising model on $\mathbf{Z}^2$ context, based on observing the spins on Λ_n. For reasons that will become clear, I'll assume we're allowed to peek at the values on $\partial(\Lambda_n)$ as well. The literal ML estimate is unavailable, because we do not have an expression for

$$l(\beta) = \Pi_\beta(X_{\Lambda_n} = x_{\Lambda_n})$$

since, as we saw in Lesture 6, β doesn't even uniquely determine Π_β.

The next best thing is to consider the conditional distribution of X_{Λ_n} given $X_{\partial(\Lambda_n}$, viewing the latter as fixed.That gives a conditional likelihood

$$l(\beta) = \frac{1}{Z} \exp\left(\sum_{\substack{\langle u,v\rangle \\ u,v\in\Lambda_n}} x_u x_v + \sum_{\substack{\langle u,v\rangle \\ u\in\Lambda_n, v\in\partial(\Lambda_n)}} x_u x_v \right).$$

This would be easy to work with, were it not for the hidden fact that Z depends on β. Treating Z by direct summation is obviously impractical (the sum as $2^{|\Lambda_n|}$ terms), but it is possible via various numerical and MCMC schemes to find an approximation, and that way get an approximate ML estimate (which can also be shown to be consistent).

That's very intricate, however, and we'll choose another way, namely to condition further on $X_{\Lambda_{n,odd}}$, where $\Lambda_{n,odd}$ is defined as the set of vertices in Λ_n whose sum of coordinates is odd. The point of doing so is that by the Markov random field property, the variables $\{X_v\}_{v\in\Lambda_{n,even}}$ become *conditionally independent* given $X_{\lambda_{n,odd}}$ and $X_{\partial(n)}$, rendering the likelihood and easy-to-handle product structure, translating into a sum in the log-likelihood. They are not *identically distributed*, however, because

$$\Pi_\beta(X_v = +1 | X_{\partial(v)} = x_{\partial(v)}) = \frac{1}{1 + \exp\left(-2\beta \sum_{w\in\partial(v)} x_w\right)}$$

which depends on the number of +1's in the neighborhood.

But if we restrict to looking at vertices in the set $\Lambda_{v\in\Lambda_{n,even},4+}$ defined as those vertices in $\Lambda_{n,even}$ that happen to have all +1 neighbors, then these become conditionally i.i.d. with

$$\Pi_\beta(X_v = +1|\text{all else}) = \frac{1}{1+e^{-8\beta}}\,,$$

so we're back in the coin tosses situation with m i.i.d. (p) binary variables, with $p = \frac{1}{1+e^{-8\beta}}$ and m is the number of vertices in $\Lambda_{n,even,4+}$. It is not hard to show that $m \to \infty$ a.s. as $n \to \infty$, which put us in a setting where we can reasonably ask for consistency. If we observe k +1's in $\Lambda_{n,even,4+}$, the log-likelihood becomes

$$L(\beta) = k\log\left(\frac{1}{1+e^{-8\beta}}\right) + (m-k)\log\left(\frac{1}{1+e^{8\beta}}\right)$$

which, as we saw for the coin tosses, is maximized when $p = \frac{k}{m}$, and solving for β in $p = \frac{1}{1+e^{-8\beta}}$ gives

$$\hat{\beta}_n = -\frac{1}{8}\log\left(\frac{m}{k} - 1\right)$$

which we take as our ML estimator.[5] Now, as $n \to \infty$, so that $m \to \infty$ the strong law of large numbers applied to the conditional distribution of $X_{\Lambda_{n,even}}$ given $X_{\Lambda_{n,odd}}$ gives that $\frac{k}{m}$ tends to $\frac{1}{1+e^{-8\beta}}$ a.s., so

$$\lim_{n\to\infty}\hat{\beta}_n = -\frac{1}{8}\log(1+e^{-8\beta}-1) = \beta\,,$$

so we have a consistent estimator of β!

But it seems somewhat wasteful to look at only those vertices with all +1 neighbors. In fact we can construct similar estimates based on vertices in $\Lambda_{n,even}$ with r neighbors taking value +1, not just for $r = 4$ but also for $r = 0, r = 1$ and $r = 3$. (We can also try $r = 2$, but that actually doesnt help, because the log-likelihood turns out to be $k\log(\frac{1}{2}) + (m-k)\log(\frac{1}{2})$ which is independent of β.) This gives four different log-likelihoods L_0, L_1, L_3, L_4 with

$$L_r(\beta) = k\log\left(1+e^{-(4r-8))\beta}-1\right) + (m-k)\log\left(1+e^{(4r-8))\beta}-1\right)$$

[5]Note that this estimator can turn out to be negative! Here we have two choices. Either we can *allow* $\beta < 0$ in the Ising moedl, leading to the so-called *Ising antiferromagnet*, or we can insist on $\beta \geq 0$ which, in accordance with the previous footnote, gives $\hat{\beta}_n = 0$ in case $-\frac{1}{8}\log\left(\frac{m}{k}-1\right)$ turns out negative.

each giving an ML estimator $\hat{\beta}_{r,n}$ which tends to β as $n \to \infty$.

Well, surely we can be more efficient and and add upp the loglikelihoods to get

$$L_{even}(\beta) = L_0(\beta) + L_1(\beta) + L_3(\beta) + L_4(\beta)\,,$$

no? (Including $L_2(\beta)$ is harmless but pointless.) Since each $L_r(\beta)$ leads to a consistent ML-estimate, surely the same holds for L_{even}?

Yes we can, and yes it does! Here's why. We calculate

$$\frac{d}{d\beta}L_r(\beta) = (4r-8)\left(k - \frac{me^{(4r-8)\beta}}{1+e^{(4r-8)\beta}}\right)$$

and

$$\frac{d^2}{d\beta^2}L_r(\beta) = (4r-8)^2 m\frac{e^{(4r-8)\beta}}{(1+e^{(4r-8)\beta})^2}) < 0$$

so each $L_r(\beta)$ is concave in β.

Fix $\epsilon > 0$, and write β^* for the true value of β. By consistency of each of the $L_r(\beta)$ estimators, we can a.s. find some (random) N such that for all $n \geq N$ that each $\hat{\beta}_{r,n}$ is in the interval $\beta^* \pm \epsilon$. Then for $r = 0, 1, 3, 4$, we have

$$\frac{d}{d\beta}L_r(\beta) \begin{cases} > 0 & \text{for all } \beta < \beta^* - \epsilon \\ < 0 & \text{for all } \beta > \beta^* + \epsilon\,. \end{cases}$$

The same conclusion follows for $L_{even}(\beta)$, so it must have its maximum at some point $\hat{\beta}_{even,n}$ somewhere in $(\beta^* - \epsilon, \beta^* + \epsilon$, and since $\epsilon > 0$ was arbitrary we have shown that $\hat{\beta}_{even,n}$ is a consistent estimator.

We can of course do the same thing with $L_{odd}(\beta)$ and get another consistent estimator. And in the same way as when we added the $L_r(\beta)$ loglikelihoods, we can also add $L_{even}(\beta)$ and $L_{odd}(\beta)$ and get another "log-likelihood" $L_{pseudo}(\beta)$ which by the same agruments yields another consistent estimators.

I think it's fair to use the term **conditional log-likelihood** for all of $L_r(\beta)$, $L_{even}(\beta)$ and $L_{odd}(\beta)$. For $L_{pseudo}(\beta)$ this term seems less appropriate, because here we are in a sense conditioning on *everything*. Instead, $L_{pseudo}(\beta)$ is called a **pseudolikelihood**, the maximization of which is a standard device in parameter estimation in Markov random fields.

11 Friday, May 9

Up to know, we have given the Ising model the energy

$$H(x) = -\beta \sum_{\langle u,v \rangle} x_u x_v$$

where $\beta \geq 0$ is called the inverse temperature parameter. Today we'll see what happens when we generalize the moedl and introduce another parameter – the so-called **external field** h – to get

$$H(x) = -\beta \left(\sum_{\langle u,v \rangle} x_u x_v + h \sum_{u \in S} x_u \right) . \tag{22}$$

Setting $h = 0$ gives back the old model. Setting $h \neq 0$, say $h > 0$ for concreteness, breaks the ± 1 symmstry in the model and favors spin configurations with many $+1$'s and disfavors those with many -1's. This has dramatic consequences.

Recall the theorem from Lecture 6 saying that the Ising model on $\mathbf{Z}^d, d \geq 2$ (without external field) has a critical value $\beta_c \in (0, \infty)$ such that there is a unique Gibbs measure if $\beta < \beta_c$ and multiple Gibbs measures if $\beta > \beta_c$.

In contrast, when $h \neq 0$ there is, on $\mathbf{Z}^d$, *always* only one Gibbs measure. I will not show the full result to this extent, but be content with a reasonably strong partial result in this direction (Proposition h below).

The Ising model on $\mathbf{Z}^d$ with external field can be defined analogously to what i did for the $h \neq 0$ case in Lecture 4, in terms of conditional probabilities on finite sets. Much of the stochastic domination machinery from Lecture 5 goes through with hardly any change in the arguments. Very briefly:

Lemma GHM 4.13 goes through for the $h \neq 0$ case, leading to the stochastically decreasing sequence

$$\Pi_1^{\beta,h,+} \succeq_{\mathcal{D}} \Pi_2^{\beta,h,+} \succeq_{\mathcal{D}} \Pi_3^{\beta,h,+} \succeq_{\mathcal{D}} \cdots$$

of probability measures $\Pi_n^{\beta,h,+}$ on $\{-1,+1\}^{\mathbf{Z}^d}$ corresponding to putting all $+1$'s outside Λ_n, and then picking the spins on Λ_n according to the conditional distribution corresponding to the energy function in (22) and all $+1$ boundary. The limiting measure $\Pi^{\beta,h,+}$ is a Gibbs measure for the Ising model on $\mathbf{Z}^d$ with the given parameters, it is translation invariant, and it stochastically dominates all other such Gibbs measures. We can analogously construct $\Pi^{\beta,h,-}$ with similar properties and the result that any further Gibbs measure $\Pi^{\beta,h}$ with the same parameter values is sandwiched between $\Pi^{\beta,h,-}$ and $\Pi^{\beta,h,+}$ in the sense of stochastic domination. Hence, Gibbsian uniqueness is equivalent to

$$\Pi^{\beta,h,-} = \Pi^{\beta,h,+} .$$

What we would now most like to show (and what is in fact true) is the result that

$$\Pi^{\beta,h,-} = \Pi^{\beta,h,+} \text{ whenever } h \neq 0$$

but will for reasons of space and time settle for the following weaker result.

Proposition h: *For fixed $\beta \geq 0$, we have*

$$\Pi^{\beta,h,-} \neq \Pi^{\beta,h,+}$$

for at most countably many values of h.

Depending our mood, we can either view this result as very weak (because it does not settle the uniqueness issue for any given (β, h)) or very strong (because it proves that uniqueness holds for Lebesgue-almost all (β, h)).

As a preparation for the proof of Proposition h, fix β, define $M^{\beta,+}(h)$ as the $\Pi^{\beta,h,+}$-expectation of the spin value at $s \in \mathbf{Z}^d$ (which by translation invariance is independent of s), and define $M^{\beta,-}(h)$ as the $\Pi^{\beta,h,-}$-expectation of the same quantity. (M is for **magnetization**.) A straightforward adaptation of the proof of Lemma GHM 4.13 shows that if $h_1 \leq h_2$, then $\Pi_n^{\beta,h_1,+} \preceq_{\mathcal{D}} \Pi_n^{\beta,h_2,+}$, so that $\Pi^{\beta,h_1,+} \preceq_{\mathcal{D}} \Pi^{\beta,h_2,+}$ and

$$M^{\beta,+}(h_1) \leq M^{\beta,+}(h_2)\,.$$

Similarly we get

$$M^{\beta,-}(h_1) \leq M^{\beta,-}(h_2)$$

and, for any $h \in \mathbf{R}$,

$$M^{\beta,-}(h) \leq M^{\beta,+}(h)\,.$$

These three inequalities for the magnetization should all be intuitively obvious, because increasing h should favor having more $+1$'s regardless of whether we're in the plus measure or the minus measure, while going from the minus measure to the plus measure while keeping the parameters constant should have the same effect.

A much less obvious issue is, again for $h_1 \leq h_2$, $M^{\beta,+}(h_1)$ compares to $M^{\beta,-}(h_2)$. The boundary condition wants the inequality to go one way, while the external field wants the other. It turns out that the external field wins:

Lemma h: *Whenever $h_1 < h_2$, we have*

$$M^{\beta,+}(h_1) \leq M^{\beta,-}(h_2)\,.$$

Intuitively, the reason that this is true is that when we look at the effect of boundary condition and external field on a box Λ_n, the former acts on the boundary while the latter acts in the entire box, and since the surface-to-volume ratio goes to 0 as $n \to \infty$ the latter wins no matter how small

$h_2 - h_1$ is. Turning this intuition into a proof is, as we shall see, nontrivial but doable.

Before proving the lemma, let's show how it implies Proposition h. Suppose that

$$\Pi^{\beta,h,-} \neq \Pi^{\beta,h,+} . \tag{23}$$

Then $M^{\beta,-}(h) < M^{\beta,+}(h)$ (because otherwise in the coupling witnessing their stochastic domination we'd never see any disrepancy, contradicting (23)). Write $\delta > 0$ for the difference between these magnetizations. Lemma h ensures that for any $\epsilon > 0$ we have

$$M^{\beta,+}(h-\epsilon) \leq M^{\beta,-}(h) = M^{\beta,+}(h) - \delta$$

so that

$$\lim_{\epsilon \searrow 0} M^{\beta,+}(h-\epsilon) < M^{\beta,+}(h)$$

and $M^{\beta,+}$ thus exhibits a discontinuity at h. But an increasing function can have at most uncountably many discontinuities (because each discontinuity skips some rational number, and there are only countably many rationals), so (23) can hold for at most countably many h, and Proposition h follows.

What remains is to prove the lemma:

Proof of Lemma h: We will couple two $\{-1,+1\}^{\mathbf{Z}^d}$-valued random objects $X \sim \Pi^{\beta,h_1,+}$ and $X' \sim \Pi^{\beta,h_2,-}$ in the simplest possible way: independently.

Our first claim is that, with probability 1,

$$\lim_{n\to\infty} \frac{1}{|\Lambda_n|} \sum_{s\in\Lambda_n} X_s = M^{\beta,+}(h_1) . \tag{24}$$

Existence of the limit follows from translation invariance, and if the limit were nontrivially random it would have to exceed $M^{\beta,+}(h_1)$ with positive probability. Conditioning on doing so would give again a Gibbs measure (because conditioning on anything in the tail σ-field preserves the defining conditional distributions on finite sets), and one with a higher magnetization, contradicting what we already know about $\Pi^{\beta,h_1,+}$ stochastically dominating all other Gibbs measures with the same parameters. Hence (24). And by the same argument,

$$\lim_{n\to\infty} \frac{1}{|\Lambda_n|} \sum_{s\in\Lambda_n} X'_s = M^{\beta,-}(h_2) . \tag{25}$$

Assume now that $M^{\beta,+}(h_1) \neq M^{\beta,-}(h_2)$ (otherwise we're done). At this point, I don't want to commit to circular reasoning by presupposing which

of them is bigger, but define

$$M_{max} = \max\{M^{\beta,+}(h_1), M^{\beta,-}(h_2)\}$$

and

$$M_{min} = \min\{M^{\beta,+}(h_1), M^{\beta,-}(h_2)\}$$

as well as

$$\Delta = M_{max} - M_{min} > 0.$$

Now imagine you're at a Game Show, where, for some large n, the host reveals to you the following information:

(i) $X_{\mathbf{Z}^d \setminus \Lambda_n} = x_{\mathbf{Z}^d \setminus \Lambda_n}$

(ii) $X'_{\mathbf{Z}^d \setminus \Lambda_n} = x'_{\mathbf{Z}^d \setminus \Lambda_n}$

(iii) two configurations $\hat{x}_{\Lambda_n}$ and $\check{x}_{\Lambda_n}$, **but no information on which of them is X_{Λ_n} and which is X'_{Λ_n}.**

Your job is to guess which is which.

First I'll (sort of) tell you how large n will be. Fix $\delta > 0$ small, and pick n large enough so that

$$P\left(\frac{1}{\Lambda_n}\sum_{s\in\Lambda_n}\hat{x}_{\Lambda_n} - \frac{1}{\Lambda_n}\sum_{s\in\Lambda_n}\check{x}_{\Lambda_n} \geq \Delta/2\right) \geq 1-\delta\,. \tag{26}$$

Suppose the event in (26) happens. Should you guess that ($X_{\Lambda_n} = \hat{x}_{\Lambda_n}, X'_{\Lambda_n} = \check{x}_{\Lambda_n}$) or vice versa? Well, let's calculate

$$\begin{aligned}
&\frac{P(X_{\Lambda_n} = \hat{x}_{\Lambda_n}, X'_{\Lambda_n} = \check{x}_{\Lambda_n}|\text{(i)},\text{(ii)},\text{(iii)})}{P(X_{\Lambda_n} = \check{x}_{\Lambda_n}, X'_{\Lambda_n} = \hat{x}_{\Lambda_n}|\text{(i)},\text{(ii)},\text{(iii)})} \\
&= \frac{\frac{1}{Z_x}\exp\left(\sum_{\substack{\langle s,t\rangle \\ s,t\in\Lambda_n}}\hat{x}_s\hat{x}_t + \sum_{\substack{\langle s,t\rangle \\ s\in\Lambda_n, t\in\partial(\Lambda_n)}}\hat{x}_s x_t + h_1\sum_{x\in\Lambda_n}\hat{x}_s\right)}{\frac{1}{Z_x}\exp\left(\sum_{\substack{\langle s,t\rangle \\ s,t\in\Lambda_n}}\check{x}_s\check{x}_t + \sum_{\substack{\langle s,t\rangle \\ s\in\Lambda_n, t\in\partial(\Lambda_n)}}\check{x}_s x_t + h_1\sum_{x\in\Lambda_n}\check{x}_s\right)} \\
&\quad\times\frac{\frac{1}{Z_x}\exp\left(\sum_{\substack{\langle s,t\rangle \\ s,t\in\Lambda_n}}\check{x}_s\check{x}_t + \sum_{\substack{\langle s,t\rangle \\ s\in\Lambda_n, t\in\partial(\Lambda_n)}}\check{x}_s x'_t + h_1\sum_{x\in\Lambda_n}\check{x}_s\right)}{\frac{1}{Z_x}\exp\left(\sum_{\substack{\langle s,t\rangle \\ s,t\in\Lambda_n}}\hat{x}_s\hat{x}_t + \sum_{\substack{\langle s,t\rangle \\ s\in\Lambda_n, t\in\partial(\Lambda_n)}}\hat{x}_s x'_t + h_1\sum_{x\in\Lambda_n}\hat{x}_s\right)} \\
&\leq \exp\left(\beta\left(4|\partial(\Lambda_n)| + \frac{\Delta}{2}(h_1-h_2)|\Lambda_n|\right)\right)
\end{aligned}$$

which tends to 0 as $n \to \infty$ due to the fact that $\frac{|\partial(\Lambda_n)|}{|\Lambda_n|} \to 0$.

So with probability at least $1 - \delta$, X' has a.s. the higher total spin in large enough boxes, and since $\delta > 0$ was arbitrary it in fact holds a.s. Hence the limit in (24) is smaller than the limit in (25), and the lemma is proved. $\diamond$

That $\frac{|\partial(\Lambda_n)|}{|\Lambda_n|} \to 0$ is in fact crucial here, and if $\mathbf{Z}^d$ is replaced by some lattice for which no sequence of finite subsets with vanishing surface-to-volume ratio exists (so-called **nonamenable** lattices), Proposition h *fails*, as discussed in [JS].

12 Friday, May 16

No deep deriavtions in today's final lecture, just brief expositions of two other Markov random field models we have't had time to discuss before: the Potts model, and Markov random fields.

*

The Potts model is a natural extension of the Ising model to larger state spaces: $\{-1, +1\}$ is replaced by $\{1, \ldots, q\}$. Taking $q = 2$ just gives back the Ising model with new symbols, whereas $q \geq 3$ gives something genuinely different.

For finite S and $\beta \geq 0$, let $H : \{1, \ldots, q\}^S \to \mathbf{R}$ be given by

$$H(x) = -2\beta \sum_{\langle s,t \rangle} \mathbf{1}_{\{x_s = x_t\}}$$

and probability measure Π on $\{1, \ldots, q\}^S$ given by, as usual,

$$\Pi(x) = \frac{1}{Z} \exp(-H(x)) \,.$$

(Why the 2 in the formula for $H(x)$? It's just a matter of definition, of course, but it's there to harmonize with the Ising model, where the summands $x_s x_t$ vary between two values differing by 2, whereas here the summands vary just between 0 and 1, differing by 1.)

The extension to $\mathbf{Z}^d$ works the same way as for the Ising model as far as definitions go, but some of the arguments for existence and uniqueness of Gibbs measures become harder, because the $q \geq 3$ Potts model does not enjoy quite the same stochastic domination properties as the Ising moedl. Still, the main theorem quoted in Lecture 6 concerning phase transition in

the $\mathbf{Z}^d$ Ising model goes through in the Potts model: for fixed $q \geq 2$ and $d \geq 2$, there is a critical value $\beta_c \in (0, \infty)$ such that the q-state Potts model on $\mathbf{Z}^d$ has a unique Gibbs measure when $\beta < \beta_c$ and multiple Gibbs measures when $\beta > \beta_c$.

One difference between phase transition behavior the Ising and the $q \geq 3$ Potts cases is that, while in both cases the magnetization (suitably defined and normalized) is 0 for $\beta < \beta_c$ and positive for $\beta > \beta_c$, it takes off continuously at $\beta = \beta_c$ in the Ising case, and has a jump discontinuity in the $q \geq 3$ Potts cases; this is of great interest in statistical mechanics.

A major tool for studying the Potts model is the so-called **random-cluster representation**, defined as follows.

Fix S finite, and neighborhood system ∂, and define the **edge set**

$$E = \{\langle s, t\rangle \in S^2 : s, t \text{ neighbors}\}.$$

Fix q and β , let $p = 1 - e^{-2\beta}$, and do as follows:

1. Let $X \in \{1, \ldots, q\}^S$ be i.i.d. uniform on $\{1, \ldots, q\}$.

2. Independently of the first step, let $Y \in \{0, 1\}^E$ be i.i.d. with each edge having probability p of taking value 1 (interpreted as "retained") and probability $1 - p$ of taking value 0 ("deleted").

3. Condition on the event that $X_s = X_t$ for all $s, t \in S$ such that $\langle s, t\rangle \in E$ and $Y_{\langle s,t\rangle} = 1$.

It turns out that if we do this, then X has distribution Π (the Potts model with parameters q and β). The random edge configuration q gets a distribution ν on $\{0, 1\}^E$ which is known as the **random-cluster model** with parameter p and q, characterized by

$$\nu(Y = y) = \frac{1}{Z} q^{k(y)} \prod_{e \in E} p^{y(e)} (1 - p)^{1 - y(e)}$$

where $k(y)$ is the number of connected components in the edge configuration y.

The distributions of X and Y both have intricate dependencies, but it turns out that in this coupling (the so-called **Edwards–Sokal coupling**) the conditional distribution of X given Y, as well as the conditional distribution of Y given X are both very simple. The former is that on each connected component, a spin value is chosen uniformly from $\{1, \ldots, q\}$ to be assigned to all vartiecs in the component, and this is done inedpendently for

different components. The latter is that given X, the edge variables are independent, with $\langle s,t\rangle$ having probability p of taking value 1 if $X(s) = X(t)$, and probability 0 otherwise.

This beautiful dependence structure can be exploited for at least tow purposes:

(i) To reduce difficult questions about dependencies in the Potts moedl to comparatively easier questions about connectivity probabilities in the random-cluster model; Chapter 6 of [GHM] contanis extensive discussion of this.

(ii) To simulate Π by going back and forth between X and Y in Gibbs sampler style. This is the so-called Swendsen–Wang algorithm, which turns out in practice to be more efficient (although less flexible) than the single-site Gibbs sampler discussed in Lecture 6.

*

Next, **Gaussian Markov random fields**. Let S and E be as before (finite), and let $B = \{s_1, \ldots, s_m\}$ be a subset of S. Fix $b_1, \ldots, b_m \subset \mathbf{R}$ (the boundary condition) and $\sigma^2 > 0$ (the variance parameter), and pick $X \in \mathbf{R}^S$ as follows.

First let $X_{s_i} = b_i$ for each $s_i \in B$. Then pich $X_{S\setminus B}$ according to density

$$\frac{1}{Z} \exp\left(-\frac{1}{2\sigma^2} \left(\sum_{\substack{\langle s,t\rangle \\ s,t\in S\setminus B}} (x_s - s_t)^2 + \sum_{\substack{\langle s,t\rangle \\ s\in S\setminus B, t\in B}} (x_s - x_t)^2 \right) \right).$$

It then turns out

(a) that X is a Markov random field in the obvious analogous sense to the discrete case: the distribution of X_A given $X_{S\setminus A}$ depends only on $X_{\partial(A)}$, and

(b) that $X_{S\setminus B}$ is **Multivariate Gaussian**.

The model is na a sense isomorphic, in a way that is both mathematically beautiful and useful, to random walks and electrical networks. The connections are outlined in Section 9.4 of [J].

For instance, calculating $E[S_s]$ for $s \in S \setminus B$ is equivalent to either of the following two:

(i) Run a simple random walk on the network (S, E) starting at s, and calculate the expected value of b_i at the first site in B encountered by the random walk.

(ii) Consider the electrical network om (S, E) with σ^2-ohm resistors on the edges, and voltages $b_1, \ldots, b_m$ applied at B, and calculate the resulting voltage at s.

The relation between $Var[X_s]$ and the random walk and electrical network formulations are even more interersting, and involve effective resistances and return probabilities. The fact that simple random walk on $\mathbf{Z}^d$ is recurrent for $d = 1, 2$ and transient for $d \geq 3$ is essentially the same thing as the following fact for Gaussian Markov random fields. If we consider the Gaussian Markov random field on $\Lambda_n \cup \partial(\Lambda_n)$ with the usual neighborhood structure, $B = \partial(\Lambda_n)$ and b_i identically 0, then the variance X_0 at the origin tends to ∞ with n for $d = 1$ and 2, but remains bounded. This means that an important limiting object known as the **discrete Gaussian free field** can be directly defined by this limiting procedure for $d \geq 3$, but requires a different formalism for $d = 1$ and 2.

Bibliography

[GHM] Georgii, H.-O., Häggström, O. and Maes, C. (2001) The random geometry of equilibrium phases, *Phase Transitions and Critical Phenomena, Volume 18* (C. Domb and J.L. Lebowitz, eds), pp 1-142, Academic Press, London.

[J] Janson, S. (1997) *Gaussian Hilbert Spaces*, Cambridge University Press.

[JS] Jonasson, J. and Steif, J. (1998) Amenability and phase transition in the Ising model, *J. Theor. Probab* **12**, 549–559.

[W] Winkler, G. (1995) *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods: A Mathematical Introduction*, Springer, Berlin.