On Möbius Transforms and Boltzmann Machines

Jason K. Johnson

August, 2006

Abstract

We consider the exponential family representation of Boltzmann machines (a collection of n binary-valued random variables) and study the algebraic structure of this family with respect to the Möbius transform. Also, we introduce a generalized class of Möbius transforms and introduce a "fast" algorithm for computing these transforms in $(n-1)2^{(n-1)}$ calculations (rather than $\mathcal{O}(2^{2n})$ calculations). We consider possible implications of this algorithm for inference and learning in Boltzmann machines.

1 Introduction

Throughout this note we will be concerned with functions that are either defined on the set of all subsets of $\{1, ..., n\}$ or that are defined on $\{0,1\}^n$. Both of these domains have cardinality 2^n and can therefore be identified with the set $\{0, 1, \ldots, 2^n - 1\}$.

To each subset $x \subseteq \{1, ..., n\}$ we identify a bit-vector $x_n \ldots x_1 \in \{0,1\}^n$ with $x_i = 1$ if $i \in x$ and $x_i = 0$ otherwise. To each $x \in \{0,1\}^n$ we can also identify an integer $x = \sum_{i=1}^n x_i 2^{i-1}$. Thus, $x_n \ldots x_1$ is just the binary expansion of the integer x. Employing this convention, we may view x as specifying either an integer, a bit-vector or a subset. For example, with $n = 2$ we have the following "states":

$$\begin{aligned} 0 &= 00 = \emptyset \\ 1 &= 01 = \{1\} \\ 2 &= 10 = \{2\} \\ 3 &= 11 = \{1, 2\} \end{aligned} \tag{1}$$

We will write $x \preceq y$ (equivalently, $y \succeq x$) if $x \subseteq y$ (equivalently, $x_i \leq y_i$ for $i = 1, ..., n$).

A set-valued function $f(x)$ (equivalently, a function of n binary variables $f(x_1, ..., x_n)$) can then be represented as a vector in $\mathbb{R}^{2^n}$ by defining the x-th element of the vector representation of f (with indices running from 0 to $2^n - 1$) to be $f(x)$. Let $\delta(x)$ denote the standard basis vector of $\mathbb{R}^{2^n}$, e.g. for $n = 2$:

$$\begin{aligned} \delta(0) &= (1, 0, 0, 0)^T \\ \delta(1) &= (0, 1, 0, 0)^T \\ \delta(2) &= (0, 0, 1, 0)^T \\ \delta(3) &= (0, 0, 0, 1)^T \end{aligned} \tag{2}$$

Employing this convention, we may rewrite $f(x)$ as $f^T \delta(x)$.

2 Möbius Transforms

We consider a generalized class of Möbius transforms. Let $f \in \mathbb{R}^{2^n}$. We define the ω-transform by:

$$(L_n^\omega f)(x) = \sum_{y \preceq x} \omega^{|x \setminus y|} f(y) \tag{3}$$

where $|x \setminus y|$ is this number of elements of x that are not contained in y (the number of bits that are 1 in x and 0 in y). Note that this defines a family $\{L_n^\omega, \omega \in \mathbb{C}\}$ of linear operators on $\mathbb{R}^{2^n}$. The usual Möbius transform is given by $L_n \equiv L_n^1$ and the inverse Möbius transform by L_n^{-1}.

For example, let $n = 2$, $f \in \mathbb{R}^4$ and $g = L_2^\omega f$:

$$\begin{pmatrix} g(00) \\ g(01) \\ g(10) \\ g(11) \end{pmatrix} = \begin{pmatrix} 1 & & & \\ \omega & 1 & & \\ \omega & 0 & 1 & \\ \omega^2 & \omega & \omega & 1 \end{pmatrix} \begin{pmatrix} f(00) \\ f(01) \\ f(10) \\ f(11) \end{pmatrix} \tag{4}$$

We now summarize some interesting properties of the ω-transform.

Proposition 1 *For $n = 0$, $L_0^\omega = 1$ for all ω. For $n \geq 1$:*

$$L_n^\omega = \begin{pmatrix} L_{n-1}^\omega & 0 \\ \omega L_{n-1}^\omega & L_{n-1}^\omega \end{pmatrix} \tag{5}$$

Proof. Let $g = L_n^\omega f$, $f = (f_1, f_2)$ and $g = (g_1, g_2)$ with $f, g \in \mathbb{R}^{2^n}$ and $f_i, g_i \in \mathbb{R}^{2^{n-1}} (i = 1, 2)$. We must show $g_1 = L_{n-1}^\omega f_1$ and $g_2 = \omega L_{n-1}^\omega f_1 + L_{n-1}^\omega f_2$. For $x, y \in \{0, \ldots, 2^n - 1\}$, $x \leq y$ implies that $x \preceq y$ (bitwise). Thus,

$$g_1(x) = \sum_{y : y < 2^{n-1}, y \preceq x} \omega^{|x \setminus y|} f_1(y) = (L_{n-1}^\omega f_1)(x) \tag{6}$$

Let $\pi_n(x) = x \mod 2^{n-1}$. For $x \in \{2^{n-1}, \ldots, 2^n - 1\}$ this has the effect of setting the n-th bit of x to zero, i.e. $\pi_n(x_n x_{n-1} \ldots x_1) = 0 x_{n-1} \ldots x_1$. Now, for $x \geq 2^{n-1}$ we have:

$$\begin{aligned} g_2(x) &= \sum_{y : y < 2^{n-1}, y \preceq x} \omega^{|x \setminus y|} f_1(y) + \sum_{y : y \geq 2^{n-1}, y \preceq x} \omega^{|x \setminus y|} f_2(\pi_n(y)) \\ &= \sum_{y : y < 2^{n-1}, y \preceq \pi_n(x),} \omega^{|\pi_n(x) \setminus y)| + 1} f_1(y) + \sum_{y : y \geq 2^{n-1}, \pi_n(y) \preceq \pi_n(x)} \omega^{|\pi_n(x) \setminus \pi_n(y)|} f_2(\pi_n(y)) \\ &= \omega (L_{n-1}^\omega f_1)(x) + (L_{n-1}^\omega f_2)(x) \end{aligned} \tag{7}$$

which proves the proposition. □

This result provides the basis for computing fast ω-transforms. Given $f = (f_1, f_2) \in \mathbb{R}^{2^n}$ with $f_1, f_2 \in \mathbb{R}^{2^{n-1}}$ we can compute the transform of f as $\tilde{f} = (\tilde{f}_1, \omega \tilde{f}_1 + \tilde{f}_2)$ where $\tilde{f}_1$ and $\tilde{f}_2$ are the ω-transforms of f_1 and f_2. Computing these recursively, we obtain an algorithm to apply the operator L_n^ω which requires $(n-1)2^{n-1}$ computations rather than $\mathcal{O}(2^{2n})$.

Proposition 2 *The ω-transforms form a commutative group:*

1. $L_n^\alpha L_n^\beta = L_n^{\alpha+\beta} = L_n^\beta L_n^\alpha$.
2. $L_n^0 = I_{2^n}$.
3. $(L_n^\omega)^{-1} = L_n^{-\omega}$.

Proof. We show (1) as follows. First, for $n = 0$ it trivially holds that $L_0^\alpha L_0^\beta = 1 = L_0^{\alpha+\beta}$. Then, by Proposition 1 and induction on n we have:

$$\begin{aligned} L_{n+1}^\alpha L_{n+1}^\beta &= \begin{pmatrix} L_n^\alpha & 0 \\ \alpha L_n^\alpha & L_n^\alpha \end{pmatrix} \begin{pmatrix} L_n^\beta & 0 \\ \beta L_n^\beta & L_n^\beta \end{pmatrix} \\ &= \begin{pmatrix} L_n^\alpha L_n^\beta & 0 \\ (\alpha+\beta) L_n^\alpha L_n^\beta & L_n^\alpha L_n^\beta \end{pmatrix} \\ &= \begin{pmatrix} L_n^{\alpha+\beta} & \\ (\alpha+\beta) L_n^{\alpha+\beta} & L_n^{\alpha+\beta} \end{pmatrix} \\ &= L_{n+1}^{\alpha+\beta} \end{aligned} \tag{8}$$

The remaining points are then self-evident. □

There also is an "upper" ω-transform defined as:

$$(U_n^\omega f)(x) = \sum_{y \succeq x} \omega^{|y \setminus x|} f(y) \tag{9}$$

Note that the sum is now over all supersets of a set rather than the subsets. However, it turns out it is just the transpose of the "lower" ω-transform, i.e. $U_n^\omega = (L_n^\omega)^T$. There also is a fast $(n-1)2^{(n-1)}$ implementation of upper ω-transforms.

3 Boltzmann Machines

We now consider the family of probability distributions on $\{0,1\}^n$ which may be parameterized in the form of an exponential family:

$$p(x) = p(x_1, \ldots, x_n) = \exp\{\theta^T \phi(x)\} \tag{10}$$

with exponential parameters $\theta \in \mathbb{R}^{2^n}$ and sufficient statistics $\phi(x) = (\phi_a(x), a \subseteq \{1, ..., n\})$ defined by $\phi_a(x) = \prod_{i \in a} x_i$. Note that $\phi_a(x) = 1$ if $x_i = 1$ for all $i \in a$ and is zero otherwise. The moment parameters of the family are defined by $\eta = \sum_x p(x)\phi(x)$.

Both η and θ are naturally viewed as vectors in $\mathbb{R}^{2^n}$ indexed by subsets of $\{1, ..., n\}$. Parameter $\theta(s)$ is the multiplier of $\phi_s(x)$. The moment $\eta(s)$ is the probability that $x_i = 1$ for all $i \in s$. Likewise, the probability mass function p may also be viewed as an element of $\mathbb{R}^{2^n}$ indexed by joint states $(x_1, \ldots, x_n)$ of the n binary variables.

Next, we show that these three vector representations p, η and θ are connected by the Möbius transform:

Proposition 3 *We have the following Möbius transform relations:*

1. $\phi(x) = L_n^T \delta(x)$ *and* $\delta(x) = L_n^{-T} \phi(x)$.
2. $\eta = L_n^T p$ *and* $p = L_n^{-T} \eta$.
3. $p = \exp(L_n \theta)$ *and* $\theta = L_n^{-1} \log p$.

Proof. (1) Element s of $\phi(x)$ is one if $s \preceq x$ (bitwise) and is zero otherwise. On the other hand, we have

$$(L_n^T \delta(x))(s) = \sum_{t \succeq s} \delta(x)(t) \tag{11}$$

which also is one if $s \preceq x$ and is zero otherwise. (2) $\eta = \mathbb{E}\{\phi(x)\} = \mathbb{E}\{L_n^T \delta(x)\} = L_n^T \mathbb{E}\{\delta(x)\} = L_n^T p$. (3) $\log p(x) = (\log p)^T \delta(x) = \theta^T \phi(x) = \theta^T L_n^T \delta(x) = (L_n \theta)^T \delta(x)$ for all x. Hence,

$\log p = L_n\theta$. $\square$

Thus, inference (computing η from θ) corresponds to the map:

$$\Lambda(\theta) = L_n^T \exp(L_n\theta) \tag{12}$$

while learning (computing θ from η) corresponds to the inverse map:

$$\Lambda^{-1}(\eta) = L_n^{-1} \log(L_n^{-T}\eta) \tag{13}$$

Both maps require $n2^n$ calculations using the fast Möbius transform.

Normalization We should point out that all three representations are overparameterized as we have not yet imposed the normalization constraint $1^T p = 1$. In the η parameterization, normalization reduces to the requirement that $1^T L_n^{-T}\eta = (L_n^{-1}1)^T\eta = \delta(\emptyset)^T\eta = \eta(\emptyset) = 1$. In the θ parameterization, we must have:

$$-\theta(\emptyset) = \Phi(\theta_{\setminus\emptyset}) \equiv \log \sum_x \exp \sum_{s\neq\emptyset} \theta_s\phi_s(x) \tag{14}$$

The function Φ is known as the *cumulant generating function* of the exponential family (in statistical physics it is called the *log-partition function*).

Marginalization The marginal distribution on a subset of random variables s is a function on the set of all joint states of $x_s = (x_i, i \in s)$ defined as:

$$p_s(x_s) = \sum_{y:y_i=x_i\forall i\in s} p(y) \tag{15}$$

for each $x_s \in \{0,1\}^{|s|} \equiv \mathbb{R}^{2^{|s|}}$. This defines a linear map $\Sigma_s : \mathbb{R}^{2^n} \to \mathbb{R}^{2^{|s|}}$, i.e. $p_s = \Sigma_s p$.

We now consider marginalization in the context of the η parameterization. Each subset $s \subseteq \{1,\ldots,n\}$ determines an injective map $\sigma_s : \{1,\ldots,|s|\} \to \{1,\ldots,n\}$ defined by ordering the elements, i.e. $s = \{\sigma_s(k), k = 1,\ldots,|s|\}$ where $\sigma_s(1) < \sigma_s(2) < \cdots < \sigma_s(|s|)$. Thus, given $t \subseteq s$ we have that $\sigma^{-1}(t)$ is the corresponding subset of $\{1,\ldots,|s|\}$ (an element of $\{0,\ldots,2^{|s|}\}$). Now, we may define Π_s by $(\Pi_s f)(\sigma^{-1}(t)) = f(t)$ for all $t \preceq s$. Thus, Π_s just gathers the elements of f which correspond to subsets of s. We now show that $\eta_s = \Pi_s\eta$ are precisely the moment parameters associated to the marginal distribution p_s, i.e.:

Proposition 4 $L_{|s|}^T \Sigma_s = \Pi_s L_n^T$.

Proof. For $t \subseteq s$ we have:

$$\begin{aligned}
\eta(t) &= \sum_{u\in\{0,1\}^{|s|}:u_i=1,\forall i\in\sigma^{-1}(t)} \sum_{v\in\{0,1\}^{n-|s|}} p(\sigma_s(u)\cup\sigma_{\setminus s}(v)) \\
&= \sum_{u:u_i=1,\forall i\in\sigma^{-1}(t)} p_s(u) \\
&= \eta_s(\sigma^{-1}(t))
\end{aligned} \tag{16}$$

$\square$

Fisher Information We consider the second moment of the statistics $\phi(x)$:

$$\begin{aligned} K(\theta) &= \mathbb{E}_\theta\{\phi(x)\phi^T(x)\} \\ &= L_n^T \mathbb{E}_\theta\{\delta(x)\delta^T(x)\} L \\ &= L_n^T \mathrm{Diag}(p_\theta) L_n \end{aligned} \tag{17}$$

which is a symmetric positive semi-definite matrix. Also, we define the inverse matrix parameterized by η:

$$\begin{aligned} K^*(\eta) &\equiv K^{-1}(\Lambda^{-1}(\eta)) \\ &= L_n^{-1} \mathrm{Diag}(1/p_\eta) L_n^{-T} \end{aligned} \tag{18}$$

Note that both matrices submit to a "fast" implementation as a linear operator employing the fast Möbius transforms. It is easily verified that:

$$\begin{aligned} K(\theta) &= \frac{\partial \Lambda(\theta)}{\partial \theta} \\ K^*(\eta) &= \frac{\partial \Lambda^{-1}(\eta)}{\partial \eta} \end{aligned} \tag{19}$$

Hence, we expect these linear operators may prove useful in variational methods for inference and learning. In fact, these matrices are closely related to the Fisher information matrices associated with the η and θ parameterizations (imposing normalization eliminates $\theta_\emptyset$ and $\eta_\emptyset$):

$$\begin{aligned} G(\theta) &= (K(\theta) - \Lambda(\theta)\Lambda(\theta)^T)_{\setminus\emptyset,\setminus\emptyset} \\ G^*(\eta) &= K^*(\eta)_{\setminus\emptyset,\setminus\emptyset} \end{aligned} \tag{20}$$

Note also, $G(\theta)$ is the Schur complement of $K(\theta)$ obtained by eliminating the first row/column associated with the $\theta_\emptyset$ parameter since

$$K = \begin{pmatrix} 1 & \eta_{\setminus\emptyset}^T \\ \eta_{\setminus\emptyset} & K_{\setminus\emptyset,\setminus\emptyset} \end{pmatrix} \tag{21}$$

and $G = K_{\setminus\emptyset,\setminus\emptyset} - \eta_{\setminus\emptyset}\eta_{\setminus\emptyset}^T$.