Failures of the One-Step Learning Algorithm

David J.C. MacKay

Cavendish Laboratory, Madingley Road, Cambridge CB3 0HE. United Kingdom.

mackay@mrao.cam.ac.uk

August 22, 2001 — Draft 3.1

Abstract

The Hinton network (Hinton, 2001, personal communication) is a deterministic mapping from an observable space $\mathbf{x}$ to an energy function $E(\mathbf{x};\mathbf{w})$, parameterized by parameters $\mathbf{w}$. The energy defines a probability $P(\mathbf{x}|\mathbf{w}) = \exp(-E(\mathbf{x};\mathbf{w}))/Z(\mathbf{w})$.

A maximum likelihood learning algorithm for this density model takes steps $\Delta\mathbf{w} \propto -\langle\mathbf{g}\rangle_0 + \langle\mathbf{g}\rangle_\infty$ where $\langle\mathbf{g}\rangle_0$ is the average of the gradient $\mathbf{g} = \partial E/\partial\mathbf{w}$ evaluated at points $\mathbf{x}$ drawn from the data density, and $\langle\mathbf{g}\rangle_\infty$ is the average gradient for points $\mathbf{x}$ drawn from $P(\mathbf{x}|\mathbf{w})$.

If T is a Markov chain in $\mathbf{x}$-space that has $P(\mathbf{x}|\mathbf{w})$ as its unique invariant density then we can approximate $\langle\mathbf{g}\rangle_\infty$ by taking the data points $\mathbf{x}$ and hitting each of them I times with T, where I is a large integer. In the one-step learning algorithm of Hinton (2001), we set I to 1.

In this paper I give examples of models $E(\mathbf{x};\mathbf{w})$ and Markov chains T for which the true likelihood is unimodal in the parameters, but the one-step algorithm does not necessarily converge to the maximum likelihood parameters.

It is hoped that these negative examples will help pin down the conditions for the one-step algorithm to be a correctly convergent algorithm.

The Hinton network (Hinton, 2001, personal communication) is a deterministic mapping from an observable space $\mathbf{x}$ of dimension D to an energy function $E(\mathbf{x};\mathbf{w})$, parameterized by parameters $\mathbf{w}$. The energy defines a probability

$$P(\mathbf{x}|\mathbf{w}) = \frac{\exp(-E(\mathbf{x};\mathbf{w}))}{Z(\mathbf{w})}, \tag{1}$$

where

$$Z(\mathbf{w}) = \int d^D\mathbf{x}\ \exp(-E(\mathbf{x};\mathbf{w})) \tag{2}$$

is the hard-to-evaluate normalizing constant or partition function.

A maximum likelihood learning algorithm for this density model takes steps

$$\Delta\mathbf{w} \propto -\langle\mathbf{g}\rangle_0 + \langle\mathbf{g}\rangle_\infty\,, \tag{3}$$

where $\langle\mathbf{g}\rangle_0$ is the average of the gradient $\mathbf{g} = \partial E/\partial\mathbf{w}$ evaluated at points $\mathbf{x}$ drawn from the data density, and $\langle\mathbf{g}\rangle_\infty$ is the average gradient for points $\mathbf{x}$ drawn from $P(\mathbf{x}|\mathbf{w})$.

If T is a Markov chain in $\mathbf{x}$-space that has $P(\mathbf{x}|\mathbf{w})$ as its unique invariant density then we can approximate $\langle\mathbf{g}\rangle_\infty$ by taking the data points $\mathbf{x}$ and hitting each of them I times with T, where I

is a large integer, and evaluating the gradient $\mathbf{g}_I$ at each resulting transformed point; for each data point a step

$$\Delta\mathbf{w} \propto -\mathbf{g}_0 + \mathbf{g}_I \tag{4}$$

is taken. Empirically, it has been found that, even for small I, the learning algorithm converges to the maximum likelihood answer; indeed, with small I, the algorithm can converge more rapidly because the difference $\mathbf{g}_0 - \mathbf{g}_I$ can be a less noisy quantity than $\mathbf{g}_0 - \mathbf{g}_\infty$. In the one-step learning algorithm of Hinton, we set I to 1.

The fact that $-\mathbf{g}_0 + \mathbf{g}_I$ is a biased estimate of the gradient of the likelihood $-\langle\mathbf{g}\rangle_0 + \langle\mathbf{g}\rangle_\infty$ is not important as long as the algorithm converges to the maximum likelihood parameters. It would be nice to know a set of conditions for this convergence property to hold.

In this paper I give examples of models $E(\mathbf{x};\mathbf{w})$ and Markov chains T for which the true likelihood is unimodal in the parameters, but *the one-step algorithm does not necessarily converge to the maximum likelihood parameters.*

It is hoped that these negative examples will help pin down the conditions for the one-step algorithm to be a correctly convergent algorithm.

A publication by Williams *et al.* contains a closely related positive example: when the model is Gaussian and the Markov chain is Gibbs sampling, the one-step algorithm converges to the maximum likelihood parameters (Williams, 2001).

1 Toy example

The toy model I study is the axis-aligned Gaussian distribution whose energy function is

$$E(\mathbf{x};\{\mu,\sigma\}) = \sum_{d=1}^{D} \frac{1}{2}\frac{(x_d-\mu_d)^2}{\sigma_d^2}. \tag{5}$$

The number of dimensions, D, will be 2. At times it may be convenient to parameterize the standard deviation via $\tau_d \equiv 1/\sigma_d^2$.

In the examples that follow, we will fit this energy function to a data set of N points $\{\mathbf{x}^n\}$, adjusting either the means $\{\mu_d\}$ or the standard deviations $\{\sigma_d\}$ or both. The likelihood function is unimodal and the joint maximum likelihood parameters are

$$\mu_d^{\mathrm{ML}} = \frac{1}{N}\sum_n x_d \tag{6}$$

$$\sigma_d^{2\,\mathrm{ML}} = \frac{1}{N}\sum_n (x_d - \mu_d^{\mathrm{ML}})^2. \tag{7}$$

There are many transition operators T which leave the Gaussian distribution invariant.

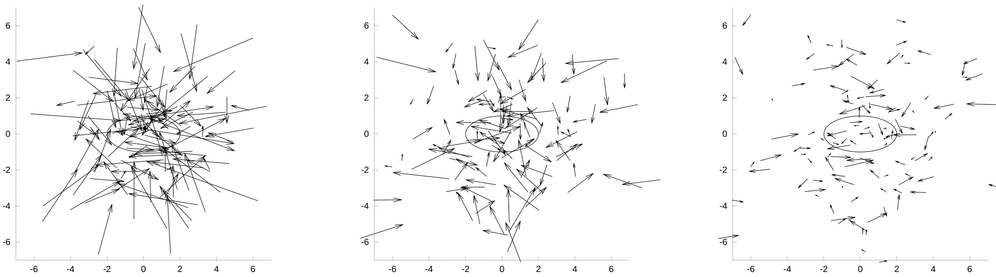
The drift-and-diffuse operator is

$$\mathbf{x} \to \mathbf{x}' = \mu + \alpha(\mathbf{x}-\mu) + \sqrt{1-\alpha^2}\mathbf{M}\mathbf{n} \tag{8}$$

where $\mathbf{n}$ is a standard normal random vector,

$$\mathbf{M} = \left[\begin{array}{cc} \sigma_1 & 0 \\ 0 & \sigma_2 \end{array}\right], \tag{9}$$

and $-1 < \alpha < 1$. This transition operator has the Gaussian as its unique invariant distribution. It also satisfies detailed balance.



$\alpha = 0.5$ $\alpha = 0.8$ $\alpha = 0.95$

In this figure the ellipse indicates the axis-aligned Gaussian to which the transitions are matched.

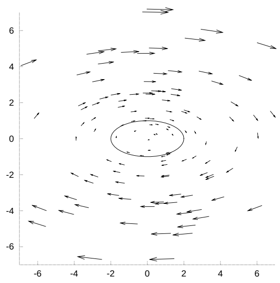
The swirl operator rotates all points $\mathbf{x}$ around the hypothesized mean:

$$\mathbf{x} \to \mathbf{x}' = \mu + \mathbf{R}(\mathbf{x} - \mu), \tag{10}$$

where

$$\mathbf{R} = \mathbf{M} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \mathbf{M}^{-1} \tag{11}$$

is the operator that stretches the Gaussian into circularly symmetric shape, rotates all the points, then turns the circular distribution back into the required unequal-variances distribution.



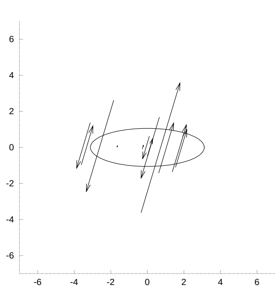
The flip operator reflects all points in a plane passing through the mean:

$$\mathbf{x} \to \mathbf{x}' = \mu + \mathbf{F}(\mathbf{x} - \mu), \tag{12}$$

where

$$\mathbf{F} = \mathbf{M} \begin{bmatrix} \cos 2\alpha & \sin 2\alpha \\ \sin 2\alpha & -\cos 2\alpha \end{bmatrix} \mathbf{M}^{-1} \tag{13}$$

is the operator that stretches the Gaussian into circularly symmetric shape, reflects in a plane at angle α, then turns the circular distribution back into the required unequal-variances distribution.



The Star Trek operator or jet operator takes points that are within a certain distance of the mean and moves them radially away from the mean, and maps points beyond that distance to points closer to the mean.

We describe the mapping $\mathbf{x} \rightarrow \mathbf{x}'$ in five steps:

1. Spherify
$$\mathbf{z} := \mathbf{M}^{-1}(\mathbf{x} - \mu), \tag{14}$$
and define polar coordinates (r, θ) in the usual way: $r = \sqrt{z_1^2 + z_2^2}$.

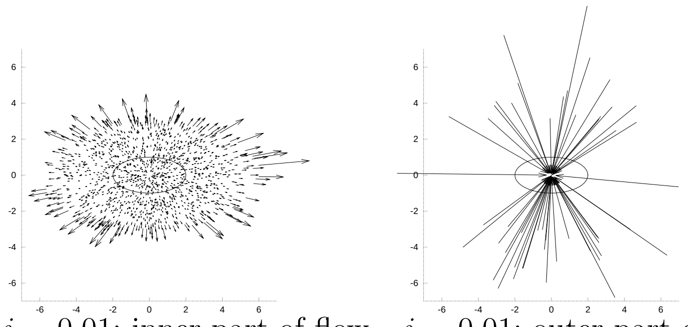
2. Extract from r the variable u
$$u := \exp(-r^2/2) \tag{15}$$
[If $\mathbf{z}$ is standard-normal then u is uniformly distributed between 0 and 1.]

3. Perturb u through a distance j, where j is the strength of the 'jet':
$$u' := (u - j) \bmod 1 \tag{16}$$

4. Recover r'
$$r' := \sqrt{2 \log 1/u'} \tag{17}$$

5. Recover $\mathbf{x}'$
$$\mathbf{x}' = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \mathbf{M} \begin{bmatrix} r' \cos\theta \\ r' \sin\theta \end{bmatrix}. \tag{18}$$

‘



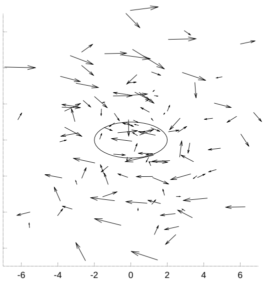
$j = 0.01$; inner part of flow $j = 0.01$; outer part of flow

The Gaussian is invariant under the swirl, flip, and Star Trek operators but these operators are not ergodic. To obtain transition operators that have the Gaussian as their unique invariant distribution, we can concatenate each operator with the drift-and-diffuse operator.

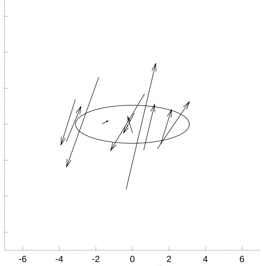
The noisy swirl operator is the concatenation of the swirl operator with the drift-and-diffuse operator.

The noisy flip operator is the concatenation of the flip operator with the drift-and-diffuse operator. The noisy flip operator satisfies detailed balance.

The noisy Star Trek operator is the concatenation of the Star Trek operator with the drift-and-diffuse operator.



Noisy swirl



Noisy flip

The swirl, flip, and Star Trek operators are a little contrived, but they are similar in character to advanced Monte Carlo operators, introduced to reduce random walk behaviour, such as the hybrid Monte Carlo method and overrelaxation. So problems found with the swirl, flip, and Star Trek operators are a warning that problems might also arise with advanced Monte Carlo operators.

2 Results

We first summarise the results that are proved in the next section.

1. Under the drift-and-diffuse operator, the one-step algorithm for learning $\{\mu\}$ and $\{\sigma\}$ converges to the maximum-likelihood parameters.

2. Under the noisy swirl operator, the one-step algorithm for learning $\{\mu\}$ *alone* converges to the maximum-likelihood parameters.

3. Under the noisy swirl operator, the one-step algorithm for learning $\{\mu\}$ and $\{\sigma\}$ does *not* necessarily converge to the maximum-likelihood parameters. Typically the algorithm has a single fixed point but the values of $\{\sigma\}$ at that fixed point are wrong; there is *no* fixed point at the maximum likelihood parameters. [Easy to confirm with a simple example: two data points at (1,1) and (-1,-1).]

4. In the special case of an infinite amount of data that come from an axis-aligned Gaussian, the noisy swirl operator's one-step algorithm for learning $\{\mu\}$ and $\{\sigma\}$ does converge to the maximum-likelihood parameters.

5. Under the noisy flip operator, the one-step algorithm for learning $\{\mu\}$ and $\{\sigma\}$ does *not* necessarily converge to the maximum-likelihood parameters. Typically the algorithm has a single fixed point but the values of $\{\sigma\}$ at that fixed point are wrong; there is *no* fixed point at the maximum likelihood parameters. [Easy to confirm with a simple example: two data points at (1,1) and (-1,-1).] This result is interesting because the noisy flip satisfies detailed balance; this proves false the conjecture that detailed balance would be a sufficient condition for convergence to the ML parameters.

6. Under the noisy Star Trek operator, the one-step algorithm for learning $\{\mu\}$ and $\{\sigma\}$ typically does *not* converge to the maximum-likelihood parameters. Typically the algorithm has many fixed points. [Found numerically.]

3 Details

3.1 Drift-and-diffuse

Under the drift-and-diffuse operator, the one-step algorithm for learning $\{\mu\}$ and $\{\sigma\}$ converges to the maximum-likelihood parameters.

We assume the step size parameter is very small and compute the expected direction of the step. The derivative of E with respect to μ is

$$\frac{\partial E}{\partial \mu_d} = \frac{1}{\sigma_d^2}(\mu_d - x_d) \tag{19}$$

and the derivative with respect to $\log \sigma_d$ is

$$\frac{\partial E}{\partial \log \sigma_d} = -\frac{1}{\sigma_d^2}(\mu_d - x_d)^2 \tag{20}$$

Each data point $\mathbf{x}$ is moved to a point

$$\mathbf{x}' = \mu + \alpha(\mathbf{x} - \mu) + \sqrt{1-\alpha^2}\mathbf{Mn} \tag{21}$$

and the learning algorithm for μ involves

$$\begin{aligned}
\delta\mu_d &\propto -\left\langle \left.\frac{\partial E}{\partial \mu_d}\right|_0 \right\rangle + \left\langle \left.\frac{\partial E}{\partial \mu_d}\right|_1 \right\rangle && (22)\\
&= \frac{1}{\sigma_d^2}\left(\langle x_d\rangle - \langle x'_d\rangle\right) && (23)\\
&= \frac{1}{\sigma_d^2}\left(\mu_d^{\mathrm{ML}} - (\mu_d + \alpha(\mu_d^{\mathrm{ML}} - \mu_d)\right) && (24)\\
&= \frac{1-\alpha}{\sigma_d^2}\left(\mu_d^{\mathrm{ML}} - \mu_d\right) && (25)
\end{aligned}$$

so μ_d decays exponentially to μ_d^{ML} with rate proportional to $(1-\alpha)/\sigma_d^2$. The mean converges correctly whether or not the standard deviations σ_d have their correct values.

The learning rule for $\log \sigma_d$ involves

$$\begin{aligned}
\delta \log \sigma_d &\propto -\left\langle \left.\frac{\partial E}{\partial \log \sigma_d}\right|_0 \right\rangle + \left\langle \left.\frac{\partial E}{\partial \log \sigma_d}\right|_1 \right\rangle && (26)\\
&= \frac{1}{\sigma_d^2}\left\langle (\mu_d - x_d)^2 - (\mu_d - x'_d)^2 \right\rangle && (27)\\
&= \frac{1}{\sigma_d^2}\left\langle (\mu_d - x_d)^2 - (\alpha(x_d - \mu_d) + \sqrt{1-\alpha^2}[\mathbf{Mn}]_d)^2 \right\rangle && (28)\\
&= \frac{1}{\sigma_d^2}(1-\alpha^2)\left[\left\langle (\mu_d - x_d)^2 \right\rangle - \sigma_d^2\right] && (29)\\
&= \frac{(1-\alpha^2)}{\sigma_d^2}\left[(\mu_d - \mu_d^{\mathrm{ML}})^2 + \left(\sigma_d^{2\,\mathrm{ML}} - \sigma_d^2\right)\right] && (30)
\end{aligned}$$

So if μ_d has converged to μ_d^{ML} then σ_d^2 converges to $\sigma_d^{2\,\mathrm{ML}}$.

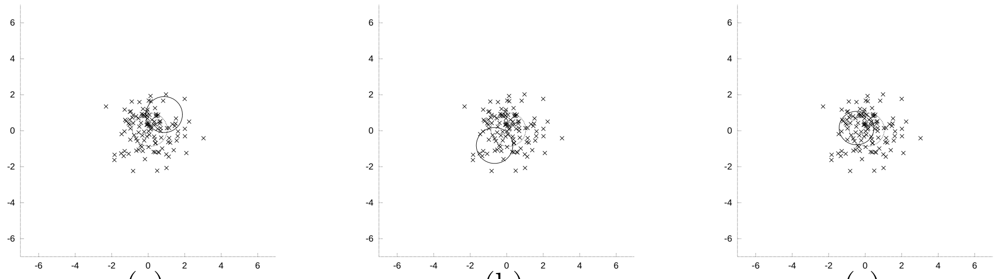
3.2 Noisy swirl I

Under the noisy swirl operator, the one-step algorithm for learning $\{\mu\}$ *alone* converges to the maximum-likelihood parameters. [Proof available.]

3.3 Noisy swirl II

Under the noisy swirl operator, the one-step algorithm for learning $\{\mu\}$ and $\{\sigma\}$ does *not* necessarily converge to the maximum-likelihood parameters. Typically the algorithm has a single fixed point but the values of $\{\sigma\}$ at that fixed point are wrong; there is *no* fixed point at the maximum likelihood parameters.

[Easy to confirm with a simple example: two data points at (1,1) and (-1,-1).]



(a) (b) (c)

Figure 1. Noisy Star Trek operator: Fixed points found with σ fixed to true values. The stable hypothesis μ is shown by the solid contour. The dotted line indicates the maximum likelihood hypothesis.

3.4 Noisy swirl III

In the special case of an infinite amount of data that come from an axis-aligned Gaussian, the noisy swirl operator's one-step algorithm for learning $\{\mu\}$ and $\{\sigma\}$ does converge to the maximum-likelihood parameters.

3.5 Noisy flip

Under the noisy flip operator, the one-step algorithm for learning $\{\mu\}$ and $\{\sigma\}$ does *not* necessarily converge to the maximum-likelihood parameters. Typically the algorithm has a single fixed point but the values of $\{\sigma\}$ at that fixed point are wrong; there is *no* fixed point at the maximum likelihood parameters. [Easy to confirm with a simple example: two data points at (1,1) and (-1,-1).] This result is interesting because the noisy flip satisfies detailed balance; this proves false the conjecture that detailed balance would be a sufficient condition for convergence to the ML parameters.

3.6 Noisy Star Trek

Under the noisy Star Trek operator, the one-step algorithm for learning $\{\mu\}$ and $\{\sigma\}$ typically does *not* converge to the maximum-likelihood parameters. The algorithm can have many fixed points.

We illustrate this empirical finding in two cases, first with the standard deviations $\{\sigma\}$ fixed, and second with $\{\mu\}$ and $\{\sigma\}$ varying.

First, with $N = 120$ datapoints from a spherical Gaussian, we learned the means $\{\mu\}$ with the standard deviations $\{\sigma\}$ fixed to their true values. The parameters of the Markov chain were $j = 0.001$, $\alpha = 0.995$. Many fixed points were found. If the initial condition had μ to one side of the data, a fixed point would be reached about one standard-deviation away on the same side as the initial condition (figure 1a,b). Initial conditions close to the maximum of the likelihood led to other fixed points (figure 1c).

Second, with the same data, we adapted both $\{\mu\}$ and $\{\sigma\}$. In this case, only one fixed point was found, having standard deviations much smaller than the maximum likelihood values (figure 2). There was no fixed point at the maximum likelihood values.

4 A one-dimensional example

All these examples of failures of the one-step algorithm work because the data distribution (a cloud of delta functions) is not precisely realisable by the model, and the sufficient statistics of the data

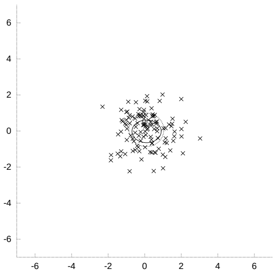


Figure 2. Noisy Star Trek operator: Fixed points found with σ free to vary. The stable hypothesis is shown by the solid contour. The dotted line indicates the maximum likelihood hypothesis.

are not invariant under the operator T.

The neatest example I have found is the following one-parameter model. A one-dimensional Gaussian distribution has mean μ. The standard deviation is fixed to $\sigma = 1$.

$$E(x; \mu) = \frac{1}{2}(x - \mu)^2 \tag{31}$$

The transition operator T works as follows: 10% of the time, x moves under the drift-and-diffuse operator, and 90% of the time, it moves under the *right-mixing* operator,

$$x \rightarrow x' = \begin{cases} x & \text{if } x \leq \mu \\ \mu + |\nu| & \text{if } x > \mu, \end{cases} \tag{32}$$

where ν is a standard normal random variate. This operator leaves points to the left of μ alone and mixes up points to the right of μ.

The operator T satisfies detailed balance.

There are two data points at ±1. The maximum likelihood mean is $\mu_{\text{ML}} = 0$. Let's assume we start from $\mu = 0$. Since the mean of the distribution of $|\nu|$ is $\sqrt{2/\pi} \simeq 0.8$, the expectation of the mean of the translated data is

$$0.9(-1 + 0.8) = 0.9 \times (-0.2). \tag{33}$$

So the one-step learning algorithm will on average move μ to the right. The maximum likelihood μ is not a fixed point. The algorithm has a fixed point at a point $\mu^* > 0$ whose precise location depends on the value of α in the drift-and-diffuse operation.

This example is not unrealistic. In a more complex problem using, say, a hybrid Monte Carlo algorithm to make moves in $\mathbf{x}$ space, it is quite possible that the characteristic step size in one part of the space will be large, and the characteristic step size in another part of the space will be much smaller.

5 Discrete state space, straightforward sampling

The preceding examples have used continuous state spaces and somewhat exotic transition operators. The next example, motivated by work on fair electoral systems (Sewell *et al.*, 2001), uses a discrete state space and the most straightforward transition operator imaginable. Let $\mathbf{x}$ be a ranking of three individuals, for example $\mathbf{x} = (A > B > C)$ or $\mathbf{x} = (C > A > B)$. Assume we wish to construct a probability distribution over such rankings of the form

$$P(\mathbf{x}|\{\alpha\}) = \frac{1}{Z} \exp\left(- \sum_{p=1}^{P} \alpha_p f_p(\mathbf{x}) \right) \tag{34}$$

where each function $f_p(\mathbf{x})$ is an indicator function for the truth of a statement of the form 'A is ranked somewhere above B'. To be precise, f_1, f_2, and f_3 are equal to one or zero if the following respective statements are true or false: 'A is ranked above B'; 'A is ranked above C'; and 'B is ranked above C'.

We are supplied with four data points as follows:

$$\begin{aligned}&\mathbf{x}^{(1)} = (A > B > C);\\ &\mathbf{x}^{(2)} = (A > B > C);\\ &\mathbf{x}^{(3)} = (A > B > C); \quad \mathbf{x}^{(4)} = (C > B > A).\end{aligned}$$

The maximum likelihood parameter values given these data are

$$\alpha_1 = 1.54, \;\; \alpha_2 = 0, \;\; \alpha_3 = 1.54.$$

We consider using one-step learning to find the parameters $\{\alpha\}$ with the operator T working as follows: first select at random between the upper two individuals in the current ranking $\mathbf{x}$ and lower two individuals; consider exchanging these two individuals. Accept the proposed interchange using the Metropolis method, *i.e.*, on the basis of the change in energy.

Now, with only one step made from the data shown above, the change in energy will never depend on the value of α_2, since a single step can never interchange A and C. Furthermore, the gradient of the energy with respect to α_2, which is given by the mean value of $f_2(\mathbf{x})$, will be the same before and after a single step, so the one-step learning rule for α_2 will certainly be

$$\Delta\alpha_2 = 0. \tag{35}$$

So α_2 will never change from its initial value.

Thus we have another example of a failure of the one-step learning algorithm.

Acknowledgements

I thank Geoff Hinton, Graeme Mitchison, Radford Neal, Seb Wills, Ed Ratzer, David Ward, Chris Williams, and Simon Osindero for invaluable discussions.

References

Sewell, R. F., MacKay, D. J. C., and McLean, I., (2001) A maximum entropy approach to fair elections. In preparation.

Williams, C. K. I., (2001) Personal communication.