1 Bayesian principles

1.1 Slide 3

The use of probability to represent uncertainty, however, is not an ad-hoc choice, but is inevitable if we are to respect common sense while making rational coherent inferences. For instance, [Cox, 1946] showed that if numerical values are used to represent degrees of belief, then a simple set of axioms encoding common sense properties of such beliefs leads uniquely to a set of rules for manipulating degrees of belief that are equivalent to the sum and product rules of probability. This provided the first rigorous proof that probability theory could be regarded as an extension of Boolean logic to situations involving uncertainty.

1.2 Slide 8

Stochastic techniques generally have the property that given infinite computational resource, they can generate exact results, and the approximation arises from the use of a finite amount of processor time. In practice, sampling methods can be computationally demanding, often limiting their use to small-scale problems. Also, it can be difficult to know whether a sampling scheme is generating independent samples from the required distribution.

Deterministic approximations can never generate exact results, and so their strengths and weaknesses are complementary to those of sampling methods.

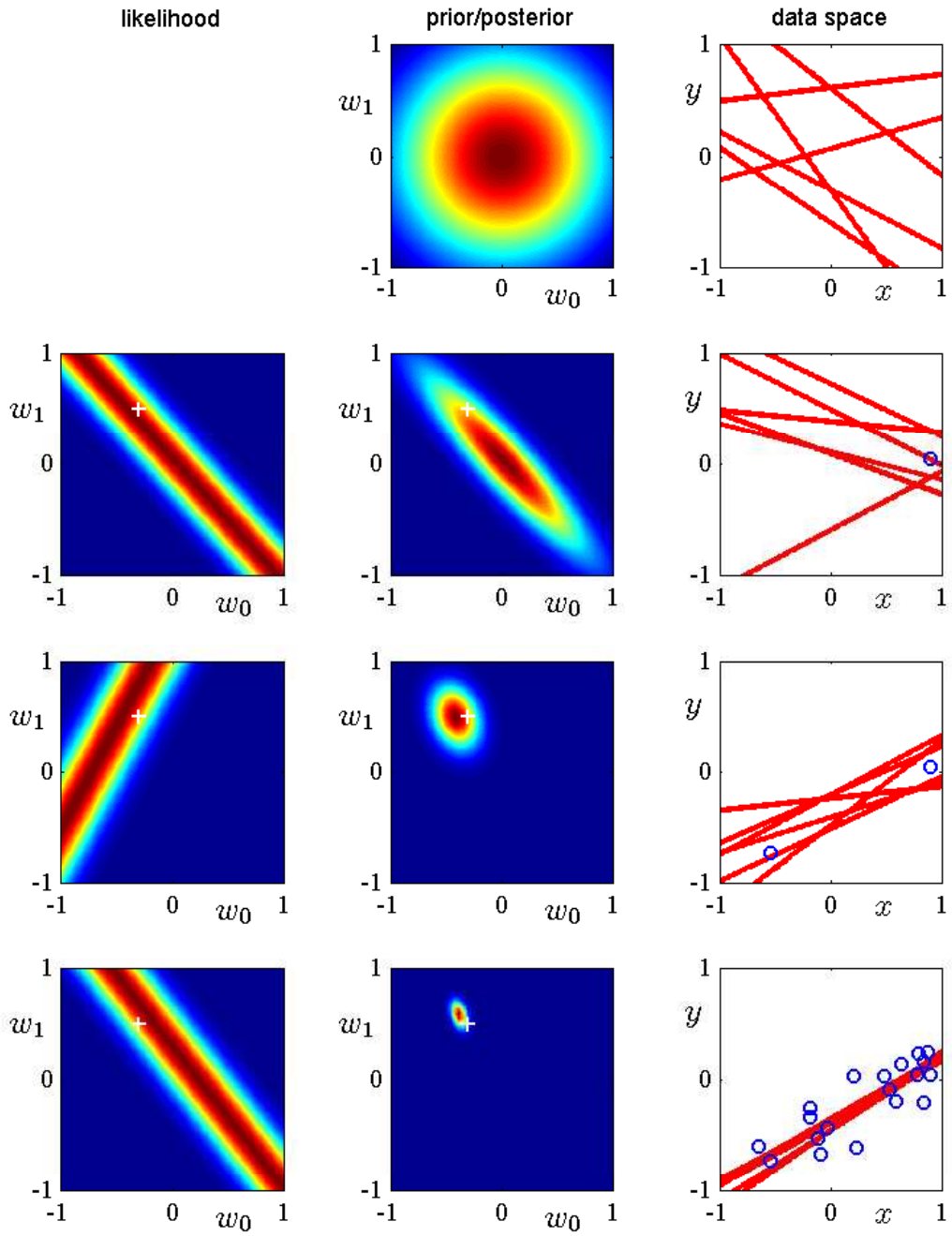
1.3 Slide 9

We shall treat the noise precision parameter β as a known constant. First note that the likelihood function $p(\mathbf{t}|\mathbf{w})$ is the exponential of a quadratic function of $\mathbf{w}$.

The log of the posterior distribution is given by the sum of the log likelihood and the log of the prior and, as a function of $\mathbf{w}$, takes the form

$$\ln p(\mathbf{w}|\mathbf{t}) = -\frac{\beta}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 - \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + \text{const.} \tag{1.1}$$

Maximization of this posterior distribution with respect to $\mathbf{w}$ is therefore equivalent to the minimization of the sum-of-squares error function with the addition of a quadratic regularization term, corresponding to the regularised linear regression with sum of squares with $\lambda = \alpha/\beta$.



The first row of this figure corresponds to the situation before any data points are observed and shows a plot of the prior distribution in $\mathbf{w}$ space together with six samples of the function $y(x, \mathbf{w})$ in which the values of $\mathbf{w}$ are drawn from the prior.

In the second row, we see the situation after observing a single data point. The location (x, t) of the data point is shown by a blue circle in the right-hand column. In the left-hand column is a plot of the likelihood function $p(t|x, \mathbf{w})$ for this data point as a function of $\mathbf{w}$. Note that the

likelihood function provides a soft constraint that the line must pass close to the data point, where close is determined by the noise precision β. For comparison, the true parameter values $a_0 = -0.3$ and $a_1 = 0.5$ used to generate the data set are shown by a white cross in the plots in the left column of the Figure. When we multiply this likelihood function by the prior from the top row, and normalize, we obtain the posterior distribution shown in the middle plot on the second row. Samples of the regression function $y(x, \mathbf{w})$ obtained by drawing samples of $\mathbf{w}$ from this posterior distribution are shown in the right-hand plot. Note that these sample lines all pass close to the data point.

The third row of this figure shows the effect of observing a second data point, again shown by a blue circle in the plot in the right-hand column. The corresponding likelihood function for this second data point alone is shown in the left plot. When we multiply this likelihood function by the posterior distribution from the second row, we obtain the posterior distribution shown in the middle plot of the third row. Note that this is exactly the same posterior distribution as would be obtained by combining the original prior with the likelihood function for the two data points. This posterior has now been influenced by two data points, and because two points are sufficient to define a line this already gives a relatively compact posterior distribution. Samples from this posterior distribution give rise to the functions shown in red in the third column, and we see that these functions pass close to both of the data points.

The fourth row shows the effect of observing a total of 20 data points. The left-hand plot shows the likelihood function for the 20^{th} data point alone, and the middle plot shows the resulting posterior distribution that has now absorbed information from all 20 observations. Note how the posterior is much sharper than in the third row. In the limit of an infinite number of data points, the posterior distribution would become a delta function centred on the true parameter values, shown by the white cross.

1.4 Slide 11

This shows that taking a (weighted) model average is the principled Bayesian approach. If we have two models that are a-posteriori equally likely and one predicts a narrow distribution around $t = a$ while the other predicts a narrow distribution around $t = b$, the overall predictive distribution will be a bimodal distribution with modes at $t = a$ and $t = b$, not a single model at $t = (a + b)/2$.

A simple approximation to model averaging is to use the single most probable model alone to make predictions. This is known as *model selection*.

1.5 Slide 13

Note that the distributions are normalized. In this example, for the particular observed data set $\mathcal{D}_0$, the model $\mathcal{M}_2$ with intermediate complexity has the largest evidence.

A simple model (for example, based on a first order polynomial) has little variability and so will generate data sets that are fairly similar to each other. Its distribution $p(\mathcal{D})$ is therefore confined to a relatively small region of the horizontal axis. By contrast, a complex model (such as a ninth order polynomial) can generate a great variety of different data sets, and so its distribution $p(\mathcal{D})$ is spread over a large region of the space of data sets. Because the distributions $p(\mathcal{D}|\mathcal{M}_i)$ are normalized,

we see that the particular data set $\mathcal{D}_0$ can have the highest value of the evidence for the model of intermediate complexity. Essentially, the simpler model cannot fit the data well, whereas the more complex model spreads its predictive probability over too broad a range of data sets and so assigns relatively small probability to any one of them.

2 Evidence approximation

2.1 Slide 2

This framework is known in the statistics literature as *empirical Bayes* [Gelman *et al.*, 2013], or *type 2 maximum likelihood* [Berger, 1985], or *generalized maximum likelihood* [Wahba, 1975], and in the machine learning literature is also called the *evidence approximation* [MacKay, 1992].

2.2 Slide 3

Here we have omitted the dependence on the input variable $\mathbf{x}$ to keep the notation uncluttered.

We shall proceed by evaluating the marginal likelihood for the linear basis function model and then finding its maxima. This will allow us to determine values for these hyperparameters from the training data alone, without recourse to cross-validation. Recall that the ratio α/β is analogous to a regularization parameter.

2.3 Slide 6

Recall that

$$\mathbf{m}_N = \beta \mathbf{S}_N \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{t} \tag{2.1}$$

$$\mathbf{S}_N^{-1} = \alpha \mathbf{I} + \beta \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi}. \tag{2.2}$$

2.4 Slide 9

We can illustrate the evidence framework for setting hyperparameters using the sinusoidal synthetic data set together with the Gaussian basis function model comprising 9 basis functions, so that the total number of parameters in the model is given by $M = 10$ including the bias. Here, for simplicity of illustration, we have set β to its true value of 11.1 and then used the evidence framework to determine α.

The left plot shows γ (red curve) and $2\alpha E_W(\mathbf{m}_N)$ (blue curve) versus $\ln \alpha$ for the sinusoidal synthetic data set. It is the intersection of these two curves that defines the optimum value for α given by the evidence procedure. The right plot shows the corresponding graph of log evidence $\ln p(\mathbf{t}|\alpha, \beta)$ versus $\ln \alpha$ (red curve) showing that the peak coincides with the crossing point of the curves in the left plot. Also shown is the test set error (blue curve) showing that the evidence maximum occurs close to the point of best generalization.

2.5 Slide 10

The eigenvalues λ_i measure the curvature of the likelihood function, and so in the Figure the eigenvalue λ_1 is small compared with λ_2 (because a smaller curvature corresponds to a greater elongation of the contours of the likelihood function.

2.6 Slide 13

The variational approach to Bayesian PCA can be found in [Bishop, 1999]

2.7 Slide 16

The effective dimensionality of the principal subspace is then determined by the number of finite α_i values, and the corresponding vectors $\mathbf{w}_i$ can be thought of as 'relevant' for modelling the data distribution. In this way, the Bayesian approach is automatically making the trade-off between improving the fit to the data, by using a larger number of vectors $\mathbf{w}_i$ with their corresponding eigenvalues λ_i each tuned to the data, and reducing the complexity of the model by suppressing some of the $\mathbf{w}_i$ vectors.

3 Variational methods

3.1 Slide 4

One way to restrict the family of approximating distributions is to use a parametric distribution $q(\mathbf{Z}|\boldsymbol{\omega})$ governed by a set of parameters $\boldsymbol{\omega}$. The lower bound $\mathcal{L}(q)$ then becomes a function of $\boldsymbol{\omega}$, and we can exploit standard nonlinear optimization techniques to determine the optimal values for the parameters. An example of this approach, in which the variational distribution is a Gaussian and we have optimized with respect to its mean and variance, is shown in the Figure.

These techniques are applied to the distribution $p(z) \propto \exp(-z^2/2)\sigma(20z + 4)$ where $\sigma(z)$ is the logistic sigmoid function defined by $\sigma(z) = (1 + e^{-z})^{-1}$. Note that the Laplace approximation fits the mode of the distribution, which is the maximum value.

3.2 Slide 6

Left-hand plot. The green contours corresponding to 1, 2, and 3 standard deviations for a correlated Gaussian distribution $p(\mathbf{z})$ over two variables z_1 and z_2, and the red contours represent the corresponding levels for an approximating distribution $q(\mathbf{z})$ over the same variables given by the product of two independent univariate Gaussian distributions whose parameters are obtained by minimization of (a) the Kullback-Leibler divergence $\mathrm{KL}(q\|p)$.

Right-hand plot. The blue contours show a bimodal distribution $p(\mathbf{Z})$ given by a mixture of two Gaussians, and the red contours correspond to the single Gaussian distribution $q(\mathbf{Z})$ that best approximates $p(\mathbf{Z})$ in the sense of minimizing the Kullback-Leibler divergence $\mathrm{KL}(q\|p)$.

3.3 Slide 7

We now illustrate the factorized variational approximation using a Gaussian distribution over a single variable x. Our goal is to infer the posterior distribution for the mean μ and precision (inverse variance) τ, given a data set $\mathcal{D} = \{x_1, \ldots, x_N\}$ of observed values of x which are assumed to be drawn independently from the Gaussian.

We use conjugate prior distributions for μ and τ given by

$$p(\mu|\tau) = \mathcal{N}\left(\mu|\mu_0, (\lambda_0\tau)^{-1}\right) \tag{3.1}$$

$$p(\tau) = \mathrm{Gam}(\tau|a_0, b_0) \tag{3.2}$$

where $\mathrm{Gam}(\tau|a_0, b_0)$ is the gamma distribution. Together these distributions constitute a Gaussian-Gamma conjugate prior distribution.

Contours of the true posterior distribution $p(\mu, \tau|D)$ are shown in green. (a) Contours of the initial factorized approximation $q_\mu(\mu)q_\tau(\tau)$ are shown in blue. (b) After re-estimating the factor $q_\mu(\mu)$. (c) After re-estimating the factor $q_\tau(\tau)$. (d) Contours of the optimal factorized approximation, to which the iterative scheme converges, are shown in red.

3.4 Slide 8

A conjugate prior is one that is conjugate to the likelihood function so that the posterior distribution of the parameter has the same functional form as the prior. For example, the conjugate prior to the Bernoulli distribution is the beta distribution. Every distribution in the exponential family has a conjugate prior.

3.5 Slide 10

The update equations for the other parameters are more complicated.

$$\mathbb{E}_{\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k}\left[(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Lambda}_k(\mathbf{x}_n - \boldsymbol{\mu}_k)\right] = D\beta_k^{-1} + \nu_k(\mathbf{x}_n - \mathbf{m}_k)^{\mathrm{T}}\mathbf{W}_k(\mathbf{x}_n - \mathbf{m}_k) \tag{3.3}$$

$$\ln \widetilde{\Lambda}_k \equiv \mathbb{E}\left[\ln|\boldsymbol{\Lambda}_k|\right] = \sum_{i=1}^{D}\psi\left(\frac{\nu_k + 1 - i}{2}\right) + D\ln 2 + \ln|\mathbf{W}_k| \tag{3.4}$$

$$\ln \widetilde{\pi}_k \equiv \mathbb{E}\left[\ln \pi_k\right] = \psi(\alpha_k) - \psi(\widehat{\alpha}) \tag{3.5}$$

3.6 Slide 11

The Figure shows the results of applying this approach to the rescaled Old Faithful data set for a Gaussian mixture model having $K = 6$ components.

The ellipses denote the one standard-deviation density contours for each of the components, and the density of red ink inside each ellipse corresponds to the mean value of the mixing coefficient for each component. The number in the top left of each diagram shows the number of iterations of variational inference. Components whose expected mixing coefficient are numerically indistinguishable from zero are not plotted.

3.7 Slide 14

Bishop uses the word 'straightforwardly' here, but he must have a different definition of that term to me!

In a similar way, we replace the true posterior distribution $p(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Lambda}|\mathbf{X})$ with its variational approximation $q(\boldsymbol{\pi})q(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ in order to compute the predictive density $p(\widehat{\mathbf{x}}|\mathbf{X})$.

3.8 Slide 17

The method described was first defined in [Corduneanu and Bishop, 2001].

4 Sampling

4.1 Slide 3

See [Press *et al.*, 1992] for examples of uniform pseudo-random number generators. It is harder than you might think to do this in a robust way.

4.2 Discrete Markov Chains

- Consider a set of random variables $X(t)$ for $t \in \mathbb{R}$. Usually we will take $t \in \mathbb{N}$. This process is Markov if
$$P(X(t_n) \leq x|X(t_1), \ldots, X(t_{n-1})) = P(X(t_n) \leq x|X(t_{n-1})),$$
for all x and $t_1 < t_2 < \cdots < t_n$.
- If X is discrete, a state i is *persistent* if return is certain: $P(X_n = i \text{ for some } n \geq 1|X_1 = i) = 1$.
- A persistent state is *non-null* if the expected return time is finite.
- A state is said to be *ergodic* if it is persistent, non-null and aperiodic (which means that the h.c.f of the return times is 1).

- An irreducible chain has a *stationary* distribution $\pi = \pi P$ (where P is the transition matrix for the chain) iff all states are non-null persistent: in that case $\pi_i = \mu_i^{-1}$ where μ_i is the expected time to return to state i. (From which we can see that the stationary distribution is unique.)

4.2.1 Stationary Distribution

- A Markov chain will always *converge* to its stationary distribution, though the length of time it takes to converge depends on the starting point; just as with optimisation, it pays to take some trouble over initialisation.

- When using MCMC it is normal to throw away the first part of the chain (known as the *burn-in period* so that samples are taken only from the stationary distribution.

- For most Markov chains there are no convergence *tests* that are both rigorous and practical, and we usually have to rely on more empirical methods to determine if the stationary distribution has been reached.

4.2.2 Reversibility

- Let $\{X(n) : -\infty < n < \infty\}$ be an ergodic Markov chain with transition kernel $\mathbf{P}$ and stationary distribution $\boldsymbol{\pi}$ and each $X(n)$ has distribution $\boldsymbol{\pi}$ for all $n \in (-\infty, \infty)$.

- Let $Y(n)$ denote the reversed chain $Y(n) = X(-n)$: we say that X is *time-reversible* if the transition matrices of X and Y are the same.

- It is easy to show that X is time-reversible if and only if

$$\pi_i p_{ij} = \pi_j p_{ji} \qquad \text{for all } i,\, j. \tag{4.1}$$

- In this case, rather than solve the eigenproblem for the transition matrix $\mathbf{P}$, it is sufficient to find a vector $\boldsymbol{\pi}$ that satisfies (4.1) and the usual conditions for a probability vector:

$$0 \leq \pi_i \leq 1 \qquad \text{and} \qquad \sum_i \pi_i = 1. \tag{4.2}$$

4.3 Gibbs Sampling

- Applicable when we want to sample from a multi-dimensional parameter vector $\theta = (\theta_1, \ldots, \theta_n)$.

- We assume that, although sampling from $Q(\theta)$ directly is impossible, it is possible to generate samples from the *conditional* distribution (under Q) of one component of θ given values for all the other components.

- This involves sampling from a *one-dimensional* distribution, and may in its turn require *importance* or *rejection* sampling.

4.3.1 Gibbs Algorithm

- We generate a Markov chain as follows. Given $\theta^{(t)}$, we generate $\theta^{(t+1)}$ with the following n steps:
  - Sample $\theta_1^{(t+1)}$ from the distribution of θ_1 given $\theta_2^{(t)}, \theta_3^{(t)}, \ldots, \theta_n^{(t)}$.
  - Sample $\theta_2^{(t+1)}$ from the distribution of θ_2 given $\theta_1^{(t+1)}, \theta_3^{(t)}, \ldots, \theta_n^{(t)}$.

    $\ldots$
  - Sample $\theta_j^{(t+1)}$ from the distribution of θ_j given $\theta_1^{(t+1)}, \ldots, \theta_{j-1}^{(t+1)}, \theta_{j+1}^{(t)}, \ldots, \theta_n^{(t)}$.

    $\ldots$
  - Sample $\theta_n^{(t+1)}$ from the distribution of θ_n given $\theta_1^{(t+1)}, \ldots, \theta_{n-1}^{(t+1)}$.
- Note that the new value for θ_j is used immediately when considering the conditional distribution for θ_{j+1}.

4.3.2 Algorithm Properties

- This transition distribution leaves the desired distribution $Q(\theta)$ invariant if all the steps making up each transition also leave Q invariant.
  1. Since step j leaves θ_k unchanged for $k \neq j$, the desired marginal distribution for these components is certainly invariant.
  2. Furthermore, the conditional distribution for θ_j in the new state given the other components is defined to be the right function.
- Together, these two properties ensure that if we started with a sample from the desired distribution, then the *joint distribution* of all the θ_j after all n of the above steps must also be the desired distribution.
- These transitions do not necessarily ensure that the Markov chain is ergodic: this must be established in each application.

4.4 Metropolis-Hastings algorithm

4.4.1 Transition Density

- To show that these transitions leave Q invariant, we first need to work out the transition density function. This density function is *singular*, since there is a non-zero point probability mass that the new state will be the same as the old state.
- Luckily the detailed balance condition need only be verified for transitions that change the state. For $\theta' \neq \theta$, the transition density for the Metropolis–Hastings algorithm is given by

$$T(\theta'|\theta) = S(\theta'|\theta) \min\left(1, \frac{Q(\theta')}{Q(\theta)}\right).$$

4.4.2 Proof of Detailed Balance

$$\begin{aligned} T(\theta'|\theta)Q(\theta) &= S(\theta'|\theta)\min\left(1, \frac{Q(\theta')}{Q(\theta)}\right)Q(\theta) \\ &= S(\theta'|\theta)\min(Q(\theta), Q(\theta')) \\ &= S(\theta|\theta')\min(Q(\theta'), Q(\theta)) \quad \text{by symmetry of } S \\ &= S(\theta|\theta')\min\left(1, \frac{Q(\theta)}{Q(\theta')}\right)Q(\theta') \\ &= T(\theta|\theta')Q(\theta'). \end{aligned}$$

Although Q is guaranteed to be invariant, we must check if the chain is ergodic. This depends on the details of Q and on the proposal distribution S.

4.5 Hybrid Monte Carlo

4.6 Properties of Hamiltonian Dynamics

1. H is constant as q and p vary:
$$\frac{dH}{d\tau} = \sum_i \left[\frac{\partial H}{\partial q_i}\frac{dq_i}{d\tau} + \frac{\partial H}{\partial p_i}\frac{dp_i}{d\tau}\right] = \sum_i \left[\frac{\partial H}{\partial q_i}\frac{\partial H}{\partial p_i} - \frac{\partial H}{\partial p_i}\frac{\partial H}{\partial q_i}\right] = 0.$$
2. If we follow how the points in some region, of volume V, move according to the dynamical equations, we find that the region where these points end up also has volume V. Compute the divergence of the motion in phase space:
$$\sum_i \left[\frac{\partial}{\partial q_i}\left(\frac{dq_i}{d\tau}\right) + \frac{\partial}{\partial p_i}\left(\frac{dp_i}{d\tau}\right)\right] = \sum_i \left[\frac{\partial H}{\partial q_i \partial p_i} - \frac{\partial H}{\partial p_i \partial q_i}\right] = 0.$$
3. The dynamics are *reversible*. After following the dynamics forwards for t time units, we can recover the original state by following the dynamics backward in time for t.

- These properties imply that the canonical distribution for q and p is invariant with respect to (deterministic) transitions that consist of following a trajectory for a certain period of time using Hamiltonian dynamics.
- The probability that we will end in some region after the transition is the same as the probability that we started in the corresponding region (of equal volume) found by reversing the dynamics.
- If this probability is given by the canonical distribution, then the probability of being in the final region will also be given by the canonical distribution, since the probabilities depend only on H, which is the same at the start and end of the trajectory.

4.6.1 Discrete Simulation

- In practice, the dynamics cannot be simulated exactly, but are approximated using time steps of finite size.

- In the *leapfrog* method, a single iteration updates approximations $\hat{q}$ and $\hat{p}$ to the true position and momentum as follows:

$$\hat{p}_i(\tau + \frac{\epsilon}{2}) = \hat{p}_i(\tau) - \frac{\epsilon}{2}\frac{\partial E}{\partial q_i}(\hat{q}(\tau)) \tag{4.3}$$

$$\hat{q}_i(\tau + \epsilon) = \hat{q}_i(\tau) + \epsilon\frac{\hat{p}_i(\tau + \frac{\epsilon}{2})}{m_i} \tag{4.4}$$

$$\hat{p}_i(\tau + \epsilon) = \hat{p}_i(\tau + \frac{\epsilon}{2}) - \frac{\epsilon}{2}\frac{\partial E}{\partial q_i}(\hat{q}(\tau + \epsilon)) \tag{4.5}$$

  This iteration consists of a half step for the $\hat{p}_i$, a full step for the $\hat{q}_i$ and another half step for the $\hat{p}_i$.

- We choose a value of ϵ that is sufficiently small and then we apply the updates defined by (4.3)–(4.5) for $L = \Delta\tau/\epsilon$ iterations.

4.6.2 HMC Implementation

Once the step size ϵ and number of iterations L are fixed, a dynamical transition consists of the following steps:

1. Randomly choose a direction $\lambda \in \{-1, +1\}$ for the trajectory, where $+1$ represents a forward trajectory. Both directions are equally likely.

2. Starting from the current state $(q, p) = (\hat{q}(0), \hat{p}(0))$, carry out L leapfrog iterations with a step size of $\lambda\epsilon$, resulting in the candidate state $(\hat{q}(\lambda\epsilon L), \hat{p}(\lambda\epsilon L)) = (q^*, p^*)$.

3. Accept the candidate state with probability

$$\min(1, \exp(-(H(q^*, p^*) - H(q, p)))).$$

   If the candidate state is rejected, then the new state will be the old state (q, p).

4.6.3 HMC Algorithm Properties

- Preservation of phase space volume by the dynamics and the random choice of λ ensure that the proposal distribution is symmetric, as required for the Metropolis algorithm.

- It is only when L is reasonably large that we obtain the principal benefit of HMC, which is the avoidance of random walk behaviour.

- In practice, the value of H oscillates along the trajectory, and the acceptance rate is almost independent of L. For step sizes greater than a certain value, the leapfrog discretisation becomes unstable and the acceptance rate drops.

- The optimal strategy is to select a step size just below the point of instability. Trajectories should be long enough so that they typically lead to states distant from their starting point, but no longer. Shorter trajectories lead to random walk behaviour, while longer trajectories would wastefully curve back on themselves.

References

[Berger, 1985] Berger, J. O. 1985. *Statistical Decision Theory and Bayesian Analysis* (second edn.). New York: Springer-Verlag.

[Bishop, 1999] Bishop, C. M. 1999. Bayesian pca. In *Advances in neural information processing systems*, pp. 382–388. MIT; 1998.

[Corduneanu and Bishop, 2001] Corduneanu, A. and C. M. Bishop 2001. Variational bayesian model selection for mixture distributions. In *Artificial intelligence and Statistics*, Volume 2001, pp. 27–34. Morgan Kaufmann Waltham, MA.

[Cox, 1946] Cox, R. T. 1946. Probability, frequency and reasonable expectation. *American Journal of Physics* **14** (1), 1–13.

[Gelman *et al.*, 2013] Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin 2013. *Bayesian data analysis*. CRC press.

[MacKay, 1992] MacKay, D. J. C. 1992. Bayesian interpolation. *Neural Computation* **4** (3), 415–447.

[Press *et al.*, 1992] Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery 1992. *Numerical Recipes in C: The Art of Scientific Computing* (second edn.). Cambridge University Press.

[Wahba, 1975] Wahba, G. 1975. Smoothing noisy data with spline functions. *Numerische mathematik* **24** (5), 383–393.