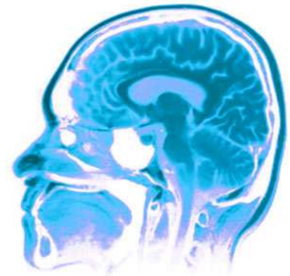
CPSC540

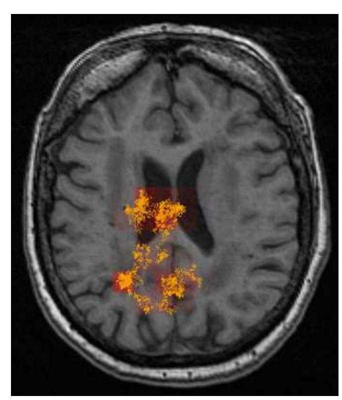


Monte Carlo

Nando de Freitas
October, 2011
University of British Columbia

Understanding the Brain

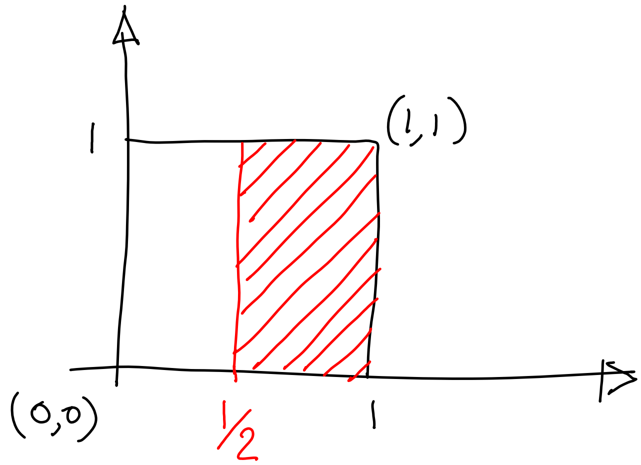
"You'd think this is crazy because engineers are always fighting to reduce the noise in their circuits, and yet here's the best computing machine in the universe—and it looks utterly random," Alex Pouget, associate professor of brain and cognitive sciences at the University of Rochester.



"We've known for several years that at the behavioral level, we're 'Bayes optimal,' meaning we are excellent at taking various bits of probability information, weighing their relative worth, and coming to a good conclusion quickly," *…"But we've always been at a loss to explain how our brains are able to conduct such complex Bayesian computations so easily."*

The idea

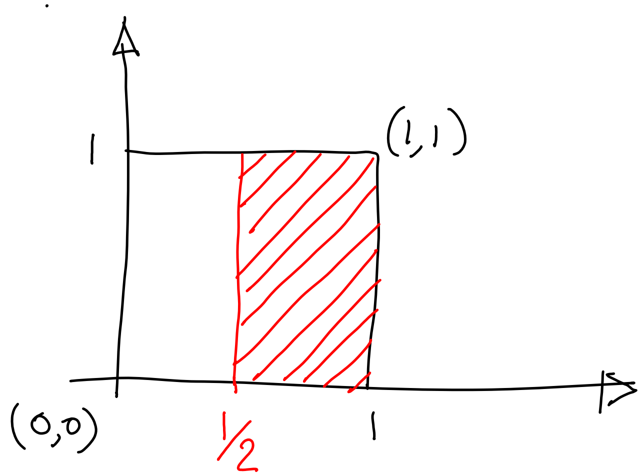
What is the probability that a dart thrown uniformly at random will hit the red area?



$$P(\text{area}) =$$

The idea

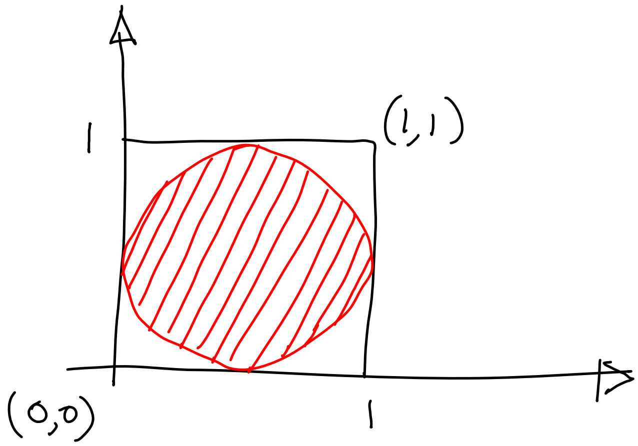
What is the probability that a dart thrown uniformly at random will hit the red area?



$$P(\text{area}) = 1/2$$

The idea

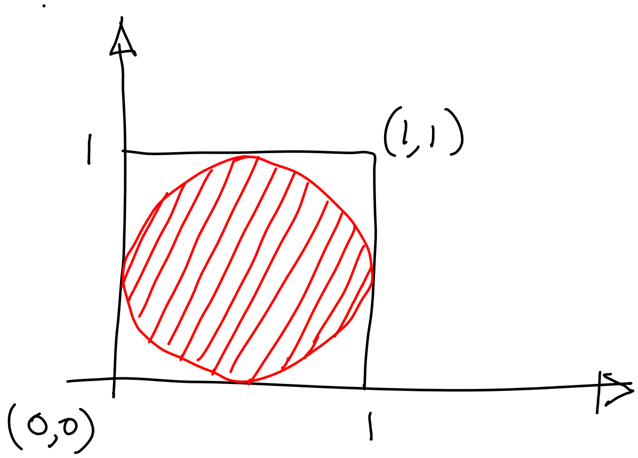
What is the probability that a dart thrown uniformly at random will hit the red area?



$$P(\text{area}) =$$

The idea

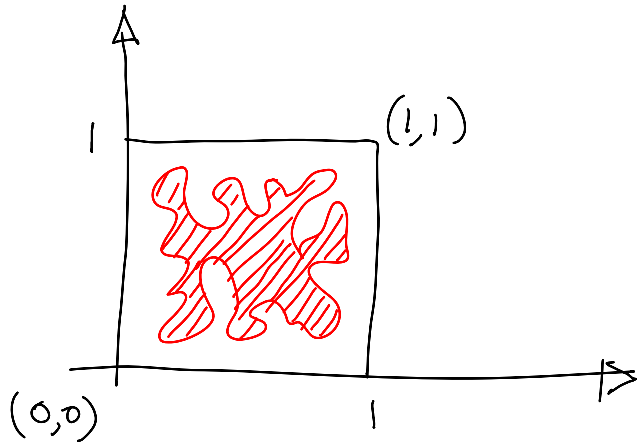
What is the probability that a dart thrown uniformly at random will hit the red area?



$$P(\text{area}) = \pi\left(\tfrac{1}{2}\right)^2 = \frac{\pi}{4}$$

The idea

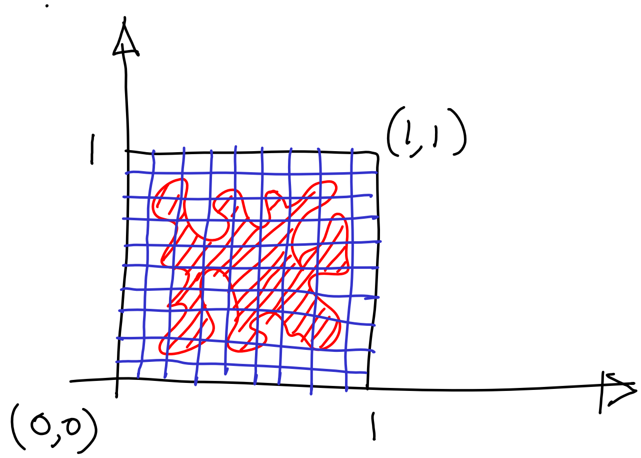
What is the probability that a dart thrown uniformly at random will hit the red area?



$$P(\text{area}) =$$

The idea

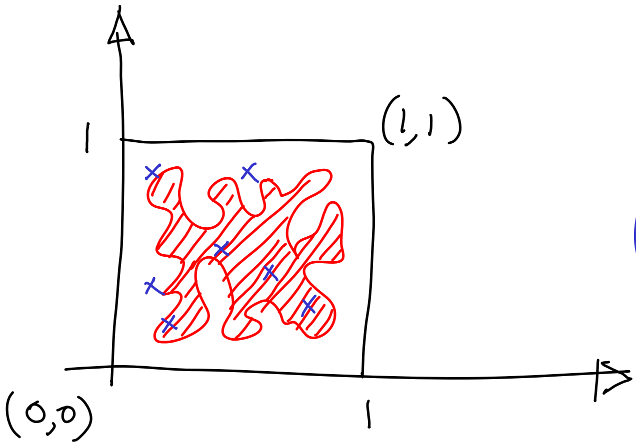
What is the probability that a dart thrown uniformly at random will hit the red area?



$$P(\text{area}) = \frac{\#\text{ red boxes}}{\#\text{ boxes}}$$

The idea

What is the probability that a dart thrown uniformly at random will hit the red area?

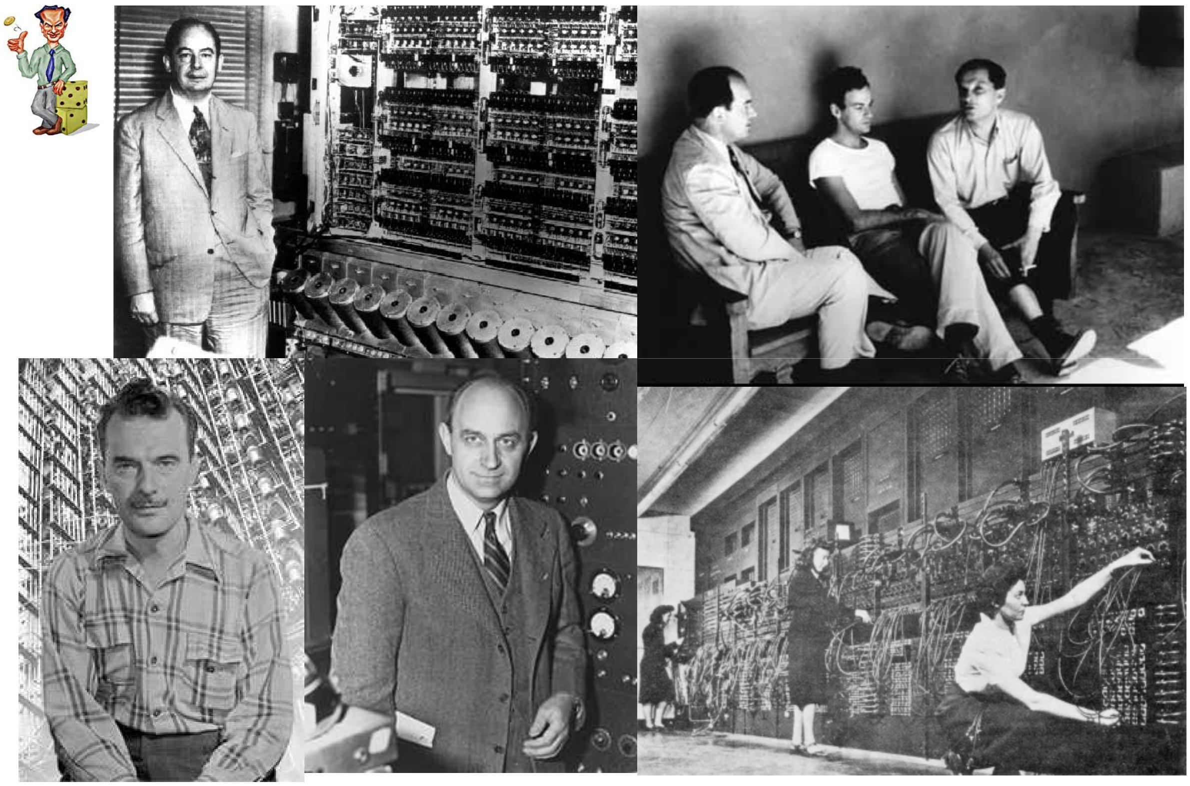


$4/7$

$$P(\text{area}) = \frac{\#\text{ darts in}}{\#\text{ darts in}}$$

History of the Monte Carlo method: The bomb and ENIAC

History of the Monte Carlo method



Integrals in Probabilistic Inference

1. *Normalisation*:

$$p(x|y) = \frac{p(y|x)p(x)}{\int_X p(y|x^\star)p(x^\star)dx^\star}$$

2. *Marginalisation*:

$$p(x|y) = \int_Z p(x, z|y)dz$$

3. *Expectation*:

$$\mathbb{E}_{p(x|y)}(f(x)) = \int_X f(x)p(x|y)dx$$

Monte Carlo Integration

Suppose we want to compute

$$I = \int f(x)\, P(x \mid data)\, dx$$

Monte Carlo Integration

Suppose we want to compute

$$I = \int f(x)\, P(x \mid data)\, dx$$

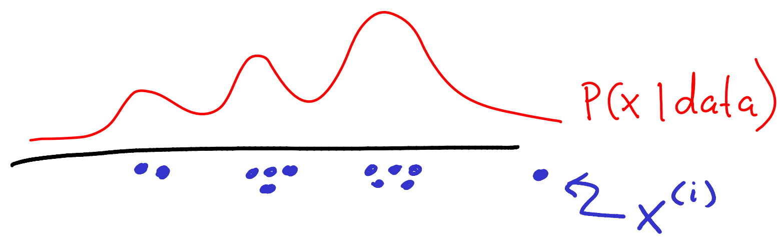
(i) Simulate $x^{(i)}\Big|_{i=1}^{N}$ from $P(x \mid data)$

Monte Carlo Integration

Suppose we want to compute

$$I = \int f(x)\, P(x \mid data)\, dx$$

(i) Simulate $x^{(i)}\Big|_{i=1}^{N}$ from $P(x \mid data)$

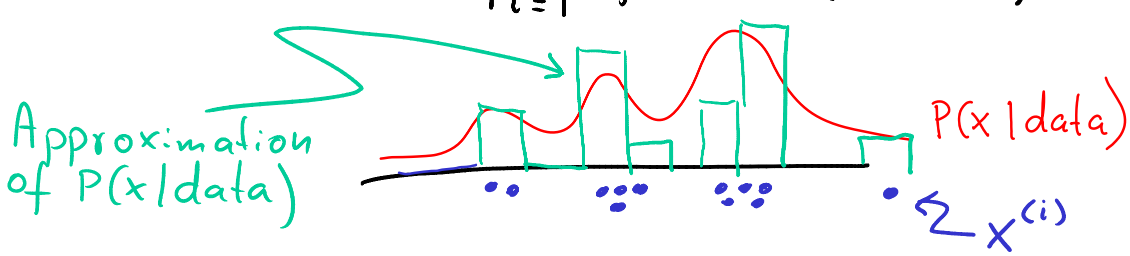


Monte Carlo Integration

Suppose we want to compute

$$I = \int f(x)\, P(x \mid data)\, dx$$

(i) Simulate $x^{(i)}\Big|_{i=1}^{N}$ from $P(x \mid data)$

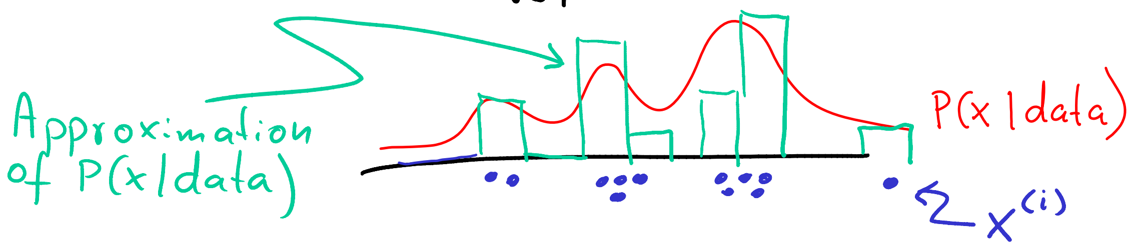


Monte Carlo Integration

Suppose we want to compute

$$I = \int f(x)\, P(x|\text{data})\, dx$$

(i) Simulate $x^{(i)}\Big|_{i=1}^{N}$ from $P(x|\text{data})$



(ii) Replace nasty integral with simple sum: $I \approx \frac{1}{N} \sum_{i=1}^{N} f(x^{(i)})$ $\blacksquare$

Density, measure and distribution

Lebesgue integral

Approximating distributions

The idea of Monte Carlo simulation is to draw an i.i.d. set of samples $\{x^{(i)}\}_{i=1}^{N}$ from a target density $p(x)$ defined on a high-dimensional space $\mathcal{X}$. These N samples can be used to approximate the target distribution with the following empirical point-mass function (think of it as a histogram):

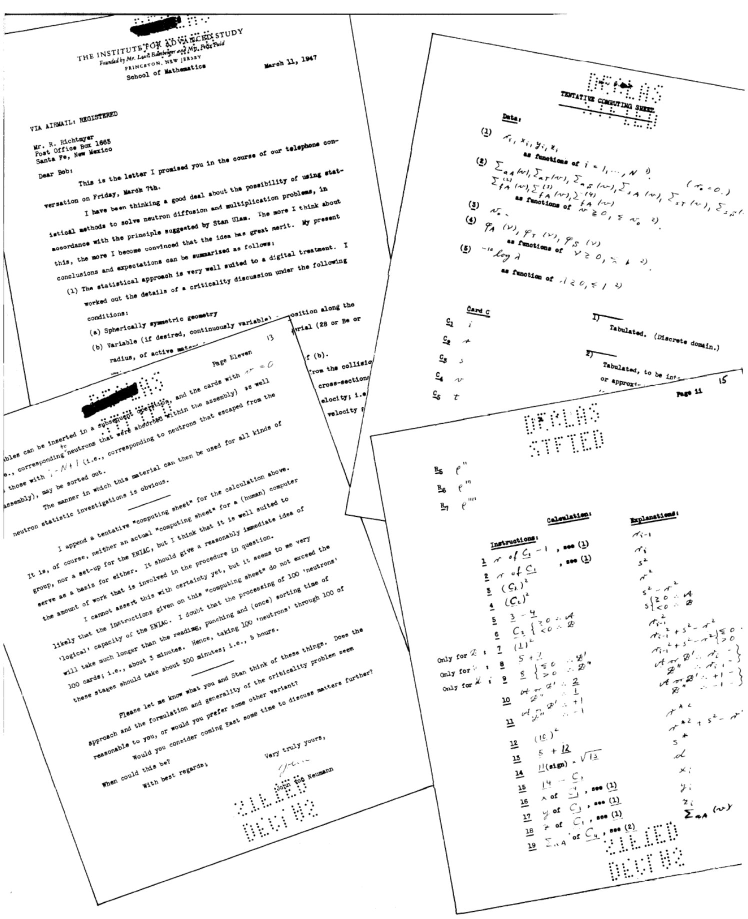
$$p_N(dx) = \frac{1}{N}\sum_{i=1}^{N} \delta_{x^{(i)}}(dx),$$

where $\delta_{x^{(i)}}(dx)$ denotes the delta-Dirac mass located at $x^{(i)}$.

Asymptotic behavior of Monte Carlo

$$I_N(f) = \frac{1}{N} \sum_{i=1}^{N} f\left(x^{(i)}\right) \xrightarrow[N\to\infty]{a.s.} I(f) = \int_{\mathcal{X}} f(x)p(x)\,dx$$

$$\sqrt{N}\left(I_N(f) - I(f)\right) \underset{N\to\infty}{\Longrightarrow} \mathcal{N}\left(0, \sigma_f^2\right)$$



[redacted]

THE INSTITUTE FOR ADVANCED STUDY
Founded by Mr. Louis Bamberger and Mrs. Felix Fuld
PRINCETON, NEW JERSEY
School of Mathematics

March 11, 1947

VIA AIRMAIL: REGISTERED

Mr. R. Richtmyer
Post Office Box 1663
Santa Fe, New Mexico

Dear Bob:

This is the letter I promised you in the course of our telephone conversation on Friday, March 7th.

I have been thinking a good deal about the possibility of using statistical methods to solve neutron diffusion and multiplication problems, in accordance with the principle suggested by Stan Ulam. The more I think about this, the more I become convinced that the idea has great merit. My present conclusions and expectations can be summarized as follows:

(1) The statistical approach is very well suited to a digital treatment. I worked out the details of a criticality discussion under the following conditions:

(a) Spherically symmetric geometry

(b) Variable (if desired, continuously variable) ... position along the radius, of active mat... ...rial (28 or Be or

...f (b).

...rom the collisio...

cross-section...

...elocity; i.e...

velocity ...

13

Page Eleven

[redacted]

...ables can be inserted in a subsequent operation, and the cards with $x = 0$ ...e., corresponding to neutrons that were absorbed within the assembly) as well ... those with $i - N + 1$ (i.e., corresponding to neutrons that escaped from the ...assembly), may be sorted out.

The manner in which this material can then be used for all kinds of neutron statistic investigations is obvious.

I append a tentative "computing sheet" for the calculation above. It is, of course, neither an actual "computing sheet" for a (human) computer group, nor a set-up for the ENIAC, but I think that it is well suited to serve as a basis for either. It should give a reasonably immediate idea of the amount of work that is involved in the procedure in question.

I cannot assert this with certainty yet, but it seems to me very likely that the instructions given on this "computing sheet" do not exceed the 'logical' capacity of the ENIAC. I doubt that the processing of 100 'neutrons' will take much longer than the reading, punching and (once) sorting time of 100 cards; i.e., about 3 minutes. Hence, taking 100 'neutrons' through 100 of these stages should take about 300 minutes; i.e., 5 hours.

Please let me know what you and Stan think of these things. Does the approach and the formulation and generality of the criticality problem seem reasonable to you, or would you prefer some other variant?

Would you consider coming East some time to discuss matters further? When could this be?

With best regards,

Very truly yours,

John von Neumann

TENTATIVE COMPUTING SHEET

Page 12

Calculation: Explanations:

Instructions:

Importance Sampling

$$I(f) = \int_{\mathcal{X}} f(x)p(x)\,dx$$

$$I\left(f\right) = \int f\left(x\right) w\left(x\right) q\left(x\right) dx$$

where $w\left(x\right) \triangleq \frac{p(x)}{q(x)}$ is known as the *importance weight.*

$$\widehat{p}_N\left(dx\right) = \frac{1}{N}\sum_{i=1}^{N} w(x^{(i)})\delta_{x^{(i)}}\left(dx\right)$$

Normalized Importance Sampling

When the normalising constant of $p(x)$ is unknown, it is still possible to apply the importance sampling method:

$$I(f) = \frac{\int f(x)w(x)q(x)\,dx}{\int w(x)q(x)\,dx}$$

Normalized Importance Sampling

The Monte Carlo estimate of $I(f)$ becomes

$$\widetilde{I}_N(f) = \frac{\frac{1}{N}\sum_{i=1}^{N} f\left(x^{(i)}\right) w(x^{(i)})}{\frac{1}{N}\sum_{j=1}^{N} w\left(x^{(i)}\right)} = \sum_{i=1}^{N} f\left(x^{(i)}\right) \widetilde{w}(x^{(i)})$$

where $\widetilde{w}(x^{(i)})$ is a normalised importance weight. For N finite, $\widetilde{I}_N(f)$ is biased (ratio of two estimates) but asymptotically, under weak assumptions, the strong law of large numbers applies, that is $\widetilde{I}_N(f) \underset{N\to\infty}{\xrightarrow{a.s.}} I(f)$.

What is the best proposal?

The IS estimator is unbiased, but has variance

$$\mathrm{var}_{q(x)}\left(\widehat{I}_N(f)\right) = \mathbb{E}_{q(x)}\left(f^2(x)w^2(x)\right) - I^2(f)$$

This variance is minimised when

$$q^\star(x) = \frac{|f(x)|p(x)}{\int |f(x)|p(x)dx}$$

What is the best proposal?

Introduce parametric proposals and adapt the parameters so as to minimise the variance

$$\theta_{t+1} = \theta_t - \alpha \frac{1}{N} \sum_{i=1}^{N} f^2(x^{(i)}) w(x^{(i)}, \theta_t) \frac{\partial w(x^{(i)}, \theta_t)}{\partial \theta_t}$$

where α is a learning rate and $x^{(i)} \sim q(x, \theta)$.

Proposal distributions that adapt to the data are also very widely used.

Example: Logistic Regression

For practical reasons, we parameterise our model. In particular, we introduce the following Bernoulli likelihood function:

$$p(y_t|x_t, \theta) = \left[\frac{1}{1+\exp(-\theta x_t)}\right]^{y_t} \left[1 - \frac{1}{1+\exp(-\theta x_t)}\right]^{1-y_t}$$

where θ are the model parameters. The logistic function $p(y_t = 1|x_t) = \frac{1}{1+\exp(-\theta x_t)}$ is conviniently bounded between 0 and 1.

Example: Logistic Regression

We also assume a Gaussian prior

$$p(\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(\theta - \mu)'(\theta - \mu)\right)$$

The goal of the analysis is then to compute the posterior distribution $p(\theta|x_{1:T}, y_{1:T})$. This distribution will enable us to classify new data as follows

$$p(y_{T+1}|x_{1:T+1}) = \int_\Theta p(y_{T+1}|x_{T+1}, \theta) p(\theta|x_{1:T}, y_{1:T}) d\theta$$

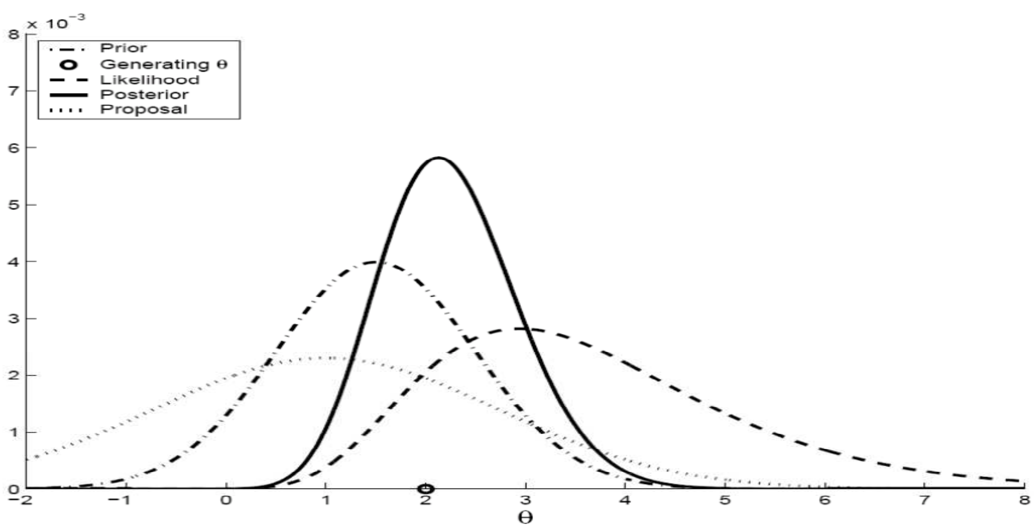
Example: Logistic Regression

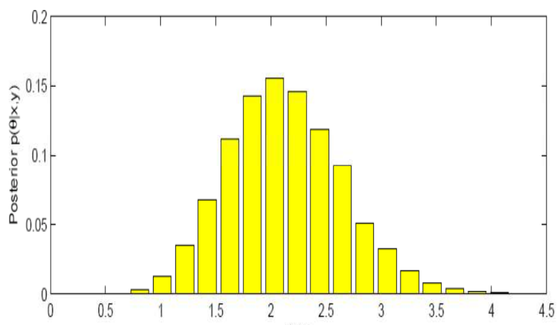
Bayes' rule gives us the following expression for the posterior

$$\begin{aligned} p(\theta|x_{1:T}, y_{1:T}) \;\propto\; & \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(\theta - \mu)'(\theta - \mu)\right) \\ & \times \prod_{t=1}^{T} \left[\frac{1}{1 + \exp\left(-\theta' x\right)}\right]^{y_t} \left[1 - \frac{1}{1 + \exp\left(-\theta' x\right)}\right]^{1-y_t} \end{aligned}$$

Example: Logistic Regression

The problem is that in this case we can't solve the normalising integral analytically. So we have to use numerical methods — in this case importance sampling — to approximate $p(\theta|x_{1:T}, y_{1:T})$. Note that we cannot sample from $p(\theta|x_{1:T}, y_{1:T})$ directly because we don't know the normalising constant. So instead we sample from a proposal distribution $q(\theta)$ (say a Gaussian) and weight the samples using importance sampling. After obtaining N samples of θ from the posterior, we can classify new data as follows

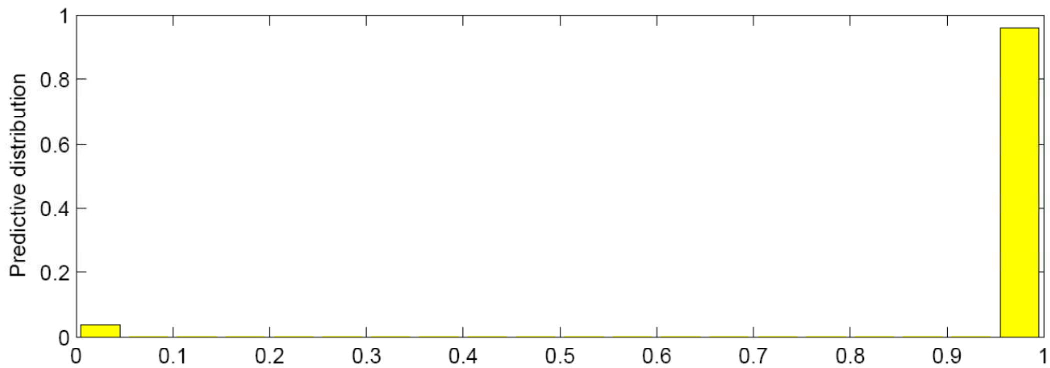




Example: Logistic Regression

$$p(y_{T+1}|x_{1:T+1}) = \int_\Theta p(y_{T+1}|x_{T+1},\theta)p(\theta|x_{1:T},y_{1:T})d\theta$$

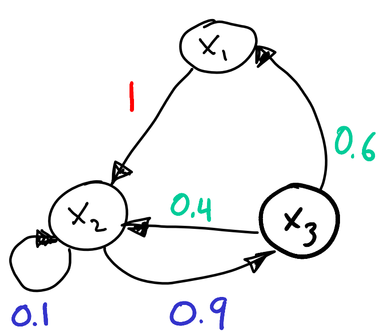
$$p(y_{T+1}|x_{1:T+1}) = \frac{1}{N}\sum_{i=1}^{N} p(y_{T+1}|x_{T+1},\theta^{(i)})$$



Markov Chain Monte Carlo

For simplicity, let's consider only 3 states:

$x_t \in \mathcal{X} = \{x_1, x_2, x_3\}$



$$T = p(x_t | x_{t-1}) = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0.1 & 0.9 \\ 0.6 & 0.4 & 0 \end{bmatrix}$$

Think of this as a webgraph. Our goal is to crawl it to find the "relevance" of each node.

Markov Chain Monte Carlo

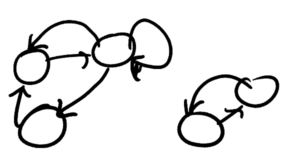
T is a stochastic matrix. As long as the graph (state space) is aperiodic and irreducible, we have that for any initial vector of probabilities ν:

$$\nu' T^t \rightarrow \pi' \quad \text{as } t \rightarrow \infty$$

where π is the invariant or stationary distribution of the chain. It is unique.

Markov Chain Monte Carlo

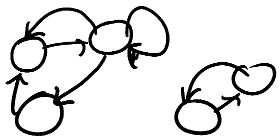
Need for irreducibility:



One cluster might never be visited!

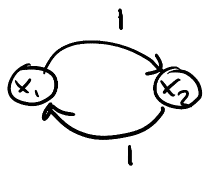
Markov Chain Monte Carlo

Need for irreducibility:



One cluster might never be visited!

Need for aperiodicity:



$$T = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

Let $\pi = [\tfrac{1}{3} \;\; \tfrac{2}{3}]$

$$\pi' T = [2/3 \;\; 1/3]$$
$$\pi' T^2 = [1/3 \;\; 2/3]$$
$$\vdots$$

Oscillation!

Markov Chain Monte Carlo

In the limit:

$$\pi' T = \pi'$$

π is the left eigenvector of T with corresponding eigenvalue 1.

Markov Chain Monte Carlo

In the limit:

$$\pi' T = \pi'$$

π is the left eigenvector of T with corresponding eigenvalue 1. Componentwise, we have:

$$\sum_{i=1}^{3} \pi_i T_{ij} = \pi_j$$

Markov Chain Monte Carlo

In the limit:

$$\pi' T = \pi'$$

π is the left eigenvector of T with corresponding eigenvalue 1. Componentwise, we have:

$$\sum_{i=1}^{3} \pi_i T_{ij} = \pi_j$$

As the state space grows:

$$\int \pi(x) \underbrace{p(y|x)}_{\text{Markov Chain Kernel}} dx = \pi(y)$$

Markov Chain Monte Carlo

Detailed Balance:

If $\pi(x_t)P(x_{t+1}|x_t) = \pi(x_{t+1})P(x_t|x_{t+1})$

Integrating over x_t yields

$$\int \pi(x_t)P(x_{t+1}|x_t) = \pi(x_{t+1})$$

Which is the ergodic behaviour we want. Now we have a sufficient condition for designing $P(x_{t+1}|x_t)$ so as to get samples from π ☒

MCMC: Metropolis-Hastings

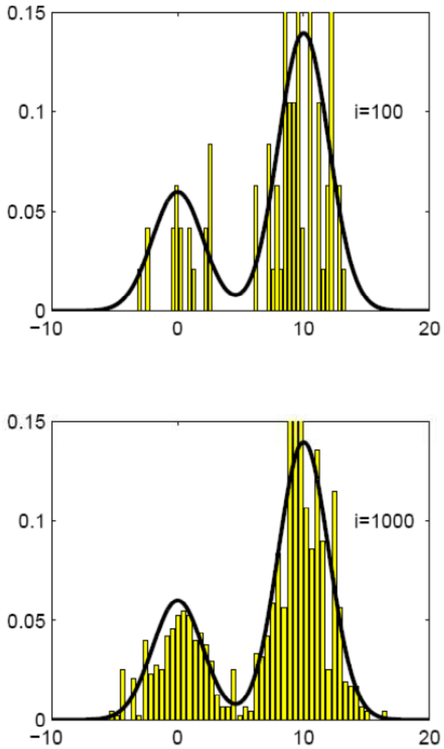
- Initialise $x^{(0)}$.
- For $i = 0$ to $N-1$
  - Sample $u \sim U_{[0,1]}$.
  - Sample $x^\star \sim q(x^\star|x^{(i)})$.
  - If $u < A(x^{(i)}, x^\star) = \min\left\{1, \frac{p(x^\star)q(x^{(i)}|x^\star)}{p(x^{(i)})q(x^\star|x^{(i)})}\right\}$

    $$x^{(i+1)} = x^\star$$

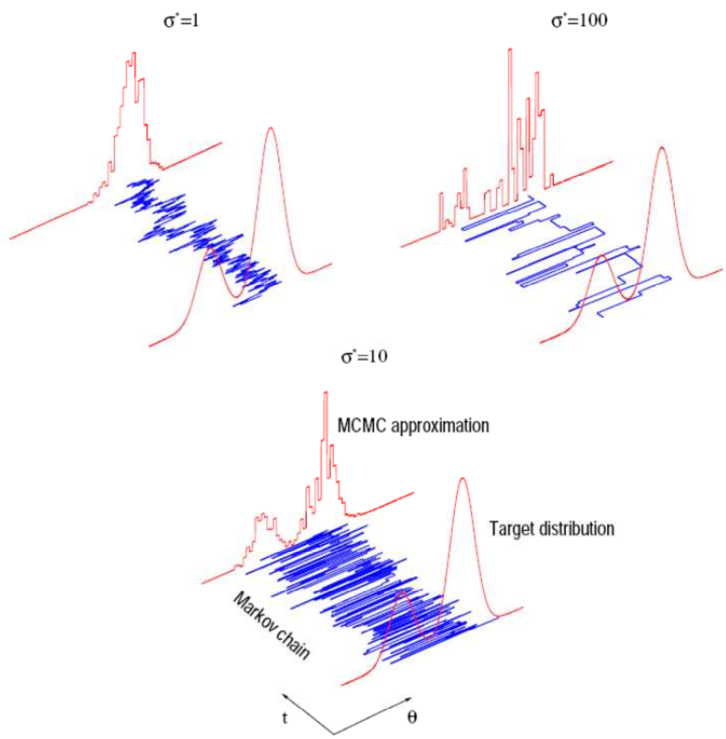
    else

    $$x^{(i+1)} = x^{(i)}$$

MCMC: Metropolis-Hastings



MCMC: Choosing the Right Proposal



MCMC: Theory

Kernel:

$$K(x,B) = \begin{cases} q(B|x)\, A(x,B) & x \notin B \\ 1 - \int_{x' \in \{\mathcal{X} \setminus B\}} q(x'|x)\, A(x,x') & x \in B \end{cases}$$

$$\therefore \quad K(x,B) = q(B|x)\, A(x,B) + \mathbb{I}_{x \in B} \left\{ \begin{array}{l} 1 - q(B|x)\, A(x,B) \\ - \int_{x' \in \{\mathcal{X} \setminus B\}} q(x'|x)\, A(x,x') \end{array} \right\}$$

$$K(x,B) = q(B|x)\, A(x,B) + \mathbb{I}_{x \in B} \left\{ 1 - \int_{x' \in \mathcal{X}} q(x'|x)\, A(x,x') \right\}$$

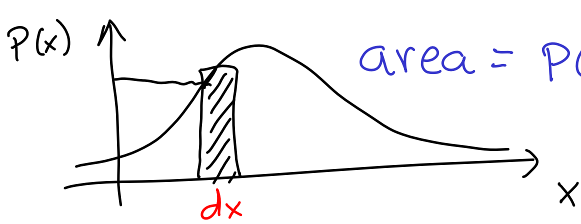
MCMC: Theory

Detailed balance:

$$\pi(A)\, K(A,B) = \pi(B)\, K(B,A)$$

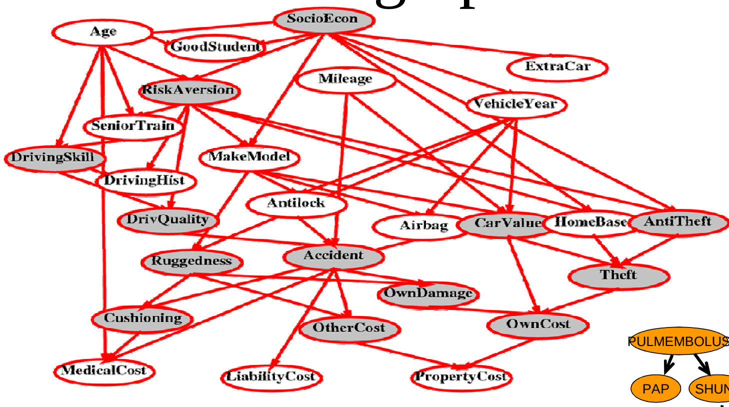
$$\int_{x \in A} \pi(dx)\, K(x,B) = \int_{y \in B} \pi(dy)\, K(y,A)$$

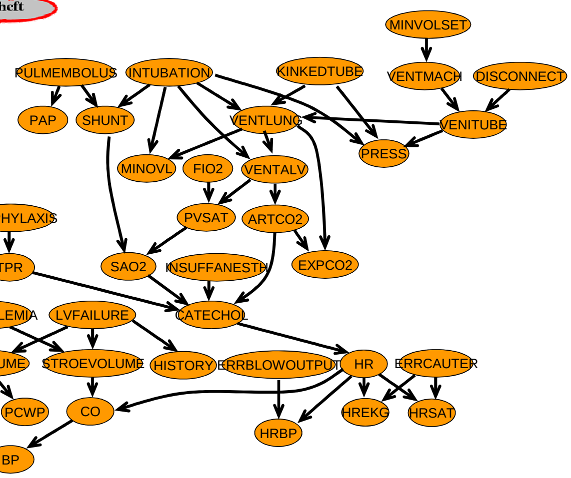
Note: $\int f(x)\, p(x)\, dx \equiv \int f(x)\, p(dx)$



area $= P(dx) = P(x)\, dx$

Extending MH to directed probabilistic graphical models





Gibbs Sampling

Choose the following proposal:

$$q(x^\star|x^{(i)}) = \begin{cases} p(x_j^\star|x_{-j}^{(i)}) & \text{If } x_{-j}^\star = x_{-j}^{(i)} \\ 0 & \text{Otherwise.} \end{cases}$$

where $x_{-j} = \{x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n\}$.

Then the acceptance is:

$$A(x^{(i)}, x^\star) \quad = \quad \min\left\{1, \frac{p(x^\star)q(x^{(i)}|x^\star)}{p(x^{(i)})q(x^\star|x^{(i)})}\right\} = 1.$$

Gibbs Sampling

- Initialise $x_{1:n}^{(0)}$.
- For $i = 0$ to $N - 1$
  - Sample $x_1^{(i+1)} \sim p(x_1 | x_2^{(i)}, x_3^{(i)}, \ldots, x_n^{(i)})$.
  - Sample $x_2^{(i+1)} \sim p(x_2 | x_1^{(i+1)}, x_3^{(i)}, \ldots, x_n^{(i)})$.

$$\vdots$$

  - Sample $x_j^{(i+1)} \sim p(x_j | x_1^{(i+1)}, \ldots, x_{j-1}^{(i+1)}, x_{j+1}^{(i)}, \ldots, x_n^{(i)})$.

$$\vdots$$

  - Sample $x_n^{(i+1)} \sim p(x_n | x_1^{(i+1)}, x_2^{(i+1)}, \ldots x_{n-1}^{(i+1)})$.

Gibbs Sampling For Graphical models

A large-dimensional joint distribution is factored into a directed graph that encodes the conditional independencies in the model. In particular, if $x_{pa(j)}$ denotes the parent nodes of node x_j, we have

$$p(x) = \prod_j p(x_j | x_{pa(j)}).$$

It follows that the full conditionals simplify as follows

$$p(x_j | x_{-j}) = p(x_j | x_{pa(j)}) \prod_{k \in ch(j)} p(x_k | x_{pa(k)})$$

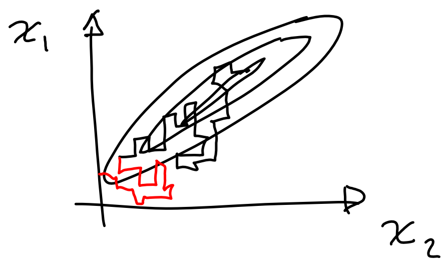
where $ch(j)$ denotes the children nodes of x_j.

MH is a Building Block

- **Idea**: Split the high dimensional vector x into blocks $\{x_{b1}, \ldots, x_{bn}\}$.

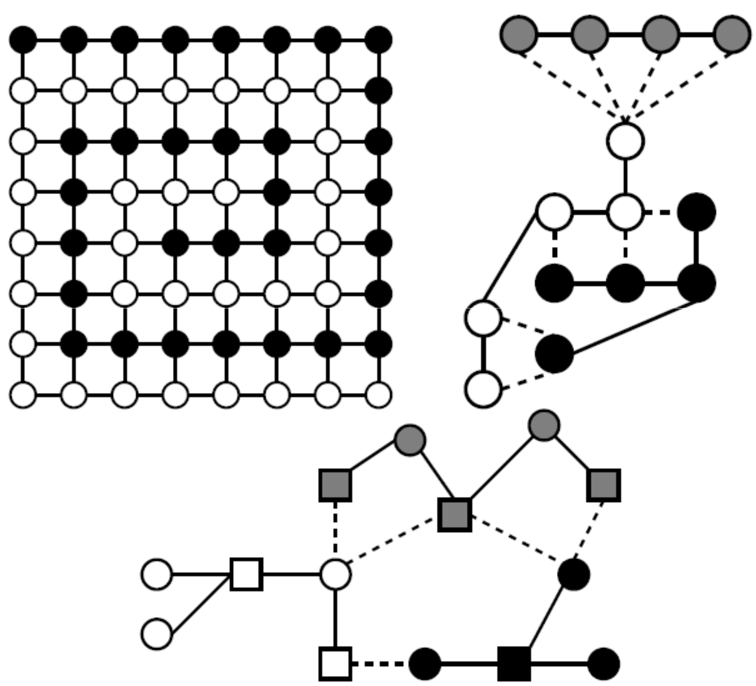
- **Cycle**: sample each block using an MH algorithm with invariant distribution $p(x_{bi}|x_{-bi})$ and proposal distribution $q(x_{bi})$, where $x_{-bi} = \{\text{All blocks except } x_{bi}\}$.

- Block highly correlated variables.

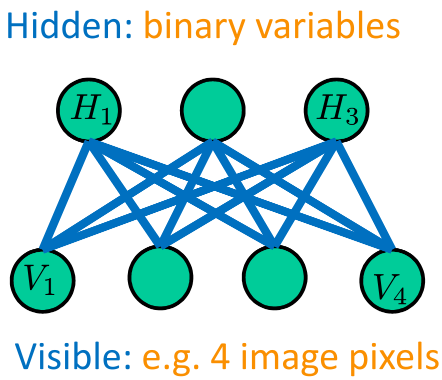


Chain can take a long time to mix when variables are correlated.

Collapsing and Blocking



Restricted Boltzmann Machines



A joint configuration (v,h) of the binary visible and hidden units has an energy given by

$$E(\mathbf{v}, \mathbf{h}) = -\sum_{i \in \text{pixels}} b_i v_i - \sum_{j \in \text{features}} b_j h_j - \sum_{i,j} v_i h_j w_{ij}$$

Auxiliary Variable Samplers

- It is often easier to sample from an augmented distribution $p(x, u)$, where u is an auxiliary variable, than from $p(x)$.
- It is possible to obtain marginal samples $x^{(i)}$ by sampling $(x^{(i)}, u^{(i)})$ according to $p(x, u)$ and, then, ignoring the samples $u^{(i)}$.
- This very useful idea was proposed in the physics literature (Swendsen and Wang, 1987).

Hybrid (Hamiltonian) Monte Carlo

- The idea is to exploit gradient information.
- Define the extended target distribution:

$$p(x, u) = p(x) N(u; 0, I_{n_x}).$$

- Introduce the gradient vector: $\Delta(x) = \partial \log p(x) / \partial x$
- Introduce the parameters ρ and L.
- Next we "leapfrog".

Hybrid Monte Carlo

- Sample $v \sim U_{[0,1]}$ and $u^\star \sim N(0, I_{n_x})$.
- Let $x_0 = x^{(i)}$ and $u_0 = u^\star + \rho \Delta(x_0)/2$.
- For $l = 1, \ldots, L$, take steps

$$x_l = x_{l-1} + \rho u_{l-1}$$
$$u_l = u_{l-1} + \rho_l \Delta(x_l)$$

where $\rho_l = \rho$ for $l < L$ and $\rho_L = \rho/2$.

- If $v < A = \min\left\{1, \frac{p(x_L)}{p(x^{(i)})} \exp\left(-\frac{1}{2}(u_L^{\mathrm{T}} u_L - u^{\star\mathrm{T}} u^\star)\right)\right\}$

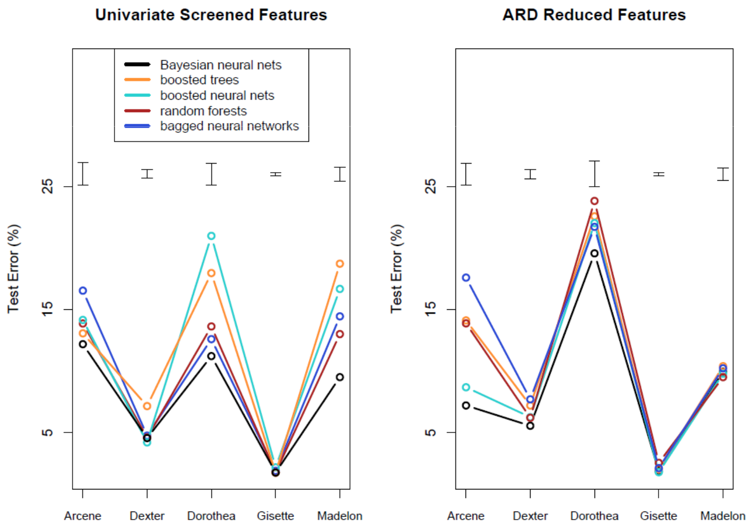
$$(x^{(i+1)}, u^{(i+1)}) = (x_L, u_L)$$

else $$(x^{(i+1)}, u^{(i+1)}) = (x^{(i)}, u^\star)$$

HMC for Bayesian NNs

$$\Pr(\theta|\mathbf{X}_{\text{tr}}, \mathbf{y}_{\text{tr}}) = \frac{\Pr(\theta)\Pr(\mathbf{y}_{\text{tr}}|\mathbf{X}_{\text{tr}}, \theta)}{\int \Pr(\theta)\Pr(\mathbf{y}_{\text{tr}}|\mathbf{X}_{\text{tr}}, \theta)d\theta}$$

$$\Pr(Y_{\text{new}}|X_{\text{new}}, \mathbf{X}_{\text{tr}}, \mathbf{y}_{\text{tr}}) = \int \Pr(Y_{\text{new}}|X_{\text{new}}, \theta)\Pr(\theta|\mathbf{X}_{\text{tr}}, \mathbf{y}_{\text{tr}})d\theta$$



[Radford Neal – Hastie, Friedman & Tibshirani]