ISTA 421 + INFO 521
Introduction to Machine Learning

Lecture 17: Estimation III: Sampling, MCMC, Metropolis-Hastings

Clay Morrison
claytonm@email.arizona.edu
Harvill 437A
Phone 621-6609

30 October 2017

1

$$g(\mathbf{w}; \mathbf{X}, \mathbf{t}, \sigma^2) = p(\mathbf{t}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\sigma^2)$$
$$p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \sigma^2) = Z^{-1} g(\mathbf{w}; \mathbf{X}, \mathbf{t}, \sigma^2)$$
$$Z^{-1} = p(\mathbf{t}|\mathbf{X}, \sigma^2) = \int p(\mathbf{t}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\sigma^2)\, d\mathbf{w}$$

Our Options (when cannot compute ***posterior*** directly)

1. Find the single value of **w** that corresponds to the highest value of the posterior. As $g(\mathbf{w}; \mathbf{X}, \mathbf{t}, \sigma^2)$ is proportional to the posterior, a maximum of *g* will also correspond to a maximum of the posterior. Z^{-1} is not a function of **w**. MAP (or ML)
2. Approximate $p(\mathbf{w}; \mathbf{X}, \mathbf{t}, \sigma^2)$ with some other density that we can compute analytically.
3. Sample directly from the posterior $p(\mathbf{w}; \mathbf{X}, \mathbf{t}, \sigma^2)$, knowing only $g(\mathbf{w}; \mathbf{X}, \mathbf{t}, \sigma^2)$

2

Method 1: ML or MAP point estimate

- While we cannot derive a direct analytic likelihood or posterior density, we can computer something proportional to it – e.g., for the posterior:

$$g(\mathbf{w}; \mathbf{X}, \mathbf{t}, \sigma^2) = p(\mathbf{t}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\sigma^2)$$

- Find the value of **w** that maximizes *g*
- Corresponds to the value at the maximum of the posterior (or just the likelihood).
- This is the most likely value $\hat{\mathbf{w}}$ under the posterior (or just the likelihood).
- In cases like logistic regression, need to estimate $\hat{\mathbf{w}}$ though an approximation method; we introduced gradient methods:
  - general first order gradient descent/ascent (Woodrow-Hoff)
  - Newton-Raphson

3

Widrow-Hoff

$$\mathbf{w} = \mathbf{w} - \alpha \frac{\partial \mathcal{L}}{\partial \mathbf{w}}$$

Gradient descent (-) for loss
Gradient ascent (+) for probability density

Newton-Raphson for MAP

$$\log g(\mathbf{w}; \mathbf{X}, \mathbf{t}) = \log p(\mathbf{t}|\mathbf{X}, \mathbf{w}) + \log p(\mathbf{w}|\sigma^2)$$

$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}$$

$$\mathbf{w}' = \mathbf{w} - \left(\frac{\partial^2 \log g(\mathbf{w}; \mathbf{X}, \mathbf{t})}{\partial \mathbf{w} \partial \mathbf{w}^\mathsf{T}} \right)^{-1} \frac{\partial \log g(\mathbf{w}; \mathbf{X}, \mathbf{t})}{\partial \mathbf{w}}$$

Point is guaranteed maximum if Hessian is negative definite
(as we showed for max likelihood)

4

Method 2:
The Laplace* Approximation

- **The Idea**: *approximate* the density of interest with a Gaussian.
- (Recall that the Gaussian is used quite often in statistics to approximate other distributions!)
- However, **keep in mind**: our predictions will only be as good as our approximation – if the true posterior is not very Gaussian, then our predictions will be easy to compute but not very useful.

*Following the note in the book: the Machine Learning community has come to refer to the method this way, but this is elsewhere referred to as **saddle-point approx.**, and in statistics, the Laplace approx. is something different.

$$p(\mathbf{t}|\mathbf{X},\mathbf{w}) = \prod_{n=1}^{N} \left(\frac{1}{1+\exp(-\mathbf{w}^\top \mathbf{x}_n)}\right)^{t_n} \left(\frac{\exp(-\mathbf{w}^\top \mathbf{x}_n)}{1+\exp(-\mathbf{w}^\top \mathbf{x}_n)}\right)^{1-t_n} \qquad p(\mathbf{w}|\sigma^2) = \mathcal{N}(0, \sigma^2 \mathbf{I})$$

Approximating *g* using the Taylor Expansion

$$g(\mathbf{w}; \mathbf{X}, \mathbf{t}, \sigma^2) = p(\mathbf{t}|\mathbf{X},\mathbf{w}) p(\mathbf{w}|\sigma^2)$$

$$\log g(\mathbf{w}; \mathbf{X}, \mathbf{t}) = \log p(\mathbf{t}|\mathbf{X},\mathbf{w}) + \log p(\mathbf{w}|\sigma^2)$$

$$\sum_{n=0}^{\infty} \frac{(\mathbf{w}-\hat{\mathbf{w}})^n}{n!} \frac{\partial^n f(\mathbf{w})}{\partial \mathbf{w}^n}\bigg|_{\hat{\mathbf{w}}}$$

Recall, the univariate Gaussian:

$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(w-\mu)^2\right\}$$

The log of the univ. G
(K is the normalizing constant):

$$\log(K) - \frac{1}{2\sigma^2}(w-\mu)^2$$

similar form!

This is the Laplace approximation!
We approximate the posterior with a Gaussian that has its
mean at the posterior **mode** ($\hat{w}$),
variance inversely proportional to the curvature of the posterior (g'') at its mode.

Univariate version:

$\mu = \hat{w}$ $\quad \sigma^2 = 1/v$

Multivariate version:

$$\boldsymbol{\mu} = \hat{\mathbf{w}}, \quad \boldsymbol{\Sigma}^{-1} = -\left(\frac{\partial^2 \log g(\mathbf{w}; \mathbf{X}, \mathbf{t})}{\partial \mathbf{w} \partial \mathbf{w}^\top}\right)\bigg|_{\hat{\mathbf{w}}}$$

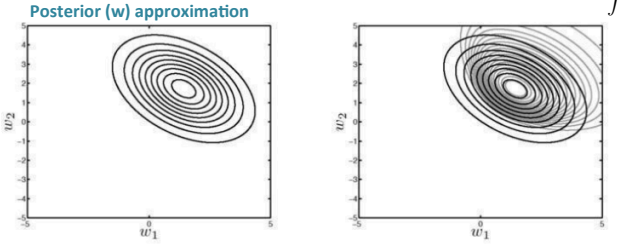
$$\log g(w; \mathbf{X}, \mathbf{t}, \sigma^2) \approx \log g(\hat{w}; \mathbf{X}, \mathbf{t}, \sigma^2) - \frac{v}{2}(w-\hat{w})^2 \qquad v = -\frac{\partial^2 \log g(w; \mathbf{X}, \mathbf{t}, \sigma^2)}{\partial w^2}\bigg|_{\hat{w}}$$

Laplace Approximation Example

Predictive distribution: $P(T_{\text{new}=1}|\mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t}, \sigma^2) = \mathbf{E}_{p(\mathbf{w}|\mathbf{X},\mathbf{r},\sigma^2)}\{P(T_{\text{new}} = 1|\mathbf{x}_{\text{new}}, \mathbf{w})\}$

$$= \int P(T_{\text{new}} = 1|\mathbf{x}_{\text{new}}, \mathbf{w}) p(\mathbf{w}|\mathbf{X}, \mathbf{r}, \sigma^2) d\mathbf{w}$$

Posterior (w) approximation



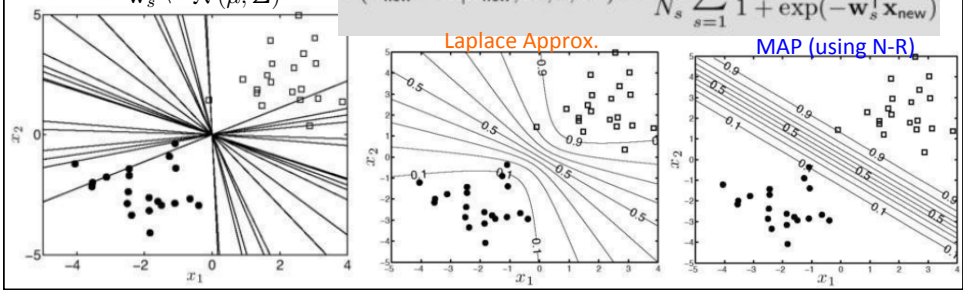
(a) Laplace approximation to the posterior

$$\mathbf{w}_s \leftarrow \mathcal{N}(\mu, \mathbf{\Sigma})$$

$$P(T_{\text{new}} = 1|\mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t}, \sigma^2) = \frac{1}{N_s} \sum_{s=1}^{N_s} \frac{1}{1 + \exp(-\mathbf{w}_s^{\mathsf{T}} \mathbf{x}_{\text{new}})}$$

Laplace Approx.

MAP (using N-R)



$$g(\mathbf{w}; \mathbf{X}, \mathbf{t}, \sigma^2) = p(\mathbf{t}|\mathbf{X}, \mathbf{w}) p(\mathbf{w}|\sigma^2)$$

$$p(\mathbf{w}|\mathbf{X}, \mathbf{t}, \sigma^2) = Z^{-1} g(\mathbf{w}; \mathbf{X}, \mathbf{t}, \sigma^2)$$

$$Z^{-1} = p(\mathbf{t}|\mathbf{X}, \sigma^2) = \int p(\mathbf{t}|\mathbf{X}, \mathbf{w}) p(\mathbf{w}|\sigma^2)\, d\mathbf{w}$$

Our Options (when cannot compute ***posterior*** directly)

1. Find the single value of **w** that corresponds to the highest value of the posterior. As $g(\mathbf{w}; \mathbf{X}, \mathbf{t}, \sigma^2)$ is proportional to the posterior, a maximum of *g* will also correspond to a maximum of the posterior. Z^{-1} is not a function of **w**. MAP
2. Approximate $p(\mathbf{w}; \mathbf{X}, \mathbf{t}, \sigma^2)$ with some other density that we can compute analytically.
3. Sample directly from the posterior $p(\mathbf{w}; \mathbf{X}, \mathbf{t}, \sigma^2)$, knowing only $g(\mathbf{w}; \mathbf{X}, \mathbf{t}, \sigma^2)$

Method 3:
Sampling from Posterior

- Interest in Posterior density is to allow us to take ***all*** the uncertainty in **w** into account when making predictions:

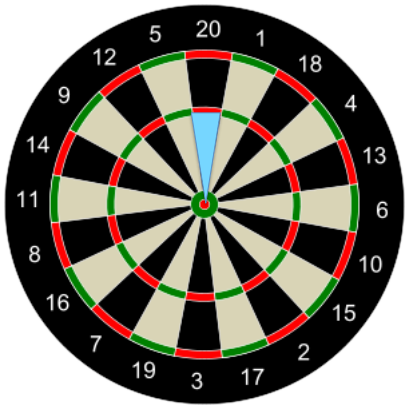
Posterior density over the parameters

$$P(T_{\text{new}=1}|\mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t}, \sigma^2) = \mathbf{E}_{p(\mathbf{w}|\mathbf{X},\mathbf{r},\sigma^2)}\{P(T_{\text{new}} = 1|\mathbf{x}_{\text{new}}, \mathbf{w})\}$$
$$= \int P(T_{\text{new}} = 1|\mathbf{x}_{\text{new}}, \mathbf{w}) p(\mathbf{w}|\mathbf{X}, \mathbf{r}, \sigma^2) d\mathbf{w}$$

- Laplace method uses a *similar* density to provide an *approximation* of the posterior; still had to sample from it to estimate the integral.
- Now we'll look at sampling ***directly*** from the ***true*** posterior.

9

The Intuition



$\mathbf{y}$: position of the dart

$\boldsymbol{\Delta}$: intended target

$p(\mathbf{y}|\boldsymbol{\Delta})$ Can be hard to compute

$T = f(\mathbf{y})$ A new random variable: T=1 : within 20 T=0 : outside of 20

$P(T = 1|\boldsymbol{\Delta})$

Can be hard to compute

$$P(T = 1|\boldsymbol{\Delta}) = \mathbb{E}_{p(\mathbf{y}|\boldsymbol{\Delta})}\{f(\mathbf{y})\} = \int f(\mathbf{y}) p(\mathbf{y}|\boldsymbol{\Delta}) \mathrm{d}\mathbf{y}$$

OR: have your friend try to hit the 20, and find average hits!

$\mathbf{y}_s \leftarrow p(\mathbf{y}|\boldsymbol{\Delta})$

$$P(T = 1|\boldsymbol{\Delta}) \simeq \frac{1}{N_s} \sum_{s=1}^{N_s} f(\mathbf{y}_s)$$

10

Expectation

$$\mathbb{E}_{p(z)}\{f(z)\} = \int f(z)p(z)\mathrm{d}z \simeq \frac{1}{N_s}\sum_{s=1}^{N_s} f(z_s) \ , \quad z_s \leftarrow p(z)$$

Monte Carlo Approximation

$$\mathbb{E}_{p(z)}\{f(z)\} = \int f(z)p(z)\mathrm{d}z \simeq \frac{1}{N_s}\sum_{s=1}^{N_s} f(z_s) \ , \quad z_s \leftarrow p(z)$$

Monte Carlo Approximation

The **predictive distribution** is a marginalization:

$$P(T_{new}=1|\mathbf{x}_{new},\mathbf{X},\mathbf{t},\sigma^2) = \int P(T_{new}=1|\mathbf{x}_{new},\mathbf{w})p(\mathbf{w}|\mathbf{X},\mathbf{t},\sigma^2)\mathrm{d}\mathbf{w}$$
$$= \mathbb{E}_{p(\mathbf{w}|\mathbf{X},\mathbf{t},\sigma^2)}\{P(T_{new}=1|\mathbf{x}_{new},\mathbf{w})\}$$

… Put another way: it's the expectation of the likelihood function under the posterior distribution

We can *estimate* this expectation through **Monte Carlo approximation**:

$$P(T_{new}=1|\mathbf{x}_{new},\mathbf{X},\mathbf{t},\sigma^2) \simeq \frac{1}{N_s}\sum_{s=1}^{N_s} P(T_{new}=1|\mathbf{x}_{new},\mathbf{w}_s)$$

$$\mathbf{w}_s \leftarrow p(\mathbf{w}|\mathbf{X},\mathbf{t},\sigma^2)$$

But how do we estimate this?

(A touch of theory on) **Markov Chains**

- A **Stochastic Process** is a collection of random variables indexed by a set T; i.e., $\{ X_t \mid t \in T \}$
  - Here, we only care about discrete stochastic processes, i.e., when T is a countable set.
  - For example, $T = \mathbf{Z}_+$ or $T = \{0,1,2,3,4,5\}$
- **Example**
  - The results of 100 coin tosses is a stochastic process, with random variables $C_1,\ldots,C_{100}$
  - The daily temperature is a stochastic process, represented by random variables T_1, T_2, …
- Stochastic processes are like any other sets of variables; we can talk about distributions:
  - $p_t(x_t)$, for any $t \in T$
  - $p(x_{t1},\ldots,x_{tn})$, for any $\{t_1, \ldots, t_n\} \subset T$

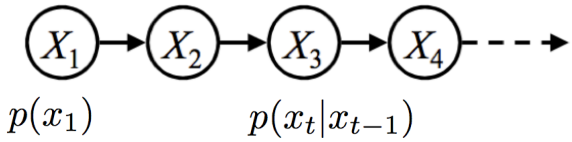
13

(A touch of theory on) **Markov Chains**

- A (first-order) **Markov chain** is a discrete(-time) stochastic process with the Markov Property:

$$p(x_t|x_{t-1}, x_{t-2}, \ldots, x_1) = p(x_t|x_{t-1})$$

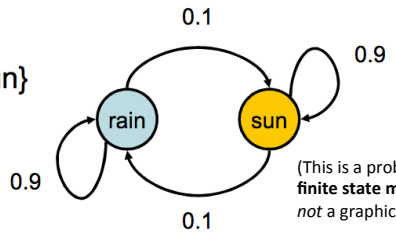
  - That is, the variable at time t only depends on the variable at time t–1.
  - Here, the conditional density $p(x_t \mid x_{t-1})$ (also called the **transition kernel**) is the **same** for all $t \in T$



14

Example Markov Chain

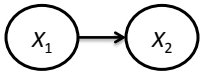
- Weather:
  - States: X = {rain, sun}
  - Transitions:



(This is a probabilistic **finite state machine**, *not* a graphical model !)

  - Initial distribution: 1.0 sun
  - What's the probability distribution after one step?

$$P(X_2 = \text{sun}) = P(X_2 = \text{sun}|X_1 = \text{sun})P(X_1 = \text{sun}) + P(X_2 = \text{sun}|X_1 = \text{rain})P(X_1 = \text{rain})$$

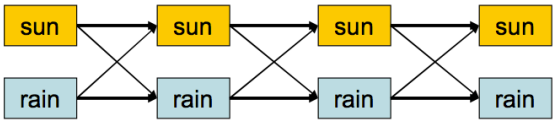


$$0.9 \cdot 1.0 + 0.1 \cdot 0.0 = 0.9$$

Join (product) of X_1 and X_2, followed by sum (marginalization) of X_1

Mini "Forward" Algorithm for Markov Chain

- Question: What's P(X) on some day t?
  - An instance of variable elimination!



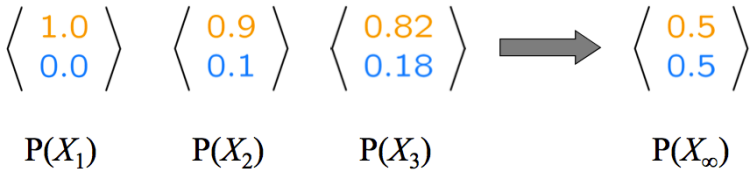
$$P(x_t) = \sum_{x_{t-1}} P(x_t|x_{t-1})P(x_{t-1})$$

$$P(x_1) = \text{known}$$

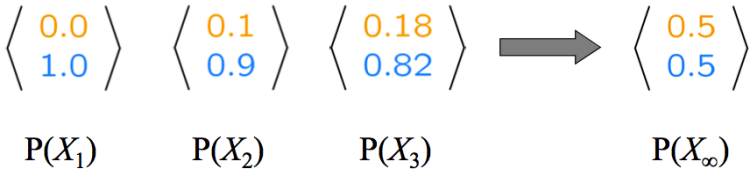
Forward simulation

Example Markov Chains

- From initial observation of sun



- From initial observation of rain



Example Markov Chains

- What if we had a different transition probability ?

$$t \begin{matrix} \\ \text{sunny} \\ \text{rainy} \end{matrix} \overset{\begin{matrix} t+1 \\ \text{sunny} \quad \text{rainy} \end{matrix}}{\begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix}}$$

- If we set $\pi_0 = [1, 0]^\top$, i.e., today is sunny:
  - $\pi_2 = [0.844, 0.156]^\top$
  - $\pi_5 = [0.834, 0.166]^\top$
  - $\pi_{20} = [0.83333, 0.16667]^\top$
  - $\pi_{50} = [0.83333, 0.16667]^\top$
  - See a pattern?

Stationary Distribution

- A distribution π is **stationary** for a Markov chain if the transition kernel for the chain preserves π; i.e., if for all $x_t \in R^d$

$$\int p(x_t|x_{t-1})\pi(x_{t-1})\mathrm{d}x_{t-1} = \pi(x_t)$$

- **Implication**: if at any time t, $p_t(x_t) = \pi(x_t)$, then the marginals from that point on will be $\pi(x_t)$, since

$$\begin{aligned}\int p_{t+1}(x_{t+1}) &= \int p(x_{t+1}|x_t)p_t(x_t)\mathrm{d}x_t \\ &= \int p(x_{t+1}|x_t)\pi(x_t)\mathrm{d}x_t \\ &= \pi(x_{t+1})\end{aligned}$$

- **Ergodicity**: guarantees a stationary distribution exists and is unique
  - **Aperiodic**: can always transition back to state *a* having just transitioned from *a* to *b* (a state *x* has a period *k* if, starting from that state, it is only possible to return to it in multiples of *k*; in that case, the x is said to be periodic.)
  - **Irreducible**: It is possible to get to any state from any state (i.e., does not end up in a sink)
- **Theorem**: If a Markov chain is irreducible and aperiodic, then it will have a unique stationary distribution
- So, how do we construct one for our posterior distribution of interest?

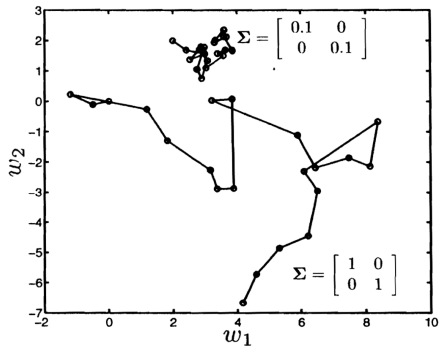
19

Named in the list of Top 10 Algorithms of the 20th Century (*SIAM News*, Vol 33, No 4, 2000)

Nicholas Metropolis-Hastings Algorithm W. Keith

physicist 1953 *statistician* 1970

Random Walk



Two desirable criteria for proposal distribution:
(a) Easy to sample from
(b) Symmetric:

$$p(\widetilde{\mathbf{w}_s}|\mathbf{w}_{s-1}, \boldsymbol{\Sigma}) = p(\mathbf{w}_{s-1}|\widetilde{\mathbf{w}_s}, \boldsymbol{\Sigma})$$

$\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{s-1}, \mathbf{w}_s, \ldots, \mathbf{w}_{N_s}$

(0) Getting started: choosing $\mathbf{w}_1$
Turns out it doesn't matter: in theory, sample long enough and guaranteed to converge
(e.g.: sample from prior!)

Generating $\mathbf{w}_s$ takes 2 steps:

(1) Propose a new sample (based on previous)

$$\widetilde{\mathbf{w}_s} \leftarrow p(\widetilde{\mathbf{w}_s}|\mathbf{w}_{s-1})$$ proposal distribution

Does **not** have to be related to target distribution!
Popular to use Gaussian centered on $\mathbf{w}_{s-1}$

$$p(\widetilde{\mathbf{w}_s}|\mathbf{w}_{s-1}, \boldsymbol{\Sigma}) = \mathcal{N}(\mathbf{w}_{s-1}, \boldsymbol{\Sigma})$$

(2) Test whether to accept or reject $\widetilde{\mathbf{w}_s}$

$$r = \frac{p(\widetilde{\mathbf{w}_s}|\mathbf{X}, \mathbf{t}, \sigma^2)}{p(\mathbf{w}_{s-1}|\mathbf{X}, \mathbf{t}, \sigma^2)} \frac{p(\mathbf{w}_{s-1}|\widetilde{\mathbf{w}_s}, \boldsymbol{\Sigma})}{p(\widetilde{\mathbf{w}_s}|\mathbf{w}_{s-1}, \boldsymbol{\Sigma})}$$

posterior at proposed sample / posterior at old sample; ratio of Proposal densities; 1

Problem?: Cannot directly compute posteriors!

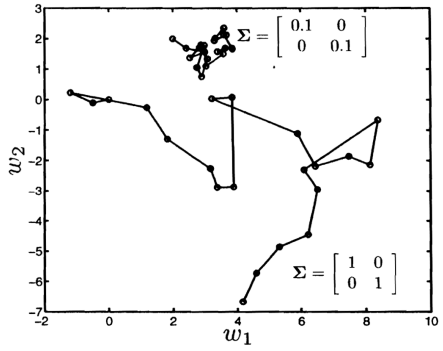
20

Named in the list of Top 10 Algorithms of the 20th Century (*SIAM News*, Vol 33, No 4, 2000)

Nicholas Metropolis-Hastings W. Keith Algorithm

physicist 1953 *statistician* 1970

Random Walk



Two desirable criteria for proposal distribution:
(a) Easy to sample from
(b) Symmetric:

$$p(\widetilde{\mathbf{w}}_s|\mathbf{w}_{s-1}, \boldsymbol{\Sigma}) = p(\mathbf{w}_{s-1}|\widetilde{\mathbf{w}}_s, \boldsymbol{\Sigma})$$

$\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{s-1}, \mathbf{w}_s, \ldots, \mathbf{w}_{N_s}$

(0) Getting started: choosing $\mathbf{w}_1$
Turns out it doesn't matter: in theory, sample long enough and guaranteed to converge
(e.g.: sample from prior!)

Generating $\mathbf{w}_s$ takes 2 steps:

(1) Propose a new sample (based on previous)

$$\widetilde{\mathbf{w}}_s \leftarrow p(\widetilde{\mathbf{w}}_s|\mathbf{w}_{s-1}) \quad \text{proposal distribution}$$

Does **not** have to be related to target distribution!
Popular to use Gaussian centered on $\mathbf{w}_{s-1}$

$$p(\widetilde{\mathbf{w}}_s|\mathbf{w}_{s-1}, \boldsymbol{\Sigma}) = \mathcal{N}(\mathbf{w}_{s-1}, \boldsymbol{\Sigma})$$

(2) Test whether to accept or reject $\widetilde{\mathbf{w}}_s$

$$r = \frac{g(\widetilde{\mathbf{w}}_s; \mathbf{X}, \mathbf{t}, \sigma^2)}{g(\mathbf{w}_{s-1}; \mathbf{X}, \mathbf{t}, \sigma^2)} = \frac{p(\widetilde{\mathbf{w}}_s|\sigma^2)}{p(\mathbf{w}_{s-1}|\sigma^2)} \frac{p(\mathbf{t}|\widetilde{\mathbf{w}}_s, \mathbf{X})}{p(\mathbf{t}|\mathbf{w}_{s-1}, \mathbf{X})}$$

Don't need to calculate posteriors directly **because the Marginal Likelihoods cancel in the ratio!**

$r \geq 1$? If **yes**: always choose best: $\mathbf{w}_s = \widetilde{\mathbf{w}}_s$

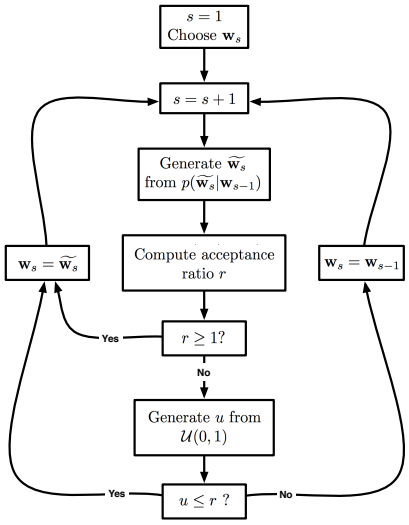
If **no**: possibly accept anyway

$u \leftarrow \mathcal{U}(0,1), \quad u \leq r$? yes / no → $\mathbf{w}_s = \mathbf{w}_{s-1}$

Named in the list of Top 10 Algorithms of the 20th Century (*SIAM News*, Vol 33, No 4, 2000)

Nicholas Metropolis-Hastings W. Keith Algorithm

physicist 1953 *statistician* 1970



$\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{s-1}, \mathbf{w}_s, \ldots, \mathbf{w}_{N_s}$

(0) Getting started: choosing $\mathbf{w}_1$
Turns out it doesn't matter: in theory, sample long enough and guaranteed to converge

Generating $\mathbf{w}_s$ takes 2 steps:

(1) Propose a new sample (based on previous)

$$\widetilde{\mathbf{w}}_s \leftarrow p(\widetilde{\mathbf{w}}_s|\mathbf{w}_{s-1}) \quad \text{proposal distribution}$$

Does **not** have to be related to target distribution!
Popular to use Gaussian centered on $\mathbf{w}_{s-1}$

$$p(\widetilde{\mathbf{w}}_s|\mathbf{w}_{s-1}, \boldsymbol{\Sigma}) = \mathcal{N}(\mathbf{w}_{s-1}, \boldsymbol{\Sigma})$$

(2) Test whether to accept or reject $\widetilde{\mathbf{w}}_s$

$$r = \frac{g(\widetilde{\mathbf{w}}_s; \mathbf{X}, \mathbf{t}, \sigma^2)}{g(\mathbf{w}_{s-1}; \mathbf{X}, \mathbf{t}, \sigma^2)} = \frac{p(\widetilde{\mathbf{w}}_s|\sigma^2)}{p(\mathbf{w}_{s-1}|\sigma^2)} \frac{p(\mathbf{t}|\widetilde{\mathbf{w}}_s, \mathbf{X})}{p(\mathbf{t}|\mathbf{w}_{s-1}, \mathbf{X})}$$

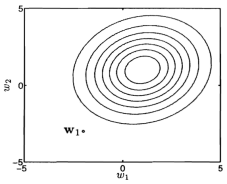
Don't need to calculate posteriors directly **because the Marginal Likelihoods cancel in the ratio!**

$r \geq 1$? If **yes**: always choose best: $\mathbf{w}_s = \widetilde{\mathbf{w}}_s$

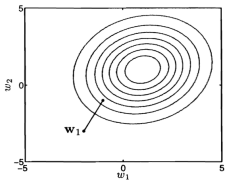
If **no**: possibly accept anyway

$u \leftarrow \mathcal{U}(0,1), \quad u \leq r$? yes / no → $\mathbf{w}_s = \mathbf{w}_{s-1}$

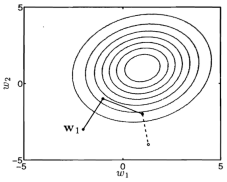
Metropolis-Hastings Example 1



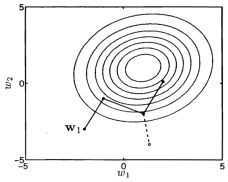
(a) Starting point



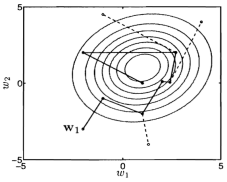
(b) After one sample



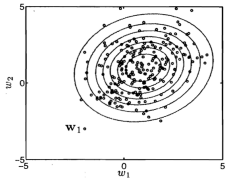
(c) After three samples. $\widetilde{\mathbf{w}}_3$ was accepted, $\widetilde{\mathbf{w}}_4$ rejected (dashed line)



(d) After four samples



(e) After 10 samples



(f) The first 300 samples

True mean and covariance

$$\boldsymbol{\mu} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} 3 & 0.4 \\ 0.4 & 3 \end{bmatrix}$$

Proposal density:

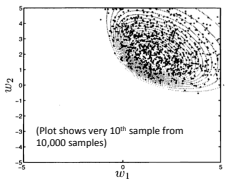
$$\mathcal{N}(\boldsymbol{\mu} = 0, \boldsymbol{\Sigma} = \mathbf{I})$$

Sample mean and covariance after 300 samples

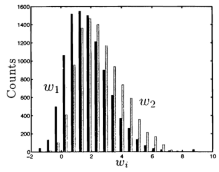
$$\boldsymbol{\mu}' = \frac{1}{N_s}\sum_{s=1}^{N_s}\mathbf{x}_s, \quad \mathbf{S}' = \frac{1}{N_s - 1}\sum_{s=1}^{N_s}(\mathbf{x}_s - \boldsymbol{\mu}')(\mathbf{x}_s - \boldsymbol{\mu}')^\top$$

$$\boldsymbol{\mu}' = \begin{bmatrix} 0.9770 \\ 1.0928 \end{bmatrix}, \quad \mathbf{S}' = \begin{bmatrix} 3.0777 & 0.4405 \\ 0.4405 & 2.8983 \end{bmatrix}$$

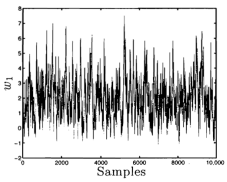
Metropolis-Hastings Example 2



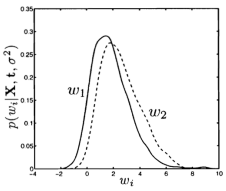
(a) 1000 of the MH samples along with the posterior contours



(b) Histograms of the samples for both w_1 (black) and w_2 (grey)



(c) All of the w_1 samples plotted against iteration, s



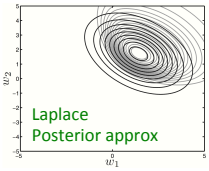
(d) Continuous densities fitted to the w_1 and w_2 samples

Burn-in

Convergence

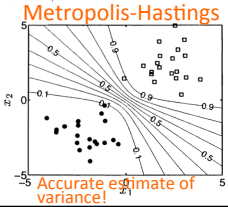
Estimating predictions from samples $\mathbf{w}_s$

$$P(T_{\text{new}} = 1|\mathbf{x}_{\text{new}}, \mathbf{X}, \mathbf{t}, \sigma^2) = \frac{1}{N_s}\sum_{s=1}^{N_s}\frac{1}{1 + \exp(-\mathbf{w}_s^{\mathrm{T}}\mathbf{x}_{\text{new}})}$$



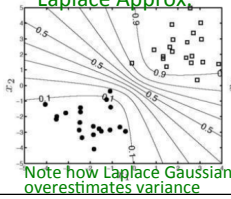
Laplace
Posterior approx

Metropolis-Hastings



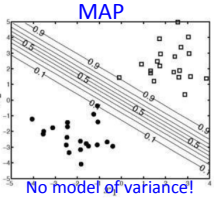
Accurate estimate of variance!

Laplace Approx



Note how Laplace Gaussian overestimates variance

MAP



No model of variance!