Bayesian Machine Learning

&

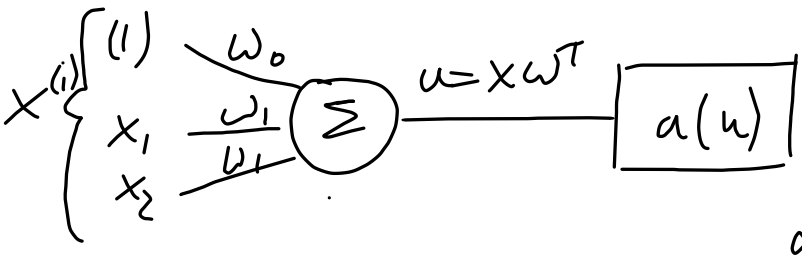
Its Application to Neural Networks

Questions

- Why Bayesian?
- Why do we need uncertainty estimates?
- What is overfitting? Where does it come from?

- What are we optimizing?
- Where does the Cost Function come from?
- How can we solve complicated Probability Distributions?

Motivating Example



$p(y|x,w) \Rightarrow 0.0, 0.1 \ldots 1.0$

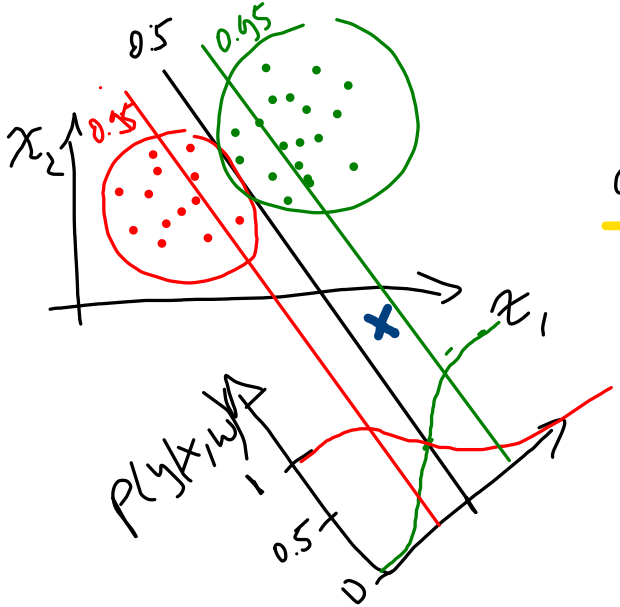
probabilities

$Q(w)$ - Loss Function

$\nabla_w Q(w)$ - Gradient

$w \leftarrow w - \alpha \nabla_w Q(w)$

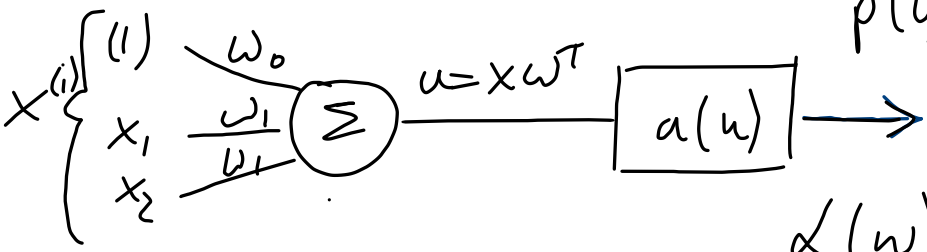
Gradient descend



Frequentist

Point Estimate

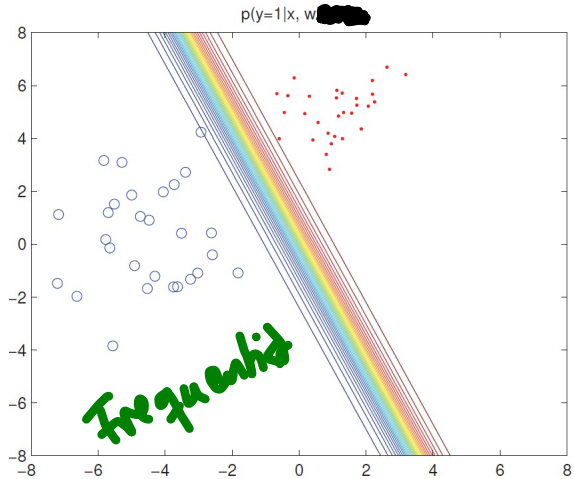
Motivating Example

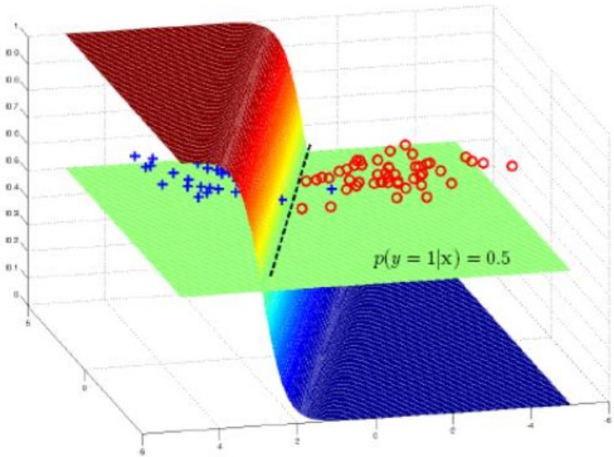


$p(y|x,w) \Rightarrow 0.0, 0.1 - 1.0$ probabilities

$\alpha(w)$ - Loss Function

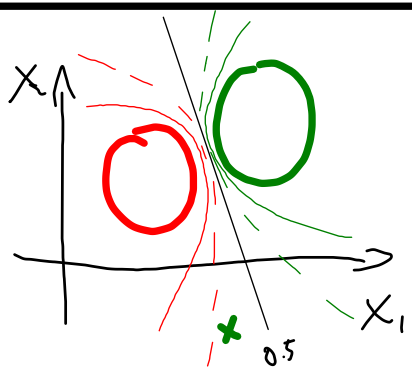
$\nabla_w \alpha(w)$ - Gradient



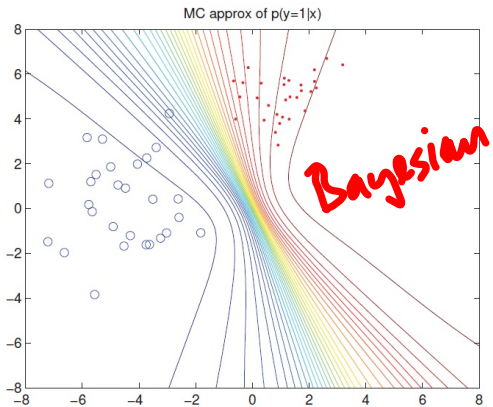


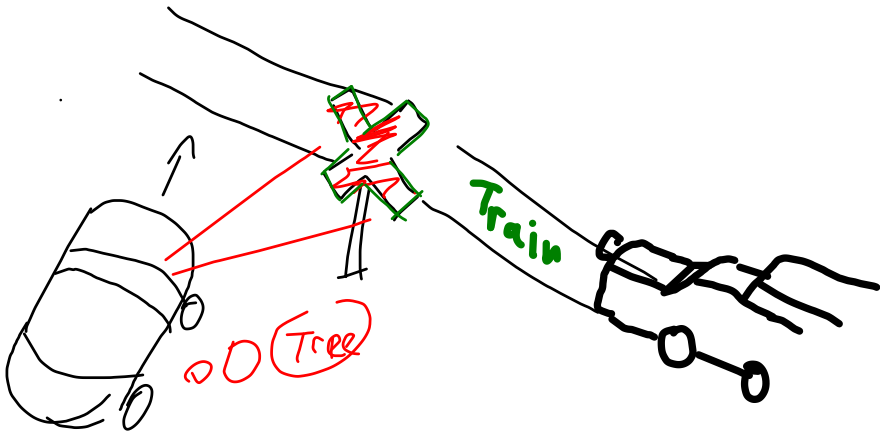
$\alpha(w)$
t descend

Uncertainty: Single Neuron → Bayesian



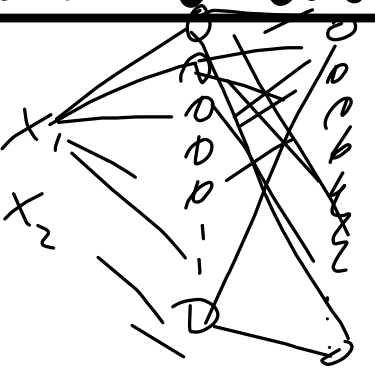
- Medical field
- Automotive

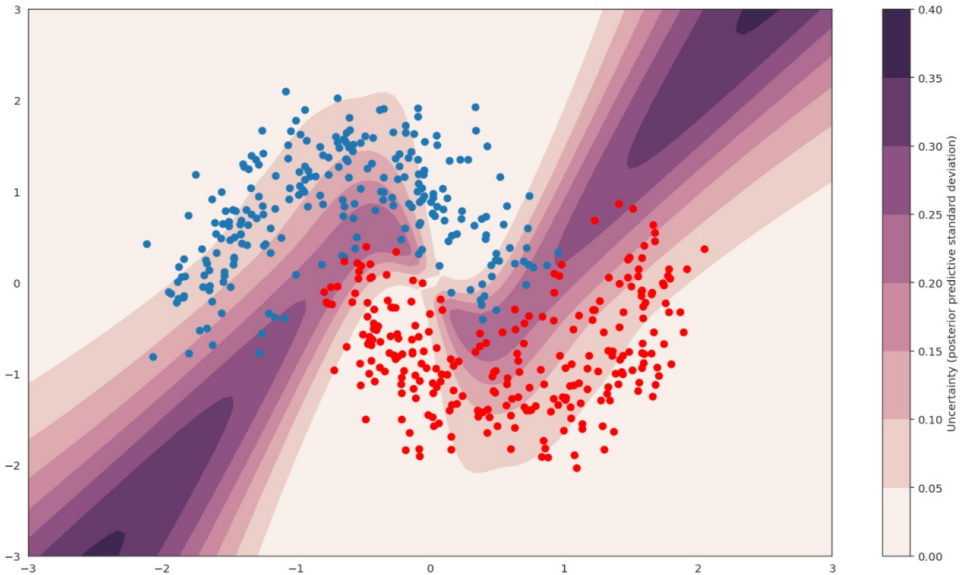




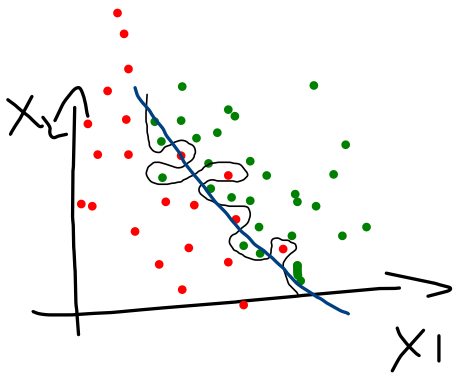
More Complex Data and Model

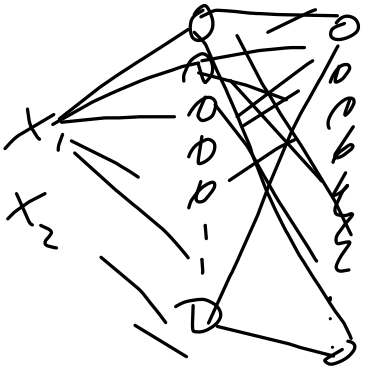
Neural Net





Over fitting



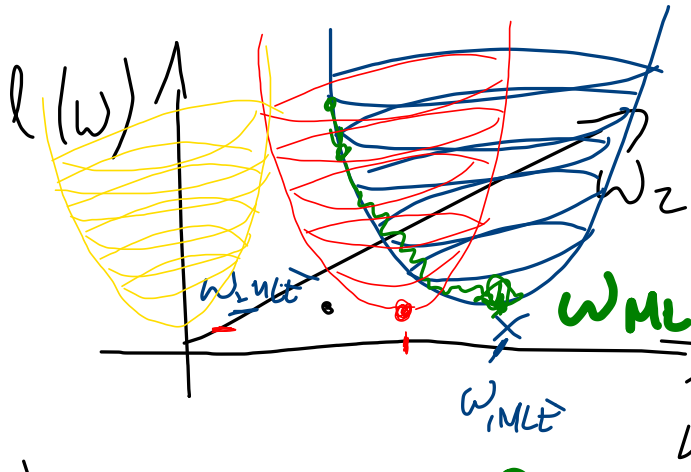


Optimazation World

$$\ell_r(\omega) = \ell(\omega) + \underbrace{\alpha\, r(\omega)}_{\text{Regularizer}}$$

Bayesian Point Estimate

Point Estimates vs. Probability Distribution over Parameters (weights)

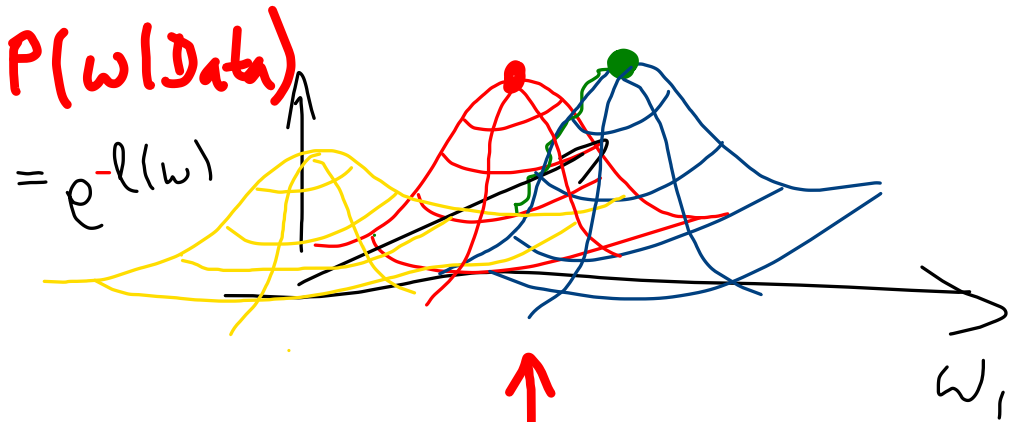


w_{MLE} ⇝ Frequentist Point Estimate

→ Bayesian Point Est.

Regularized Logistic Regression

Goal: Probability Distribution over parameters



We want the whole distribution over the weights,

instead of a point estimate

0

0

Bayesian Probability

Intuition

Biased Coin Flips with „Small Data"

bend coin

$$y = \{1, 1, 1, 1, 1, 1, 1, 1, 1, 1\} \quad (H\,H\,H\,H\,H\,H\,H\,H\,H\,H)$$

$n = 10$

probability of getting Heads

$$\theta_H = \frac{10}{10} = 1$$

Is this correct?

Can it be 100%?

Black Swan Paradox

Biased Coin Flips with „Small Data"

$y = \{1, 1, 1, 1, 1, 1, 1, 1, 1, 1\}$ $[H\ H\ H\ H\ H\ H\ H\ H\ H\ H\ H\ T]$ $\{1, 1, 0\}$

„Hallucinations"

probability of getting Heads

$$\theta_H = \frac{10+1}{10+2} = 0.833$$

Choose your „Hallucinations" according to your „PRIOR" knowledge!

Probabilistic Model of a Coin Flip

Over fitting with MLE

[H H H H H H H H H H]

$y = \{1, 1, 1, 1, 1, 1, 1, 1, 1, 1\}$

$$P(y=\{\dots\} \mid \theta) = \prod_{i=1}^{n} \text{Ber}(y^{(i)} \mid \theta) = \prod_{i=1}^{n} \theta^{y^{(i)}} \cdot (1-\theta)^{(1-y^{(i)})}$$

$$= \theta^{n_1} \cdot (1-\theta)^{n_0}$$

likelihood

$$\theta_{MLE} = \arg\max_{\theta} P(y \mid \theta)$$

Short cut:

$$\hat{\theta}_{MLE} = \frac{\#\ 1\text{'s}}{\#\ \text{flips}}$$

$$P(y \mid \theta = 0.3) = \prod_{i=1}^{10} 0.3^{1} = 0.3^{10}$$ ✗

$$P(y \mid \theta = 0.5) = 0.5^{10}$$ ✗

$$P(y \mid \theta = 1) = 1^{10} = 1$$ ✓ Max

Black Swan Paradox

Overfitting – Bayesian Fix

$y = \{1, 1, 1, 1, 1, 1, 1, 1, 1, 1\}$ (H H H H H H H H H H)

Hallucinations

$P(\theta \mid y)$

Posterior

Proportionality

$\propto \prod_{i=1}^{n} \text{Ber}(y^{(i)} \mid \theta) \cdot$ "?"

$\propto \theta^{n_1} \cdot (1-\theta)^{n_0} \cdot \theta^{\alpha-1} \cdot (1-\theta)^{\beta-1}$

$$\propto \theta^{n_1+\alpha-1} \cdot (1-\theta)^{n_0+\beta-1}$$

normalization constant $\frac{1}{z}$ is missing

$$\rightarrow \theta_{\text{Bayes}} = \frac{\#\,1\text{'s} + \alpha}{\#\,\text{flips} + \alpha + \beta}$$

$\alpha = 1, \beta = 1$

$$\theta_{\text{Bayes}} = \frac{10+1}{10+1+1} = \frac{10}{12} = 0.8\overline{3}$$

Beta-Bernoulli Model

$P(\theta|y)$ — Posterior

$$\propto \prod_{i=1}^{n} Ber(y^{(i)}|\theta) \cdot \text{"?"}$$

$$\propto \theta^{n_1} \cdot (1-\theta)^{n_2} \cdot \theta^{\alpha-1} \cdot (1-\theta)^{\beta-1}$$

$$\propto \theta^{n_1+\alpha-1} \cdot (1-\theta)^{n_0+\beta-1}$$

$$P(\theta|y) = Ber(y|\theta) \cdot Beta(\alpha,\beta)$$
$$= Beta(\alpha+n_1, \beta+n_0)$$
$$= Beta(\alpha', \beta')$$

} Conjugate Prior

Prior - Posterior

deterministic

$\theta_i \rightarrow \boxed{\arg\max_\theta Ber(y|\theta)} \rightarrow \theta_{MLE}$

dirac

$\theta_{MLE} = \delta(\theta_{MLE})$

$\theta \sim p(\theta)$

random

Prior:

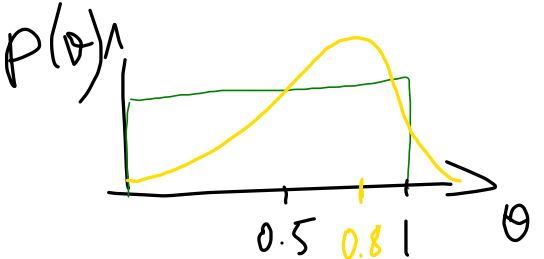
$\boxed{Ber(y|\theta)} - \boxed{\frac{1}{Z}}$

Likelihood **Norm.**

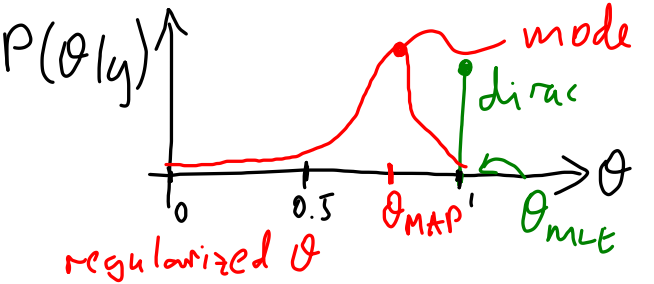
$\theta|y \sim p(\theta|y)$

Posterior

i.e. $Beta(\alpha, \beta)$



$Beta(\alpha+n_1, \beta+n_0)$



General Bayes Formula

(Model) likelihood

prior (Hallucinations)

$$p(\theta \mid y) = \frac{p(y \mid \theta)\, p(\theta)}{p(y)}$$

posterior over parameters (weights)

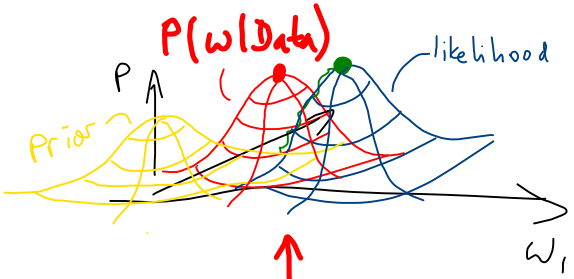
normalization constant

aka. prob. of data

→ makes it sum to one

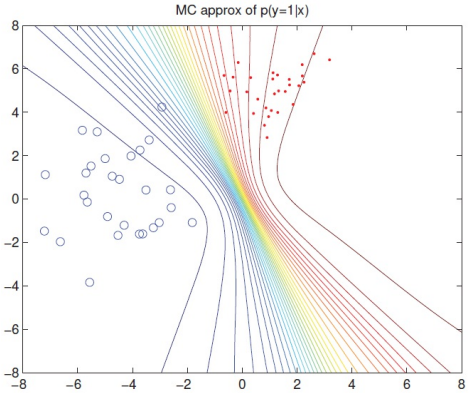
Remember?

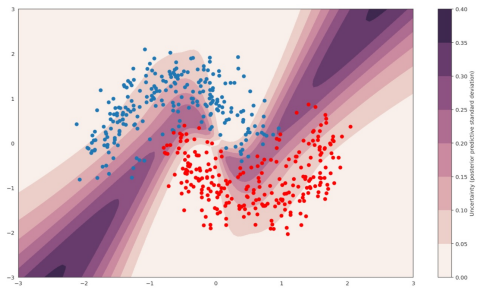
Goal: Probability Distribution over Parameters



We want the whole distribution over the weights,

instead of a point estimate





Easy Posterior vs. Nasty Posterior

Models with a Closed Form Solution aka. Conjugate Prior

Binary Data • Beta-Bernoulli

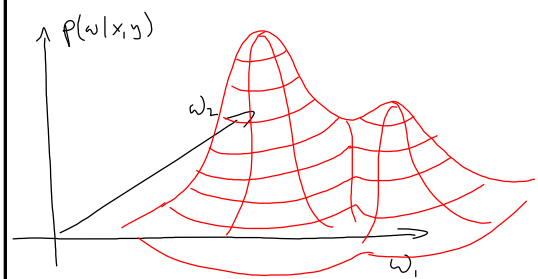
Multiclass Data • Dirichlet-Categorical

Linear Regression • Gaussian-Gaussian

⋮

Difficult/Intractable Posteriors

- Logistic Regression/ Single Neuron
- Neural Networks



How to solve Nasty Posterior Distributions

→ Approximate Methods:

- Variational Methods (i.e. VAE)
- Monte Carlo Methods

Part II

Connecting Bayes to the Optimization of a Single Neuron

General Bayes Formula – Coin Model

(Model) likelihood prior (Hallucinations)

$$p(\theta \mid y) = \frac{p(y \mid \theta)\, p(\theta)}{p(y)}$$

Posterior over parameters (weights)

normalization constant
aka. prob. of data

→ makes it sum to one

General Bayes Formula - Single Neuron

?

(Model) likelihood

?

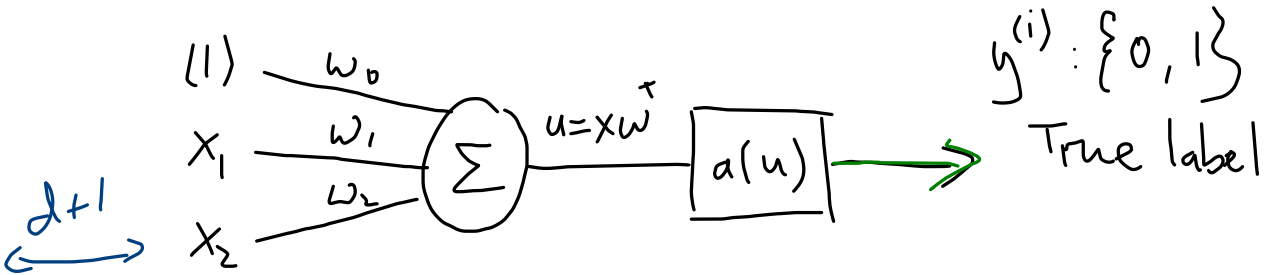
prior (Hallucinations)

$$p(\theta|x,y) = \frac{p(y|x,\theta)\,p(\theta)}{Z}$$

Also adding features x

Z: normalization const? integral of numerator

Recap: Single Neuron Optimization



$$X = \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} \\ \vdots & \vdots & \vdots \\ 1 & x_1^{(n)} & x_2^{(n)} \end{bmatrix} \quad n$$

$$w = \begin{bmatrix} w_0 & w_1 & w_2 \end{bmatrix}$$

$a(u) = \hat{y}$ Prediction

Learning as Optimization

Regularized Loss Function: $\ell_r(w) = \ell(w) + \alpha_1 r(w)$

Gradient: $\nabla_w \ell(w)$

Update: $w \leftarrow w - \alpha_2 \nabla_w \ell(w)$

From Regularized Loss Function to Bayes

$$\ell_r(w) = \ell(w) + \alpha r(w) \qquad | \exp(\ldots)$$

$$e^{-\ell_r(w)} = \boxed{e^{-\ell(w)} \cdot e^{-\alpha r(w)}}$$

(Model) likelihood prior (Hallucinations)

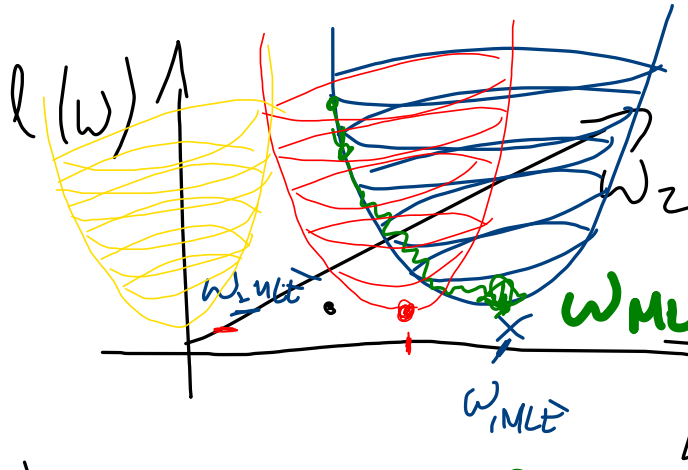
$$p(\theta | x, y) = \frac{p(y | x, \theta)\, p(\theta)}{Z}$$

Z ← normalization const.

integral of numerator

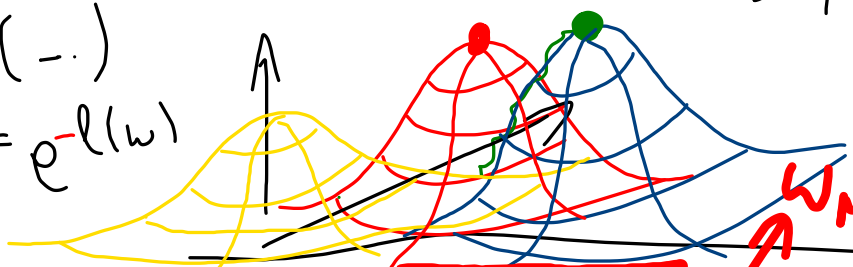
$$\int_{-\infty}^{\infty} e^{-\ell(w)} \cdot e^{-\alpha r(w)}\, dw$$

Remember this Connection?



w_{MLE} ⇝ Frequentist Point Estimate

$p(\cdot) = e^{-l(w)}$



w_{MAP} → Bayesian Point Est.

Regularizer

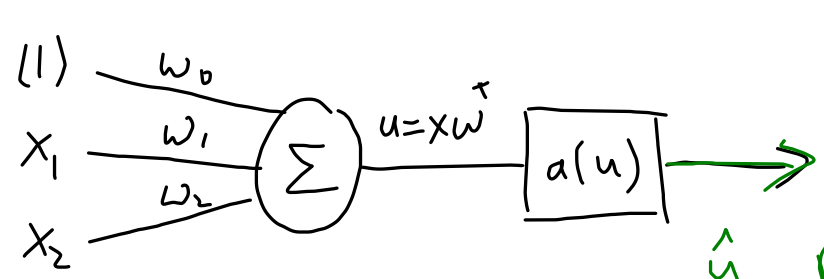
regularized probility

likelihood

Regularized Logistic Regression

Going Deeper into the Likelihood and the Prior of a Single Neuron

Mathematical Details: Single Neuron



$$X = \begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} \\ \vdots & \vdots & \vdots \\ 1 & x_1^{(n)} & x_2^{(n)} \end{bmatrix} \quad n$$

$$w = \begin{bmatrix} w_0 & w_1 & w_2 \end{bmatrix}$$

$$a(u) = \sigma(u) = \frac{1}{1 + e^{-xw^T}}$$

Learning as Optimization

$$\ell_r(w) = \ell(w) + \alpha r(w)$$

$$\ell_r(w) = -\sum_{i=1}^{n} \left[y^{(i)} \log\left(\hat{y}^{(i)}\right) + (1 - y^{(i)}) \log\left(1 - \hat{y}^{(i)}\right) \right] + \frac{\alpha}{2} w w^T$$

From Unregularized Loss to Likelihood

$$\ell(w) = \sum_{i=1}^{n} \left[y^{(i)} \cdot \log\left(\hat{y}^{(i)}\right) + (1-y^{(i)}) \log\left(1-\hat{y}^{(i)}\right) \right]$$

$$e^{-\ell(w)} = \prod_{i=1}^{n} \hat{y}^{y} \cdot (1-\hat{y})^{(1-y)}$$

Remember?

$$\text{Ber}(y|\theta) = \prod_{i=1}^{n} \theta^{y^{(i)}} \cdot (1-\theta)^{(1-y^{(i)})}$$

$$\underbrace{P(y|x,w)}_{\text{Likelihood}} = \prod_{i}^{n} \text{Ber}\left(y^{(i)}|\hat{y}^{(i)}\right)$$

$$= \prod_{i}^{n} \text{Ber}\left(y^{(i)}|\sigma(x^{(i)}w^{T})\right) = \prod_{i}^{n} \sigma(x^{(i)}w^{T})^{y^{(i)}} \cdot (1-\sigma(x^{(i)}w^{T}))^{(1-y^{(i)})}$$

Model of the NN

From Regularizer to Prior

$$\alpha r(w) = \frac{\alpha}{2} w w^T$$

$$e^{-\alpha r(w)} = e^{-\frac{\alpha}{2} w w^T}$$

Remember?

$$N(w \mid \mu=0, \sigma^2=\alpha^{-1}) = \frac{1}{Z} e^{-\frac{\alpha}{2} w \cdot w^T} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(w-\mu)(w-\mu)^T}$$

(μ and μ both $\to 0$)

$$P(\theta) = N(w \mid \mu=0, \sigma^2=\alpha^{-1})$$

Gaussian Prior

Posterior over the Weights

$$P(w \mid x, y) = \frac{\prod_{i}^{n} \mathrm{Ber}(y^{(i)} \mid \sigma(x^{(i)} w^{T}))\, \mathcal{N}(\mu_w, \sigma_w)}{P(y)}$$

w

D

Bayesian Prediction

$$P(w \mid \underbrace{x, y}_{D}) = \frac{\prod_{i}^{n} \text{Ber}(y^{(i)} \mid \sigma(x^{(i)} w^{T}))\, N(\mu_w, \sigma_w)}{P(y)}$$

$$P(y^{*}=1 \mid x^{*}, D) = \int P(y=1 \mid x^{*}, w) \cdot P(w \mid D)\, dw$$

↑ new input

$$P(y=1 \mid x^{*}, w) = \sigma(x^{*} w^{T})$$
$$P(y=0 \mid x^{*}, w) = 1 - \sigma(x^{*} w^{T})$$

So far so good...

What's the problem now?

Nasty Likelihood Dist. and Intractable Integrals

$$P(\omega \mid x, y) = \frac{\prod_{i}^{n} P(y^{(i)} \mid x, \omega) \cdot P(\omega)}{P(y)}$$

$$P(y) = \int_{-\infty}^{\infty} \prod_{i}^{n} P(y^{(i)} \mid x, \omega) \cdot P(\omega)\, d\omega$$

Problem 1

Nasty Likelihood

Problem 2

Intractable Integrals

$$P(y^{*}=1 \mid x^{*}, D) = \int P(y=1 \mid x^{*}, \omega) \cdot P(\omega \mid D)\, d\omega$$

How to solve an intractable integral?

From Integral to Sample Mean

$$P(y^*=1 \mid \mathcal{D}) = \int P(y=1 \mid x^*, w) \cdot P(w \mid \mathcal{D}) \, dw$$

$$= \mathop{E}_{w \sim P(w \mid \mathcal{D})} \left[y(x^*, w) \right]$$

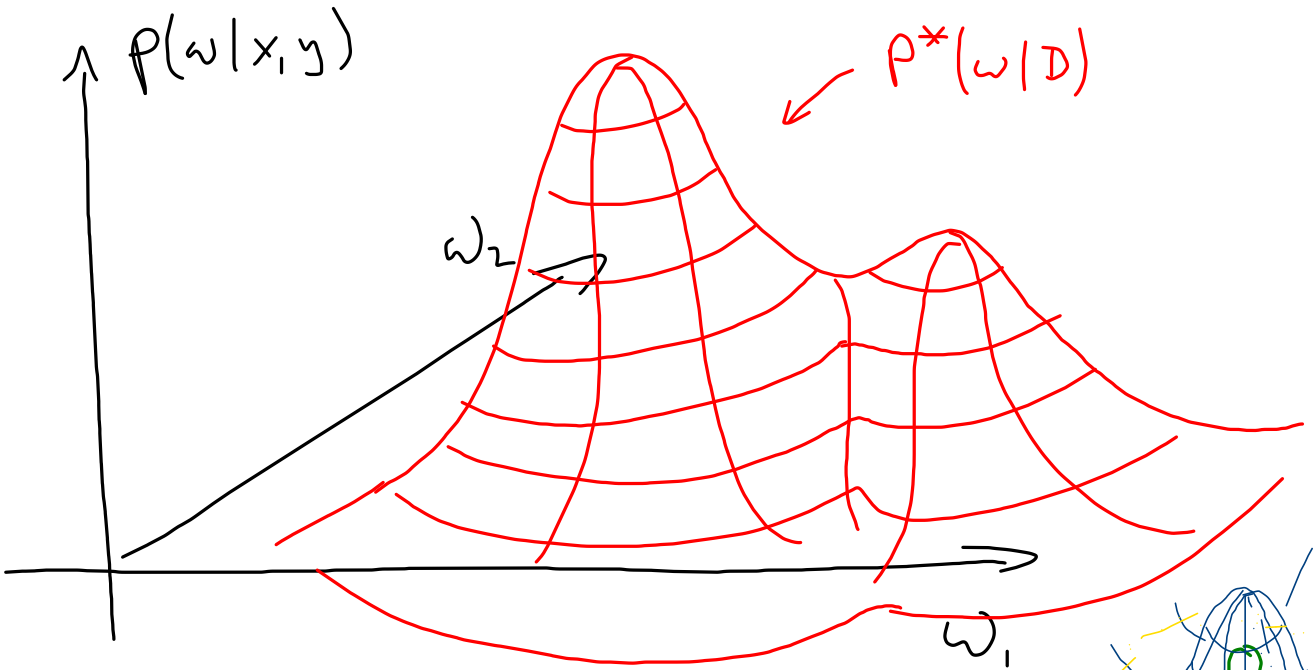
$$\approx \frac{1}{n} \sum_{i=1}^{n} y(x^*, w_i)$$

↑ given

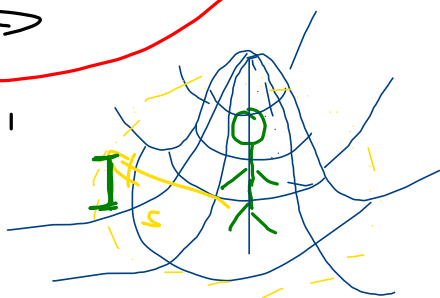
where $w_i \sim P(w \mid \mathcal{D})$

↑ needed

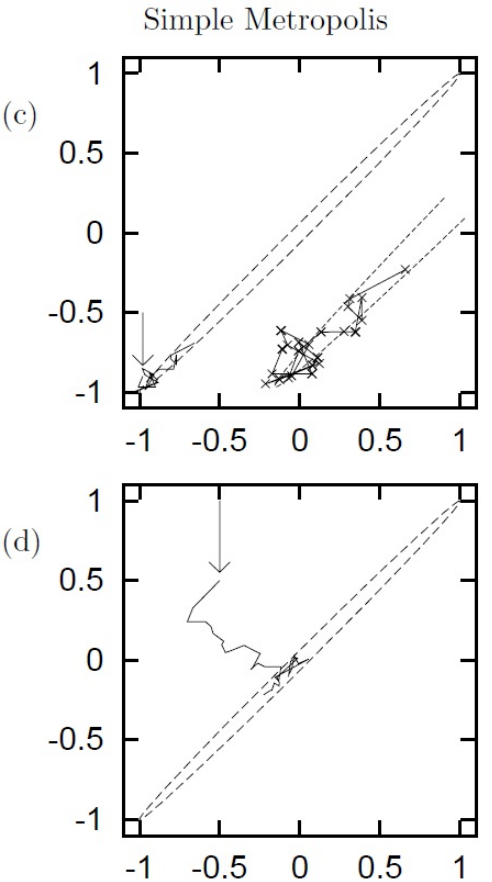
Sampling from the Posterior: MCMC



$$P(\omega|D) = \frac{p^*(\omega|D)}{Z}$$



Random Walk Visualization

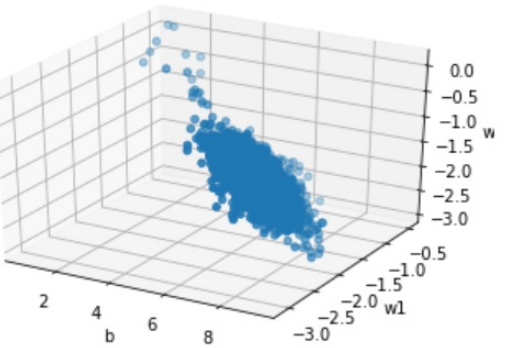


Python Code: Weight Samples

```python
def MCMC(function, X, Y, N, d, sigma=0.1):
    mu0 = [0]*d
    cov = np.diag([sigma]*d)
    w_old = np.random.multivariate_normal(mu0, cov)
    sample_list = []
    for i in range(N):
        w_new = np.random.multivariate_normal(w_old, cov)
        a = np.exp(function(X, Y, w_new) - function(X, Y, w_old))
        u = np.random.uniform(0, 1)
        if (a > u):
            sample_list.append(w_new)
            w_old = w_new
    return np.array(sample_list)
```

$\propto p^*(w_{new}|D)$

$\propto p^*(w_{old}|D)$




```python
def sigmoid_layer(X, W):
    x_biases = np.full((X.shape[0],1), 1)
    X_ = np.concatenate((x_biases, X), axis=1)
    odds = np.exp(-np.dot(X_, W.T))
    y_hat = 1/(1 + odds)
    return y_hat
def log_posterior_over_parameters_logreg(X, Y, W):
    sigma_prior = 1.5
    log_prior = (-np.dot(W, W.T)/(2*sigma_prior**2))
    y_hat = sigmoid_layer(X, W)
    full_log_likelihood = (Y.T)*np.log(y_hat) + (1 - Y.T)*np.log(1-y_hat)
    post = log_prior + np.sum(full_log_likelihood)
    return post

W_samples = (MCMC(log_posterior_over_parameters_logreg, X, Y, N=10000, d=3))
```

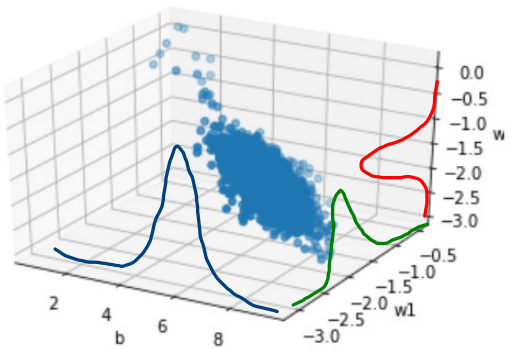
$\log\left(e^{-\frac{1}{2\sigma^2}ww^T}\right)$

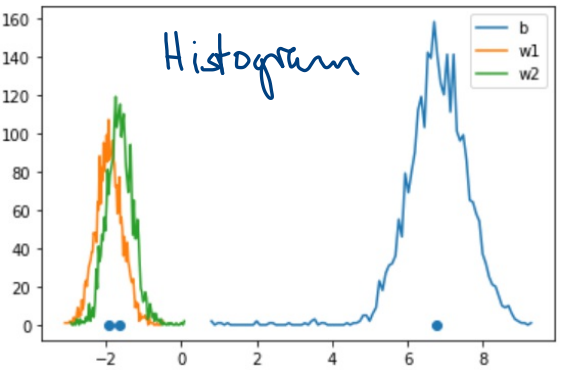
$$\hat{y} = \frac{1}{1 + e^{-xw^T}}$$

xw = odds

$$\log p^*(w|D) = \sum \log\left[Ber(y|\hat{y})\right] + \log\left[N^*(w|\mu,\sigma)\right] = \text{Loss function}$$

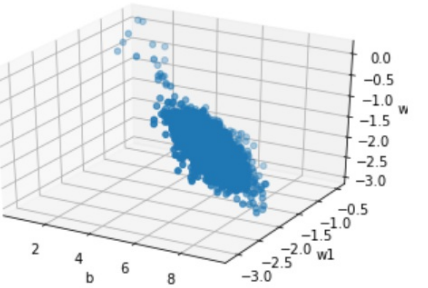
Visualization as Histogram





Python Code: Prediction

$$P(y^*|D) = \frac{1}{n}\sum_{i=1}^{n} \sigma(x^* w^{(i)})$$




```
def predict_MC(X, W_samples):
    return (np.nansum(sigmoid_layer(X, W_samples), axis=1)/W_samples.shape[0])
```

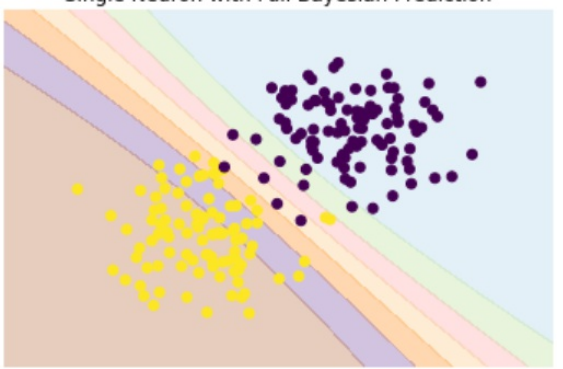
$$w_{MAP} = \frac{1}{N}\sum_{i=1}^{n} w^{(i)}$$

$$\hat{y} = \sigma(x\, w_{MAP}^{T})$$

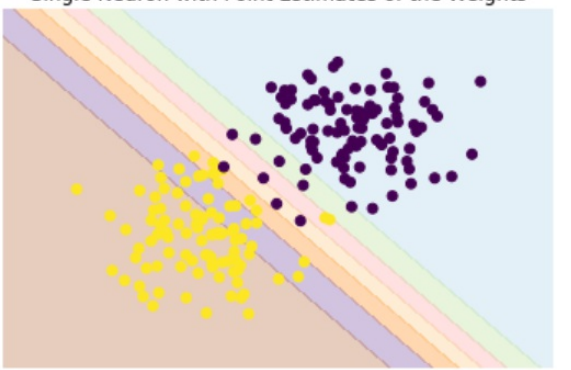
```
y_pred_full = predict_MC(X_pred, W_samples)
```

```
# POINT ESTIMATE
W_map = np.median(W_samples, axis=0).reshape(1, -1)
y_pred_map = sigmoid_layer(X_pred, W_map)
```

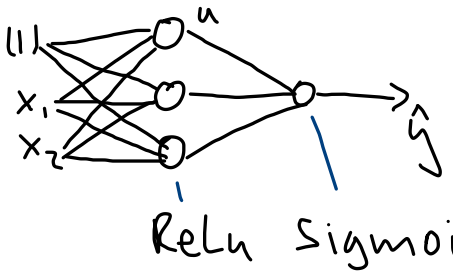
Single Neuron with Full Bayesian Prediction



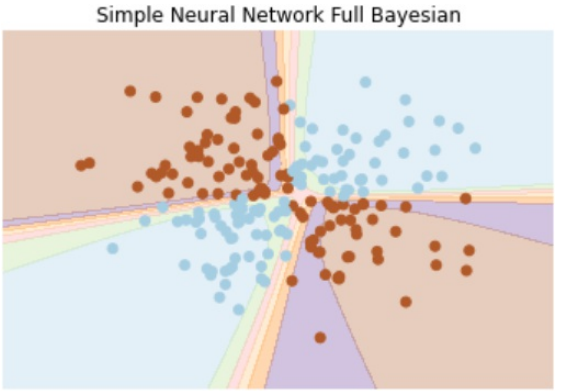
Single Neuron with Point Estimates of the Weights

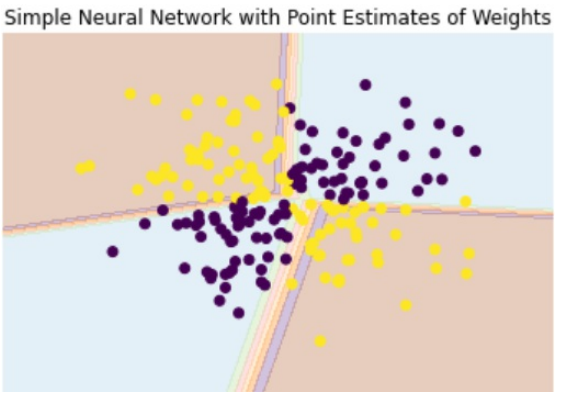


XOR Data with Simple Neural Network



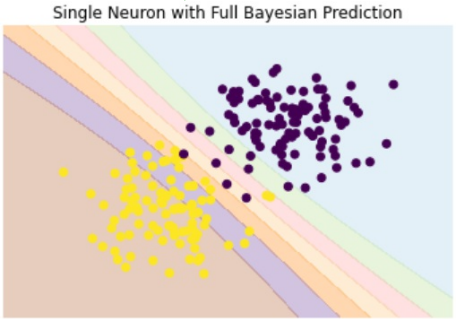
$$\hat{y} = \sigma\left(\max(0, x w_1^T) \cdot w_2^T\right)$$

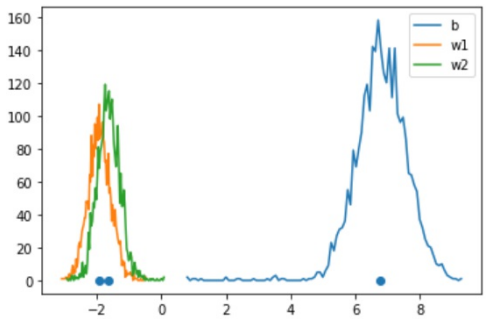




Dependence of Uncertainty on Sample Size

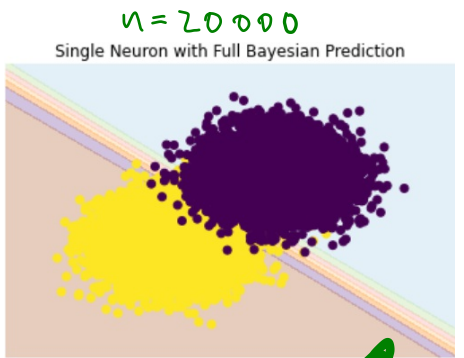
$n = 200$

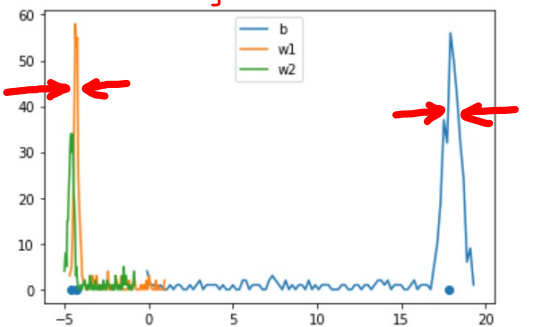




Uncertainty shrinks if $n \to \infty$

$n = 20000$

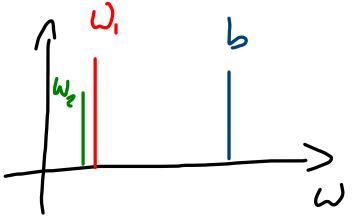




Looks like w_{MLE}

$$\lim_{n \to \infty} w_{MLE} = w_{MAP}$$

$$\lim_{n \to \infty} \sigma_w = 0$$



Disadvantages of Metropolis Algorithm

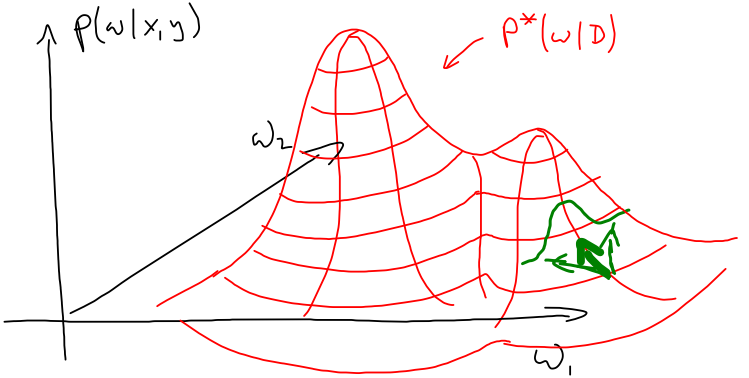
- Many rejected samples
- Can wander off in many directions

} computationally expensive → slow

Future Work

→ Hamiltonian Monte Carlo ("Hybrid")

Optimization and Random Walk Hybrid



→ Speed

Take Away

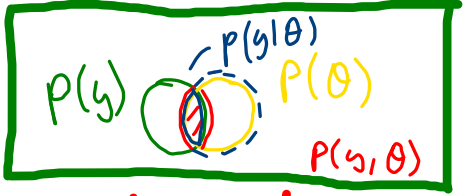
- Black Swan Paradox
- Loss $\xrightarrow{e^{-\text{loss}}}$ Probability, Probability $\xrightarrow{-\log(p)}$ Loss
- $\int f(x) \cdot p(x)\, dx = \underset{x \sim p(x)}{E}[f(x)] \approx \frac{1}{n} \sum_{i}^{n} f(x^{(i)})$
- $\lim_{n \to \infty} w_{MLE} = w_{MAP} \Rightarrow \lim_{n \to \infty} \sigma_w = 0$

Understanding Numerator and Denominator of Bayes

(Optional?)

Rules of Probability

$P(y)$



Creates a joint probability

Product Rule $P(y, \theta) = p(y|\theta)\, p(\theta)$

$$p(\theta|y) = \frac{p(y|\theta)\, p(\theta)}{p(y)} = \frac{P(y, \theta)}{P(y)}$$

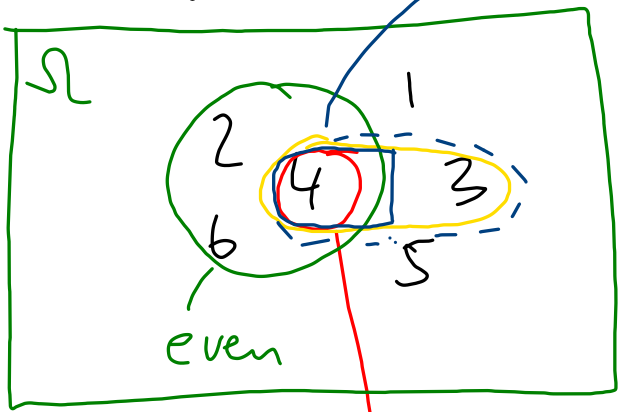
All possible y's

$P(\Omega) \rightarrow P(y)$

The reference system switches from $\Omega \rightarrow y$

Example

Die



$p(\text{even} \mid d12) = \frac{1}{2}$

divides of 12 "d12"

$p(d12) = \frac{1}{3}$

$p(\text{even}) = \frac{1}{2}$

both $P(\text{even}, d12) = \frac{1}{6}$ ← #4 ← $|\Omega|$

$$P(d12 \mid \text{even}) = \frac{P(\text{even}, d12)}{P(\text{even})} = \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2}} = \frac{1}{3}$$

$$P(\text{even} \mid d12) \cdot P(d12) + P(\text{even} \mid \text{not } d12) \cdot P(\text{not } d12)$$