# Bayesianska numeriska metoder I

T. Olofsson

## Marginalisering

En återkommende teknik inom Bayesiansk inferens är det som kallas för marginalisering. I grund och botten rör det sig om tillämpning av ett specialfall av *summaregeln*. För uteslutande utsagor $A$ och $B$ under bakgrundsinformation $C$ gäller som bekant att $P(A,B|C)=0$ och vi har då att

$$P(A+B|C)=P(A|C)+P(B|C)-P(A,B|C)=P(A|C)+P(B|C) \tag{1}$$

Marginalisering används då vi har att göra med sannolikheter för utsagor som kan kan brytas ner i ett antal, sinsemellan uteslutande delutsagor. Ofta uppträder sådana delutsagor i modeller med s.k. nuissance-parametrar ("besvärliga parametrar" eller "okynnesparametrar"), dvs modellparametrar vars värden är både okända och ointressanta för oss men som likväl påverkar utsignalen från modellen och måste hanteras på något sätt.

Betrakta till exempel sannolikheten att en begagnad bil från 1990 och som kostat 20000 kr kommer att klara sig genom nästa besiktning. En nuissance-parameter kan i detta fall vara bilmärket. Låt $I$ stå för bakgrundsinformationen "allt vi vet om bilar från 1990 som kostar 20000 kr" och $B_1$ till $B_M$ stå för utsagor om märket: $B_1$ ="märket är Alfa Romeo", $B_2$ ="märket är Aston Martin", osv till $B_M$ ="märket är Volvo". Vi antar att vi täckt in alla märken på marknaden. Utsagan $A$ ="Bilen klarar nästa besiktning", kan då brytas upp i uteslutande delutsagor $A=AB_1+AB_2+...+AB_M$ och sannolikheten $P(A)$ ges via summaregeln av

$$P(A|I)=P(AB_1|I)+...+P(AB_M|I)=P(A|B_1I)P(B_1|I)+...+P(A|B_MI)P(B_M|I) \tag{2}$$

Den sista likheten fick vi mha produktregeln. Vi har här brutit ner en sannolikhet i några beståndsdelar som troligtvis är enklare att hantera än originalproblemet.

### Marginalisering i samband med kontinuerliga fördelningar

#### Exempel 1: Parameterestimering

Marginalisering uppträder ofta naturligt i samband med parameterestimering. Anta tex att vi vet (via tex något fysikaliskt resonemang) att en serie uppmätta data $\mathcal{X}=\{x_1,...,x_N\}$ respektive $\mathcal{Y}=\{y_1,...,y_N\}$ kan väl beskrivas av en modell

$$y_n=w_1x_n^{w_0}+e_n \tag{3}$$

där vi antar $e_n\sim N(0,\sigma^2)$ (samt att alla brustermer är oberende) och där parameterarna $w_0$ och $w_1$ båda ligger i intervallet $[0,2]$. Låt oss anta här att faktorn $w_1$ inte har så mycket med något intressant fysikaliska samband att göra (en nuissance-parameter). Den kan tex beskriva en okänd förstärkning i vår mätutrustning. Den intressanta fysikaliska parametern för vår del är $w_0$.

Vad kan vi nu med hjälp av våra data säga om $w_0$, eller med andra ord, vad är $p(w_0|\mathcal{X},\mathcal{Y})$? För att ta reda på det kan vi här välja att först beräkna $p(w_0,w_1|\mathcal{X},\mathcal{Y})$ vilket som vanligt ger ett mått på sannolikheten att den sanna parameterkombinationen ligger i ett visst litet intervall kring punkten $(w_0,w_1)$. Att det sanna värdet på $w_1$ samtidigt skulle kunna anta två olika värden, säg $w_1^i$ och $w_1^j$, är förstås uteslutet vilket i sin tur innebär att parameterkombinationerna $(w_0,w_1^i)$ och $(w_0,w_1^j)$ också är uteslutande utsagor. Därför kan vi applicera summaregeln i stil med exemplet ovan fast med skillnaden att summan övergår i en integral eftersom nuissance-parametern $w_1$ här är en kontinuerlig variabel. Alltså, $p(w_0|\mathcal{X},\mathcal{Y})$ ges av

$$p(w_0|\mathcal{X},\mathcal{Y})=\int p(w_0,w_1|\mathcal{X},\mathcal{Y})dw_1. \tag{4}$$

Det återstår att räkna ut $p(w_0, w_1|\mathcal{X}, \mathcal{Y})$. Via Bayes sats får vi

$$p(w_0, w_1|\mathcal{X}, \mathcal{Y}) = \frac{p(\mathcal{Y}|w_0, w_1, \mathcal{X})p(w_1, w_2|\mathcal{X})}{p(\mathcal{Y}|\mathcal{X})}. \tag{5}$$

Vår bakgrundsinformation om parametrarna säger att de båda ligger i intervallet $[0, 2]$ och utan någon ytterligare information om eventuella samband mellan parameterna så bör vi anta att de är oberoende. Detta innebär att de är *a priori* likformigt fördelade i kvadraten med hörn i $(0, 0)$ och $(2, 2)$ vilket ger $p(w_0, w_1) = 1/4$, dvs en konstant. Nämnaren $p(\mathcal{Y}|\mathcal{X})$ är som vanligt en normeringskonstant så den enda faktorn som beror av $w_0$ och $w_1$ är alltså $p(\mathcal{Y}|w_0, w_1, \mathcal{X})$. Om vi kan räkna ut denna faktor som funktion av $w_0$ och $w_1$ så räcker det med att normera denna funktion så har vi vår a-posteriori PDF.

I vårt fall får vi att

$$p(w_0, w_1|\mathcal{X}, \mathcal{Y}) \propto p(\mathcal{Y}|w_0, w_1, \mathcal{X}) = \prod_{n=1}^{N} p(y_n|w_0, w_1, x_n) = \frac{1}{(2\pi)^{N/2}\sigma^N} e^{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - w_1 x_n^{w_0})^2} \tag{6}$$

Låt oss via en simulering se hur vår uppfattning om värdena på $w_0$ och $w_1$ ändras i takt med att vi mäter upp fler och fler datapunkter samt, som en konsekvens, hur $p(w_0|\mathcal{X}, \mathcal{Y})$ ändras. I figurerna 1-3 nedan visas resultaten för ett antal simuleringar med olika antal data, $N$. Variansen för $e_n$ är fixerad till, $\sigma^2 = 0.01$. I simuleringen har vi satt de sanna parametervärdena till $w_0 = 1.5$ och $w_1 = 0.5$. I figur 1 visas ett antal mätvärdespar (samt kurvan $y = 0.5x^{1.5}$) och i figur 2 visas den betingade PDF:en $p(w_0, w_1|\mathcal{X}, \mathcal{Y})$. Slutligen visas i figur 3 den betingade PDF:en $p(w_0|\mathcal{X}, \mathcal{Y})$ som erhålls via ekv. (4) (integrationen utförs numeriskt)

**Exempel 2: Prediktion av utsignal från modell**

En annan viktig situation där vi kan ha nytta av marginalisering är då vi vill utnyttja en modell med osäkra modellparametrar för att prediktera utsignalen, $y$, för en given insignal, $x$. Låt oss igen betrakta modellen i ekv. (3). Anta att vi, precis som ovan, har haft tillgång till uppmätta data $\mathcal{X}$ och $\mathcal{Y}$ för att uppskatta vilka värden parametrarna $w_0$ och $w_1$ har. I de fall då mängden data är stor så brukar det utkristallisera sig en kombination $\hat{w}_0, \hat{w}_1$ (MAP-skattningen) som totalt dominerar PDF:en $p(w_0, w_1|\mathcal{X}, \mathcal{Y})$. Det naturliga i ett sånt fall är förstås att anta att $\hat{w}_0, \hat{w}_1$ är de helt korrekta värdena och vi sätter därför in dessa värden direkt i modellen.[1] Vi kan i detta fall modellera ett hittills osett $y$ som

$$y = \hat{w}_1 x^{\hat{w}_0} + e. \tag{7}$$

Notera att vi tar med brustermen $e$ och att vår osäkerhet om $y$ därmed beskrivs med en betingad PDF

$$p(y|x, \hat{w}_0, \hat{w}_1) = \frac{1}{(2\pi)^{1/2}\sigma} e^{-\frac{1}{2\sigma^2}(y - \hat{w}_1 x^{\hat{w}_0})^2}. \tag{8}$$

(Vi har helt enkelt löst ut $e$ och satt in detta i PDF:en för $e$ som enligt antagandet är normalfördelad.)

Problem uppstår då $N$ är litet vilket gör $p(w_0, w_1|\mathcal{X}, \mathcal{Y})$ mer utspridd. Vi bör i så fall ta med vår osäkerhet om de korrekta parametervärdena vid beräkningen av $p(y|x, \mathcal{X}, \mathcal{Y})$. Vi kan skriva denna PDF som

$$p(y|x, \mathcal{X}, \mathcal{Y}) = \int p(y, w_0, w_1|x, \mathcal{X}, \mathcal{Y}) dw_0 dw_1 = \{ \text{ produktregeln} \} = \int p(y|x, w_0, w_1) p(w_0, w_1|\mathcal{X}, \mathcal{Y}) dw_0 dw_1 \tag{9}$$

Notera att vi i den sista integralen har strukit de faktorer som spelat ut sin roll i de betingade fördelningarna. Till exempel har vi i $p(y|x, w_0, w_1, \mathcal{X}, \mathcal{Y})$ ingen användning av informationen i $\mathcal{X}$ och $\mathcal{Y}$ eftersom vi ändå dessutom betraktar fixerade värden på $w_0$ och $w_1$. På liknande sätt så påverkar kännedom om värdet på variabeln $x$ inte vår osäkerhet om värdena på $w_0$ och $w_1$.

Den sista integralen har en enkel intuitiv tolkning. Den totala osäkerheten om $y$ får vi som ett viktat medelvärde av de osäkerheter vi har för olika modellparametrar. Vikten ges av sannolikheten för att parametervärdena är de korrekta.

[1] Ett sådant angreppssätt brukar kallas "plug-in estimate".

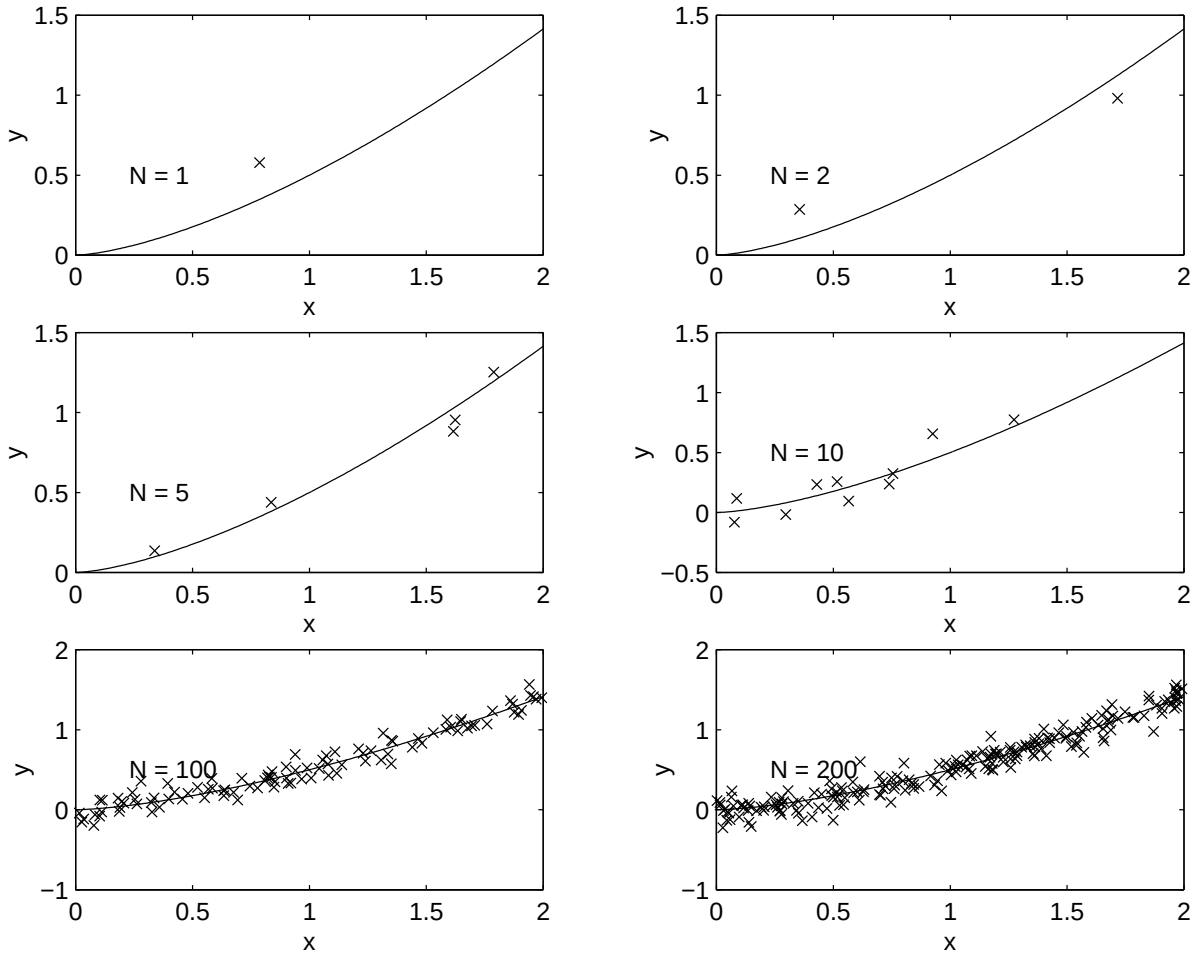


Figure 1: Simulerade data (kryss) från modellen $y = 0.5x^{1.5} + e$. Den heldragna kurvan är $g(x) = 0.5x^{1.5}$. De sanna värdena på $w_0$ och $w_1$ är alltså 1.5 respektive 0.5.

Ett specialfall har vi då träningsdata pekar entydigt på endast en tänkbar parameterkombination, dvs all sannolikhetsmassa är samlad i en punkt. Uttryckt i PDF:er blir det att $p(w_0, w_1|\mathcal{X}, \mathcal{Y}) = \delta(w_0 - \hat{w}_0, w_1 - \hat{w}_1)$, dvs en s.k deltafunktion[2] med centrum i $(\hat{w}_0, \hat{w}_1)$. Integralen ovan ges då av

$$p(y|x, \mathcal{X}, \mathcal{Y}) = \int p(y|x, w_0, w_1, \mathcal{X}, \mathcal{Y})\delta((w_0 - \hat{w}_0, w_1 - \hat{w}_1)dw_0 dw_1 = p(y|x, \hat{w}_0, \hat{w}_1) \tag{10}$$

vilket var exakt det vi kom fram till på intuition tidigare.

## Monte–Carlo-simuleringar

Betrakta återigen modellprediktionsexemplet i förra avsnittet. Vi kom fram till att för att modellera $y$ på ett bra sätt så borde vi beräkna integralen i ekv. (10). Vi kan förstås tänka oss att vi har en stor mängd modellparametrar som vi vill integrera över. Två problem som båda blir allt allvarligare ju fler dimensioner vi ska integrera över är följande:

1. Integraler över flera variabler är ytterst sällan analytiskt lösbara.

[2] egentligen Diracs deltafunktion.

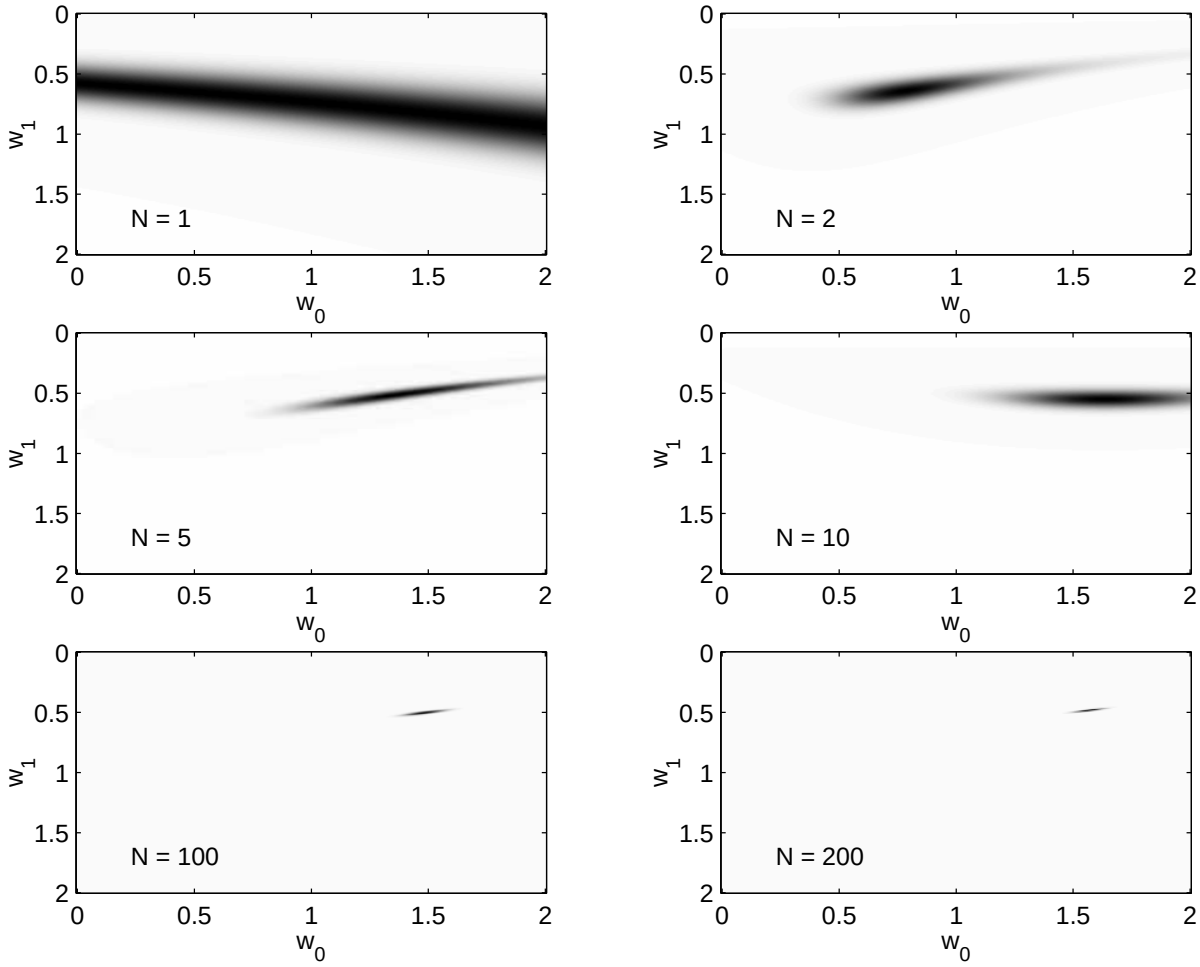


Figure 2: Den betingade fördelningen $p(w_0, w_1|\mathcal{X}, \mathcal{Y})$ illustrerat som gråskalebilder. Höga värden motsvaras av svart och låga värden av vitt. Notera att färgskalan är normerad så att maxvärdet i varje bild är maximalt svart. Det högsta värdet för de mer isolerade fördelningarna (tex de två sista) är i själva verket mycket högre än för de vida fördelningarna (tex de fyra första).

2. Numerisk lösning (tex mha Simpsons formel) av integraler över flera variabler brukar vara mycket beräkningskrävande.

I de fall där integralerna erhållits i samband med sannolikhetsberäkningar brukar de dock ofta kunna approximeras med s.k. Monte–Carlo-simuleringar. Om vi i exemplet ovan vill beräkna det betingade väntevärdet $\hat{y}(x) = E[y|x, \mathcal{X}, \mathcal{Y}] = \int yp(y|x, w_0, w_1)p(w_0, w_1|\mathcal{X}, \mathcal{Y})dw_0dw_1$, så skulle en Monte–Carlo-simulering kunna gå till enligt följande:

1. "Dra" en parameterkombination $(w_0^i, w_1^i)$ slumpmässigt ur a posteriori-fördelningen $p(w_0, w_1|\mathcal{X}, \mathcal{Y})$.
2. För ett givet $x$, generera ett slumpmässigt $y$ med hjälp av modellen i ekv. (7)
3. Upprepa från punkt 1

Med detta schema kan vi, för ett fixt $x$, generera ett antal olika $y$-värden. Fördelningen (histogrammet) för dessa värden kommer att asymptotiskt konvergera mot $p(y|x, \mathcal{X}, \mathcal{Y})$. Väntevärdet $\hat{y}(x)$ fås till sist genom att ta medelvärdet av de dragna $y$-samplen. Självklart kan vi med hjälp av dessa sampel beräkna väntevärden av i stort sett godtyckliga funktioner av $y$.

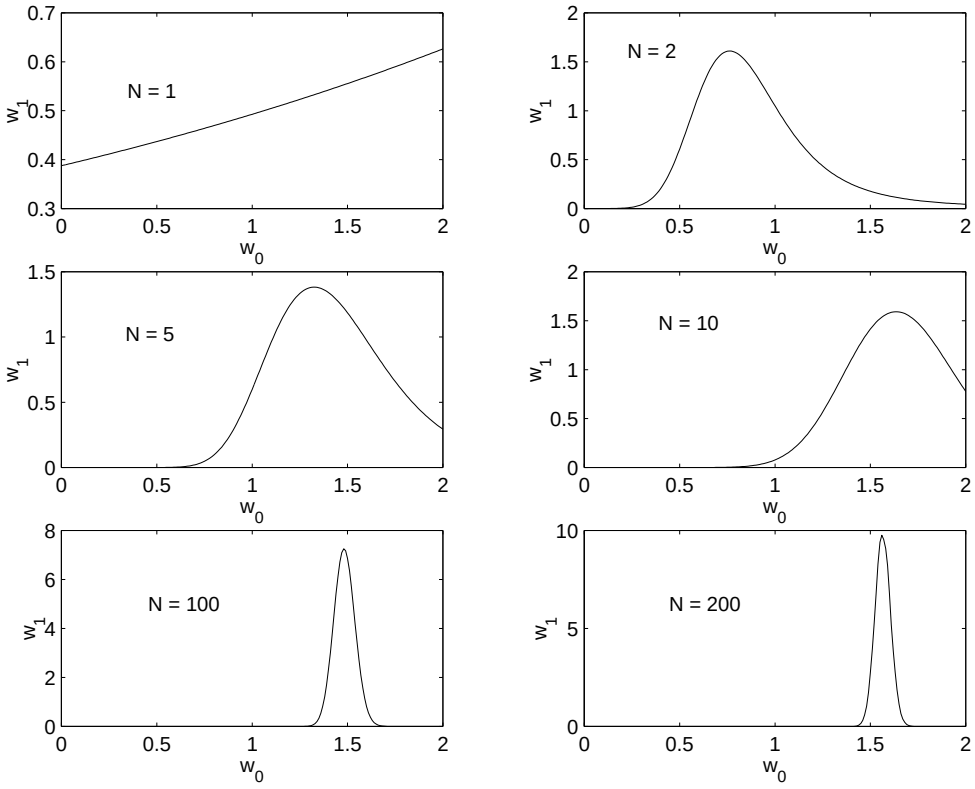


Figure 3: Fördelningen $p(w_0|\mathcal{X},\mathcal{Y}) = \int p(w_0,w_1|\mathcal{X},\mathcal{Y})dw_1$ uträknad via direkt summering över alla $w_1$-värden för fixa $w_0$. Detta följs sedan av normering.

Alltså, om vi kan hitta en metod att dra sampel ur en fördelning med någon a posteriori PDF (i det här fallet $p(\mathbf{w}|\mathcal{X},\mathcal{Y})$ ) så kan vi lösa integraler över denna PDF numeriskt. Tyvärr så är det inte alltid helt lätt att generera sampel ur en godtycklig flerdimensionell fördelning. Däremot finns det relativt allmängiltiga och någorlunda enkla och effektiva metoder för att dra sampel ur *endimensionella* fördelningar. Bara för att ni ska få en första inblick i hur sådana metoder kan fungera går vi nedan snabbt igenom den s.k. "acceptance/rejection"-metoden.

Anta att vi vill dra ett sampel ur en fördelning, $p(w)$, med PDF *proportionell mot* funktionen $f(w)$, dvs

$$p(w) \propto f(w) = \begin{cases} (4-(2w)^2) & |x|<1 \\ 0 & \text{annars} \end{cases} \tag{11}$$

Notera att vi här inte har blandat in någon normeringskonstant. Det enda vi behöver för att kunna utnyttja denna teknik är att kurvformen $f(w)$ är känd. Eventuella skalningsfaktorer kommer att försvinna vid hanteringen.

Variabeln $w$ kan här endast kan ligga mellan -1 och 1. Notera dessutom att det maximala värdet för $p(w)$ fås i $w=0$ vilket ger $f_{max}=4$. Vi vet att det är enkelt att generera likformigt fördelade slumptal (Matlabs funktion `rand`). Detta faktum kan vi här utnyttja i acceptance/rejection-metoden. Tekniken kan beskrivas kortfattat enligt följande schema:

1. Dra ett likformigt fördelat tal, $w_{prop}$ ("prop" som i "proposition") i intervallet [-1,1].
2. Beräkna kvoten $k = f(w_{prop})/f_{max} = f(w_{prop})/4$.
3. Generera ännu ett likformigt slumptal, $q$, i intervallet [0,1]. Acceptera $w_{prop}$ som det slutliga samplet om $q<k$. Detta sker med sannolikheten $P(q<k)=k$. Annars, börja om från punkt 1 och iterera tills vi får fram ett accepterat värde.

Det är relativt lätt att inse att ovanstående schema kommer att ge värden nära noll med större sannolikhet än de som ligger nära $+/-2$. Speciellt kommer vi alltid att acceptera värdet $w = 0$ om vi erhåller detta värde i första steget. Generellt kommer sannolikheten att acceptera ett sampel i ett litet intervall kring $w$ att vara proportionellt mot $f(w)/f_{max}$.

Notera att inget hindrar att vi använder andra fördelningar än den likformiga för att generera $w_{prop}$. Tekniken fungerar väl i de situationer då det går att hitta en enkel fördelning som vi kan dra $w_{prop}$ ifrån och vilken kan skalas så att den alltid är större än $f(w)$. Ju bättre dessa PDF:er är matchade desto färre sampel behöver förkastas och desto effektivare blir metoden.

I princip kan man dessutom lätt generalisera metoden så att den kan tillämpas på flerdimensionella fördelningar. Det stora problemet som gör att tekniken snabbt förlorar i praktisk användbarhet då dimensionen ökar är att flerdimensionella fördelningar generellt sett är betydligt "glesare" än endimensionella. Detta innnebär acceptance/rejection-algoritmen slösar bort större delen av tiden på att förkasta $\mathbf{w}_{prop}$.

En tänkbar lösning på detta problem får vi via s.k. Markov Chain Monte Carlo-metoder, tex *Gibb's sampling* eller *Metropolis–Hastings*.

Fortsättning följer!