Tentamen – Bayesiansk inferens och maskininlärning (TIF385)

Tid och plats:	16 augusti, 2023, fm, Johanneberg.
Hjälpmedel:	Physics Handbook, Beta Mathematics Handbook, typgodkänd kalkylator.
Examinator:	Christian Forssén.
Jourhavande lärare:	Christian Forssén (031–772 3261).

Tentamen består av sex uppgifter som kan ge maximalt 60 poäng totalt. För att bli godkänd med betyg 3 krävs 24 poäng, för betyg 4 krävs 36 poäng och för betyg 5 krävs 48 poäng.

Rättningsprinciper: Alla svar skall motiveras (om ej annat anges) och införda storheter skall förklaras. Lösningarna förväntas vara välstrukturerade och begripligt presenterade. Skriv och rita tydligt. Vid tentamensrättning gäller följande allmänna principer:

- För maximal (10) poäng krävs fullständigt korrekt och välmotiverad lösning.
- Mindre fel ger 1-3 poängs avdrag. Gäller även mindre brister i presentationen.
- Lösningar som inte går att följa (t.ex. avsaknad av relevant figur, ej definierade variabler, svårtydd, etc) renderar poängavdrag även om svaret verkar vara korrekt.
- Allvarliga principiella fel ger fullt poängavdrag.
- Allvarliga fel som leder till orimliga resultat kan ge lägre poängavdrag om orimligheten pekas ut.
- Även skisserade lösningar kan ge delpoäng.

Lycka till!

---

1. Svara på följande delfrågor:

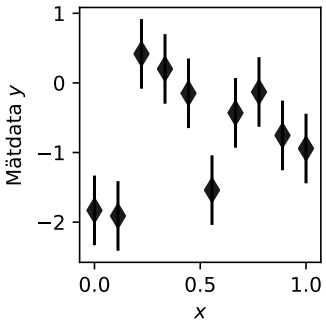
   (a) Använd produktregeln för sannolikheter för att härleda Bayes teorem
   $$\mathbb{P}\left(M \mid \mathcal{D}, I\right) = \frac{\mathbb{P}\left(\mathcal{D} \mid M, I\right)\mathbb{P}\left(M \mid I\right)}{\mathbb{P}\left(\mathcal{D} \mid I\right)}$$
   (2 poäng)

   (b) Parameterestimering är beteckningen på uppgiften att utföra inferens av modellparametrarna $\boldsymbol{\theta}$ för en modell $M(\boldsymbol{\theta})$ givet data $\mathcal{D}$. Vid parameterestimering slipper man att evaluera en av termerna i Bayes teorem. Vilken och varför? *(2 poäng)*

(c) Parametrarna för en kubisk modell

$$\hat{y} = \sum_{i=0}^{3} \theta_i x^i,$$

skall estimeras givet data (se figur).

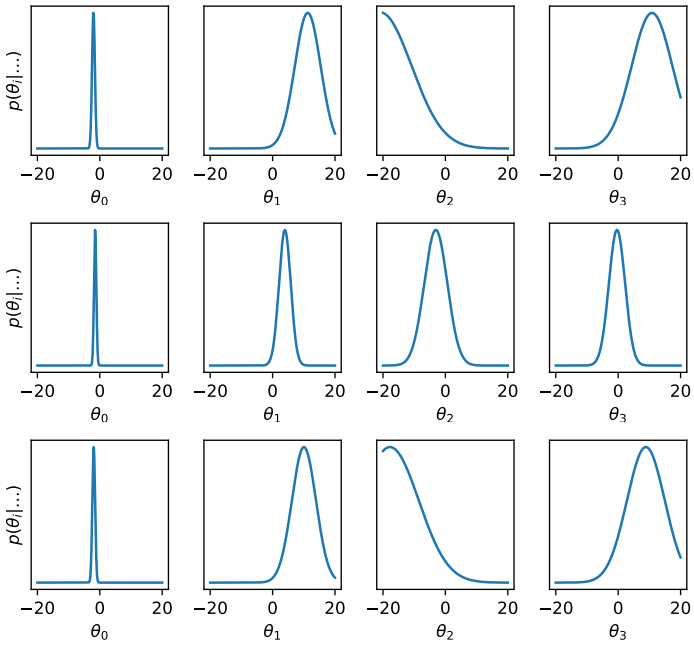


Tre olika fysiker använder var sin à-priorifördelning vid parameterestimeringen. Dessa är:

(A) $p(\theta_i \mid I) = \mathcal{U}([-20, +20])$ (likformig),
(B) $p(\theta_i \mid I) = \mathcal{N}(\mu = 0, \sigma^2 = 5^2)$ (normal),
(C) $p(\theta_i \mid I) = \mathcal{N}(\mu = 0, \sigma^2 = 30^2)$ (normal).

Varje rad i figuren nedan visar marginalfördelningarna för den à-posteriorifördelning som erhålls av någon av fysikerna (notera att fördelningarna bara visas för intervallet $[-20, 20]$). Para ihop fysiker (bokstav) med rätt rad (siffra).

(1)
(2)
(3)



(varje rätt par ger +2 poäng och varje fel ger -2 poäng; totalpoängen begränsas dock till intervallet 0–6 poäng)

2. En fysiker får reda på att hon är gravid med tvillingar. Hon frågar läkaren om sannolikheten att det är identiska (enäggs) tvillingar. Läkaren svarar att 1/3 av alla tvillinggraviditeter är enäggs medan 2/3 är tvåäggs (dvs den efterfrågade sannolikheten är 1/3).

   Senare under graviditeten avslöjar ett ultraljud att fostren har samma kön. Borde denna nya information leda till en uppdatering av den efterfrågade sannolikheten, och isf hur? Motivera ditt svar. *(10 poäng)*

   [Notera att identiska tvillingar alltid har samma kön medan det inte finns någon tydlig korrelation mellan könen hos tvåäggstvillingar.]

3. Vi vill propagera osäkerheten för slumpvariabeln Z, vilken i sin tur ges av sambandet $Z = X/Y$ där X, Y är två andra slumpvariabler vars förväntansvärden ($\mathbb{E}[X] = x_0, \mathbb{E}[Y] = y_0$) och varianser ($\mathrm{Var}(X) = \sigma_x^2$, $\mathrm{Var}(Y) = \sigma_y^2$) har uppmätts. Konsultation av en tabell för felpropagering ger uttrycken

$$\mathbb{E}[Z] = x_0/y_0, \quad \mathrm{Var}(Z) = \left(\frac{x_0}{y_0}\right)^2 \left[\left(\frac{\sigma_x}{x_0}\right)^2 + \left(\frac{\sigma_y}{y_0}\right)^2\right],$$

   för dess förväntansvärde och varians uttryckt i de uppmätta storheterna.

   (a) Härled dessa uttryck. För full poäng skall du betona vilka antaganden som ligger bakom. *(5 poäng)*

   (b) Under vissa omständigheter gäller att variansen hos Z är mycket mindre än uppskattningen ovan. Vilket antagande i härledningen ovan kan antas vara fel för att detta skall gälla. Vi kan anta att x_0, y_0 är positiva och ungefär lika stora. *(5 poäng)*

4. Betrakta en maskininlärningsmodell som ger förutsägelsen $\hat{y}_i$ att jämföra med träningsdata y_i (för $i = 1 \ldots N$). Vi definierar "träningsfelet"

$$E_{\text{train}} = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 .$$

   (a) Förklara varför E_{train} är ett dåligt mått för att uppskatta prediktionsfelet i en maskininlärningsmodell med ett stort antal parametrar. *(3 poäng)*

   (b) Ge ett exempel på ett bättre mått om man vill få en uppskattning av prediktionsfelet (du behöver inte beskriva den detaljerad algoritmen utan det räcker med en till två beskrivande meningar). *(3 poäng)*

(c) Betrakta ett scenario där E_{train} är betydligt större än den noggrannhet som du siktar på för din maskininlärningsmodell. I denna situation är det bortkastat att lägga beräkningsresurser på att faktiskt uppskatta prediktionsfelet. Varför då? *(4 poäng)*

5. Beskriv ingrediensen "aktiveringsfunktion" i konstruktionen av ett neuralt nätverk för maskininlärning. Var och hur används den? För full poäng skall du även ge matematiska uttryck för två olika aktiveringsfunktioner och betona någon skillnad som kan dyka upp i praktiken när den ena eller den andra används. *(10 poäng)*

6. I denna uppgift betraktas en uppsättning med kontinuerliga variabler $\boldsymbol{\theta}$ med en fördelning som ges av sannolikhetstätheten $p(\boldsymbol{\theta})$.

   (a) Formulera kortfattat (en eller två meningar) syftet med Metropolisalgoritmen i termer av begreppen Markovkedja och gränsdistribution. *(2 poäng)*

   (b) Metropolisalgoritmen utgår från att övergångstätheten $T(\boldsymbol{\theta}, \boldsymbol{\theta}')$ för en Markovkedja kan skrivas som en produkt

   $$T(\boldsymbol{\theta}, \boldsymbol{\theta}') = A(\boldsymbol{\theta}, \boldsymbol{\theta}')S(\boldsymbol{\theta}, \boldsymbol{\theta}'). \tag{1}$$

   Förklara med ord vad dessa två faktorer beskriver. *(2 poäng)*

   (c) Härled ett uttryck för övergångsmatrisen (1) utgående från kravet att detaljerad balans $p(\boldsymbol{\theta})T(\boldsymbol{\theta}, \boldsymbol{\theta}') = p(\boldsymbol{\theta}')T(\boldsymbol{\theta}', \boldsymbol{\theta})$ skall uppfyllas. *(6 poäng)*

Tentamen – Bayesiansk inferens och maskininlärning (TIF385)

Tid och plats:	16 augusti, 2023, fm, Johanneberg.
Hjälpmedel:	Physics Handbook, Beta Mathematics Handbook, typgodkänd kalkylator.
Lösningsskiss:	Christian Forssén.

Tentamen består av sex uppgifter som kan ge maximalt 60 poäng totalt. För att bli godkänd med betyg 3 krävs 24 poäng, för betyg 4 krävs 36 poäng och för betyg 5 krävs 48 poäng.

Rättningsprinciper: Alla svar skall motiveras (om ej annat anges) och införda storheter skall förklaras. Lösningarna förväntas vara välstrukturerade och begripligt presenterade. Skriv och rita tydligt. Vid tentamensrättning gäller följande allmänna principer:

- För maximal (10) poäng krävs fullständigt korrekt och välmotiverad lösning.
- Mindre fel ger 1-3 poängs avdrag. Gäller även mindre brister i presentationen.
- Lösningar som inte går att följa (t.ex. avsaknad av relevant figur, ej definierade variabler, svårtydd, etc) renderar poängavdrag även om svaret verkar vara korrekt.
- Allvarliga principiella fel ger fullt poängavdrag.
- Allvarliga fel som leder till orimliga resultat kan ge lägre poängavdrag om orimligheten pekas ut.
- Även skisserade lösningar kan ge delpoäng.

Detta är enbart en skiss av den fullständiga lösningen. Det kan innebära att vissa mellansteg i uträkningarna, som egentligen är nödvändiga för en komplett lösning, inte redovisas.

---

1. Svara på följande delfrågor:

   (a) Använd produktregeln för sannolikheter för att härleda Bayes teorem

   $$\mathbb{P}(M \mid \mathcal{D}, I) = \frac{\mathbb{P}(\mathcal{D} \mid M, I)\,\mathbb{P}(M \mid I)}{\mathbb{P}(\mathcal{D} \mid I)}$$

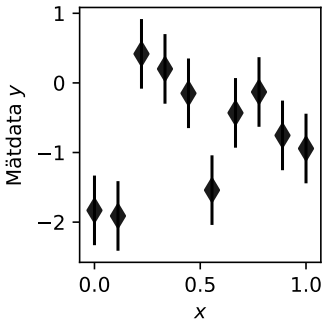
   (2 poäng)

   (b) Parameterestimering är beteckningen på uppgiften att utföra inferens av modellparametrarna $\boldsymbol{\theta}$ för en modell $M(\boldsymbol{\theta})$ givet data $\mathcal{D}$. Vid parameterestimering slipper man att evaluera en av termerna i Bayes teorem. Vilken och varför? *(2 poäng)*

(c) Parametrarna för en kubisk modell

$$\hat{y} = \sum_{i=0}^{3} \theta_i x^i,$$

skall estimeras givet data (se figur).

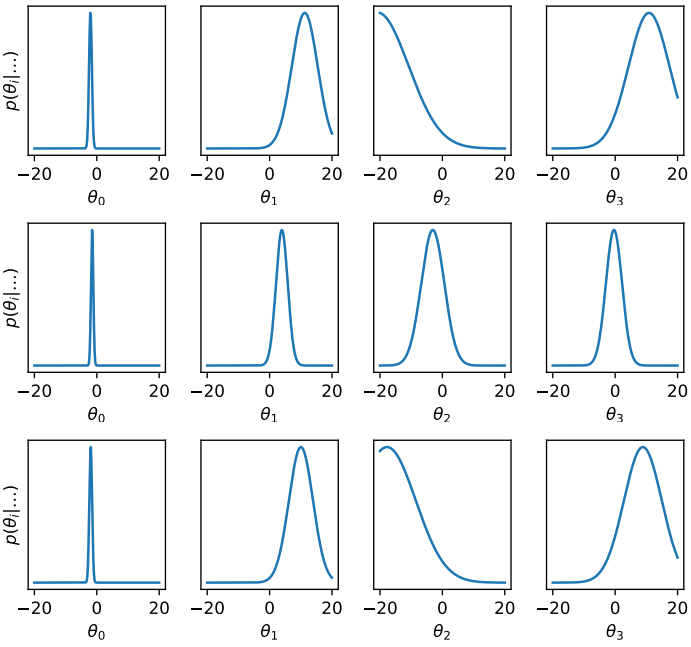


Tre olika fysiker använder var sin à-priorifördelning vid parameterestimeringen. Dessa är:

(A) $p(\theta_i \mid I) = \mathcal{U}([-20, +20])$ (likformig),
(B) $p(\theta_i \mid I) = \mathcal{N}(\mu = 0, \sigma^2 = 5^2)$ (normal),
(C) $p(\theta_i \mid I) = \mathcal{N}(\mu = 0, \sigma^2 = 30^2)$ (normal).

Varje rad i figuren nedan visar marginalfördelningarna för den à-posteriorifördelning som erhålls av någon av fysikerna (notera att fördelningarna bara visas för intervallet $[-20, 20]$). Para ihop fysiker (bokstav) med rätt rad (siffra).

(1)
(2)
(3)



(varje rätt par ger +2 poäng och varje fel ger -2 poäng; totalpoängen begränsas dock till intervallet 0–6 poäng)

Lösning:

(a) Se till exempel ekvation 7.1 och 7.3 i kurskompendiet.

(b) Nämnaren $\mathbb{P}(\mathcal{D} \mid I)$ beror ej på modellen och därmed inte heller på dess parametrar. Eftersom målet med parameterestimering blir att bestämma en PDF för parametrarna kommer denna term bara att ge en normaliseringskonstant.

(c) A1, B2, C3. Trolighetsfunktionen är densamma i de tre olika analyserna. Men multiplikation med någon av de normalfördelade apriori-fördelningarna medför att toppen för den resulterande aposteriori-fördelningen skiftas mot noll. De kommer också att påverka dessa bredd. Effekterna syns tydligast i de breda marginalfördelningarna (eftersom trolighetsfunktionen har mindre inflytande på dessa). T.ex. marginalfördelningen för θ_2 eller θ_3 .

---

2. En fysiker får reda på att hon är gravid med tvillingar. Hon frågar läkaren om sannolikheten att det är identiska (enäggs) tvillingar. Läkaren svarar att 1/3 av alla tvillinggraviditeter är enäggs medan 2/3 är tvääggs (dvs den efterfrågade sannolikheten är 1/3).

   Senare under graviditeten avslöjar ett ultraljud att fostren har samma kön. Borde denna nya information leda till en uppdatering av den efterfrågade sannolikheten, och isf hur? Motivera ditt svar. *(10 poäng)*

   [Notera att identiska tvillingar alltid har samma kön medan det inte finns någon tydlig korrelation mellan könen hos tvääggstvillingar.]

   Lösning:

   Ja, sannolikheten skall uppdateras eftersom vi har fått ny information. Låt H_1 vara propositionen att det är enäggs- och H_2 att det är tvääggstvillingar. Observationen att fostren har samma kön betecknas D_S medan läkarens tidigare erfarenhet av tvillingfödslar betecknas I. Vi studerar kvoten av de betingade sannolikheterna

$$r \equiv \frac{\mathbb{P}(H_1 \mid D_S, I)}{\mathbb{P}(H_2 \mid D_S, I)},$$

   och använder Bayes teorem för att skriva om både nämnaren och täljaren

$$r = \frac{\mathbb{P}(D_S \mid H_1, I)\,\mathbb{P}(H_1 \mid I)}{\mathbb{P}(D_S \mid H_2, I)\,\mathbb{P}(H_2 \mid I)},$$

där vi konstaterar att faktorn $\mathbb{P}(D_S \mid I)$ dyker upp i både nämnaren och täljaren och kan förkortas bort.

Läkarens tidigare erfarenhet ger a-priori-sannolikheterna $\mathbb{P}(H_1 \mid I) = 1/3$ och $\mathbb{P}(H_2 \mid I) = 2/3$. Vi har även troligheterna $\mathbb{P}(D_S \mid H_1, I) = 1$ och $\mathbb{P}(D_S \mid H_2, I) = 1/2$ för den nya observationen. Sammantaget blir $r = 1$, dvs den betingade sannolikheten att det är enäggstvillingar givet all information är lika med den betingade sannolikheten att det är tvåäggstvillingar.

Eftersom $\sum_{i=1,2} \mathbb{P}(H_i \mid D_S, I) = 1$ blir den eftersökta sannolikheten $\mathbb{P}(H_1 \mid D_S, I) = 1/2$.

---

3. Vi vill propagera osäkerheten för slumpvariabeln Z, vilken i sin tur ges av sambandet $Z = X/Y$ där X, Y är två andra slumpvariabler vars förväntansvärden ($\mathbb{E}[X] = x_0, \mathbb{E}[Y] = y_0$) och varianser ($\mathrm{Var}(X) = \sigma_x^2$, $\mathrm{Var}(Y) = \sigma_y^2$) har uppmätts. Konsultation av en tabell för felpropagering ger uttrycken

$$\mathbb{E}[Z] = x_0/y_0, \quad \mathrm{Var}(Z) = \left(\frac{x_0}{y_0}\right)^2 \left[\left(\frac{\sigma_x}{x_0}\right)^2 + \left(\frac{\sigma_y}{y_0}\right)^2\right],$$

för dess förväntansvärde och varians uttryckt i de uppmätta storheterna.

(a) Härled dessa uttryck. För full poäng skall du betona vilka antaganden som ligger bakom. *(5 poäng)*

(b) Under vissa omständigheter gäller att variansen hos Z är mycket mindre än uppskattningen ovan. Vilket antagande i härledningen ovan kan antas vara fel för att detta skall gälla. Vi kan anta att x_0, y_0 är positiva och ungefär lika stora. *(5 poäng)*

Lösning:

(a) Vi gör en Taylorexpansion av Z runt punkten (x_0, y_0)

$$Z = Z(X,Y) = Z(x_0,y_0) + \left.\frac{\partial Z}{\partial X}\right|_{x_0,y_0} (X - x_0) + \left.\frac{\partial Z}{\partial Y}\right|_{x_0,y_0} (Y - y_0) + \ldots$$

Om vi enbart betraktar ett litet område runt (x_0, y_0) kan vi försumma kvadratiska termer och högre givet att motsvarande högre ordningens derivator är små. Giltigheten för en sådan första ordningens Taylor-approximation av $Z(X, Y)$ är det ena antagandet. Notera att kvadratiska termer och uppåt hade varit exakt noll om funktionen hade varit linjär i både X och Y, men det är den inte i detta fall.

Förväntansvärdet blir då

$$\mathbb{E}[Z] = \mathbb{E}[Z(x_0, y_0)] + \mathbb{E}\left[\left.\frac{\partial Z}{\partial X}\right|_{x_0,y_0}(X - x_0)\right] + \mathbb{E}\left[\left.\frac{\partial Z}{\partial Y}\right|_{x_0,y_0}(Y - y_0)\right] = x_0/y_0,$$

där vi har använt att flera faktorer är konstanta och att $\mathbb{E}[X - x_0] = x_0 - x_0 = 0$ och analogt for $Y - y_0$.

Variansen får genom att ta förväntansvärdet på kvadraten av ovanstående vilket leder till följande uttryck (termer som är linjära i $(X - x_0)$ eller $(Y - y_0)$ blir noll)

$$\begin{aligned}\mathrm{Var}(Z) &= \mathbb{E}\left[\left.\frac{\partial Z}{\partial X}\right|^2_{x_0,y_0}(X - x_0)^2\right] + \mathbb{E}\left[\left.\frac{\partial Z}{\partial Y}\right|^2_{x_0,y_0}(Y - y_0)^2\right] \\ &\quad + \mathbb{E}\left[2\left.\frac{\partial Z}{\partial X}\right|_{x_0,y_0}\left.\frac{\partial Z}{\partial Y}\right|_{x_0,y_0}(X - x_0)(Y - y_0)\right] \\ &= \frac{1}{y_0^2}\sigma_x^2 + \frac{x_0^2}{y_0^4}\sigma_y^2 - 2\frac{x_0}{y_0^3}\mathrm{Cov}(X, Y).\end{aligned}$$

Detta motsvarar uttrycket i uppgiften under antagandet att X och Y är oberoende så att deras kovarians är noll.

(b) I härledningen antog vi att X och Y var oberoende vilket innebar att kovariansstermen var noll. Denna term skulle annars bidra negativt till uttrycket för $\mathrm{Var}(Z)$. Det vill säga, en positiv, nollskild kovarians mellan X och Y skulle kunna förklara en minskad varians för Z. Det vill säga, en möjlig orsak till minskad osäkerhet i Z skulle vara att X och Y är positivt korrelerade.

---

4. Betrakta en maskininlärningsmodell som ger förutsägelsen $\hat{y}_i$ att jämföra med träningsdata y_i (för $i = 1 \dots N$). Vi definierar "träningsfelet"

$$E_{\text{train}} = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2 \,.$$

   (a) Förklara varför E_{train} är ett dåligt mått för att uppskatta prediktionsfelet i en maskininlärningsmodell med ett stort antal parametrar. *(3 poäng)*

   (b) Ge ett exempel på ett bättre mått om man vill få en uppskattning av prediktionsfelet (du behöver inte beskriva den detaljerad algoritmen utan det räcker med en till två beskrivande meningar). *(3 poäng)*

   (c) Betrakta ett scenario där E_{train} är betydligt större än den noggrannhet som du siktar på för din maskininlärningsmodell. I denna situation är det bortkastat att lägga beräkningsresurser på att faktiskt uppskatta prediktionsfelet. Varför då? *(4 poäng)*

Lösning:

   (a) E_{train} kommer alltid att bli lägre när vi ökar modellens komplexitet (genom att större flexibilitet ger större möjlighet att anpassa till träningsdata). Måttet säger alltså ingenting om modellens träffsäkerhet när det gäller ny data.

   (b) Vi kan dela upp data i träningsdata och valideringsdata och använda den senare till att uppskatta hur väl modellen generaliserar till ny data. Modellen tränas enbart på träningsdata och vi definierar sedan ett felmått (t.ex. MSE) och beräknar det för valideringsdata.

   (c) I allmänhet har vi att modellfelet är större för ny data än för träningsdata som modellen har anpassats till. Vår uppskattning av prediktionsfelet lär därför bli större än träningsfelet, som redan hade konstaterats vara för stort. Det är bättre att träna modellen mer, eller att förbättra modellen och träna om den på nytt.

---

5. Beskriv ingrediensen "aktiveringsfunktion" i konstruktionen av ett neuralt nätverk för maskininlärning. Var och hur används den? För full

poäng skall du även ge matematiska uttryck för två olika aktiveringsfunktioner och betona någon skillnad som kan dyka upp i praktiken när den ena eller den andra används. *(10 poäng)*

Lösning:

- Se t.ex. avsnitten 18.2 och 18.5 i föreläsningsanteckningarna.
- Aktiveringsfunktionen beskriver sambandet mellan aktiveringen z (som beror på en nods insignaler och dess vikter) samt nodens utsignal. De är alltså viktiga för hur signaler propagerar genom ett nätverk.
- Aktiveringsfunktioner dyker också upp i det sista lagret och bestämmer då modellens utsignal.
- Två vanliga exempel är sigmoid och ReLU (ekvarion 18.11 och 18.13).
- Den förstnämnda är alltid nollskild (kan vara väldigt liten, men alltid nollskild) och innebär alltså att nätverket har någon form av signal från varje nod. Det blir ett s.k. brusigt nätverk ("noisy network").
- Den sistnämnda är delvis konstant vilker ger noll derivata. Noder där detta sker tenderar att inte påverkas under träningen (eftersom utsignalen inte ändras när vikterna ändras) och kallas för döda noder ("dying neurons").

---

6. I denna uppgift betraktas en uppsättning med kontinuerliga variabler $\boldsymbol{\theta}$ med en fördelning som ges av sannolikhetstätheten $p(\boldsymbol{\theta})$.

   (a) Formulera kortfattat (en eller två meningar) syftet med Metropolisalgoritmen i termer av begreppen Markovkedja och gränsdistribution. *(2 poäng)*

   (b) Metropolisalgoritmen utgår från att övergångstätheten $T(\boldsymbol{\theta}, \boldsymbol{\theta}')$ för en Markovkedja kan skrivas som en produkt

   $$T(\boldsymbol{\theta}, \boldsymbol{\theta}') = A(\boldsymbol{\theta}, \boldsymbol{\theta}')S(\boldsymbol{\theta}, \boldsymbol{\theta}'). \tag{1}$$

   Förklara med ord vad dessa två faktorer beskriver. *(2 poäng)*

   (c) Härled ett uttryck för övergångsmatrisen (1) utgående från kravet att detaljerad balans $p(\boldsymbol{\theta})T(\boldsymbol{\theta}, \boldsymbol{\theta}') = p(\boldsymbol{\theta}')T(\boldsymbol{\theta}', \boldsymbol{\theta})$ skall uppfyllas. *(6 poäng)*

Lösning:

(a) Metropolisalgoritmen syftar till att skapa en Markovkedja vars gränsdistribution är lika med en eftersökt sannolikhetsfördelning. Stickprov från den konvergerade Markovkedjan ger då stickprov från den sökta fördelningen.

(b) S beskriver en stegfördelning. Stickprov ur denna fördelning kommer ge algoritmen förslag på nya punkter att besöka i utfallsrummet. A är en acceptansfunktion som ger sannolikheten att nästa steg i Markovkedjan skall vara den föreslagna punkten.

(c) Se avsnitt 14.4 i föreläsningsanteckningarna.