Probabilistic Artificial Intelligence
Problem Set 4

Nov 4, 2019 Please check out the Moodle webpage for exam-like questions on the following

link: moodle-app2.let.ethz.ch/course/view.php?id=11902

1. Particle filter

Suppose that you have a robot, which is moving randomly through an 1-dimensional environment. You want to track the robot's position, x, which is discretized to integer values, $x \in \mathbb{Z}$. The robot's movement is modeled as a random walk,

$$x_{t+1} = x_t + \epsilon_t, \tag{1}$$

where ϵ_t is uniformly distributed and can take integer values in $[-3, 3]$. To track the robot, a sensor that measures the distance to the robot has been placed at the origin. The measurement model is

$$y_t = (x_t + \eta_t)^2, \tag{2}$$

where η_t is distributed according to

$$P(\eta_t) = \begin{cases} 0.6 & \text{if } \eta_t = 0 \\ 0.2 & \text{if } |\eta_t| = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

You want to use a particle filter with six particles to track the robot's position. At initial time, the robot is at the origin, $x_0 = 0$. Hence, the particles are initialized to $x_i = 0$, $i \in \{0, 1, 2, 3, 4, 5\}$.

(i) You draw samples from the distribution of ϵ_0 and obtain $(-1, -1, 0, 1, 2, 3)$. What is the position of the particles after the prediction update?

(ii) You obtain a measurement, $y_1 = 1$. What are the weights of the individual particles?

(iii) Are five particles enough to accurately estimate the state? Why/Why not?

(iv) Why would a Kalman filter not work reliably in this case?

Solution

(i) Using the movement model,

$$x_1 = x_0 + \epsilon_0, \tag{4}$$

we obtain $\mathbf{x}' = (-1, -1, 0, 1, 2, 3)$.

(ii) From the measurement model we obtain $\eta_t = \pm\sqrt{y_t} - x'_t$, and consequently the measurement probability distribution

$$P(y_{t+1}|x'_t) = P(\eta_t = \pm\sqrt{y_t} - x'_t) \tag{5}$$

The particle weights are computed as $w_i = \frac{1}{Z}P(y_{t+1}|x'_t)$

$$\begin{aligned}
n=0, \quad & P(y_1 = 1|x'_0 = -1) = P(\eta_1 = 0) && + P(\eta_1 = 2) && = 0.6 + 0.0 = 0.6 && (6)\\
n=1, \quad & P(y_1 = 1|x'_0 = -1) = P(\eta_1 = 0) && + P(\eta_1 = 2) && = 0.6 + 0.0 = 0.6 && (7)\\
n=2, \quad & P(y_1 = 1|x'_0 = 0) && = P(\eta_1 = -1) + P(\eta_1 = 1) && = 0.2 + 0.2 = 0.4 && (8)\\
n=3, \quad & P(y_1 = 1|x'_0 = 1) && = P(\eta_1 = 0) && + P(\eta_1 = -2) = 0.6 + 0.0 = 0.6 && (9)\\
n=4, \quad & P(y_1 = 1|x'_0 = 2) && = P(\eta_1 = -1) + P(\eta_1 = -3) = 0.2 + 0.0 = 0.2 && (10)\\
n=5, \quad & P(y_1 = 1|x'_0 = 3) && = P(\eta_1 = -2) + P(\eta_1 = -4) = 0.0 + 0.0 = 0.0 && (11)
\end{aligned}$$

$$Z = \sum_{i=0}^{N} P(y_1 = 1|x'_i) = 0.6 + 0.6 + 0.4 + 0.6 + 0.2 + 0.0 = \frac{24}{10} \tag{12}$$

Consequently, we can calculate the weights with $w_i = \frac{1}{Z}P(y_1 = 1 \mid x'_i)$

$$w_0 = \frac{6}{24},\ w_1 = \frac{6}{24},\ w_2 = \frac{4}{24}.\ w_3 = \frac{6}{24},\ w_4 = \frac{2}{24},\ w_5 = 0 \tag{13}$$

(iii) No, because we cannot even capture the probability distribution of the movement prediction accurately (uniform distribution). We need more samples to accurately estimate the state.

(iv) A Kalman filter can only describe Gaussian distributions (unimodal). Here, the noise is not Gaussian and the measurements are nonlinear. Furthermore, the distance measurements cannot break the symmetry in the problem, so that the posterior state distribution after one step is bimodal.

2. Bayesian networks and Markov chains

Consider the query $P(R|S = t, W = t)$ in the following Bayesian network, and how Gibbs sampling can answer it.

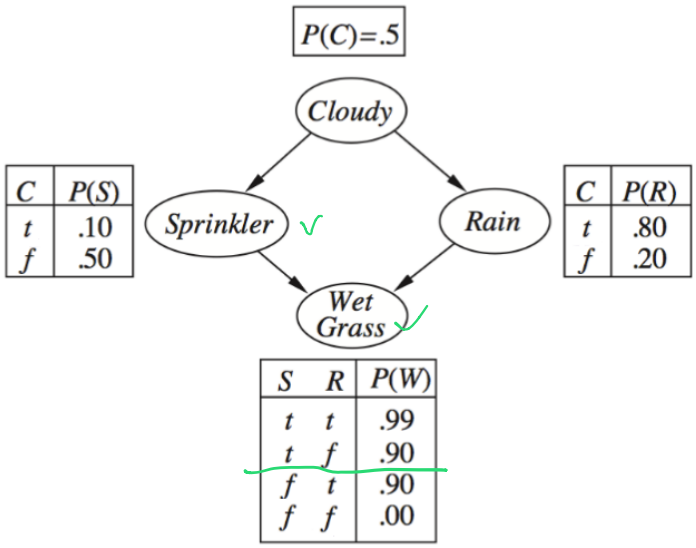

Figure 1: Bayesian Network

(i) How many states does the Markov chain have?

(ii) Calculate the transition matrix T containing $P(X_{t+1} = y \mid X_t = x)$ for all x, y.

(iii) What does T^2, the square of the transition matrix, represent?

(iv) What about T^n as $n \to \infty$?

(v) Explain how to do probabilistic inference in Bayesian networks, assuming that T^n is available. Is this a practical way to do inference?

Solution

(i) The Bayesian Network has two free binary variables R and C, therefore there are 4 states in the Markov Chain. In general, there are $|X|^N$ states, where $|X|$ is the cardinality of the variables (in this case binary) and N is the number of free variables.

(ii) First, we compute the sampling distribution for each variable:

$$\begin{aligned}
P(C|r,s,W) &= P(C|r,s) = \frac{1}{Z}P(C)P(s|C)P(r|C) \\
&= \frac{1}{Z}\langle 0.5, 0.5\rangle\langle 0.1, 0.5\rangle\langle 0.8, 0.2\rangle = \frac{1}{Z}\langle 0.04, 0.05\rangle = \langle 4/9, 5/9\rangle \\
P(C|\neg r,s,W) &= P(C|\neg r,s) = \frac{1}{Z}P(C)P(s|C)P(\neg r|C) \\
&= \frac{1}{Z}\langle 0.5, 0.5\rangle\langle 0.1, 0.5\rangle\langle 0.2, 0.8\rangle = \frac{1}{Z}\langle 0.01, 0.2\rangle = \langle 1/21, 20/21\rangle \\
P(R|c,s,w) &= \frac{1}{Z}P(R|c)P(w|s,R) \\
&= \frac{1}{Z}\langle 0.8, 0.2\rangle\langle 0.99, 0.9\rangle = \frac{1}{Z}\langle 0.792, 0.18\rangle = \langle 22/27, 5/27\rangle \\
P(R|\neg c,s,w) &= \frac{1}{Z}P(R|\neg c)P(w|s,R) \\
&= \frac{1}{Z}\langle 0.2, 0.8\rangle\langle 0.99, 0.9\rangle = \frac{1}{Z}\langle 0.198, 0.72\rangle = \langle 11/51, 40/51\rangle
\end{aligned}$$

Strictly speaking, the transition matrix is only well-defined for the variant of MCMC in which the variable to be sampled is chosen randomly[1]. (In the variant where the variables are chosen in a fixed order, the transition probabilities depend on where we are in the ordering.)

For the transition matrix:

- Entries on the diagonal correspond to self-loops. Such transitions can occur by sampling *either* variable. For example, for the self-loop on (c, r), we obtain:

  $$t((c,r) \to (c,r)) = 0.5P(c|r,s) + 0.5P(r|c,s,w) = 17/27,$$

  where the two factors of 0.5 are corresponding to the probability that the variables to be sampled are C and R, respectively.
- Entries where one variable is changed must sample that variable. For example,

  $$t((c,r) \to (c,\neg r)) = 0.5P(\neg r|c,s,w) = 5/54$$

- Entries where both variables change cannot occur. For example,

  $$t((c,r) \to (\neg c,\neg r)) = 0$$

This gives us the following transition matrix T, where the transition is from the state given by the row label to the state given by the column label:

$$\begin{array}{c} \\ (c,r) \\ (c,\neg r) \\ (\neg c,r) \\ (\neg c,\neg r) \end{array}
\begin{array}{c}
\begin{array}{cccc} (c,r) & (c,\neg r) & (\neg c,r) & (\neg c,\neg r) \end{array} \\
\begin{pmatrix}
17/27 & 5/54 & 5/18 & 0 \\
11/27 & 22/189 & 0 & 10/21 \\
2/9 & 0 & 59/153 & 20/51 \\
0 & 1/42 & 11/102 & 310/357
\end{pmatrix}
\end{array}$$

[1] Slide 15 of https://las.inf.ethz.ch/courses/pai-f19/slides/pai-06-mcmc-annotated.pdf

(iii) T^2 represents the probability of going from each state to each state in two steps.

(iv) T^n (as $n \to \infty$) represents the stationary probability of being in each state starting in each state.

(v) For Ergodic Markov Chains these probabilities are independent of the starting state, so every row of T is the same and represents the posterior distribution over states given the evidence. So we can use any row of the matrix T^n to do inference. However, We can produce very large powers of T with very few matrix multiplications. For example, we can get T^2 with one multiplication, T^4 with two, and T^{2^k} with k. Unfortunately, in a network with n non-event Boolean variables, the matrix is of size $2^n \times 2^n$, so each multiplication takes $O(2^{3n})$ operations.

3. Assumed Density Filters

Let $p(x)$ be any distribution on $\mathbf{R}^n$. Let $q(x)$ be a multivariate normal distribution with mean μ and covariance Σ, i.e.:

$$q(x) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \exp(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu))$$

Prove that the forward KL divergence between $p(x)$ and $q(x)$, $D(p||q)$ is minimzed by moment matching, i.e.: $\mu = \mathbf{E}_{x\sim p(\cdot)}[x]$ and $\Sigma = \mathbf{E}_{x\sim p(\cdot)}\left[(x-\mu)(x-\mu)^\top\right]$

Hint: Use formulas 57, 61, and 86 of the Matrix Cookbook http://www.math.uwaterloo.ca/~hwolkowi//matrixcookbook.pdf

$D(p||q) = \int p(x) \log \frac{p(x)}{q(x)} dx$

$= \int p(x) \log p(x) dx + \frac{n}{2}\log(2\pi) + \frac{1}{2}\log\det(\Sigma) + \frac{1}{2}\int p(x)(x-\mu)^\top\Sigma^{-1}(x-\mu)dx\big)$

Solution

The KL divergence is:

$$D(p||q) = \int p(x) \log \frac{p(x)}{q(x)} dx$$
$$= \int p(x) \log p(x) dx + \frac{n}{2}\log(2\pi) + \frac{1}{2}\log\det(\Sigma) + \frac{1}{2}\int p(x)(x-\mu)^\top\Sigma^{-1}(x-\mu)dx$$

$\frac{\partial D}{\partial \mu} = \int p(x)\Sigma^{-1}(x-\mu)dx = \Sigma^{-1}\int p(x)(x-\mu)dx$

$= \Sigma^{-1} E_{x\sim p(x)}[x] - \mu = 0$

Taking derivatives w.r.t. μ and Σ and setting them to zero:

$\frac{\partial D}{\partial \Sigma} = \frac{1}{2\det(\Sigma)} +$

$$\frac{\partial D(p||q)}{\partial \mu} = \int p(x)\Sigma^{-1}(x-\mu)dx$$
$$= \Sigma^{-1}\left(\int xp(x)dx - \int \mu p(x)dx\right)$$
$$= \Sigma^{-1}\left(\mathbf{E}_{x\sim p(\cdot)}[x] - \mu\right)$$
$$\Rightarrow \frac{\partial D(p||q)}{\partial \mu} = 0 \Leftrightarrow \mu = \mathbf{E}_{x\sim p(\cdot)}[x]$$

Likewise:

$$\begin{aligned}\frac{\partial D(p||q)}{\partial \Sigma} &= \frac{1}{2}\Sigma^{-1} - \frac{1}{2}\int p(x)\Sigma^{-1}(x-\mu)(x-\mu)^\top \Sigma^{-1} dx \\ &= \frac{1}{2}\Sigma^{-1} - \frac{1}{2}\Sigma^{-1}\int p(x)(x-\mu)(x-\mu)^\top dx \Sigma^{-1} \\ &= \frac{1}{2}\Sigma^{-1}\left(\Sigma - \int (x-\mu)(x-\mu)^\top p(x)dx\right)\Sigma^{-1} \\ \Rightarrow \frac{\partial D(p||q)}{\partial \Sigma} &= 0 \Leftrightarrow \Sigma = \mathbf{E}_{x\sim p(\cdot)}\left[(x-\mu)(x-\mu)^\top\right]\end{aligned}$$

4. Assumed Density Filters for Object Tracking in the Presence of Clutter

We assume an object is located at a position $\boldsymbol{\theta} \in \mathbb{R}^d$. At time step t for $t = 1, \ldots, n$ we are able to observe the noisy position of this object $\mathbf{x}_t \in \mathbb{R}^d$ through the model given below:

$$p(x_t|\theta) = (1-w)\mathcal{N}(x_t; \theta, I_d) + w\mathcal{N}(x_t; 0, 10I_d) \tag{14}$$

Where $\mathcal{N}(.; \mu, \Sigma)$ denotes the standard d-dimensional, normal distribution with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ and $0 < w < 1$ is a known mixing parameter. We let $\boldsymbol{\theta} \in \mathbb{R}^d$ have a prior distribution:

$$p(\theta) = \mathcal{N}(\theta; 0, I_d)$$

Therefore the joint distribution of $\boldsymbol{\theta}$ and n independent observations $\boldsymbol{D} \triangleq \{\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}\}$ is given by:

$$p(\theta, D) = p(\theta)\prod_{t=1}^{n} p(x_t|\theta)$$

The goal here is to approximate the true posterior $p(\theta|x_1, \ldots, x_t)$ at time step $1 \leq t \leq n$. As a new data point x_{t+1} is received at time step $t+1$ we would like to update our posterior to accommodate for the new data point.

We assume the posterior has an (approximate) distribution that is spherical Gaussian parametrized by its mean and variance. More precisely at time step $t \in [n]$,we approximate the true posterior by:

$$q^{(t)}(\theta) = \mathcal{N}(\theta; \mu^{(t)}, \sigma^{(t)2} I_d)$$

Where $\mu^{(t)} \in \mathbb{R}^d$ and $\sigma \in \mathbb{R}^+$. We set $q^{(0)}(t) = p(\theta)$ (equal to the prior).

(i) Draw a Bayesian network corresponding to this model. What are the hidden and observed random variables at time t?

(ii) For $t = 1, \ldots, n$ and assuming that $q^{(t-1)}(\theta) = p(\theta|x_1, \ldots, x_{t-1})$ i.e. $q^{(t-1)}(\theta)$ is exactly equal to the posterior at time $(t-1)$, show that the posterior at time t is given by:

$$\hat{p}(\theta) \triangleq p(\theta|x_1, \ldots, x_t) = \frac{q^{(t-1)}(\theta)p(x_t|\theta)}{\int_\theta q^{(t-1)}(\theta)p(x_t|\theta)d\theta}$$

(iii) In this step we approximate the true posterior $\hat{p}(\theta)$ with $q^{(t)}(\theta)$ as follows:

$$\mu^{(t)}, \sigma^{(t)} = \underset{\mu^t, \sigma^t}{\operatorname{argmin}} KL(\hat{p}(\theta)||q^{(t)}(\theta))$$

Show that:

$$\mu^{(t)} = \mathbb{E}_{\boldsymbol{\theta} \sim q^{(t)}}[\boldsymbol{\theta}] = \mathbb{E}_{\boldsymbol{\theta} \sim \hat{p}}[\boldsymbol{\theta}] \tag{15}$$

$$\sigma^{(t)2} d + \mu^{(t)\top} \mu^{(t)} = \mathbb{E}_{\boldsymbol{\theta} \sim q^{(t)}}[\boldsymbol{\theta}^\top \boldsymbol{\theta}] = \mathbb{E}_{\boldsymbol{\theta} \sim \hat{p}}[\boldsymbol{\theta}^\top \boldsymbol{\theta}] \tag{16}$$

Hint: Take derivatives of the KL with respect to $\mu^{(t)}$ and $\sigma^{(t)}$ and set them to zero as in question 4

(iv) In this step we show how one can compute $\mathbb{E}_{\boldsymbol{\theta} \sim \hat{p}}[\boldsymbol{\theta}]$ and $\mathbb{E}_{\boldsymbol{\theta} \sim \hat{p}}[\boldsymbol{\theta}^\top \boldsymbol{\theta}]$. Let us define:

$$Z(\mu^{(t-1)}, \sigma^{(t-1)}) = \int_\theta p(x_t|\theta) q^{(t-1)}(\theta) d\theta$$

By differentiating $\log Z$ w.r.t $\mu^{(t-1)}$ and $\sigma^{(t-1)}$ show that (we drop the dependence of $\mu^{(t-1)}$ and $\sigma^{(t-1)}$ on $(t-1)$ for notational convenience) :

$$\mathbb{E}_{\boldsymbol{\theta} \sim \hat{p}}[\boldsymbol{\theta}] = \mu + \sigma^2 \nabla_\mu \log(Z(\mu, \sigma)) \tag{17}$$

$$\mathbb{E}_{\boldsymbol{\theta} \sim \hat{p}}[\boldsymbol{\theta}^\top \boldsymbol{\theta}] - \mathbb{E}_{\boldsymbol{\theta} \sim \hat{p}}[\boldsymbol{\theta}]^\top \mathbb{E}_{\boldsymbol{\theta} \sim \hat{p}}[\boldsymbol{\theta}] = \sigma^2 d - \sigma^4 (\nabla_\mu^\top \nabla_\mu - \sigma^{-1} \nabla_\sigma) \log Z(\mu, \sigma) \tag{18}$$

Where $\nabla_\mu^\top \nabla_\mu \log Z = \|\nabla_\mu \log Z\|^2$

(v) By inserting $p(x_t|\theta)$ from equation (14) derive update equations for $\mu^{(t)}$ and $\sigma^{(t)}$ given $\mu^{(t-1)}$ and $\sigma^{(t-1)}$ and x_t.

Solution

(i) From the joint distribution the Bayesian network is shown in Figure 2.

(ii) Applying Bayes' rule yields:

$$p(\theta|x_1, \ldots, x_t) = \frac{p(x_t|x_1, \ldots, x_{t-1}, \theta) p(\theta|x_1, \ldots, x_{t-1})}{\int_\theta p(x_t|x_1, \ldots, x_{t-1}, \theta) p(\theta|x_1, \ldots, x_{t-1}) d\theta}$$

$\boldsymbol{x_t}$ is conditionally independent of $\boldsymbol{x_1}, \ldots, \boldsymbol{x_{t-1}}$ given $\boldsymbol{\theta}$, hence $p(x_t|x_1, \ldots, x_{t-1}, \theta) = p(x_t|\theta)$. By assumption $q^{(t-1)}(\theta) = p(\theta|x_1, \ldots, x_{t-1})$. Hence:

$$p(\theta|x_1, \ldots, x_t) = \frac{p(x_t|\theta) q^{(t-1)}(\theta)}{\int_\theta p(x_t|\theta) q^{(t-1)}(\theta) d\theta}$$

(iii) For convenience we drop the dependencies on t.

$$\mu, \sigma = \underset{\mu, \sigma}{\operatorname{argmin}} KL(\hat{p}(\theta)||q(\theta)) = \underset{\mu, \sigma}{\operatorname{argmin}} \int_\theta \hat{p}(\theta) \log(\frac{\hat{p}(\theta)}{q(\theta)}) d\theta = \underset{\mu, \sigma}{\operatorname{argmin}} \int_\theta \hat{p}(\theta) \log(\frac{1}{q(\theta)}) d\theta$$

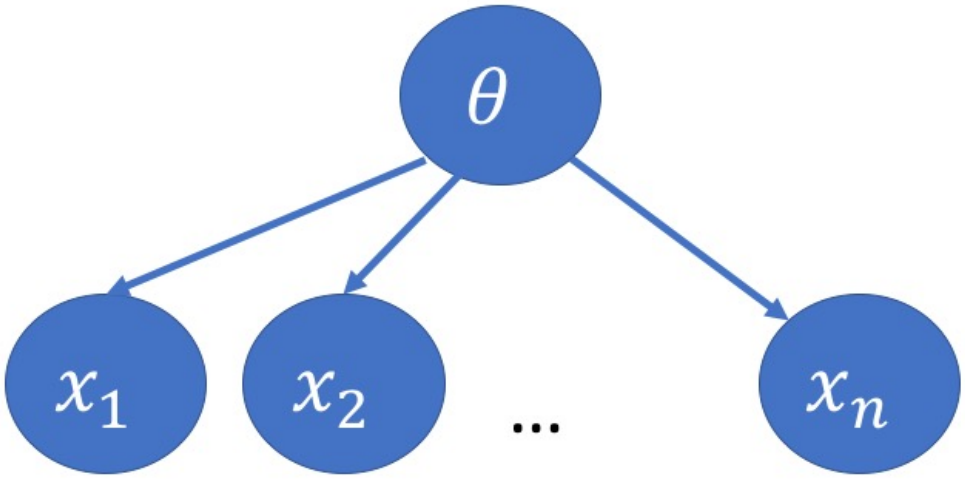

Figure 2: Bayesian Network

We replace $q(\theta)$ by its true format i.e. $q(\theta) = \frac{1}{(\sqrt{2\pi}\sigma)^d} e^{-\frac{\|\theta-\mu\|^2}{2\sigma^2}}$

$$\mu, \sigma = \text{argmin} \int_\theta \hat{p}(\theta) \log((\sqrt{2\pi}\sigma)^d) d\theta + \int_\theta \hat{p}(\theta) \frac{\|\theta-\mu\|^2}{2\sigma^2} d\theta$$

Differentiating with respect to μ we get that:

$$\begin{aligned}\frac{\partial}{\partial \mu}\left(\int_\theta \hat{p}(\theta) \log((\sqrt{2\pi}\sigma)^d) d\theta + \int_\theta \hat{p}(\theta) \frac{\|\theta-\mu\|^2}{2\sigma^2} d\theta\right) &= \int_\theta \hat{p}(\theta) \frac{\partial}{\partial \mu}(\frac{\|\theta-\mu\|^2}{2\sigma^2}) d\theta \\ = \frac{1}{\sigma^2}\int_\theta \hat{p}(\theta)(\mu-\theta) d\theta &= \frac{1}{\sigma^2}\left(\mu - \mathbb{E}_{\boldsymbol{\theta}\sim\hat{p}}[\boldsymbol{\theta}]\right)\end{aligned}$$

Setting this equation to zero yields:

$$\mu = \mathbb{E}_{\boldsymbol{\theta}\sim\hat{p}}[\boldsymbol{\theta}]$$

Differentiating with respect to σ we get that:

$$\begin{aligned}&\frac{\partial}{\partial \sigma}\left(\int_\theta \hat{p}(\theta) \log((\sqrt{2\pi}\sigma)^d) d\theta + \int_\theta \hat{p}(\theta) \frac{\|\theta-\mu\|^2}{2\sigma^2} d\theta\right) \\ &= \int_\theta d\hat{p}(\theta) \frac{\sqrt{2\pi}}{\sqrt{2\pi}\sigma} d\theta - \int_\theta \hat{p}(\theta) \frac{\|\theta-\mu\|^2}{\sigma^3} d\theta \\ &= \frac{d}{\sigma} - \frac{1}{\sigma^3}\mathbb{E}_{\boldsymbol{\theta}\sim\hat{p}}[\|\boldsymbol{\theta}-\mu\|^2] = \frac{d}{\sigma} - \frac{1}{\sigma^3}\mathbb{E}_{\boldsymbol{\theta}\sim\hat{p}}[\|(\boldsymbol{\theta}-\mu)^\top(\boldsymbol{\theta}-\mu)] \\ &= \frac{d}{\sigma} - \frac{1}{\sigma^3}\left(\mathbb{E}_{\boldsymbol{\theta}\sim\hat{p}}[\boldsymbol{\theta}^\top\boldsymbol{\theta}] - 2\mu^\top\mathbb{E}_{\boldsymbol{\theta}\sim\hat{p}}[\boldsymbol{\theta}] + \mu^\top\mu\right)\end{aligned}$$

Likewise, setting it to zero (and considering that $\sigma > 0$) yields:

$$d\sigma^2 + \mu^\top\mu = \mathbb{E}_{\boldsymbol{\theta}\sim\hat{p}}[\boldsymbol{\theta}^\top\boldsymbol{\theta}]$$

We still need to show that $\mathbb{E}_{\boldsymbol{\theta}\sim q}[\boldsymbol{\theta}^\top\boldsymbol{\theta}] = d\sigma^2 + \mu^\top\mu$ we have:

$$\mathbb{E}_{\boldsymbol{\theta}\sim q}[\boldsymbol{\theta}^\top\boldsymbol{\theta}] = \mathrm{Tr}\left[\mathbb{E}_{\boldsymbol{\theta}\sim q}[\boldsymbol{\theta}^\top\boldsymbol{\theta}]\right] = \mathbb{E}_{\boldsymbol{\theta}\sim q}\left[\mathrm{Tr}[\boldsymbol{\theta}^\top\boldsymbol{\theta}]\right] = \mathbb{E}_{\boldsymbol{\theta}\sim q}\left[\mathrm{Tr}[\boldsymbol{\theta}\boldsymbol{\theta}^\top]\right] = \mathrm{Tr}\left[\mathbb{E}_{\boldsymbol{\theta}\sim q}[\boldsymbol{\theta}\boldsymbol{\theta}^\top]\right]$$

where we used that the expectation and the trace are linear operators and commute and the cyclic property of the trace. As q is a gaussian we know that:

$$\mathbb{E}_{\boldsymbol{\theta}\sim q}[\boldsymbol{\theta}\boldsymbol{\theta}^\top] = \sigma^2 I_d + \mathbb{E}_{\boldsymbol{\theta}\sim q}[\boldsymbol{\theta}]\mathbb{E}_{\boldsymbol{\theta}\sim q}[\boldsymbol{\theta}]^\top = \sigma^2 I_d + \mu\mu^\top$$

Hence,

$$\mathbb{E}_{\boldsymbol{\theta}\sim q}[\boldsymbol{\theta}^\top\boldsymbol{\theta}] = \mathrm{Tr}\left[\mathbb{E}_{\boldsymbol{\theta}\sim q}[\boldsymbol{\theta}\boldsymbol{\theta}^\top]\right] = \mathrm{Tr}\left[\sigma^2 I + \mu\mu^\top\right] = d\sigma^2 + \mu^\top\mu = \mathbb{E}_{\boldsymbol{\theta}\sim\hat{p}}[\boldsymbol{\theta}^\top\boldsymbol{\theta}]$$

(iv) By the chain rule we have that:

$$\nabla_\mu \log(Z(\mu,\sigma)) = \frac{\nabla_\mu Z}{Z} = \frac{\int_\theta p(x_t|\theta)\nabla_\mu q^{(t-1)}(\theta)}{\int_\theta p(x_t|\theta)q^{(t-1)}(\theta)}$$

We compute $\nabla_\mu q^{(t-1)}(\theta)$:

$$\nabla_\mu q^{(t-1)}(\theta) = \nabla_\mu \frac{1}{(\sqrt{2\pi}\sigma)^d} e^{-\frac{\|\theta-\mu\|^2}{2\sigma^2}} = \frac{1}{(\sqrt{2\pi}\sigma)^d} e^{-\frac{\|\theta-\mu\|^2}{2\sigma^2}} \frac{1}{\sigma^2}(\theta-\mu) = q^{(t-1)}(\theta)\frac{1}{\sigma^2}(\theta-\mu)$$

Hence,

$$\sigma^2\nabla_\mu \log(Z(\mu,\sigma)) = \frac{1}{Z}\int_\theta p(x_t|\theta)q^{(t-1)}(\theta)(\theta-\mu)d\theta = \int_\theta \frac{1}{Z}p(x_t|\theta)q^{(t-1)}(\theta)\theta d\theta - \frac{1}{Z}Z\mu$$

By re-ordreing the terms we get that:

$$\mathbb{E}_{\boldsymbol{\theta}\sim\hat{p}}[\boldsymbol{\theta}] = \mu + \sigma^2\nabla_\mu \log(Z(\mu,\sigma)) \tag{19}$$

(as desired)

Now we compute the derivative with respect to σ. Again by the chain rule we have that:

$$\nabla_\sigma \log(Z(\mu,\sigma)) = \frac{\nabla_\sigma Z}{Z} = \frac{\int_\theta p(x_t|\theta)\nabla_\sigma q^{(t-1)}(\theta)}{\int_\theta p(x_t|\theta)q^{(t-1)}(\theta)} \tag{20}$$

We compute $\nabla_\sigma q^{(t-1)}(\theta)$:

$$\nabla_\sigma q^{(t-1)}(\theta) = \nabla_\sigma \frac{1}{(\sqrt{2\pi}\sigma)^d} e^{-\frac{\|\theta-\mu\|^2}{2\sigma^2}} = q^{(t-1)}(\theta)\left(\frac{\|\theta-\mu\|^2}{\sigma^3} - \frac{d}{\sigma}\right)$$

Inserting this into equation (20) yields:

$$\sigma^3\nabla_\sigma \log(Z(\mu,\sigma)) = \mathbb{E}_{\boldsymbol{\theta}\sim\hat{p}}[\|\boldsymbol{\theta}-\mu\|^2] - d\sigma^2$$

By expanding $\|\boldsymbol{\theta}-\mu\|^2 = \boldsymbol{\theta}^\top\boldsymbol{\theta} - 2\boldsymbol{\theta}^\top\mu + \mu^\top\mu$ and using equation (19) we get:

$$\mathbb{E}_{\boldsymbol{\theta}\sim\hat{p}}[\boldsymbol{\theta}^\top\boldsymbol{\theta}] - \mathbb{E}_{\boldsymbol{\theta}\sim\hat{p}}[\boldsymbol{\theta}]^\top\mathbb{E}_{\boldsymbol{\theta}\sim\hat{p}}[\boldsymbol{\theta}] = \sigma^2 d - \sigma^4(\nabla_\mu^\top\nabla_\mu - \sigma^{-1}\nabla_\sigma)\log Z(\mu,\sigma)$$

(as desired)

(v) We combine the two previous parts to get:

$$\mu^{(t)} = \mathbb{E}_{\boldsymbol{\theta}\sim\hat{p}}[\boldsymbol{\theta}] = \mu^{(t-1)} + \sigma^{(t-1)2}\nabla_\mu \log(Z(\mu, \sigma_{t-1}))|_{\mu=\mu_{t-1}}$$

and,

$$\sigma^{(t)2}d = \mathbb{E}_{\boldsymbol{\theta}\sim\hat{p}}[\boldsymbol{\theta}^\top\boldsymbol{\theta}] = \sigma^2 d - \sigma^4(\nabla_\mu^\top\nabla_\mu - \sigma^{-1}\nabla_\sigma)\log Z(\mu,\sigma)|_{\mu=\mu_{t-1},\sigma=\sigma_{t-1}}$$

We leave it as an (rather tedious but very straight forward) exercise to the reader to compute the exact expressions for $\nabla_\mu \log(Z(\mu, \sigma_{t-1}))|_{\mu=\mu_{t-1}}$ and $\nabla_\sigma \log Z(\mu,\sigma)|_{\mu=\mu_{t-1},\sigma=\sigma_{t-1}}$ in order to derive the exact update equations. This involves inserting the distribution of the observation model $p(x_t|\theta)$.

Conceptually the current derivations suffice to showcase this new method.