Computer Engineering Department

Probabilistic Inference (Preliminaries)

Mohammad Hossein Rohban, Ph.D.

Hosein Hasani

Spring 2023

Outline

- Bayesian inference principles
- Probabilistic latent variable models
  - Expectation maximization
  - Variational inference
  - Hidden markov model

Recap: Uncertainty Estimation

Types of uncertainties:

- **Aleatoric uncertainty**:

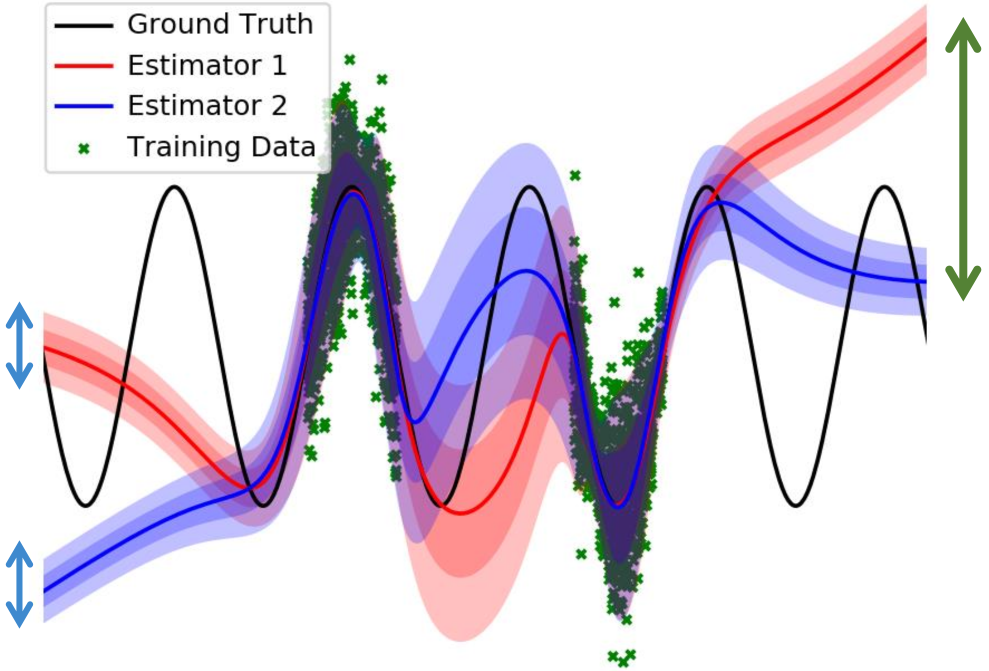
Inherent of the observations.

Both Bayesian and Frequentist Paradigms

- **Epistemic uncertainty**:

lack of knowledge in model hypothesis.

Just in Bayesian Paradigm!



[image credit: Chua, et al. nips 2018.]

Bayes' theorem

$$p(hypothesis|data) = \frac{p(data|hypothesis)p(hypothesis)}{p(data)}$$

- Bayes' rule provides a way of doing inference about hypothesis (uncertain quantities) from data (measured quantities).
- Learning and prediction can be seen as forms of inference.

Reverend Thomas Bayes (1702-1761)

Bayesian Learning and Prediction

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}$$

$\mathcal{D} = \{x_1, x_2, \dots, x_N\}$ dataset

$p(\mathcal{D}|\theta)$ likelihood of $\mathcal{D}$ with θ

$p(\theta)$ prior probability of θ

$p(\theta|\mathcal{D})$ posterior of θ given data $\mathcal{D}$

Posterior predictive:

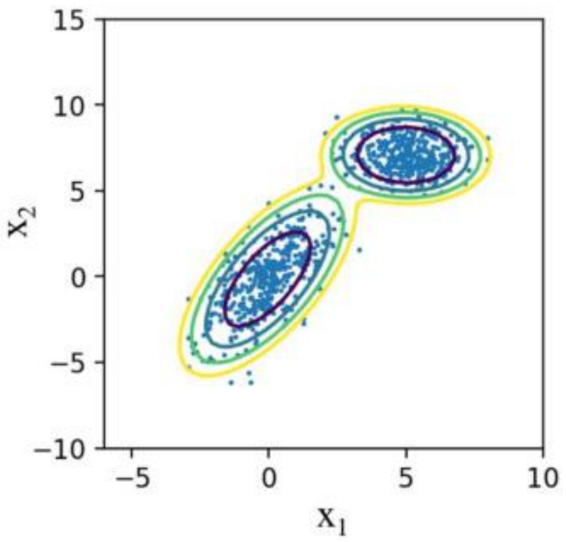
$$p(x|\mathcal{D}) = \int p(x,\theta|\mathcal{D})d\theta = \int p(x|\mathcal{D},\theta)p(\theta|\mathcal{D})d\theta = \int p(x|\theta)p(\theta|\mathcal{D})d\theta$$

marginalization chain rule conditional independence

Latent Variables

- Set of variables that are **unobservable (hidden)**, but influence observed data distributions.

- Examples of probabilistic latent variable models:
  - Gaussian Mixture Model (GMM)
  - Variational Autoencoder (VAE)
  - Hidden Markov Model (HMM)



GMM

Incomplete Likelihood

likelihood function:

$$L(\theta; x) = p(x|\theta) = \int p(x, z|\theta) dz$$

marginalization

$$= \int p(x|z, \theta) p(z|\theta) dz$$

chain rule

$$= \int p\left(x \middle| z, \theta_{x|z}\right) p(z|\theta_z) dz$$

conditional independence

Evidence Lower Bound

log-likelihood function:

$$\ell(\theta; x) = logp(x|\theta) = log \int p(x, z|\theta)dz$$

$$= log \int q(z) \frac{p(x, z|\theta)}{q(z)} dz$$

$$\geq \int q(z) log \frac{p(x, z|\theta)}{q(z)} dz$$

Jensen inequality

q(z): an arbitrary distribution!

$$= \int q(z) logp(x, z|\theta)dz - \int q(z) logq(z)dz$$

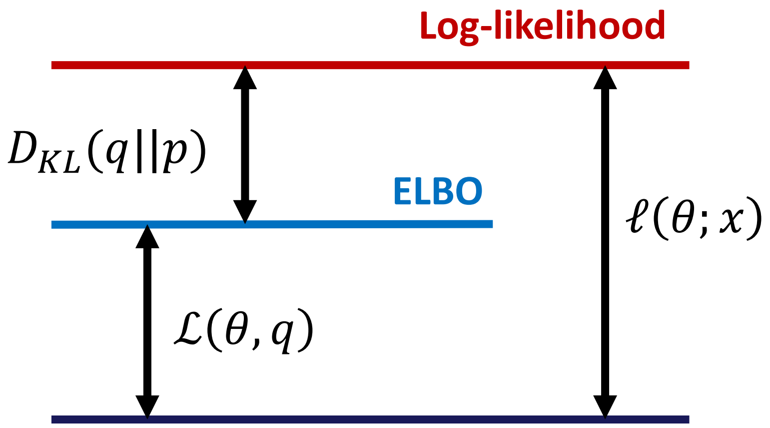
$$= \mathbb{E}_{z \sim q(z)}[\log p(x, z|\theta)] + \mathcal{H}(q)$$

Evidence Lower Bound

$\ell(\theta; x) \geq \mathbb{E}_{z \sim q(z)}[log\, p(x, z|\theta)] + \mathcal{H}(q) = \mathcal{L}(\theta, q)$ → evidence lower bound (ELBO)

We can show that:

$$\ell(\theta; x) = \mathcal{L}(\theta, q) + \underbrace{D_{KL}(q(z)||p(z|x, \theta))}_{\geq 0}$$



closer $q(z)$ to $p(z|x, \theta)$ → tighter lower bound ($\mathcal{L}(\theta, q)$) for $\ell(\theta; x)$

EM Algorithm

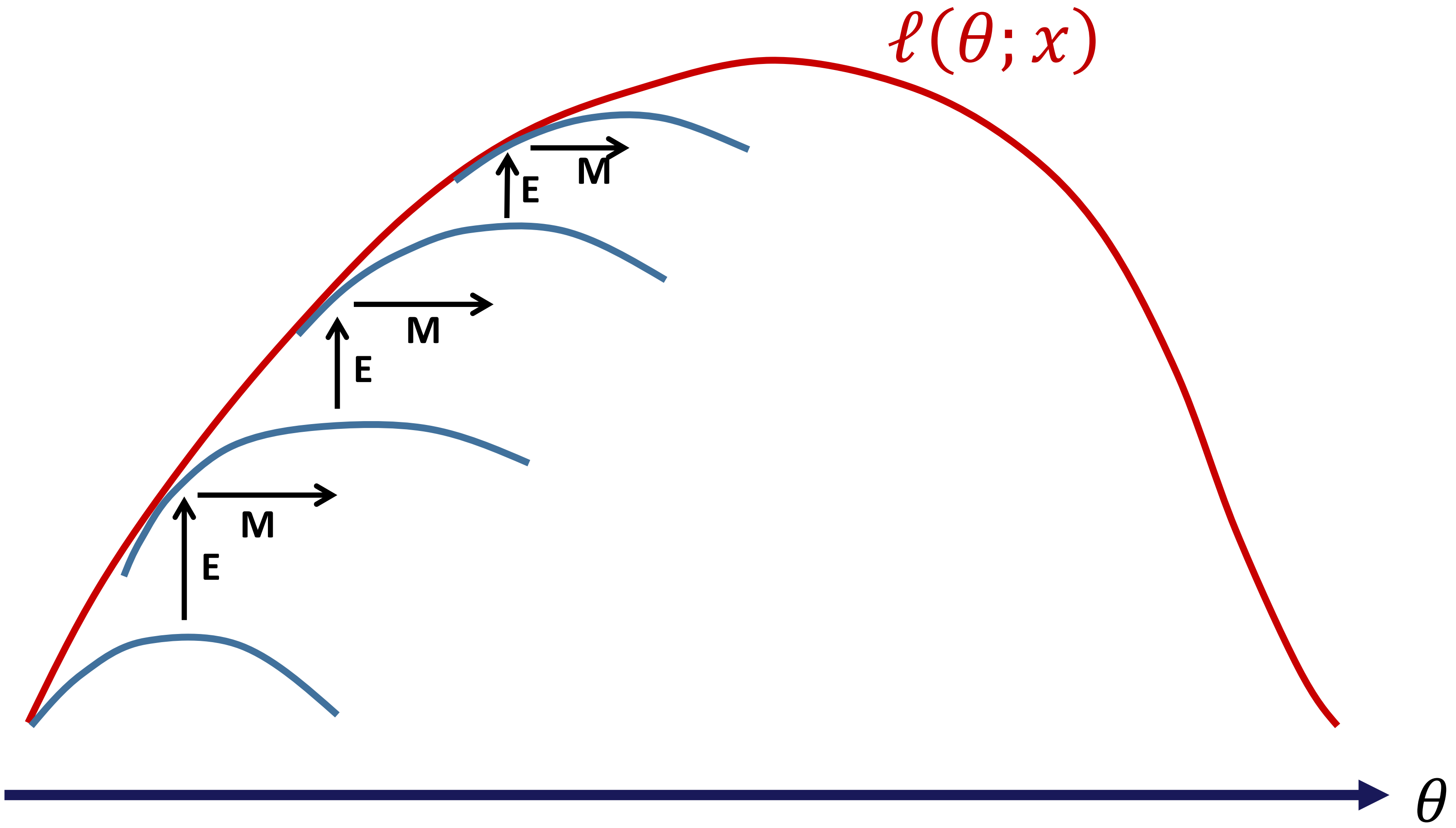
- Expectation step (E-step):
  Fill in **discrete** latent variables (z) given current parameters (θ) and data (x).

- Maximization step (M-step):
  Maximize likelihood as if latent variables were not hidden.

EM Algorithm

Initialize θ^1 with arbitrary values, and $t \leftarrow 1$
Iterate until convergence:
    E-step: calculate $p(z|x, \theta^t)$
    M-step: $\theta^{t+1} = argmax_\theta \mathbb{E}_{p(z|x,\theta^t)}[logp(x, z|\theta)]$
    $t \leftarrow t + 1$

EM Algorithm



EM for Continuous Latent Variables

Problem of E-step with Continuous latent variables:

$$p_\theta(z|x) = \frac{p_\theta(x|z)p_\theta(z)}{p_\theta(x)} = \frac{p_\theta(x|z)p_\theta(z)}{\boxed{\int p_\theta(x|z)p_\theta(z)}}$$

Intractable for many distributions!

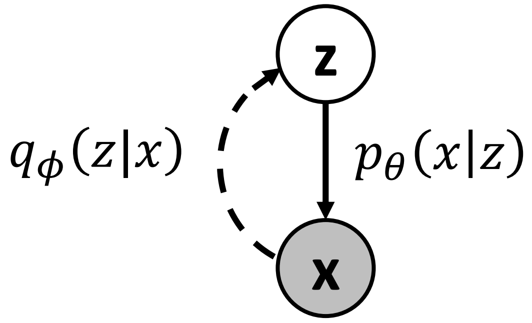
Variational Inference

Solution:

Use **variational distribution** $q_\phi(z|x)$ to minimize **ELBO:**

$$\color{red}{\ell(\theta;x)} \geq \mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x,z)] + \mathcal{H}\left(q_\phi(z|x)\right)$$

$$= \mathbb{E}_{z\sim q_\phi(z|x)}[log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)||p(z))$$



Variational Inference: Distribution Approximations

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{z \sim q_\phi(z|x)}[log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)||p(z))$$

- One common choice for $p(z)$: $\mathcal{N}(0, I)$

- One common choice for $q_\phi(z|x)$:

multivariate normal distribution parameterized by a neural network,

$$q\phi(z|x) = \mathcal{N}(z|\mu_\phi(x), \sigma_\phi(x))$$

Variational Inference: Optimization

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{z \sim q_\phi(z|x)}[log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)||p(z))$$

Optimization

for each x_i (or mini-batch):

calculate $\nabla_\theta \mathcal{L}(\theta, \phi)$:

sample $z \sim q_\phi(z|x_i)$

$\nabla_\theta \mathcal{L}(\theta, \phi) \approx \nabla_\theta log p_\theta(x_i|z)$

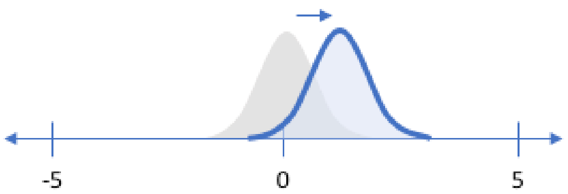
$\theta \leftarrow \theta + \nabla_\theta \mathcal{L}(\theta, \phi)$

$\phi \leftarrow \phi + \nabla_\phi \mathcal{L}(\theta, \phi)$

Variational Inference: Optimization

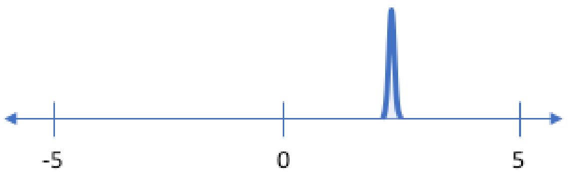
$$\mathcal{L}(\theta, \phi) = \underbrace{\mathbb{E}_{z\sim q_\phi(z|x)}[log p_\theta(x|z)]}_{\text{reconstruction loss}} - \underbrace{D_{KL}(q_\phi(z|x)||p(z))}_{\text{regularization loss}}$$

Penalizing reconstruction loss encourages the distribution to describe the input



Our distribution deviates from the prior to describe some characteristic of the data

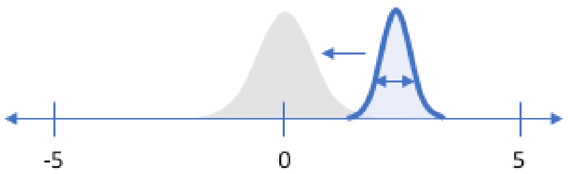
Without regularization, our network can "cheat" by learning narrow distributions



With a small enough variance, this distribution is effectively only representing a single value

Penalizing KL divergence acts as a regularizing force

Attract distribution to have zero mean



Ensure sufficient variance to yield a smooth latent space

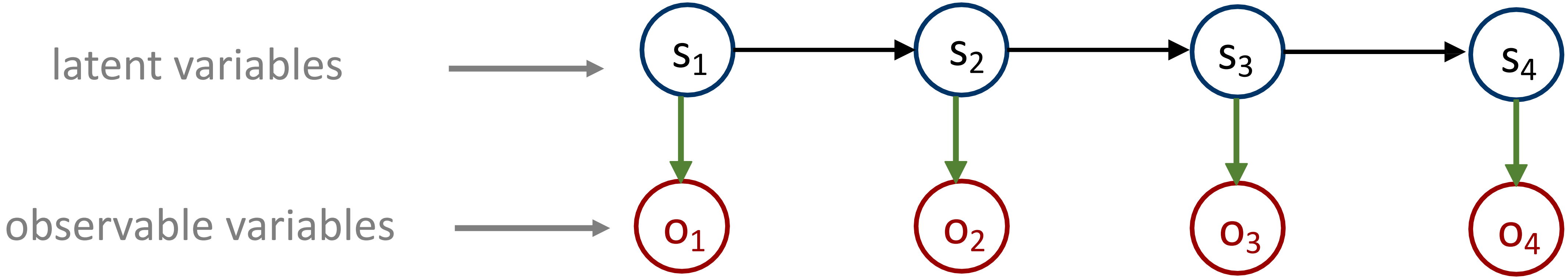
image credit: Jeremy Jordan

Hidden Markov Model (HMM)

Temporal generative model with discrete latent variables.

Applications:

- Speech recognition
- Robot localization
- Communication systems



Hidden Markov Model (HMM)

Assumptions for latent space:

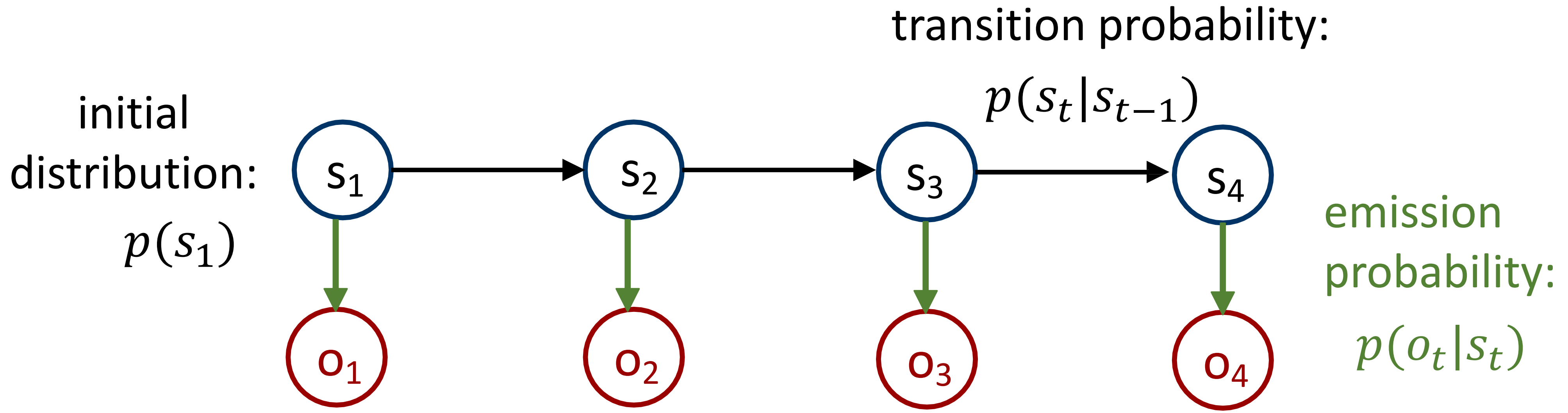
s_t: state (latent variable)
o_t: observation

- Markov assumption (first order):

$$p(s_t | s_{1:t-1}) = p(s_t | s_{t-1})$$

- Stationary:

$$p(s_t = i | s_{t-1} = j) = p(s_{t+1} = i | s_t = j) \ \ \forall t, i, j$$



Type of Inference Problems in HMM

- Monitoring (filtering): $p(s_t|o_{1:t})$
- Prediction: $p(s_{t+k}|o_{1:t})$
- Smoothing: $p(s_t|o_{1:T})$ where $t < T$
- Most likely explanation:

$$argmax_{s_{1:T}} p(s_{1:T}|o_{1:T})$$

Monitoring (Filtering)

Recursive computation:

$$p(s_t|o_{1:t}) \propto p(o_t|s_t, o_{1:t-1})p(s_t|o_{1:t-1})$$ Bayes' rule

$$= p(o_t|s_t)p(s_t|o_{1:t-1})$$ conditional independence

$$= p(o_t|s_t) \sum_{s_{t-1}} p(s_t, s_{t-1}|o_{1:t-1})$$ marginalization

$$= p(o_t|s_t) \sum_{s_{t-1}} p(s_t|s_{t-1}, o_{1:t-1})\, p(s_{t-1}|o_{1:t-1})$$ chain rule

$$= p(o_t|s_t) \sum_{s_{t-1}} p(s_t|s_{t-1})\, p(s_{t-1}|o_{1:t-1})$$ conditional independence

Forward Algorithm

Compute $p(s_t|o_{1:t})$ recursively by forward computation:

Forward Algorithm

$p(s_1|o_1) \propto p(o_1|s_1)p(s_1)$
for $k = 2$ to t do
$\quad p(s_k|o_{1:k}) \propto p(o_k|s_k) \sum_{s_{k-1}} p(s_k|s_{k-1})\, p(s_{k-1}|o_{1:k-1})$

Smoothing

$p(s_t|o_{1:T})$ where $t < T$

$$p(s_t|o_{1:T}) \propto p(o_{t+1:T}|s_t)p(s_t|o_{1:t})$$

Bayes' rule

computed by backward algorithm

computed by forward algorithm

Backward Probabilities

$$p(o_{t+1:T}|s_t) = \sum_{s_{t+1}} p(o_{t+1:T}, s_{t+1}|s_t)$$ marginalization

$$= \sum_{s_{t+1}} p(o_{t+1:T}|s_{t+1}, s_t)p(s_{t+1}|s_t)$$ chain rule

$$= \sum_{s_{t+1}} p(o_{t+1:T}|s_{t+1})p(s_{t+1}|s_t)$$ conditional independence

$$= \sum_{s_{t+1}} p(o_{t+1}, o_{t+2:T}|s_{t+1})p(s_{t+1}|s_t)$$

$$= \sum_{s_{t+1}} p(o_{t+2:T}|s_{t+1})p(o_{t+1}|s_{t+1})p(s_{t+1}|s_t)$$ conditional independence

Backward Algorithm

Compute $p(o_{t+1:T}|s_t)$ recursively by backward computation:

Backward Algorithm

$p(o_{k>T}|s_T) = 1$

from $\mathrm{k} = T - 1$ to t do

$$p(o_{k+1:T}|s_k) = \sum_{s_{k+1}} p(\, o_{k+2:T}|s_{k+1})p(o_{k+1}|s_{k+1})p(s_{k+1}|s_k)$$

Prediction

$$p(s_{t+k}|o_{1:t}) = \sum_{s_{t+k-1}} p(s_{t+k}, s_{t+k-1}|o_{1:t})$$

marginalization

$$= \sum_{s_{t+k-1}} p(s_{t+k}|\, s_{t+k-1}, o_{1:t}) p(s_{t+k-1}|o_{1:t})$$

chain rule

$$= \sum_{s_{t+k-1}} p(s_{t+k}|s_{t+k-1})\, p(s_{t+k-1}|o_{1:t})$$

conditional independence