Assorted Topics in Probabilistic ML (3)

CS772A: Probabilistic Machine Learning

Piyush Rai

Plan for today

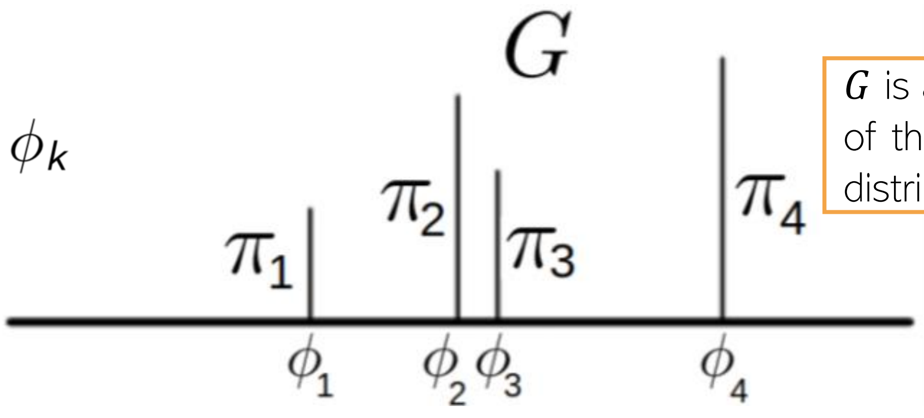
- Assorted Topics
  - Nonparametric Bayesian methods (contd)
  - Probabilistic Models for Sequential Data
    - A brief idea
  - Probabilistic Numerics

Being Nonparametric using Models that have a Shrinkage Effect

Mixture Models: Another Construction

- Consider a finite mixture model with K components with params $(\mu_k, \Sigma_k)_{k=1}^K$

$(\mu_k, \Sigma_k) = \phi_k$



G is a representation of this mixture distribution

Defined by K locations or "atoms" with parameters $\{\phi_k\}_{k=1}^K$ with respective selection probabilities $\{\pi_k\}_{k=1}^K$

$$G = \sum_{k=1}^{K} \pi_k \delta_{\phi_k}$$

- In the finite case, we can assume $\boldsymbol{\pi} = [\pi_1, \dots, \pi_K]$ and $\boldsymbol{\pi} \sim \text{Dirichlet}\left(\frac{\alpha}{K}, \dots, \frac{\alpha}{K}\right)$
- We can make it a nonparametric model by making $\boldsymbol{\pi}$ an infinite-dimensional vector

In practice, only a finite of these will have nonzero values, and others will shrink to very small (or zero), as we will see

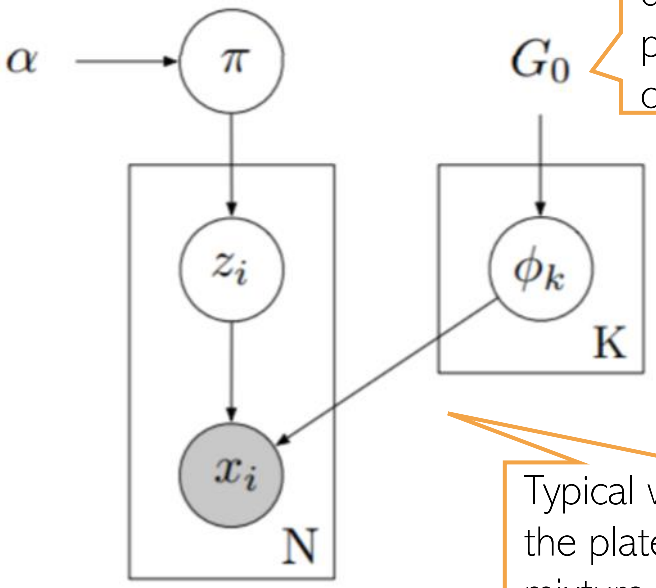
$$\pi_1, \pi_2, \pi_3, \dots, \qquad \sum_{k=1}^{\infty} \pi_k = 1$$

Indeed. Called a "Dirichlet Process"

Related: "Stick-breaking Process"

- How to construct such a vector? Is there an infinite dimensional Dirichlet distribution?

Mixture Models: Two Equivalent Views



Prior (a.k.a. "base distribution" for the parameters of each mixture component

Example: G_0 can be NIW if each component is a Gaussian and $\phi_k = (\mu_k, \Sigma_k)$

Typical way of showing the plate notation of a mixture model

$$\boldsymbol{\pi} \sim \text{Dirichlet}\left(\frac{\alpha}{K}, \dots, \frac{\alpha}{K}\right)$$

$$\phi_k \sim G_0 \qquad k = 1,2, \dots, K$$

$$z_i \sim \text{multinoulli}(\boldsymbol{\pi}) \qquad i = 1,2, \dots, N$$

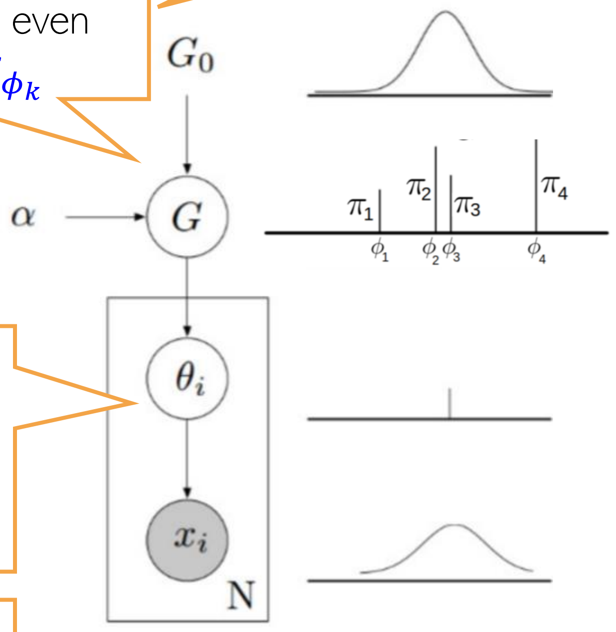
$$x_i \sim p(x|\phi_{z_i}) \qquad i = 1,2, \dots, N$$

Similar representation even when $G = \sum_{k=1}^{\infty} \pi_k \, \delta_{\phi_k}$

But how to construct such a G distribution with potentially infinite components?

No explicit cluster ids; instead, θ_i denotes the param of the distribution which will generate x_i

Since G is discrete, there will at most be K distinct θ_i's, thereby achieving clustering



$$\boldsymbol{\pi} \sim \text{Dirichlet}\left(\frac{\alpha}{K}, \dots, \frac{\alpha}{K}\right)$$

$$\phi_k \sim G_0 \qquad k = 1,2, \dots, K$$

$$G = \sum_{k=1}^{K} \pi_k \, \delta_{\phi_k}$$

$$\theta_i \sim G \qquad i = 1,2, \dots, N$$

$$x_i \sim p(x|\theta_i) \qquad i = 1,2, \dots, N$$

Stick-Breaking Process (Sethuraman'94)

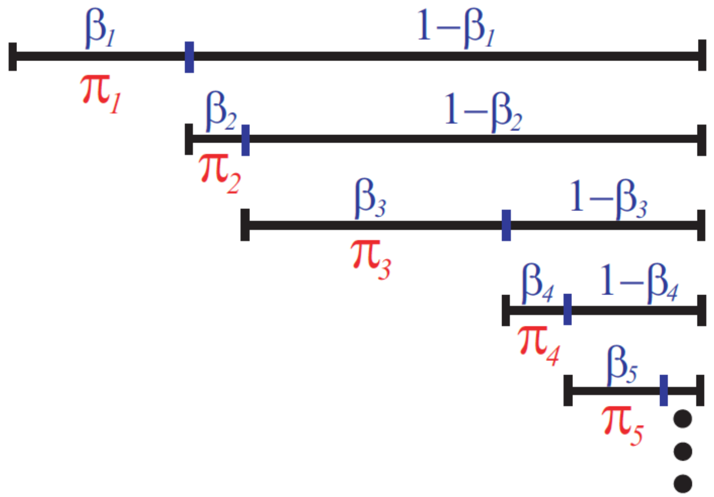
SBP gives us a way to construct infinite dimensional Dirichlet distribution known as the "Dirichlet Process"

- Recursively break a length 1 stick into two pieces
- Assume breaking point in each round is drawn from a Beta distribution

$$\beta_k \sim \text{Beta}(1,\alpha) \qquad k = 1,\ldots,\infty$$

$$\pi_1 = \beta_1$$

$$\pi_k = \beta_k \prod_{\ell=1}^{k-1}(1-\beta_\ell) \qquad k = 2,\ldots,\infty$$



- Can show that $\sum_{k=1}^{\infty}\pi_k - 1 \rightarrow 0$ which is what we want
- We can now have a "nonparametric/infinite" mixture distribution $G = \sum_{k=1}^{\infty}\pi_k\,\delta_{\phi_k}$
- "Location/atoms" ϕ_k can be drawn from a "base" distr G_0, say NIW if $\phi_k = (\mu_k, \Sigma_k)$
- We basically replaced the Dirichlet prior on $\boldsymbol{\pi}$ by a Stick-Breaking Process (SBP) prior

Infinite Dimensional Dirichlet

- Drawing from an infinite-dim Dirichlet would give an infinite-dim prob. vector

$$\boldsymbol{\pi} = [\pi_1, \pi_2, \pi_3, \ldots]$$

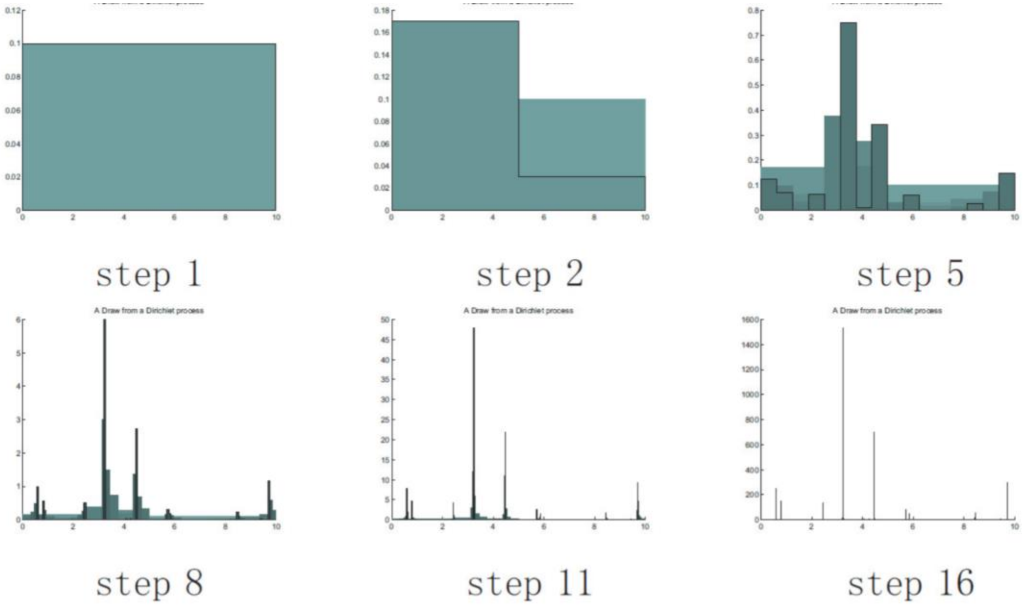
- We can construct this vector to have very few entries as nonzero
- Consider recursively drawing from a Dirichlet as defined below

$$\begin{aligned}
1 &\sim \text{Dirichlet}(\alpha) \\
(\pi_1, \pi_2) &\sim \text{Dirichlet}(\alpha/2, \alpha/2) \\
(\pi_1\pi_{11}, \pi_1\pi_{12}, \pi_2\pi_{21}, \pi_2\pi_{22}) &\sim \text{Dirichlet}(\alpha/4, \alpha/4, \alpha/4, \alpha/4)
\end{aligned}$$

As the concentration parameter gets smaller and smaller, the split of values in LHS get more and more skewed

Therefore, after doing the above a few times, the π vector will only have a very few entries as nonzero and in the infinite-sized π, there will only be a finite many nonzero entries, and rest will be zero

This is basically what happens in the case of Dirichlet Process / Stick-Breaking Process



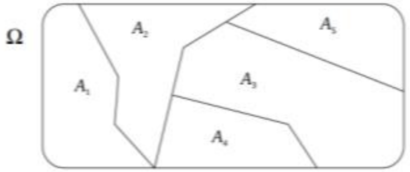
Dirichlet Process - Formally

SBP gives an explicit way to construct "Dirichlet Process"

- A Dirichlet Process DP(α, G_0) defines a distribution over distributions
  - So $G \sim$ DP(α, G_0) will give us a distribution
  - α : concentration param, G_0: base distribution (=mean of DP)
  - Large α means $G \to G_0$
- **Fact 1:** If $G \sim$ DP(α, G_0) then any finite dim. marginal of G is Dirichlet distributed

$$[G(A_1), \ldots, G(A_K)] \sim \text{Dirichlet}(\alpha G_0(A_1), \ldots, \alpha G_0(A_K))$$

for any finite partition $A_1, \ldots, A_K$ of the space Ω (Ferguson, 1973)



ϕ_k's are i.i.d. draws from the base distribution G_0

- **Fact 2:** Any G drawn from DP(α, G_0) will be of the form $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$ (Sethuraman, 1994)
- **Fact 3:** G is a discrete dist, i.e., only a few π_k's will be significant

Another NPBayes Prior: Multiplicative Gamma Process

- Consider the SVD-style probabilistic model with an *a priori* unbounded K

$$\mathbf{X} = \sum_{k=1}^{\infty} \lambda_k \boldsymbol{u}_k \boldsymbol{v}_k^\top$$

- Consider the following prior on each "singular values" λ_k

$$\begin{aligned}\lambda_k &\sim \mathcal{N}(0, \tau_k^{-1}) \\ \tau_k &= \prod_{\ell=1}^{k} \delta_\ell \\ \delta_\ell &\sim \text{Gamma}(\alpha, 1) \quad \text{where } \alpha > 1\end{aligned}$$

Precision keeps on getting larger and larger as k grows (thus variance keeps getting small and smaller)

Thus $\mathbb{E}[\delta_\ell] = \alpha$ (greater than 1 in expectation)

- In practice we can set K to be a sufficiently very large
  - Due to the shrinkage property, only a finite many λ_k will be nonzero
  - The nonzero λ_k's will dictate the effective K

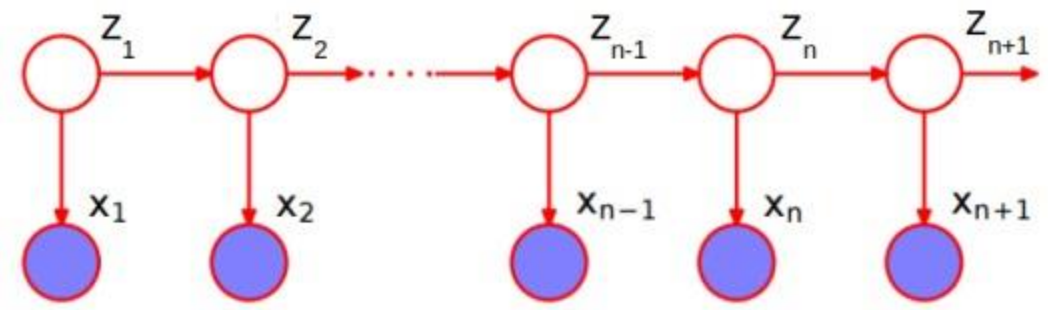
Summary of NPBayes

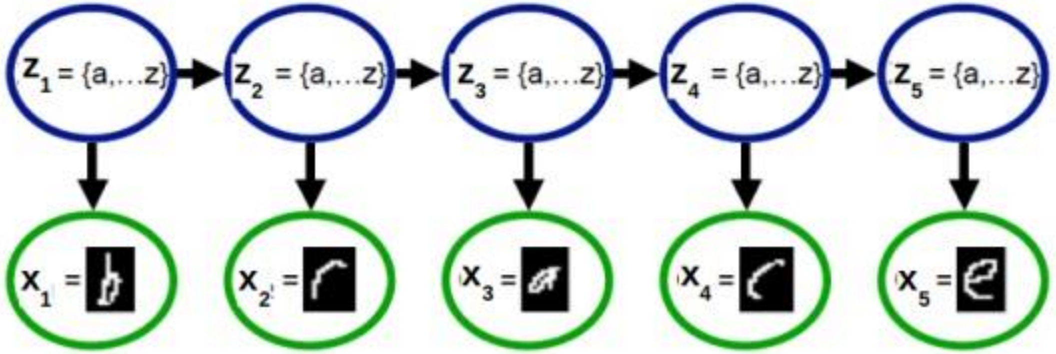
- We saw some nonparametric Bayesian models (mainly used in unsup learning)
  - CRP/Dirichlet Process: For clustering problems
  - Multiplicative Gamma Process: For SVD-like matrix factorization
- Many applications of these models to solve a wide range of problems
- Also saw GP which is another example of a nonparametric Bayesian model
  - GPs are used for function approximation problems (both supervised and unsup. learning)
- These are only some of the examples of nonparametric Bayesian models
  - Many other such nonparametric Bayesian models for other problems in machine learning
  - ”A tutorial on Bayesian nonparametric models“ Gershman and Blei, 2011) is a nice survey
- Rich theory based on stochastic processes (beyond the scope of this course)
- Inspired other non-probabilistic algos, e.g., Using Dirichlet Process Mixture Model to get a K-means like clustering algorithm (DP-means) which doesn’t require K

Probabilistic Models for Sequential Data

Latent Variable Models for Sequential Data

- Task: Given a sequence of observations, infer the latent state of each observation





Observation model

$$\boldsymbol{x}_n|\boldsymbol{z}_n \sim p(\boldsymbol{x}_n|\boldsymbol{z}_n) \quad \text{(i.i.d. draws of } \boldsymbol{x}_n \text{ given } \boldsymbol{z}_n\text{)}$$

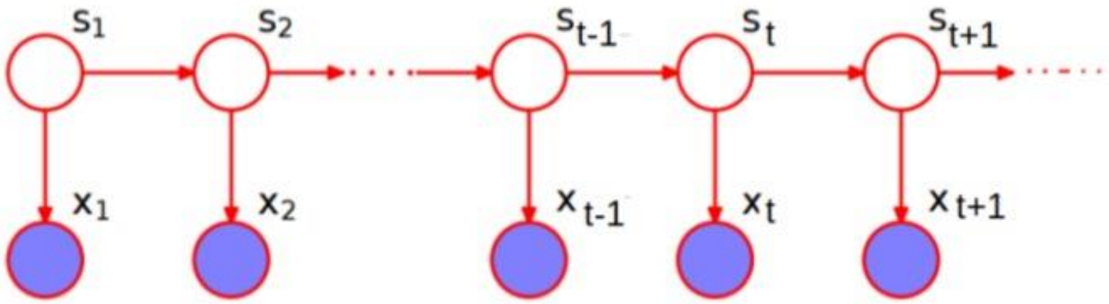
State-transition model

$$\boldsymbol{z}_n|\boldsymbol{z}_{n-1} \sim p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1}) \quad \text{(first-order dependence b/w } \boldsymbol{z}_n\text{'s)}$$

- If z_n's are discrete, we have a hidden Markov model (HMM) $p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1} = \ell) = \text{multinoulli}(\boldsymbol{\pi}_\ell)$
- If z_n's are real-valued, we have a state-space model (SSM) $p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1}) = \mathcal{N}(\mathbf{A}\boldsymbol{z}_{n-1}, \mathbf{I}_K)$

State-Space Models

- In the most general form, the state-transition and observation models of an SSM



Using ‘**s**’ instead of ‘**z**’ to refer to states

Using ‘t’ to denote the ‘time-step’

HMM is similar to SSM except the state-transition model is a discrete distribution

g_t, h_t can be linear or nonlinear functions

$$\begin{aligned} \boldsymbol{s}_t|\boldsymbol{s}_{t-1} &= g_t(\boldsymbol{s}_{t-1}) + \epsilon_t && \text{(must be a cont. dist. over } \boldsymbol{s}_t\text{)} \\ \boldsymbol{x}_t|\boldsymbol{s}_t &= h_t(\boldsymbol{s}_t) + \delta_t && \text{(can be any dist. over } \boldsymbol{x}_t\text{)} \end{aligned}$$

- Assuming Gaussian noise in the state-transition and observation models

This is a Gaussian SSM

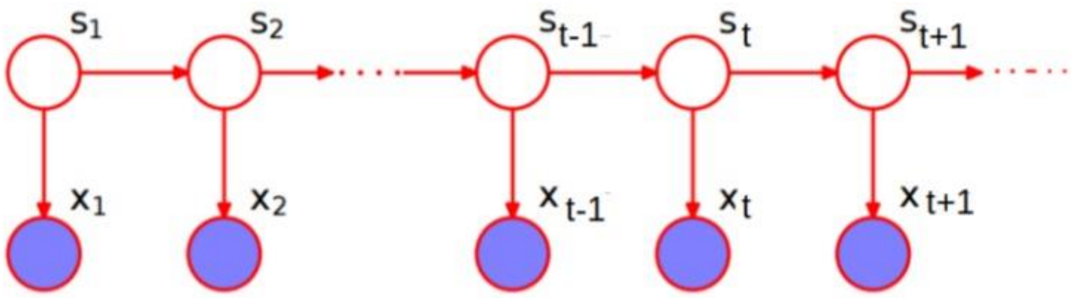
$$\begin{aligned} \boldsymbol{s}_t|\boldsymbol{s}_{t-1} &\sim \mathcal{N}(\boldsymbol{s}_t|g_t(\boldsymbol{s}_{t-1}), \mathbf{Q}_t) \\ \boldsymbol{x}_t|\boldsymbol{s}_t &\sim \mathcal{N}(\boldsymbol{x}_t|h_t(\boldsymbol{s}_t), \mathbf{R}_t) \end{aligned}$$

If g_t, h_t, Q_t, R_t are independent of t then it is called a stationary model

g_t, h_t, Q_t, R_t may be known or can be learned

Typical Inference Task for Gaussian SSM

- One of the key tasks: Given sequence $x_1, x_2, \ldots, x_T$, infer latent $s_1, s_2, \ldots, s_T$



- Usually two ways of inferring the latent states
  - Infer $p(s_t|x_1, x_2, \ldots, x_t)$: Called the "filtering" problem

$$p(\mathbf{s}_t|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t) \propto \underbrace{p(\mathbf{x}_t|\mathbf{s}_t)}_{\mathcal{N}(\mathbf{x}_t|\mathbf{B}\mathbf{s}_t, \mathbf{R})} \int \underbrace{p(\mathbf{s}_t|\mathbf{s}_{t-1})}_{\mathcal{N}(\mathbf{s}_t|\mathbf{A}\mathbf{s}_{t-1}, \mathbf{Q})} p(\mathbf{s}_{t-1}|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{t-1}) d\mathbf{s}_{t-1}$$

Turns out to be another Gaussian

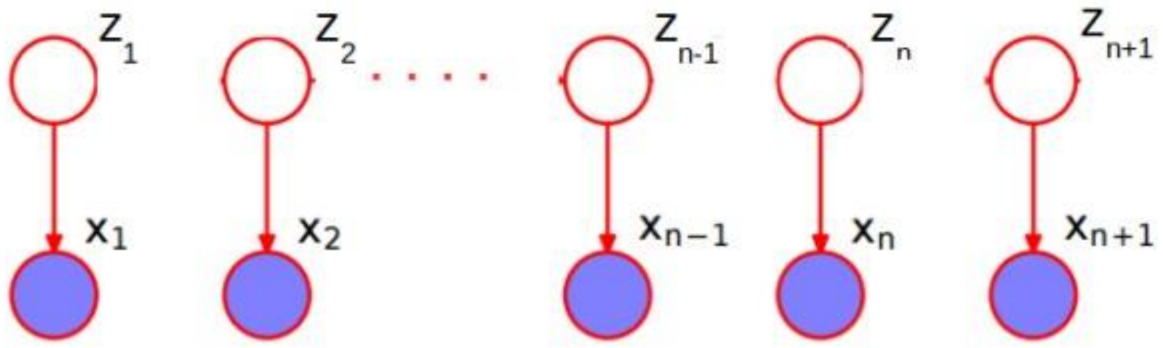
A Gaussian

Kalman Filtering is a popular algorithm for a linear Gaussian SSM

  - Infer $p(s_t|x_1, x_2, \ldots, x_t, \ldots, x_T)$: Called the "smoothing" problem
- Some other tasks one can solve for using an SSM
  - Predicting future states $p(s_{t+h}|x_1, x_2, \ldots, x_t)$ for $h \geq 1$, given observations thus far
  - Predicting future observations $p(x_{t+h}|x_1, x_2, \ldots, x_t)$ for $h \geq 1$, given observations thus far

A Special Case

- What if we have i.i.d. latent states, i.e.,. $p(z_n|z_{n-1}) = p(z_n)$?



- Discrete case (HMM) becomes a simple mixture model $p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1} = \ell) = p(\boldsymbol{z}_n) = \text{multinoulli}(\boldsymbol{\pi})$
- Real-valued case (SSM) becomes a PPCA model $p(\boldsymbol{z}_n|\boldsymbol{z}_{n-1}) = p(\boldsymbol{z}_n) = \mathcal{N}(\mathbf{0}, \mathbf{I_K})$ or $\mathcal{N}(\boldsymbol{\mu}, \Psi)$
- Inference algos for HMM/SSM are thus very similar to that of mixture models/PPCA
  - Only main difference is how the latent variables z_n's are inferred since they aren't i.i.d.
  - E.g., if using EM, only E step needs to change (Bishop Chap 13 has EM for HMM and SSM)

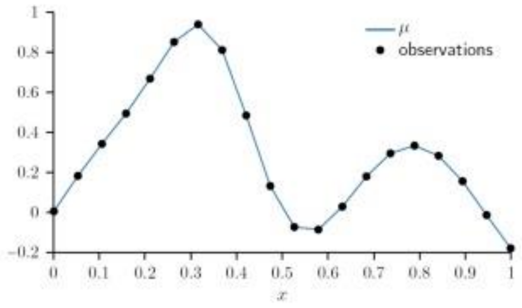
Some other topics (not covered in the course)

- Reinforcement Learning
- Probabilistic Numerics: Treating numerical problem as one of statistical inference
  - An Example: Numerical integration

How to perform expensive/intractable integrals

$$\int_0^1 \exp\left(-\frac{(x-0.35)^2}{2(0.1)^2}\right) + \frac{\sin(10x)}{3}\, dx$$

Where to do the function evaluations when using numerical approximations



What's our uncertainty in the estimate of the integral

$$\int_0^1 f(x)\, dx \approx 0.3104$$

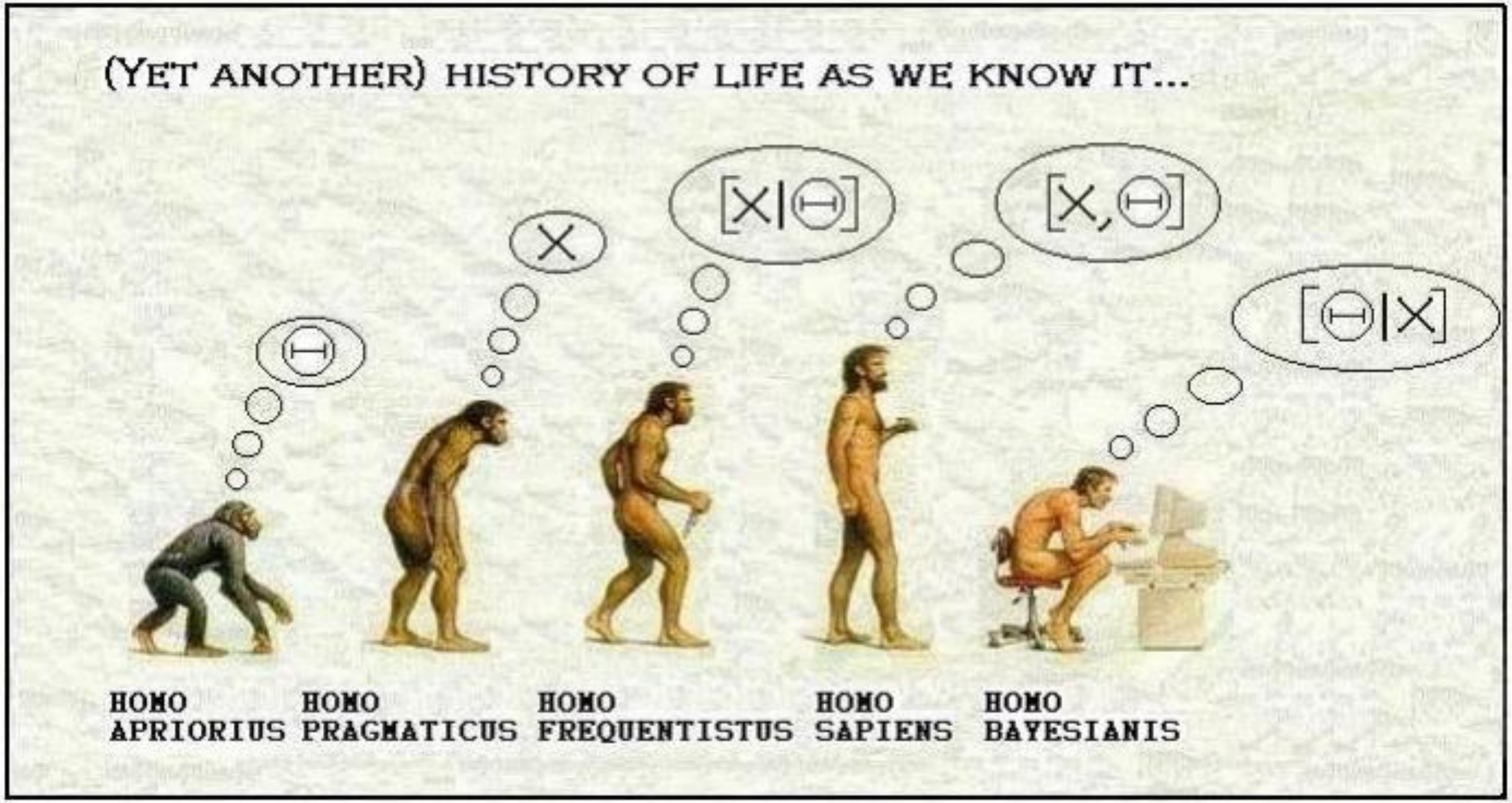
  - Many others: optimization, solution of ODE/PDE, solution of linear systems, eigenvalue problems

https://www.probabilistic-numerics.org/

Conclusion

- Probabilistic modeling provides a natural way to think about models of data
- Many benefits as compared to non-probabilistic approaches
  - Easier to model and leverage uncertainty in data/parameters
  - Principle of marginalization while making prediction
  - Easier to encode prior knowledge about the problem (via prior/likelihood distributions)
  - Easier to handle missing data (by marginalizing it out if possible, or by treating as latent variable)
  - Easier to build complex models can be neatly combining/extending simpler probabilistic models
  - Easier to learn the "right model" (hyperparameter estimation, nonparametric Bayesian models)
  - .. and various other benefits as we saw during this course
- Fast-moving field, lots of recent advances on new models and inference methods
  - The course is an attempt to guide you into exploring the area further

Thank You!



(YET ANOTHER) HISTORY OF LIFE AS WE KNOW IT...
[X|Θ]
[X,Θ]
[Θ|X]
HOMO APRIORIUS
HOMO PRAGMATICUS
HOMO FREQUENTISTUS
HOMO SAPIENS
HOMO BAYESIANIS