高级机器学习

课程简介

赵静
jzhao@cs.ecnu.edu.cn

Outline

- Advanced Models
  - GP Related Models
  - Sequential Models
  - Deep Neural Networks
- Approximate Inference and Optimization
  - Variational Inference
  - Sampling Methods
  - Stochastic Optimization

课程形式

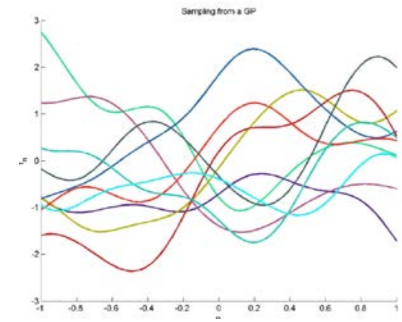
- 授课（前15周）+学生分享（后3周）
- 考核：平时成绩40%+最终报告60%
- 平时成绩
  - 两次平时作业20%
  - 课堂报告20%：学生自选内容，PPT介绍（约20分钟，根据选课人数调整），同学之间相互打分
- 最终报告：以论文形式提交
  - 机器学习相关
  - 综述论文（参考文献30篇+）/近5年顶会刊算法复现（需增加数据集）/研究报告（有创新性）

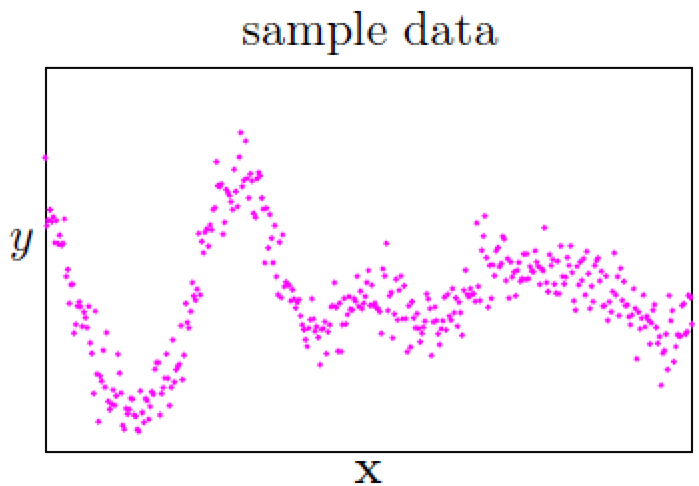
Advanced Models (I)

- GP Related Models
  - Gaussian processes (regression or classification)
  - Gaussian process latent variable models (dimensionality reduction)
  - Deep Gaussian processes (deep model)
  - Multi-view Gaussian processes (multi-view)
  - Mixtures of Gaussian process (multi-modal)

Gaussian Process Regression

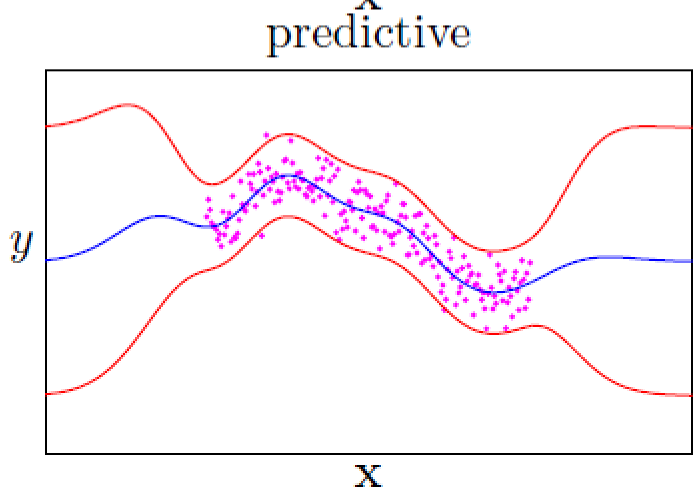
Gaussian observation noise: $y_n = f_n + \epsilon_n$, where $\epsilon_n \sim \mathcal{N}(0, \sigma^2)$





marginal likelihood

$$p(\mathbf{y}|\mathbf{X}) = \mathcal{N}(\mathbf{0}, \mathbf{K}_N + \sigma^2 \mathbf{I})$$



predictive distribution

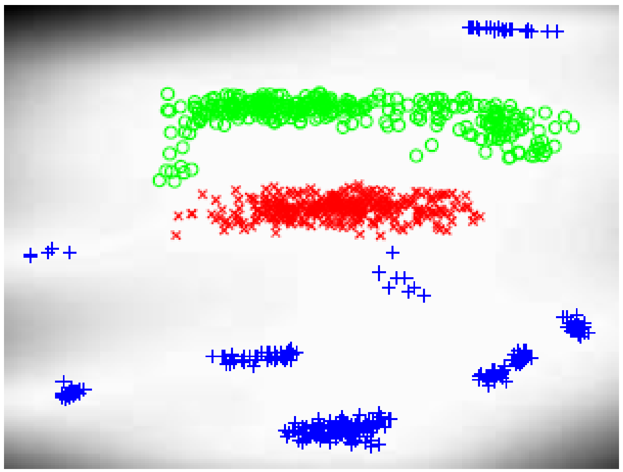
$$p(y_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N}(\mu_*, \sigma_*^2)$$

$$\mu_* = \mathbf{K}_{*N}(\mathbf{K}_N + \sigma^2 \mathbf{I})^{-1}\mathbf{y}$$

$$\sigma_*^2 = K_{**} - \mathbf{K}_{*N}(\mathbf{K}_N + \sigma^2 \mathbf{I})^{-1}\mathbf{K}_{N*} + \sigma^2$$

Gaussian Process Latent Variable Models

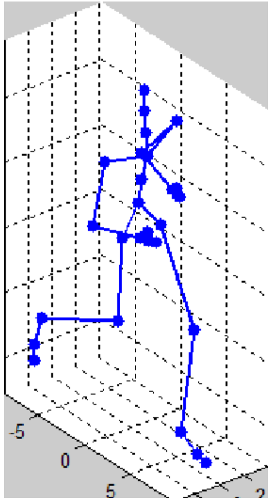
- Low dimensional visualization
- Predict missing values



Bayesian GP-LVM, $q = 10$ (2D projection)

Gaussian Process Latent Variable Models

- Gaussian process dynamical systems
- Sequential data modeling
- Prediction
- classification



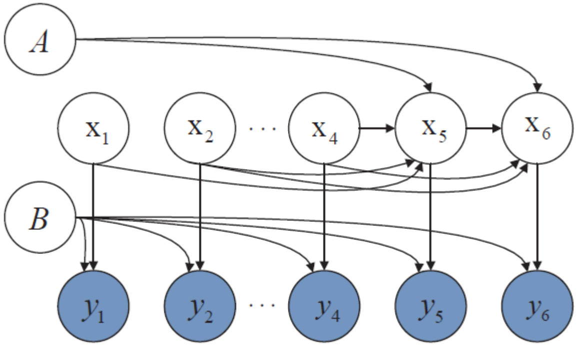
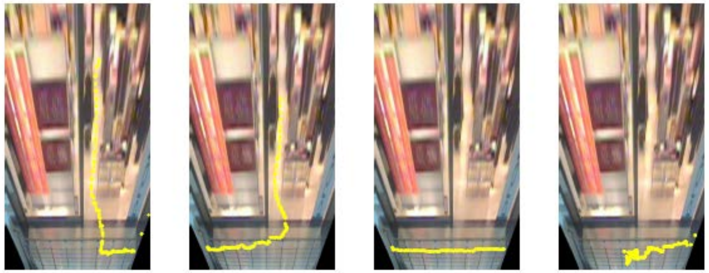


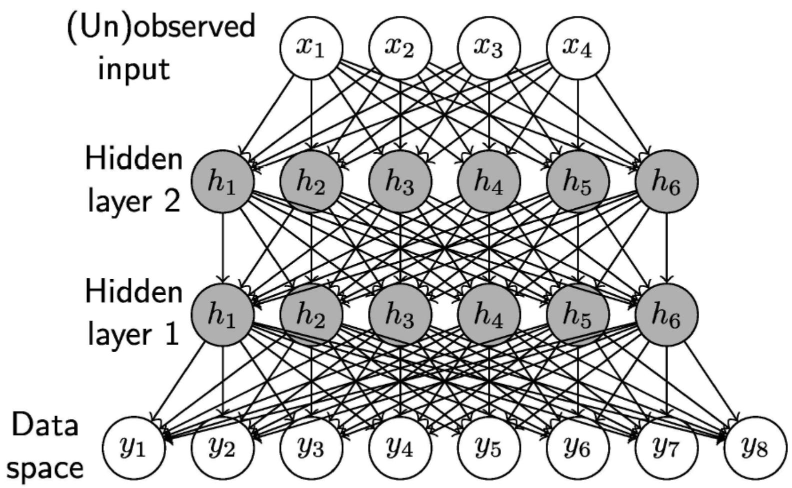
Figure: 四阶动态系统的示意图



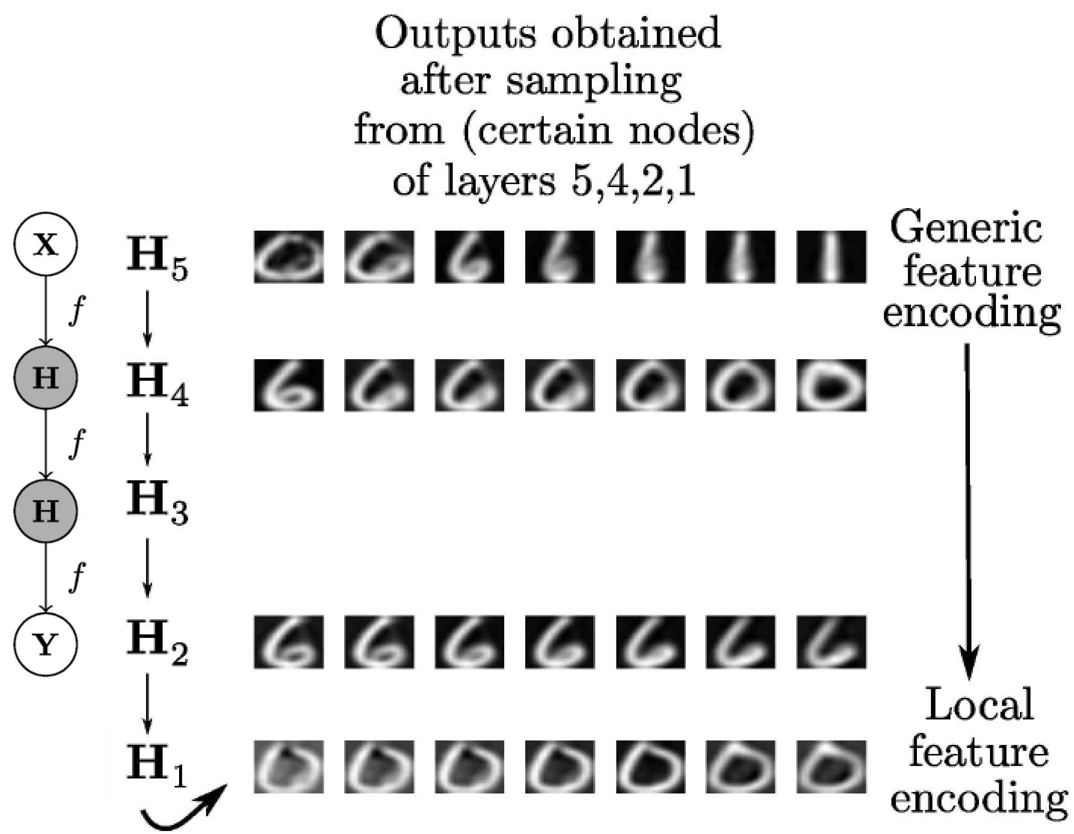
(a) (b) (c) (d)

Deep Gaussian Processes

- Deep representation learning
- Robust
- Easy sampling

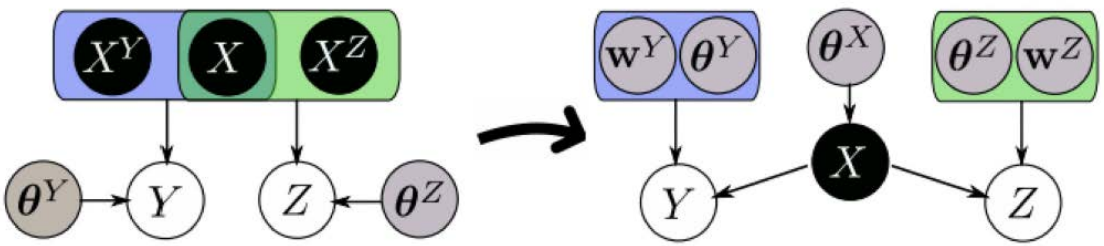


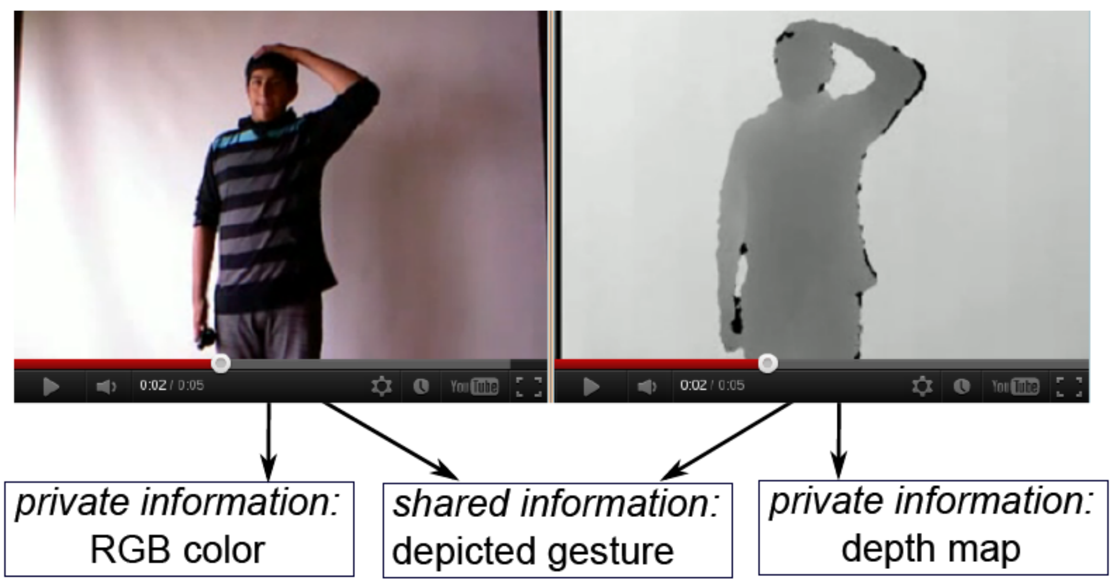
$\mathbf{Y} = f(f(\cdots f(\mathbf{X})))$

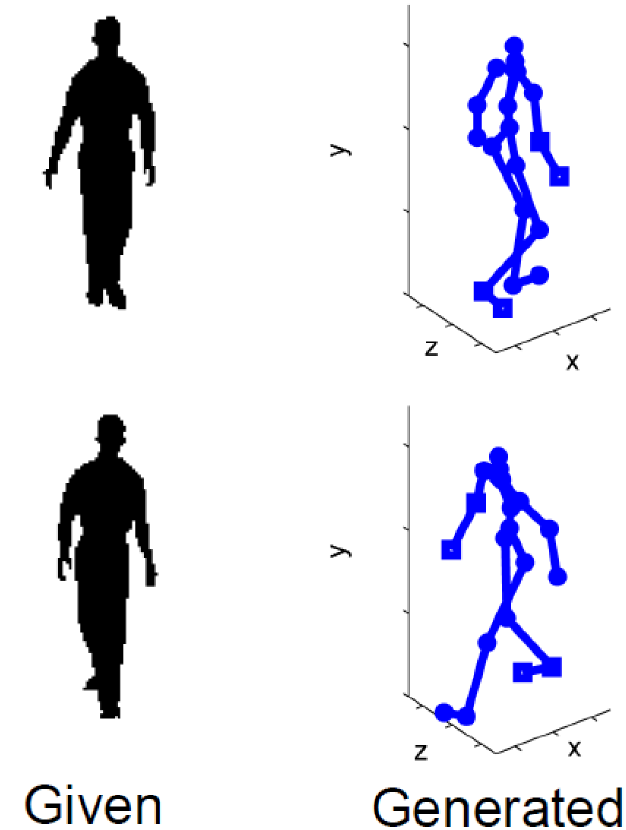


Multi-view Gaussian Processes

- Multi-view data regression/classification
- View generation

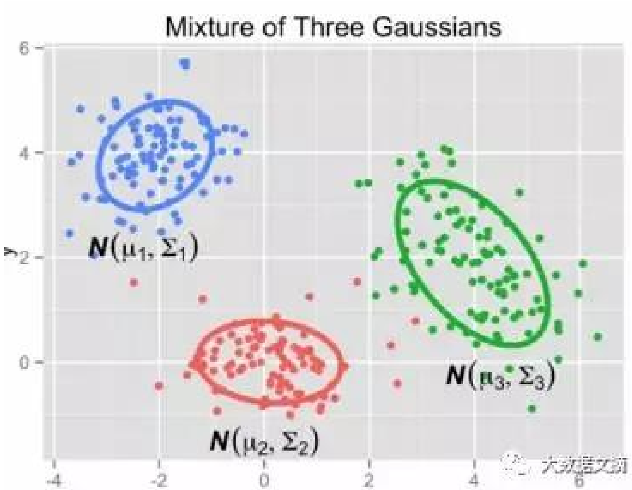


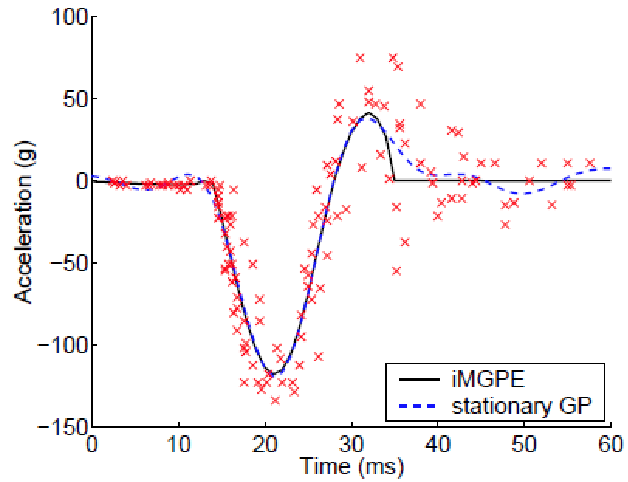


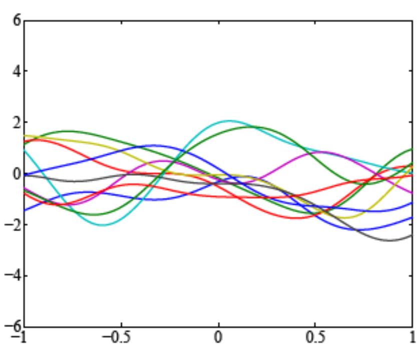


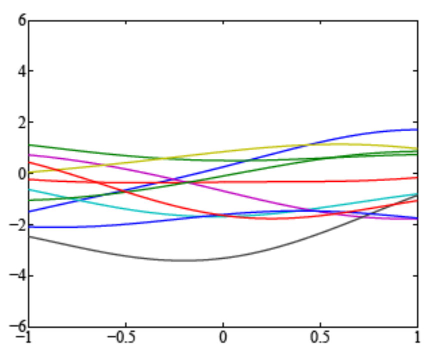
Mixtures of Gaussian Process

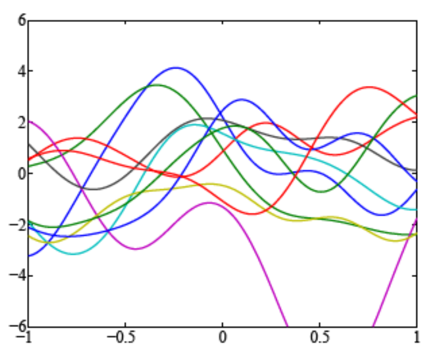
- Recall GMM: GMM vs. MGP
- Better fitting multi-modal data





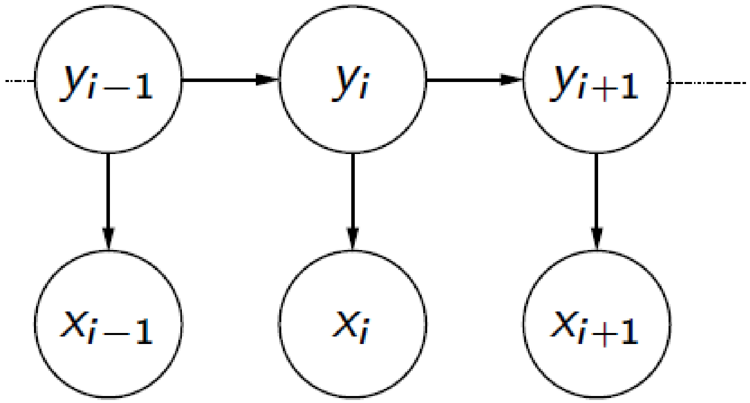




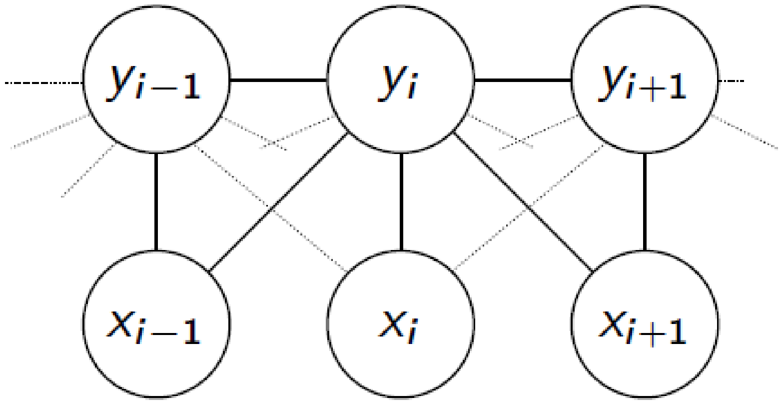


Advanced Models (II)

- Sequential Models
  - Hidden Markov Model (HMM)
  - Conditional Random Field (CRF)



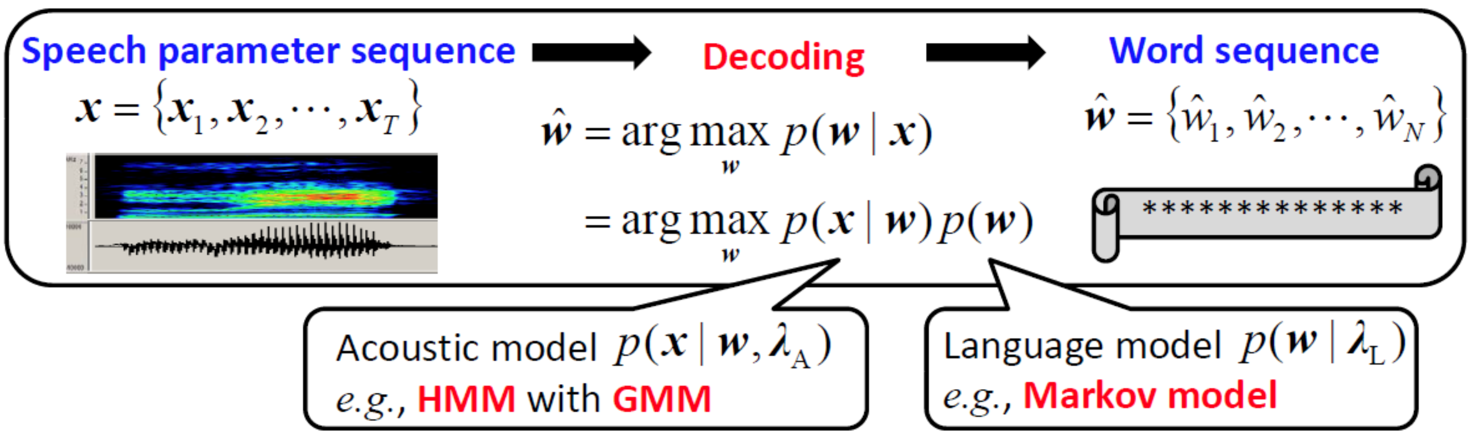
HMM



CRF

Hidden Markov Model (HMM)

- App: Automatic speech recognition (i.e., conversion from speech into text)



- Sequence annotation/classification/generation

Conditional Random Field (CRF)

- App: Part-Of-Speech tagging; Named entity recognition; Image annotation
- Structural prediction: sequence, tree, grid

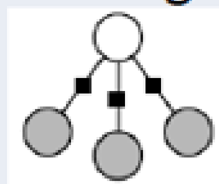
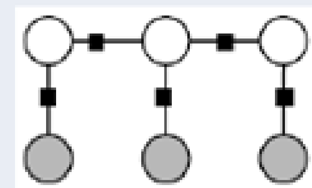
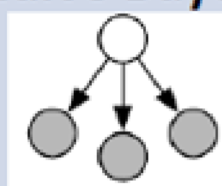
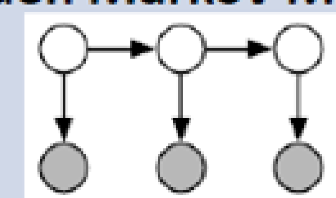
CRF model definition

$$\begin{aligned} p(\underline{y}|\underline{x};\boldsymbol{\theta}) &= \frac{1}{Z(\underline{x},\boldsymbol{\theta})} \exp \sum_{j=1}^{D} \theta_j F_j(\underline{x},\underline{y}) \\ &= \frac{1}{Z(\underline{x},\boldsymbol{\theta})} \Psi(\underline{x},\underline{y};\boldsymbol{\theta}); \quad \boldsymbol{\theta} = \{\theta_1, \cdots, \theta_D\}. \end{aligned}$$

- Defining label constraints using feature functions

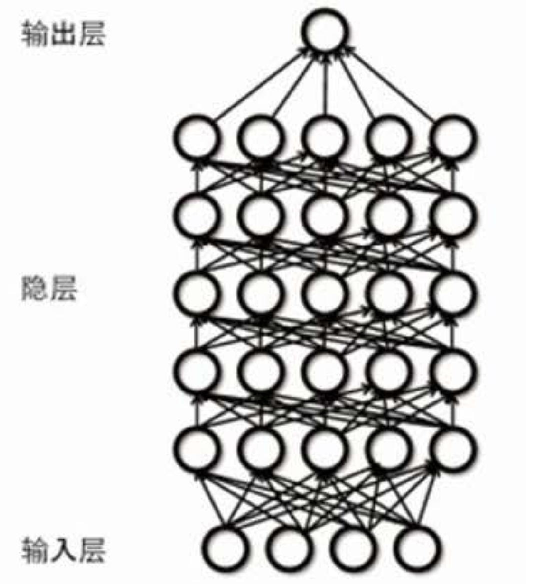
$$F_j(\underline{x},\underline{y}) = \sum_{i=1}^{n} f_j(y_{i-1}, y_i, \underline{x}, i)$$

Sequential Models

	Classifier	Structured Prediction (Sequential Modeling)
Probabilistic Discriminative Model	**Logistic Regression**  *Predict class y; modeling P(y|x)*	**Conditional Random Fields**  *Predict sequence y; modeling P(y|x)*
Probabilistic Generative Model	**Naïve Bayes**  *Predict class y; modeling P(y,x)*	**Hidden Markov Models**  *Predict sequence y; modeling P(y,x)*

Advanced Models (III)

- Deep neural networks
  - Neural network
  - Challenge of deep neural networks
    - gradient vanish
    - Local minima
    - overfitting
  - Convolutional Neural Network (CNN)
  - Recurrent Neural Network (RNN)
  - Variational auto-encoder (VAE)
  - Generative Adversarial Networks (GAN)

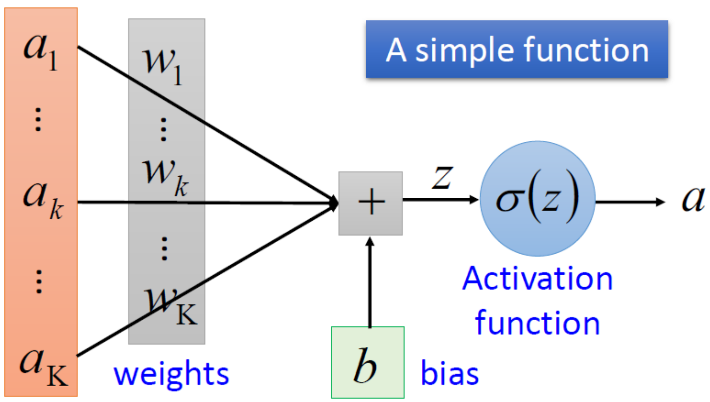


含多个隐层的深度学习模型

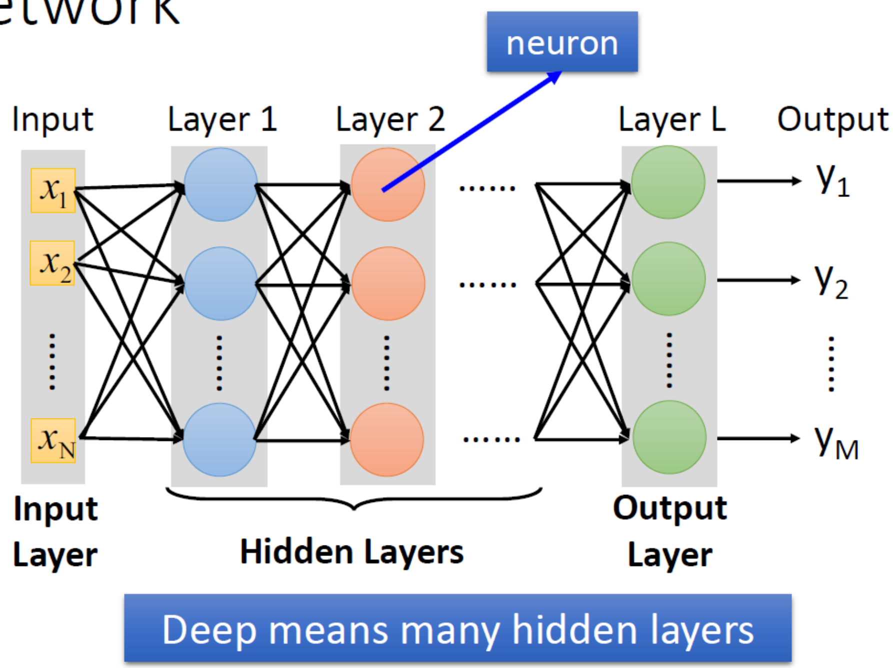
Neural Network

- neuron

$$z = a_1 w_1 + \cdots + a_k w_k + \cdots + a_K w_K + b$$



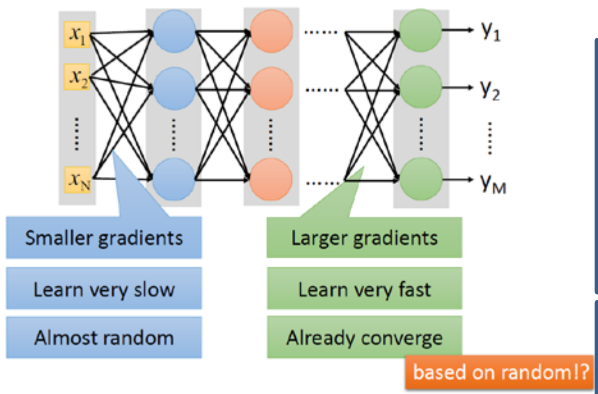
Fully Connect Feedforward Network



Challenge of deep neural networks

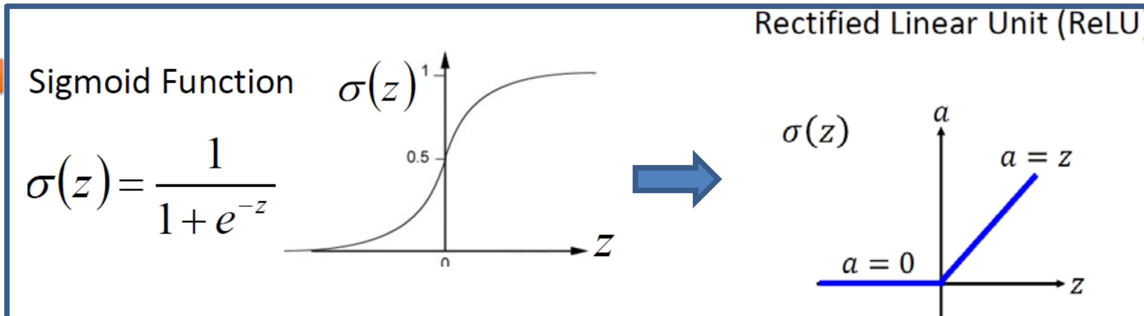
- gradient vanish

Vanishing Gradient Problem



In 2006, people used RBM pre-training.
In 2015, people use ReLU.

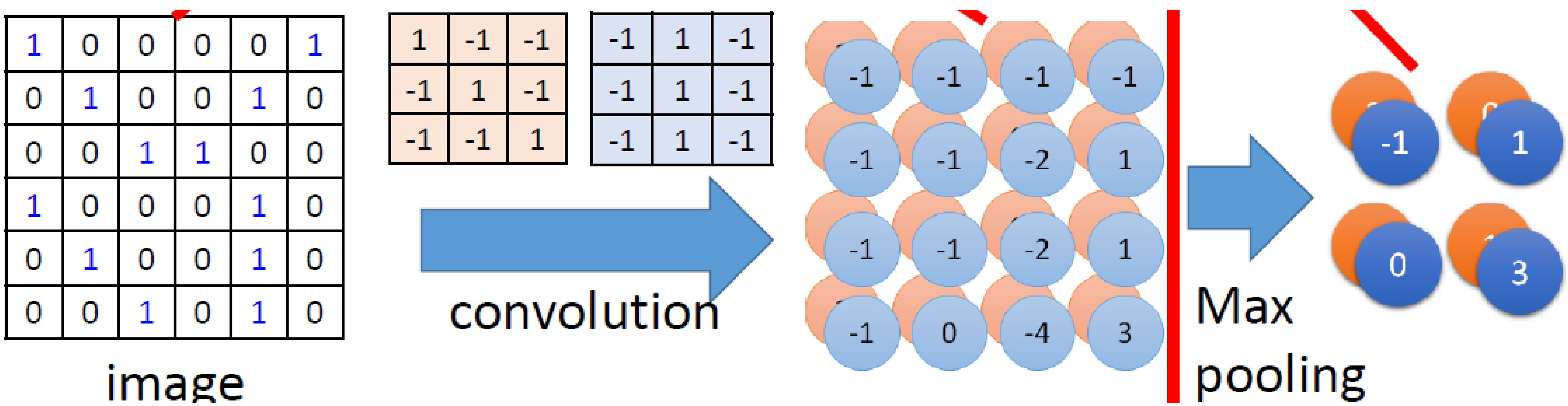
layer-wise pre-training: unsupervised network



- Local minima
  - Optimization technique
- Overfitting
  - Early stopping, regularization, dropout, design network

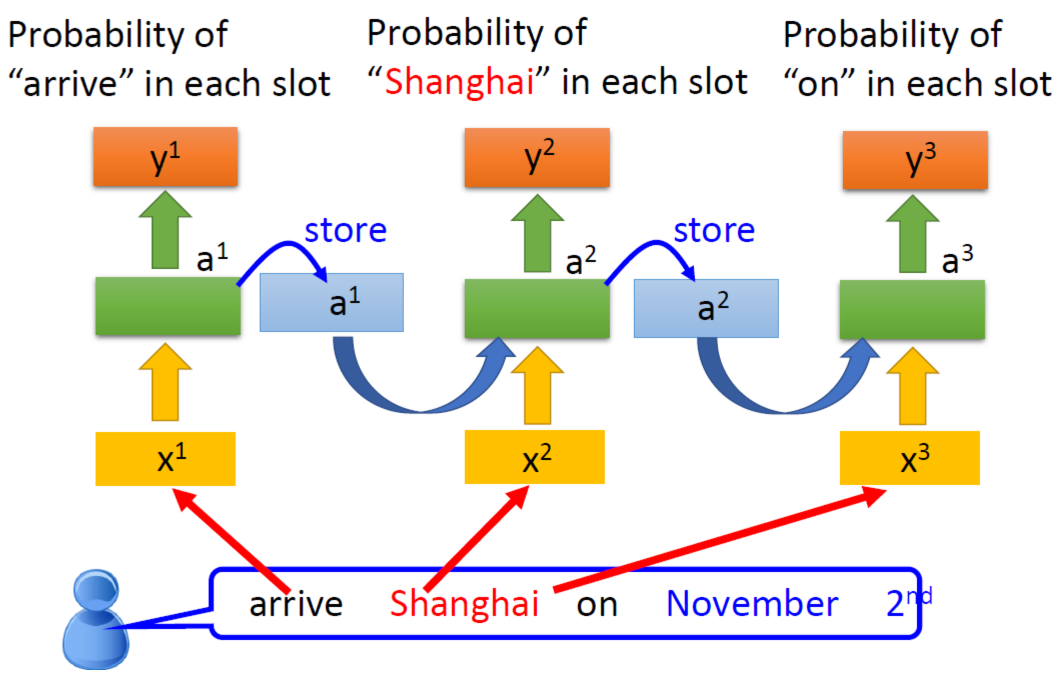
Convolutional Neural Network (CNN)

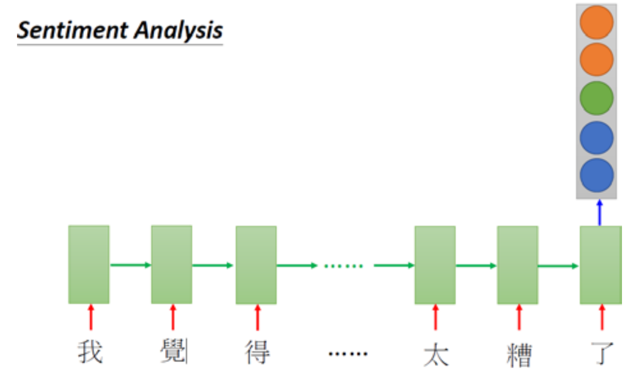
- Some patterns are much smaller than the whole image & The same patterns appear in different regions => convolution
- Subsampling the pixels will not change the object => pooling

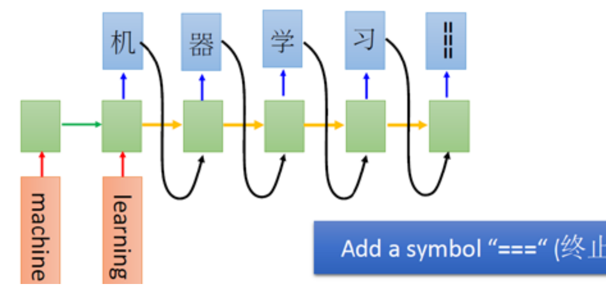


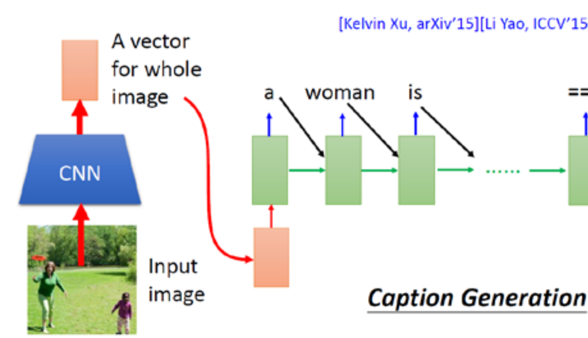
Recurrent Neural Network (RNN)

- Sequences not iid data
- Neural network needs memory







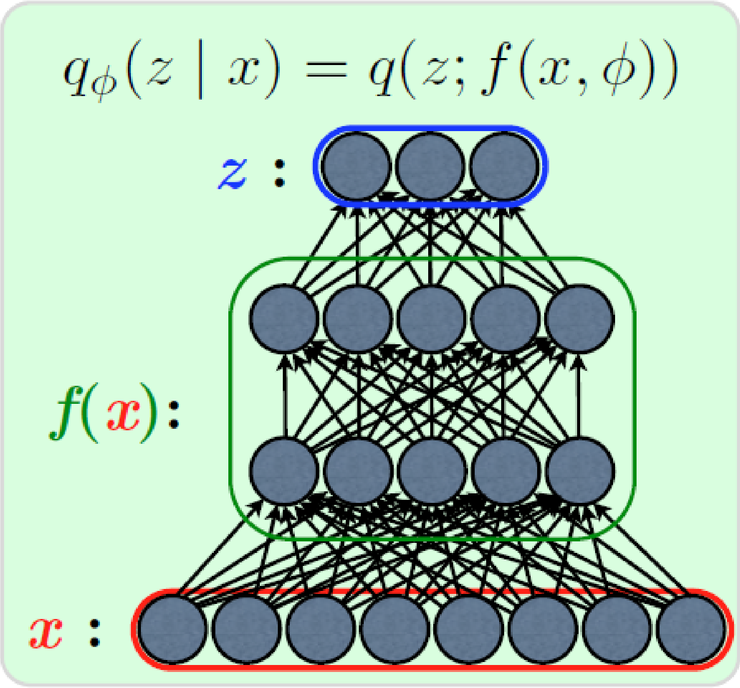


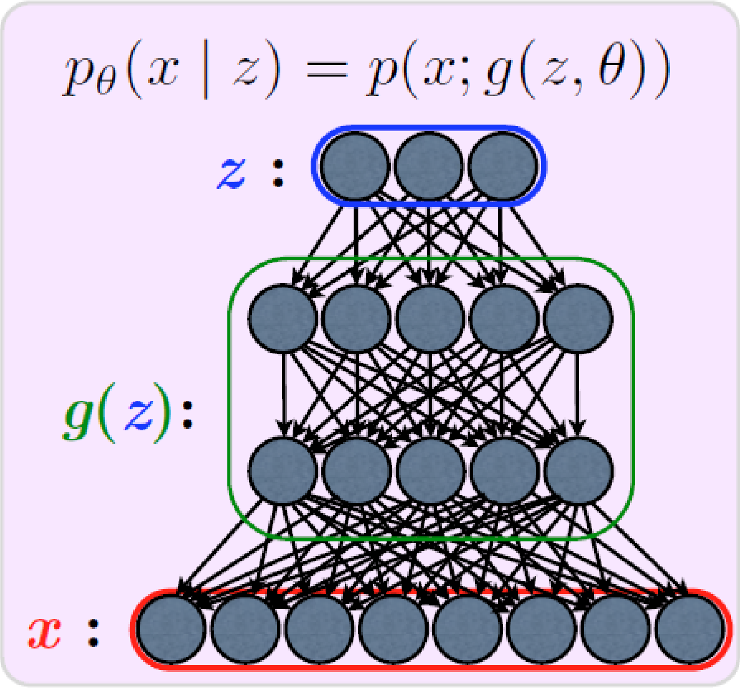
Variational auto-encoder (VAE)

- The **VAE approach**: introduce an inference model $q_\phi(z \mid x)$ that **learns** to approximates the intractable posterior $p_\theta(z \mid x)$ by optimizing the variational lower bound:

$$\mathcal{L}(\theta, \phi, x) = -D_{\mathrm{KL}}\left(q_\phi(z \mid x) \| p_\theta(z)\right) + \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x \mid z)\right]$$

- We parameterize $q_\phi(z \mid x)$ with another neural network:





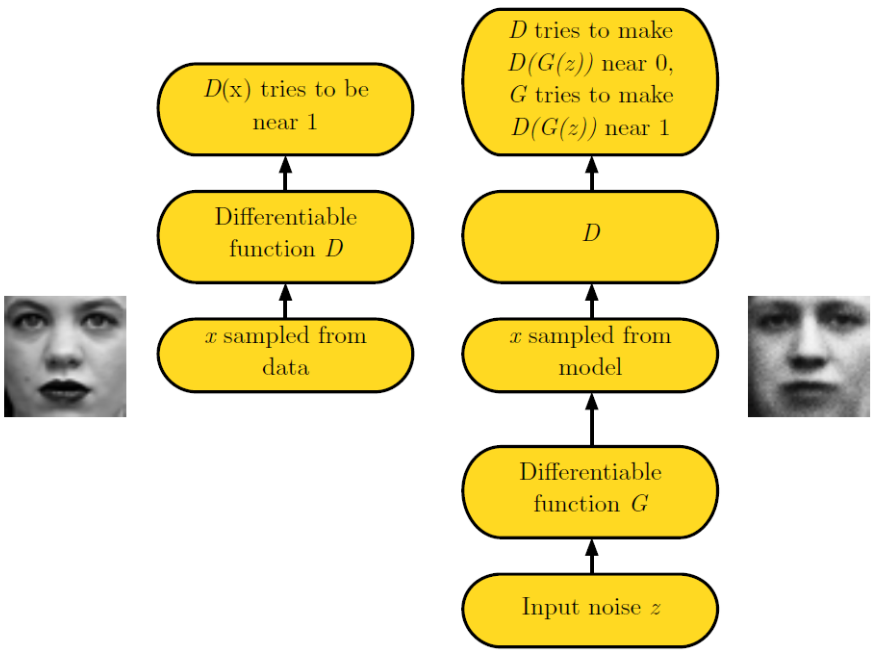
Generative Adversarial Network (GAN)

- Optimize the discriminator D(x) and generator G(z) :

$$J^{(D)} = -\frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}} \log D(\boldsymbol{x}) - \frac{1}{2}\mathbb{E}_{\boldsymbol{z}} \log\left(1 - D\left(G(\boldsymbol{z})\right)\right)$$

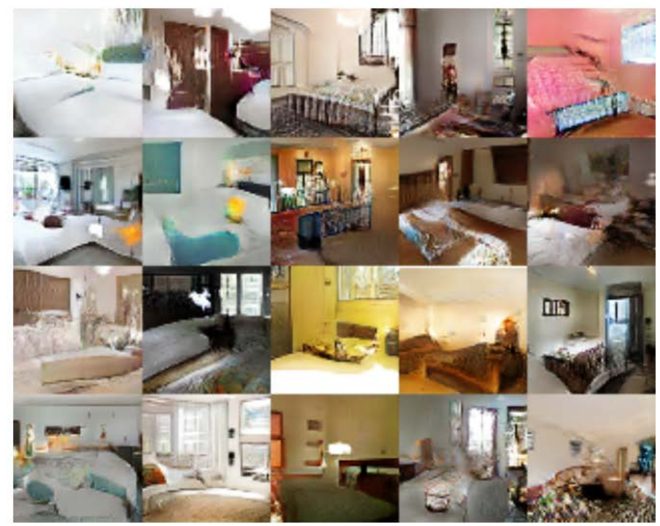
$$J^{(G)} = -J^{(D)}$$

Adversarial Nets Framework



(Goodfellow 2016)

DCGANs for LSUN Bedrooms



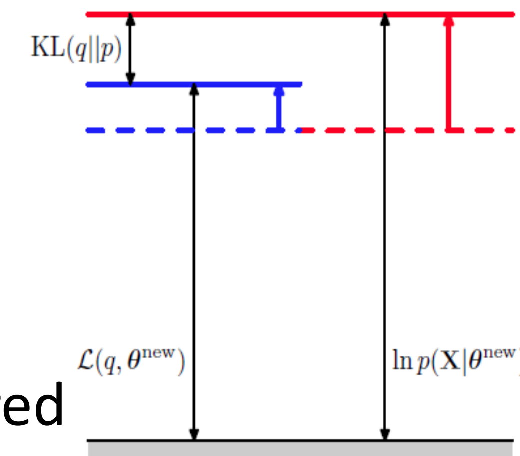
(Radford et al 2015)

Approximate Inference (I)

- Stochastic Variational Inference (SVI)
  - Variational inference
  - Exponential family
  - Natural gradient
  - SVI and its app. on LDA

Stochastic Variational Inference

- VI vs EM
  - VI optimizes q to make blue near red
  - EM
    - calculates q to make blue cover current red
    - optimizes θ to make red higher

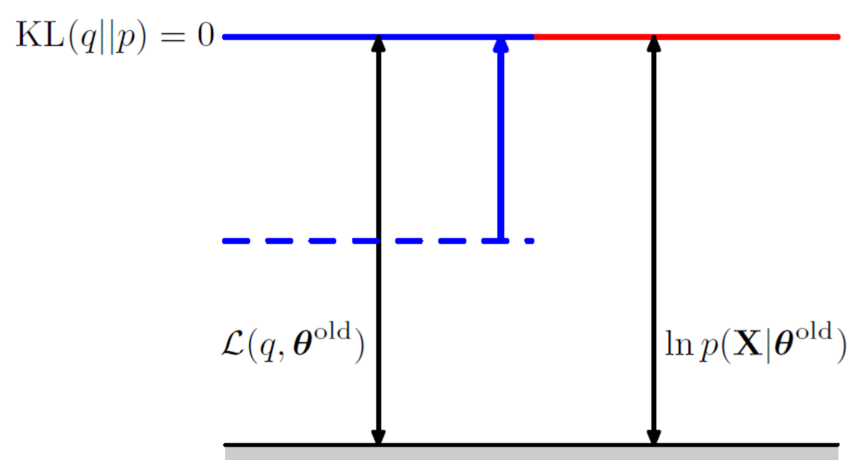


$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q||p)$$

fined

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

$$\mathrm{KL}(q||p) = -\int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}.$$



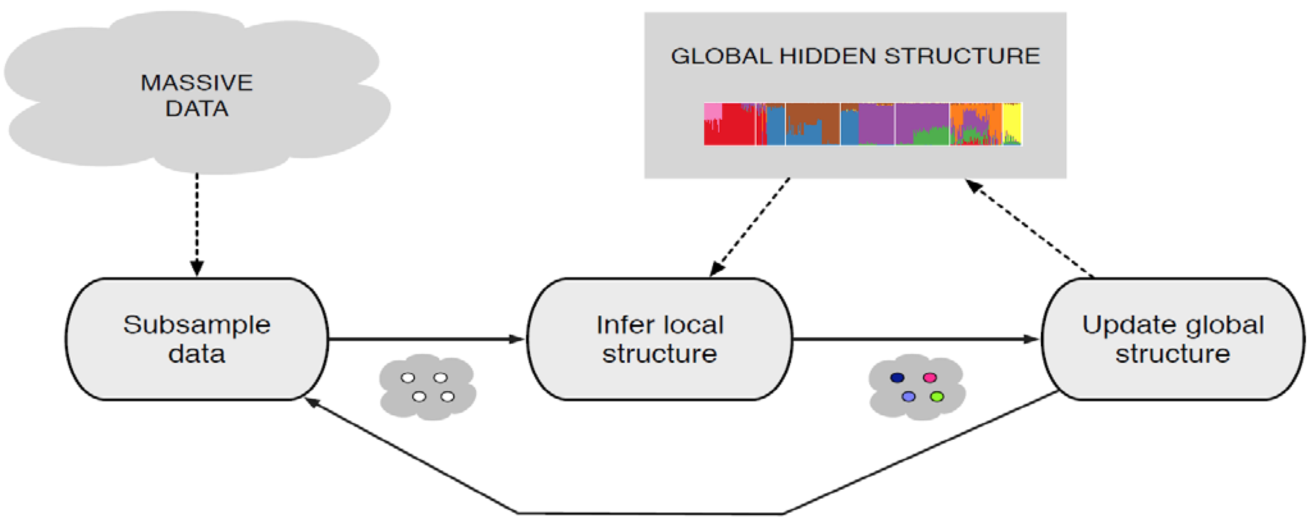
- Often Variational EM or Variational Bayes

Stochastic Variational Inference

- Decompose Linto a global term and a sum of local terms

$$\mathcal{L}(\lambda) = \mathbb{E}_q[\log p(\beta)] - \mathbb{E}_q[\log q(\beta)] + \sum_{n=1}^{N} \max_{\phi_n} (\mathbb{E}_q[\log p(x_n, z_n \mid \beta)] - \mathbb{E}_q[\log q(z_n)]).$$

$$\hat{\nabla} \mathcal{L}_i = \mathbb{E}_q \left[\eta_g \left(x_i^{(N)}, z_i^{(N)}, \alpha \right) \right] - \lambda$$



"Stochastic variational inference" [Hoffman et al., 2013, JMLR]

Approximate Inference (II)

- Sampling Methods
  - Motivation: Integration
    - Expectation, Normalization, Marginalization
  - Sampling methods:
    - Importance
    - Rejection
    - Metropolis-Hastings
    - Gibbs
    - Slice
    - Hybrid Monte Carlo

$$\int f(\theta)\,\pi(\theta)\,\mathrm{d}\theta \;=\; \text{“average over } \pi \text{ of } f\text{”}$$

$$\approx \frac{1}{S}\sum_{s=1}^{S} f(\theta^{(s)}), \quad \theta^{(s)} \sim \pi$$

Sampling Methods Applications

- Prediction

$$p(y_* \mid x_*, \mathcal{D}) = \int p(y_* \mid x_*, \theta)\, p(\theta \mid \mathcal{D})\, \mathrm{d}\theta$$

$$\approx \frac{1}{S}\sum_s p(y_* \mid x_*, \theta^{(s)}), \quad \theta^{(s)} \sim p(\theta \mid \mathcal{D})$$

- Inference

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

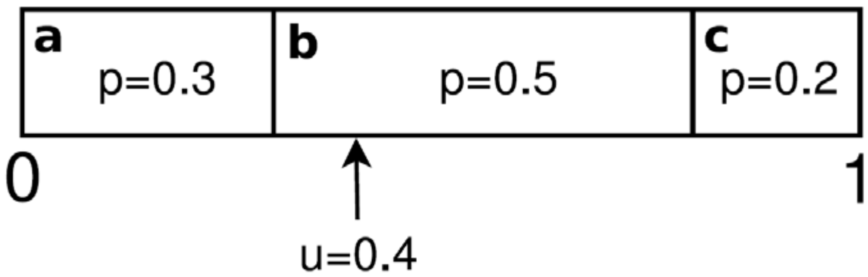
- Marginalization

Interested in particular parameter θ_i

$$p(\theta_i \mid \mathcal{D}) = \int p(\theta \mid \mathcal{D})\, \mathrm{d}\theta_{\setminus i}$$

Sampling Methods

Sampling discrete values



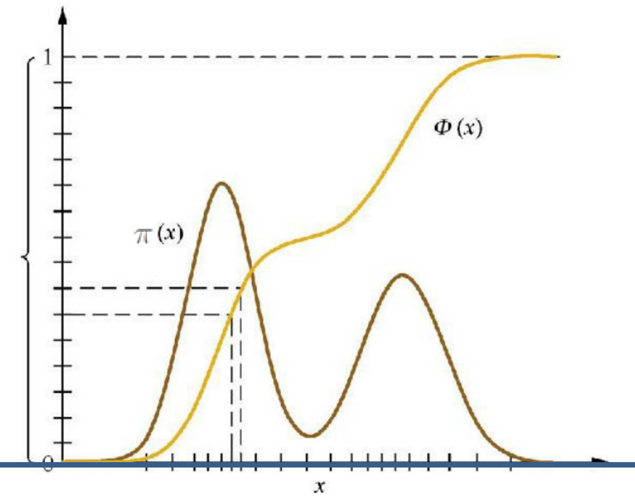
$u \sim \mathrm{Uniform}[0,1]$

$u = 0.4 \quad \Rightarrow \quad \theta = \mathrm{b}$

Math: $A^{(s)} \sim \mathrm{Uniform}[0,1], \quad \theta^{(s)} = \Phi^{-1}(A^{(s)})$

where cdf $\Phi(\theta) = \int_{-\infty}^{\theta} \pi(\theta')\, \mathrm{d}\theta'$

Geometry:

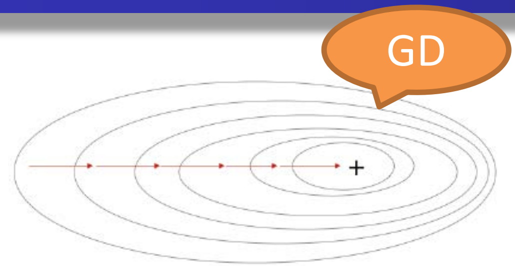


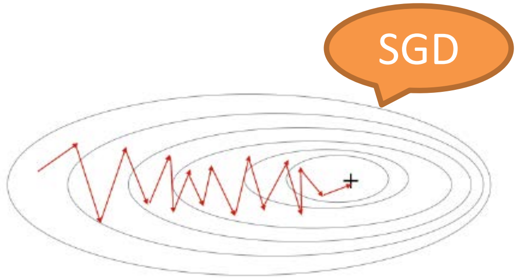
- What if probabilistic density
  - Not normalized
  - not well known
  - cumulative distribution not reversible
  - w.r.t. multivariable

Optimization Methods

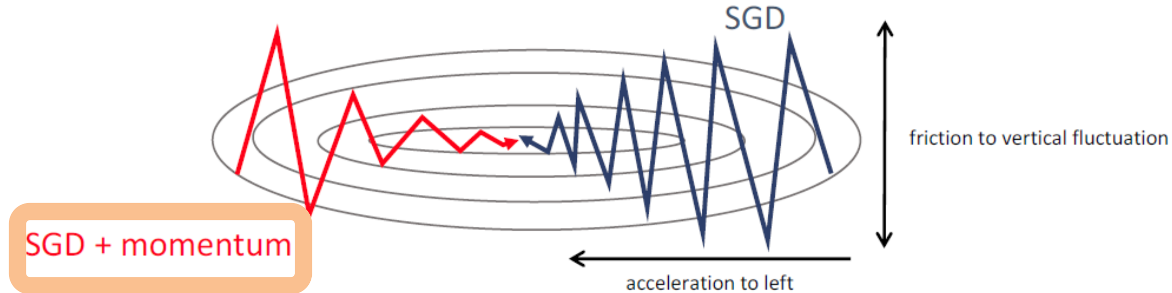
- Stochastic optimization
  - Stochastic gradient descent
  - SGD with momentum
  - Nesterov accelerated gradient
  - AdaGrad
  - Adadelta
  - RMSprop
  - Adam
  - variance reduction techniques

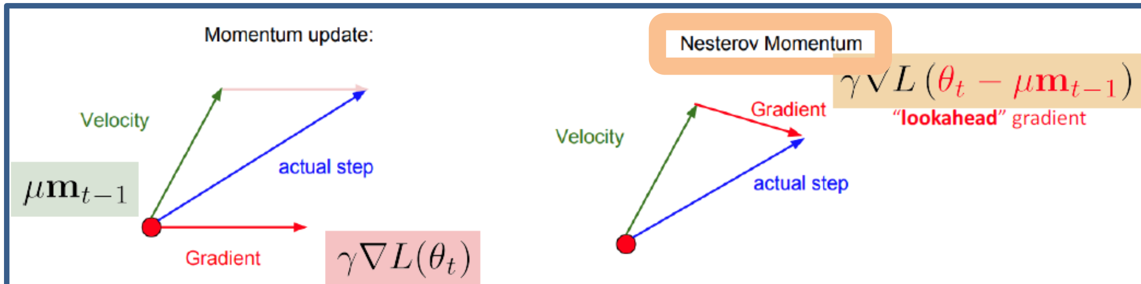
Stochastic Optimization





(+) Momentum reduces the oscillation and accelerates the convergence.





Adaptively changing learning rate (AdaGrad, RMSProp

Combination of momentum and adaptive learning rate

- **Adam** ADAptive Moment estimation) [Kingma' 2015]

momentum

$$\mathbf{m}_{t+1} \leftarrow \mu_1 \mathbf{m}_t + (1-\mu_1)\nabla L(\theta_t)$$

$$\theta_{t+1} \leftarrow \theta_t - \frac{\gamma}{\sqrt{v_t}} \mathbf{m}_t$$

$$v_{t+1} \leftarrow \mu_2 v_t + (1-\mu_2)\nabla L(\theta_t)^2$$

参考资料

- 孙仕亮，赵静. 模式识别与机器学习. 北京：清华大学出版社，2020.
- Bishop C M. Pattern Recognition and Machine Learning. New York, Springer, 2006.

- Carl Edward Rasmussen and Christopher K. I. Williams. Gaussian Process for Machine Learning. Cambridge, MA: MIT Press, 2006.
- Slim Essid. Telecom ParisTech. a tutorial on conditional random fields with applications to music analysis. November 2013.
- Hoffman, Matthew D., et al. “Stochastic variational inference.” Journal of Machine Learning Research 14.5, 2013.
- An Introduction to MCMC for Machine Learning, Machine Learning, 2003.
- Aaron Courville. Variational Autoencoder and Extensions. Deep Learning Summer School 2015.
- Ian Goodfellow, Generative Adversarial Networks (GANs), NIPS 2016 tutorial.