高级机器学习

课程简介

赵静
jzhao@cs.ecnu.edu.cn

课程形式

- 教师授课（前13周）+学生分享（后4周）
- 考核：平时成绩40%+最终报告60%
- 平时成绩
  - 按小组介绍2021年ICML、NeurIPS、IJCAI、AAAI Tutorial
  - 每组45分钟，每周两组
  - 按小组打分，学生互评

- 期末考核：
  - 机器学习相关的论文
  - 按照正规会议期刊格式撰写，提供模板，6页上限
    - 综述论文（60-80分）
    - 近5年顶会刊算法复现，需增加原文之外的数据集（80-90分）
    - 创新性论文（90-100分）

Outline

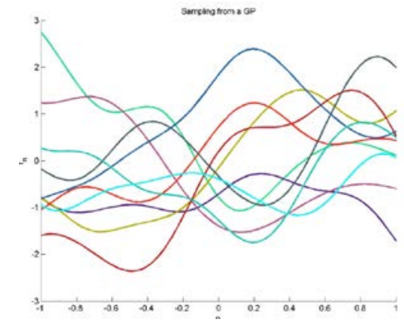
- Review
  - Bayes, LR, SVM, EM, VI, PCA, LDA, Cluster...
- Advanced Models
  - GP Related Models
  - Sequential Models
  - Deep Neural Networks
- Approximate Inference and Optimization
  - Sampling Methods
  - Stochastic Optimization

Advanced Models (I)

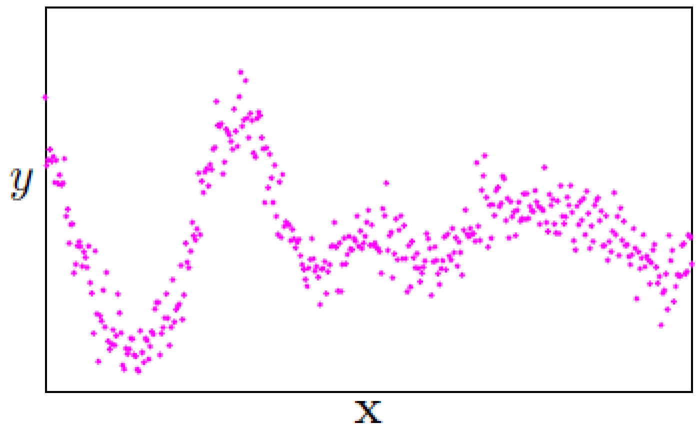
- GP Related Models
  - Gaussian processes (regression or classification)
  - Gaussian process latent variable models (dimensionality reduction)
  - Deep Gaussian processes (deep model)
  - Multi-view Gaussian processes (multi-view)
  - Mixtures of Gaussian process (multi-modal)

Gaussian Process Regression

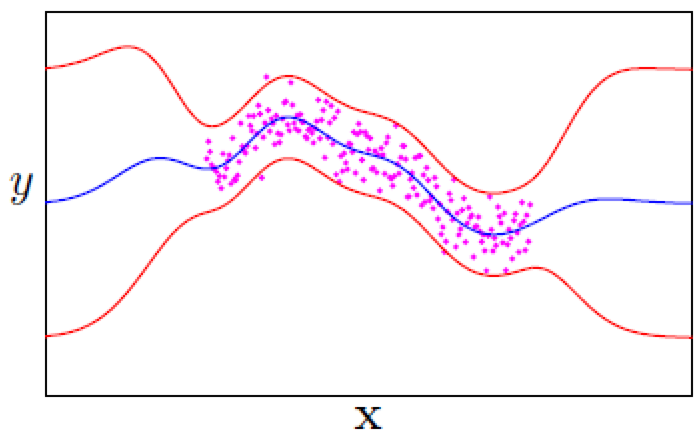
Gaussian observation noise: $y_n = f_n + \epsilon_n$, where $\epsilon_n \sim \mathcal{N}(0, \sigma^2)$



sample data



predictive



marginal likelihood

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

predictive distribution

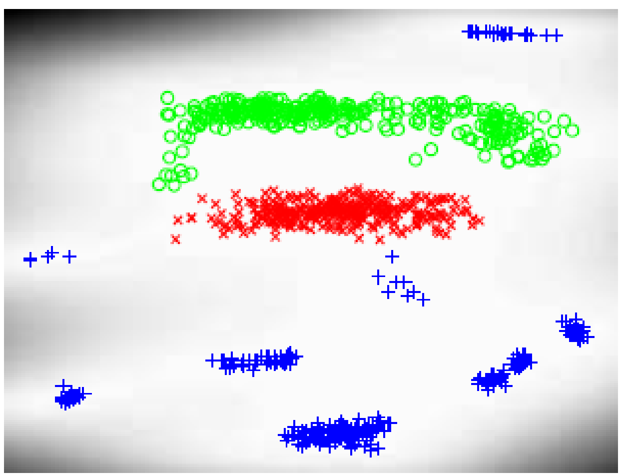
$$p(y_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N}(\mu_*, \sigma_*^2)$$

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

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

Gaussian Process Latent Variable Models

- Low dimensional visualization
- Predict missing values



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

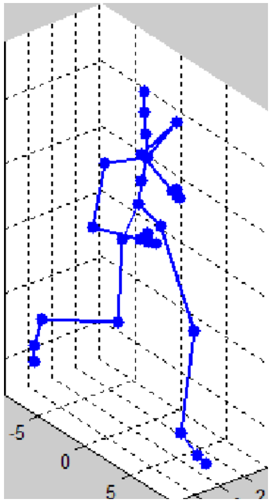
True

Test

Recon.

Gaussian Process Latent Variable Models

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



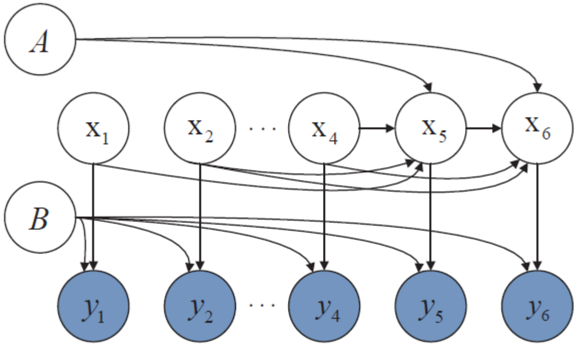
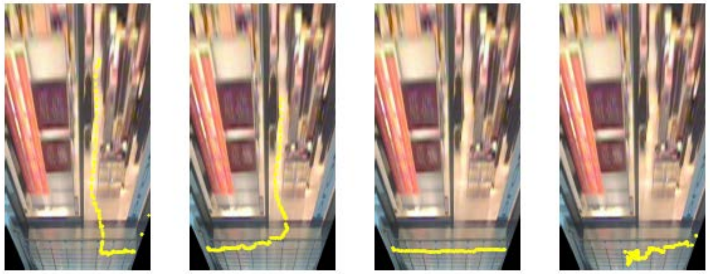


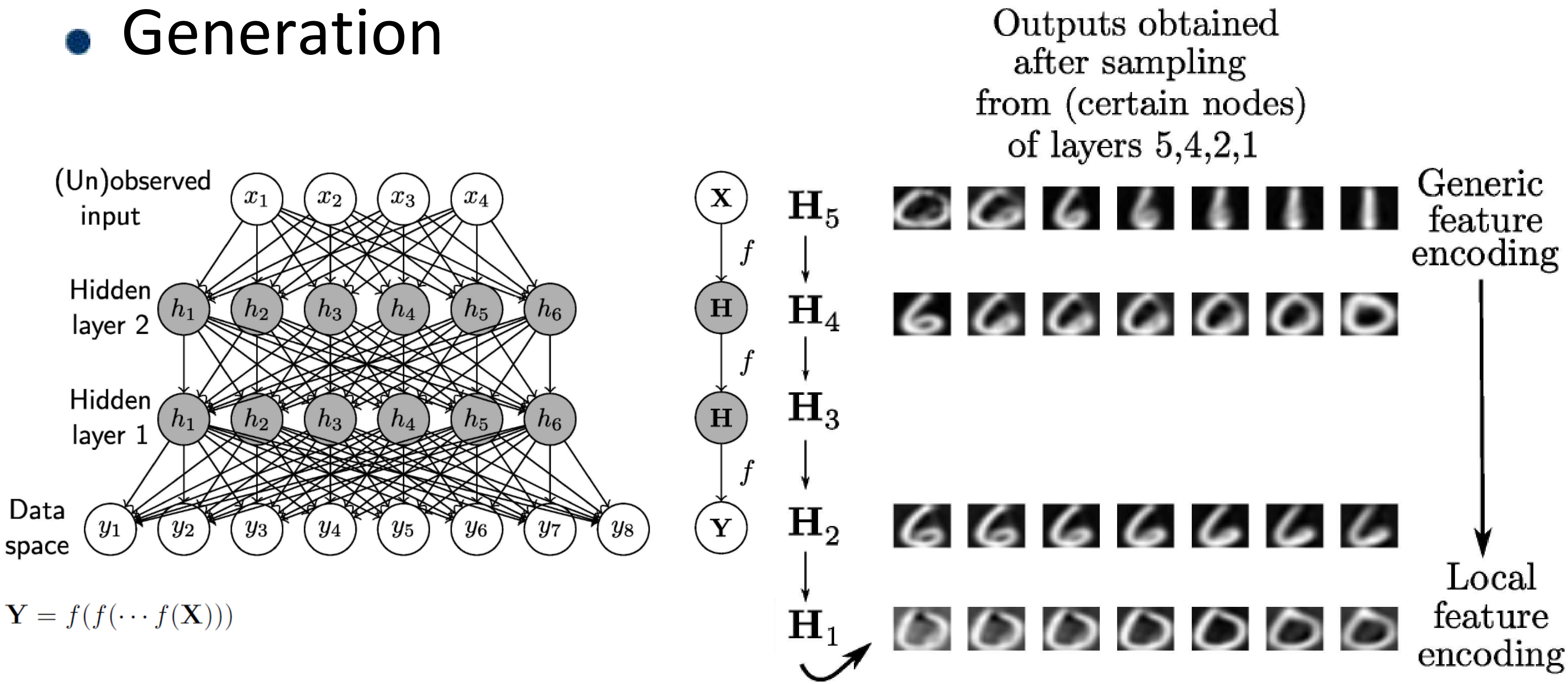
Figure: 四阶动态系统的示意图



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

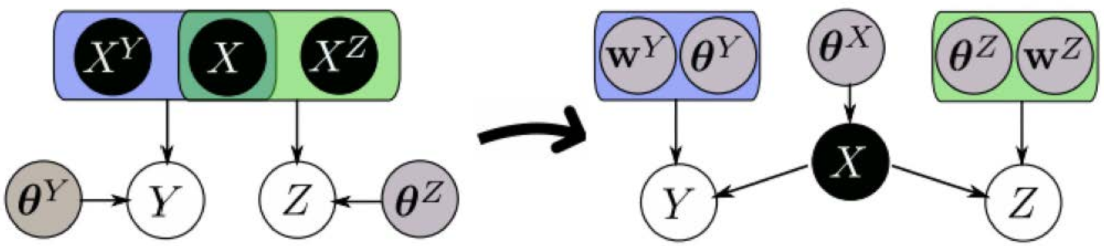
Deep Gaussian Processes

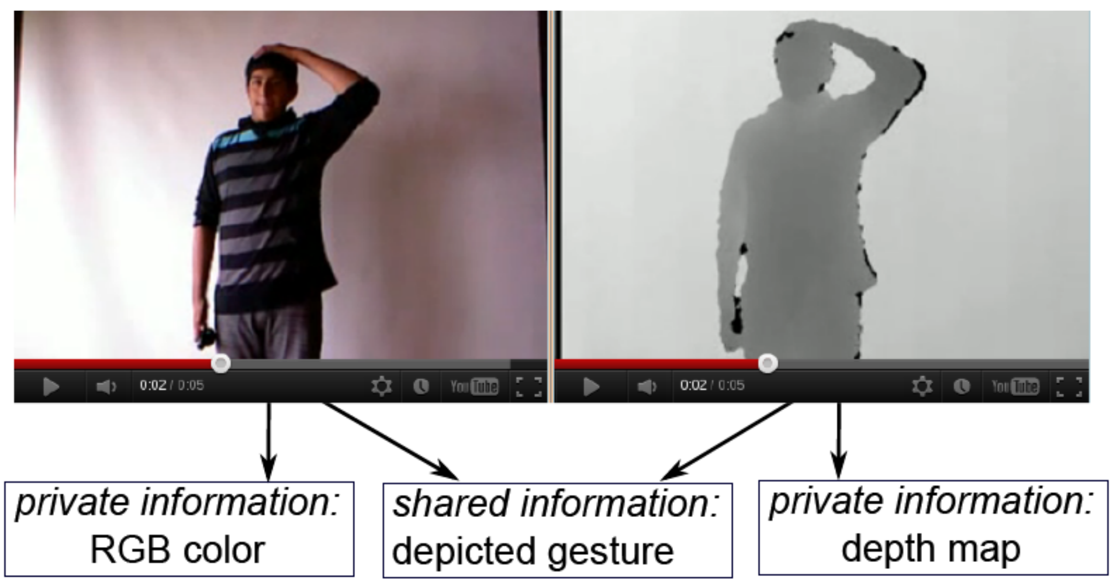
- Deep representation learning
- Robust
- Generation

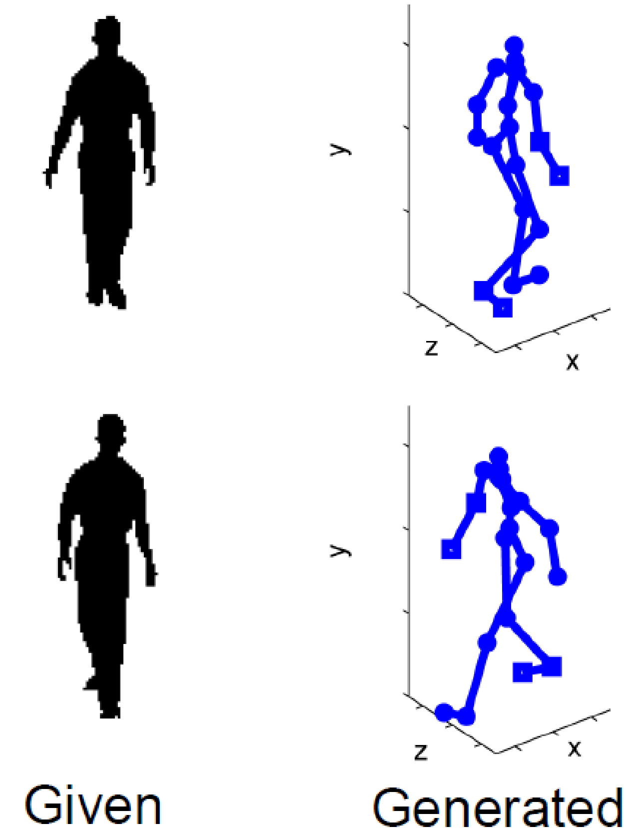


Multi-view Gaussian Processes

- Multi-view data regression/classification
- View generation

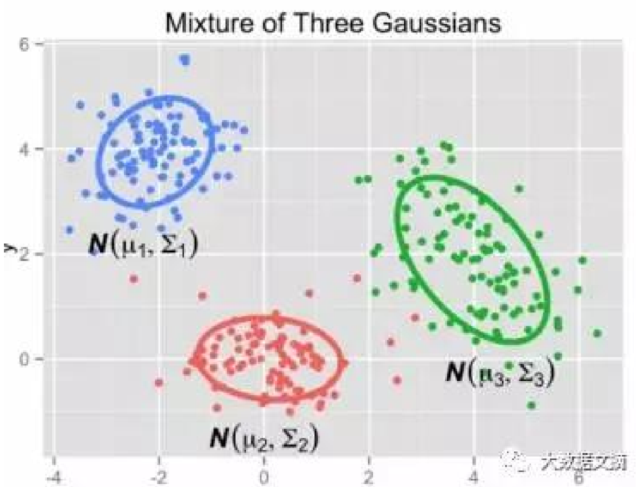


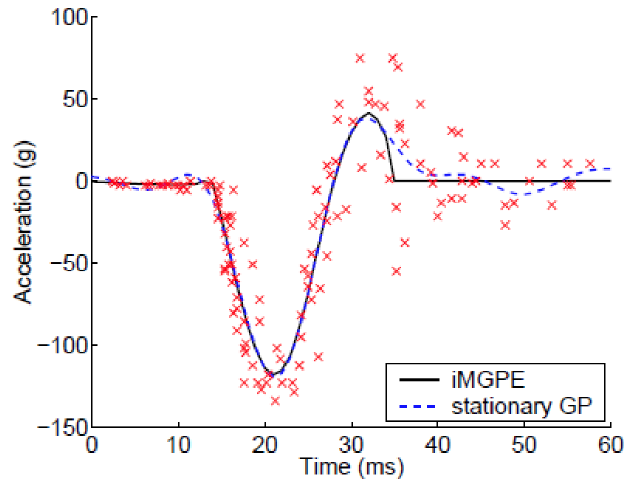


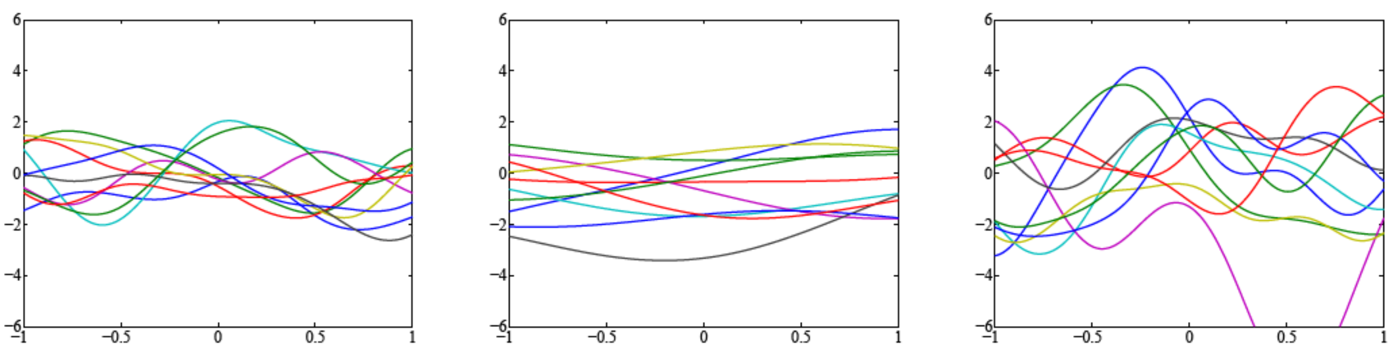


Mixtures of Gaussian Process

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

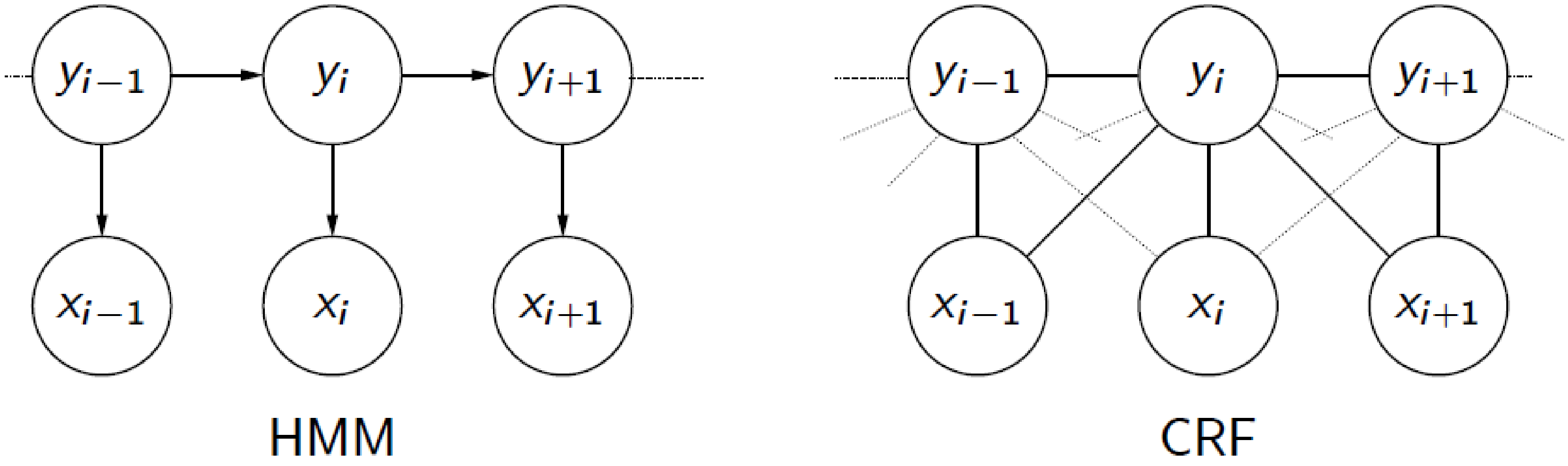






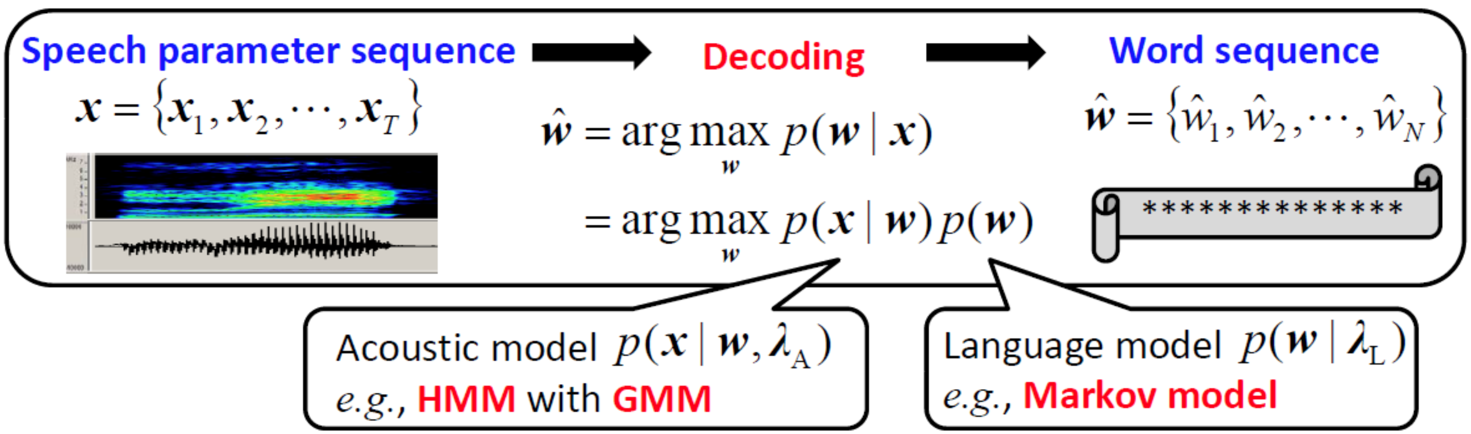
Advanced Models (II)

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



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

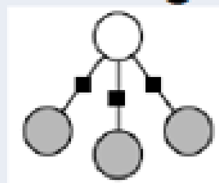
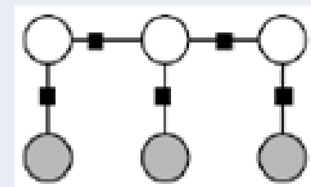
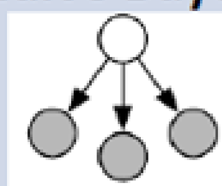
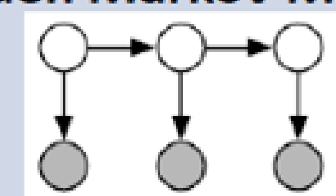
CRF model definition

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

- Defining label constraints using feature functions

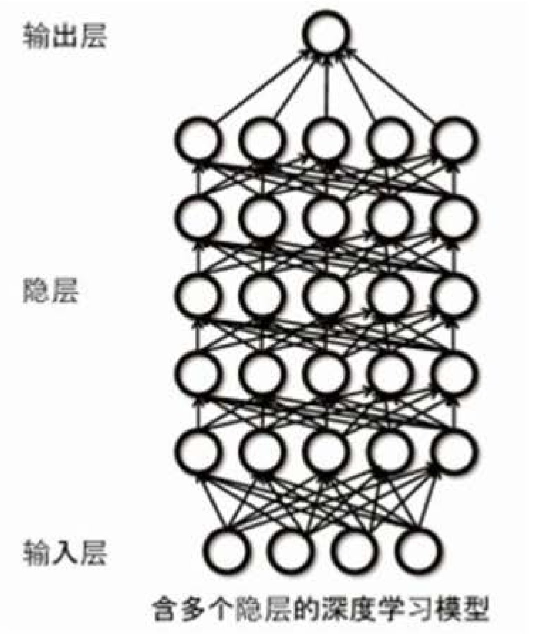
$$
F_j(\underline{x},\underline{y}) = \sum_{i=1}^{n} f_j(y_{i-1},y_i,\underline{x},i)
$$

Sequential Models

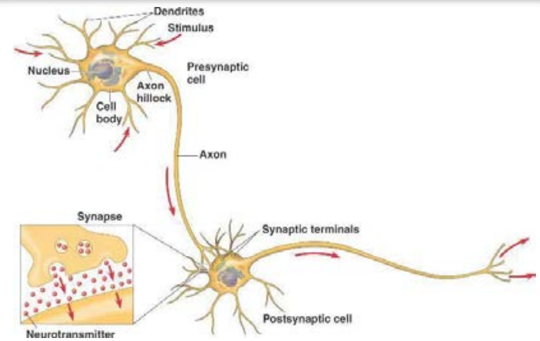
	Classifier	Structured Prediction (Sequential Modeling)
Probabilistic Discriminative Model	**Logistic Regression**  *Predict class y; modeling P(y|x)*	**Conditional Random Fields**  *Predict sequence y; modeling P(y|x)*
Probabilistic Generative Model	**Naïve Bayes**  *Predict class y; modeling P(y,x)*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

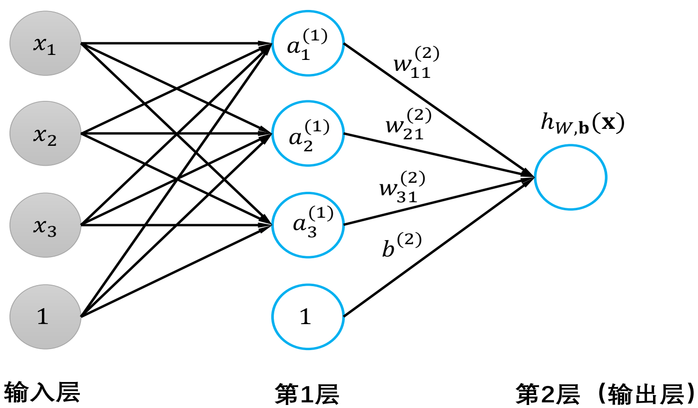
Advanced Models (III)

- 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)



- NN/MLP



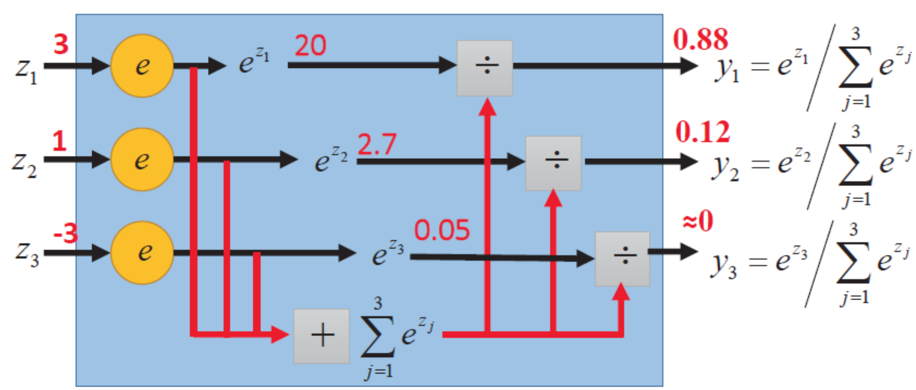


$$
\begin{cases}
a_1^{(1)} = f(w_{11}^{(1)}x_1 + w_{21}^{(1)}x_2 + w_{31}^{(1)}x_3 + b_1^{(1)}), \\
a_2^{(1)} = f(w_{12}^{(1)}x_1 + w_{22}^{(1)}x_2 + w_{32}^{(1)}x_3 + b_2^{(1)}), \\
a_3^{(1)} = f(w_{13}^{(1)}x_1 + w_{23}^{(1)}x_2 + w_{33}^{(1)}x_3 + b_3^{(1)}), \\
h_{W,\mathbf{b}}(\mathbf{x}) = a_1^{(2)} = f(w_{11}^{(2)}a_1^{(1)} + w_{21}^{(2)}a_2^{(1)} + w_{31}^{(2)}a_3^{(1)} + b^{(2)}),
\end{cases}
$$

- Softmax layer as the output layer

Probability:
- $1 > y_i > 0$
- $\sum_i y_i = 1$

Softmax Layer



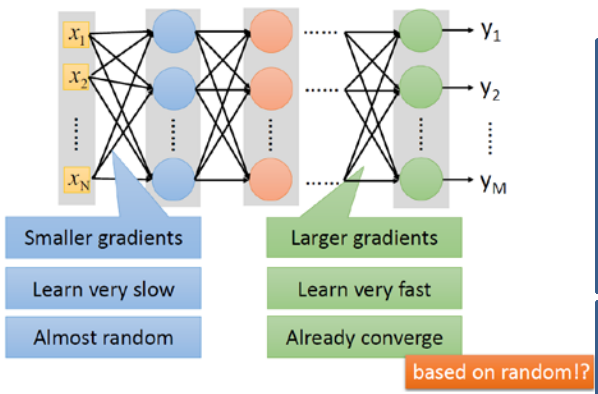
Active function

- 回归：恒等函数
- 二类：sigmoid
- 多个独立二类：sigmoid
- 多类：softmax

Challenge of deep neural networks

- gradient vanish

Vanishing Gradient Problem



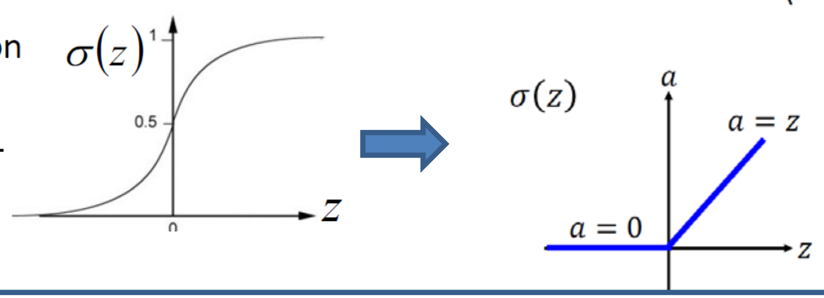
In 2006, people used RBM pre-training.
In 2015, people use ReLU.

layer-wise pre-training: unsupervised network

Sigmoid Function

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

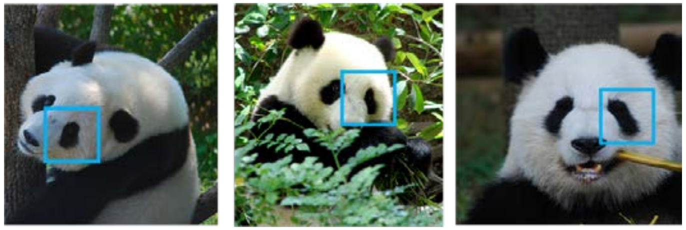
Rectified Linear Unit (ReLU)



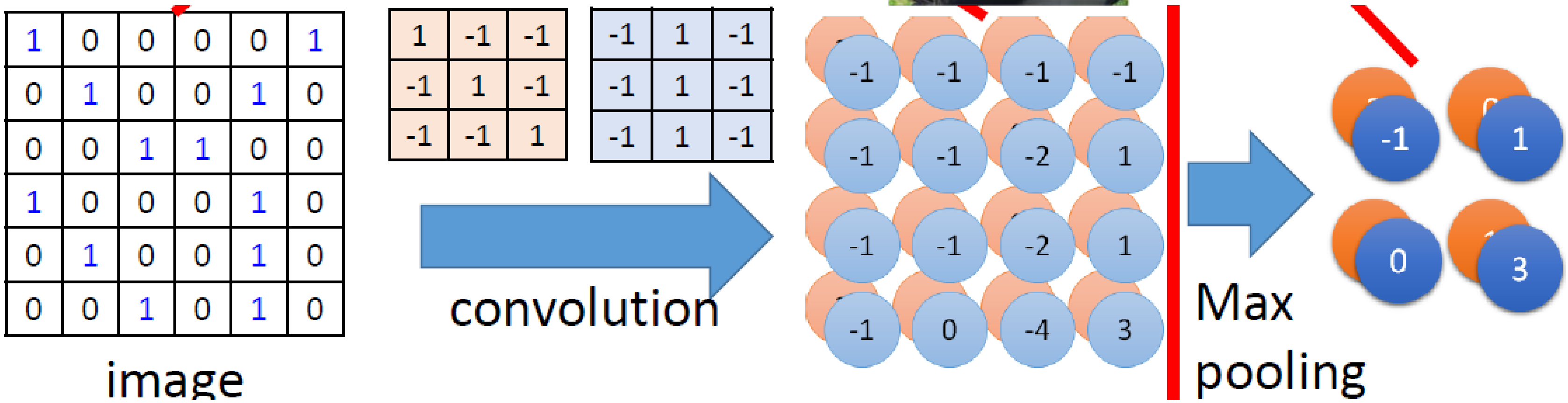
- Local minima
  - Optimization technique
- Overfitting
  - Early stopping, regularization, dropout, design network

Convolutional Neural Network (CNN)

- Some patterns are much smaller than the whole image & The same patterns appear in different regions => convolution

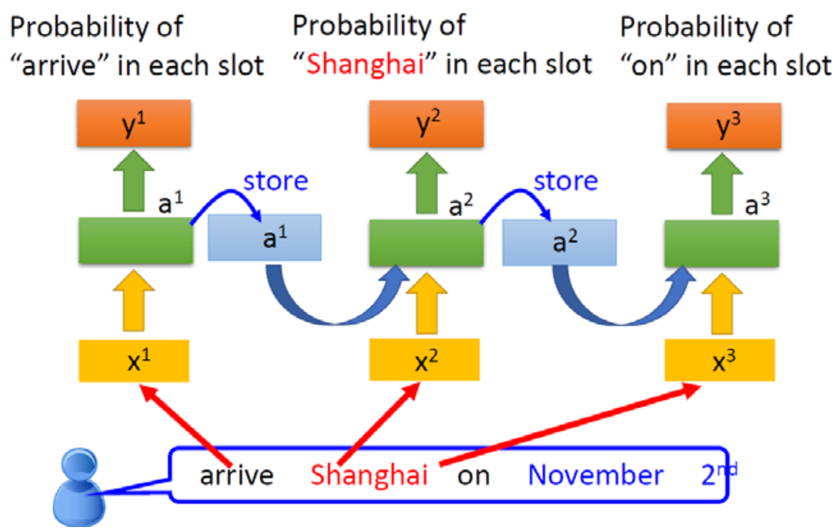


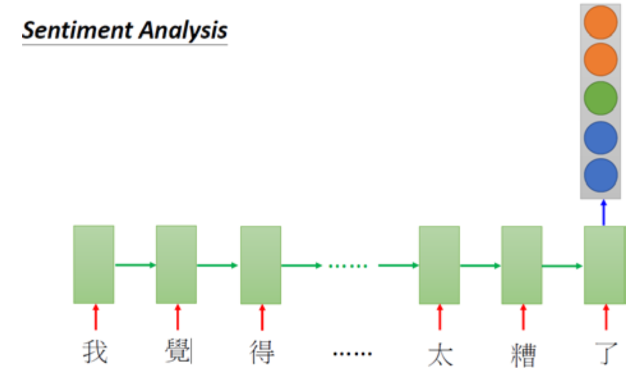
- Subsampling the pixels will not change the object => pooling

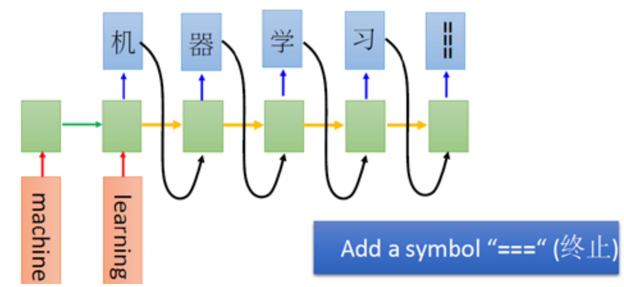


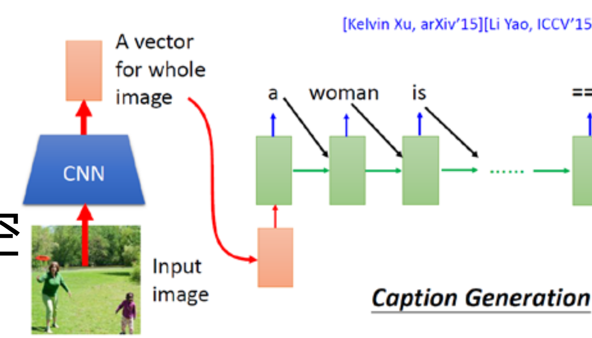
Recurrent Neural Network (RNN)

- Sequences not iid data
- Neural network needs memory









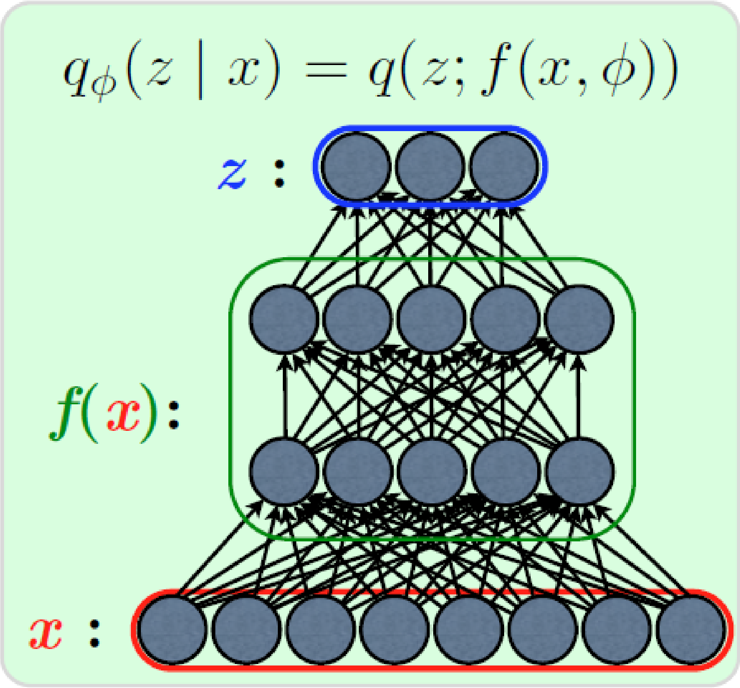
- ➢ 多对一：时序分类，如情感分析，行为识别
- ➢ 一对多：时序生成，如图像描述
- ➢ 多对多（对齐）：时序标注，如实体识别，填空
- ➢ 多对多（非对齐）：机器翻译

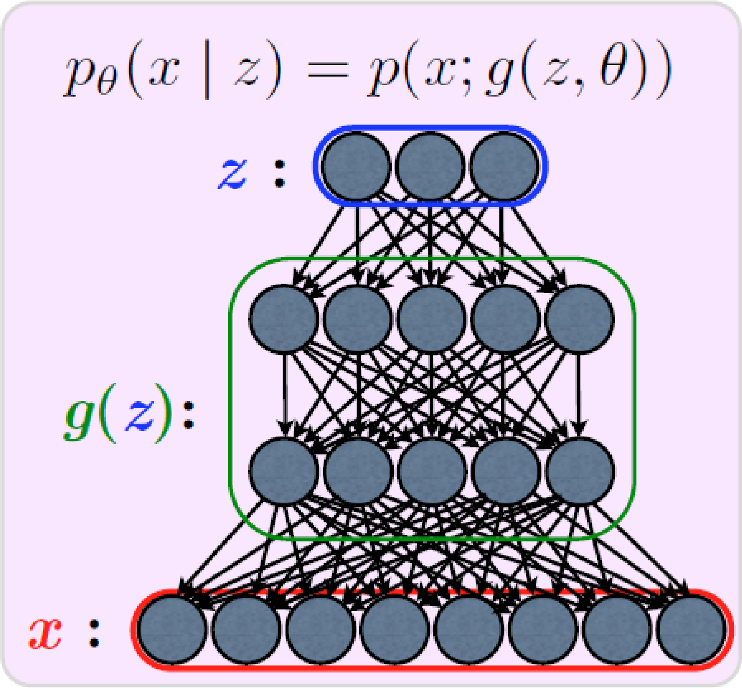
Variational auto-encoder (VAE)

- 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:





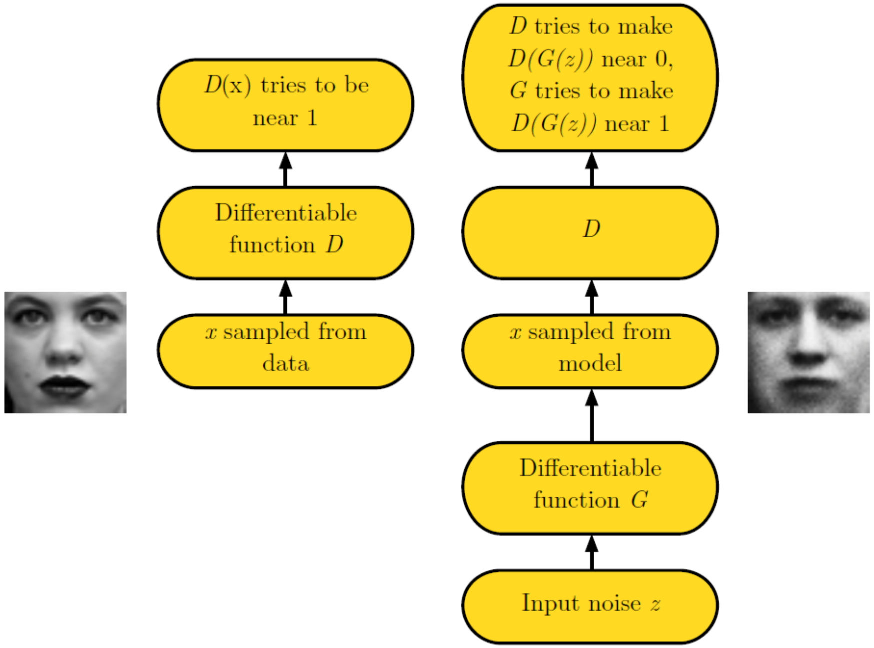
Generative Adversarial Network (GAN)

- 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 (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

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

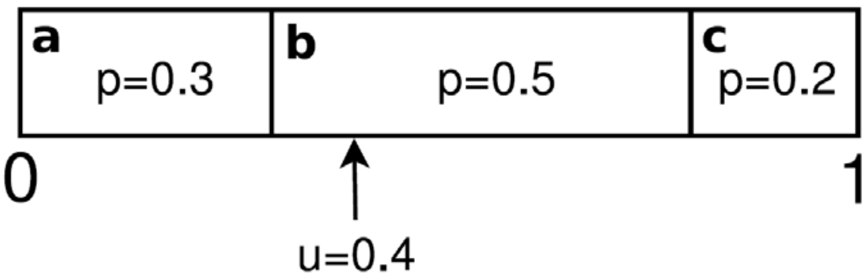
- Marginalization

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



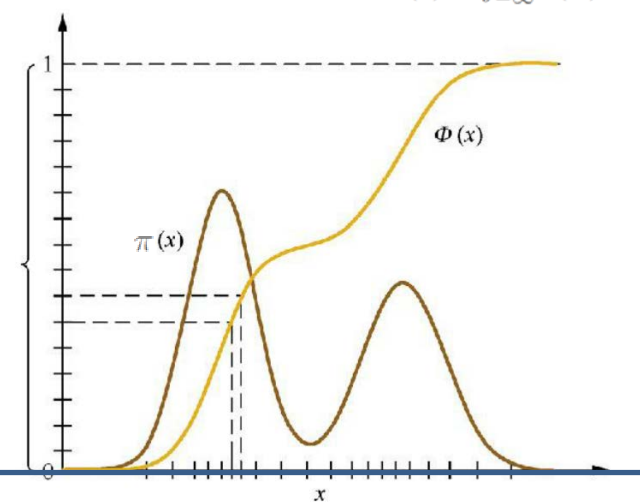
$u \sim \text{Uniform}[0,1]$

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

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

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

Geometry:

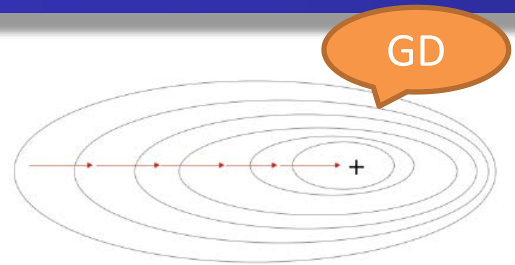


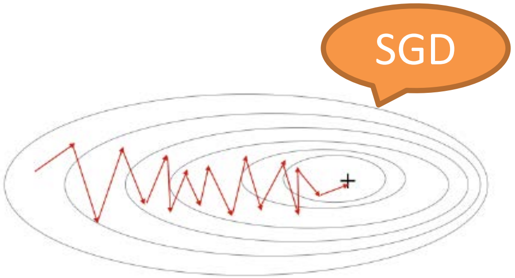
- What if probabilistic density
  - Not normalized
  - not well known
  - cumulative distribution not reversible
  - w.r.t. multivariable

Optimization Methods

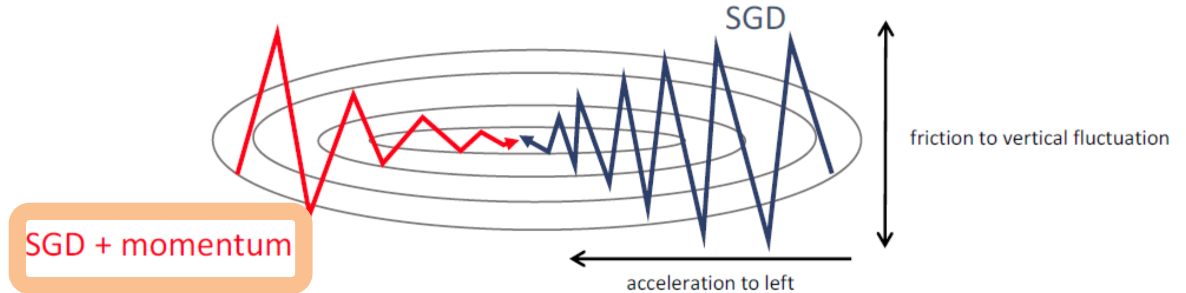
- Stochastic optimization
  - Stochastic gradient descent
  - SGD with momentum
  - Nesterov accelerated gradient
  - AdaGrad
  - Adadelta
  - RMSprop
  - Adam
  - variance reduction techniques

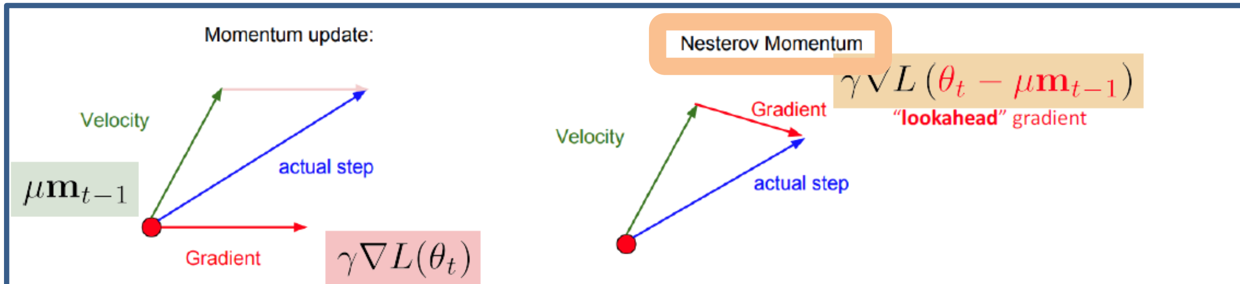
Stochastic Optimization





(+) 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

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

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

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

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.