# NEURAL NETWORKS

chapter 5

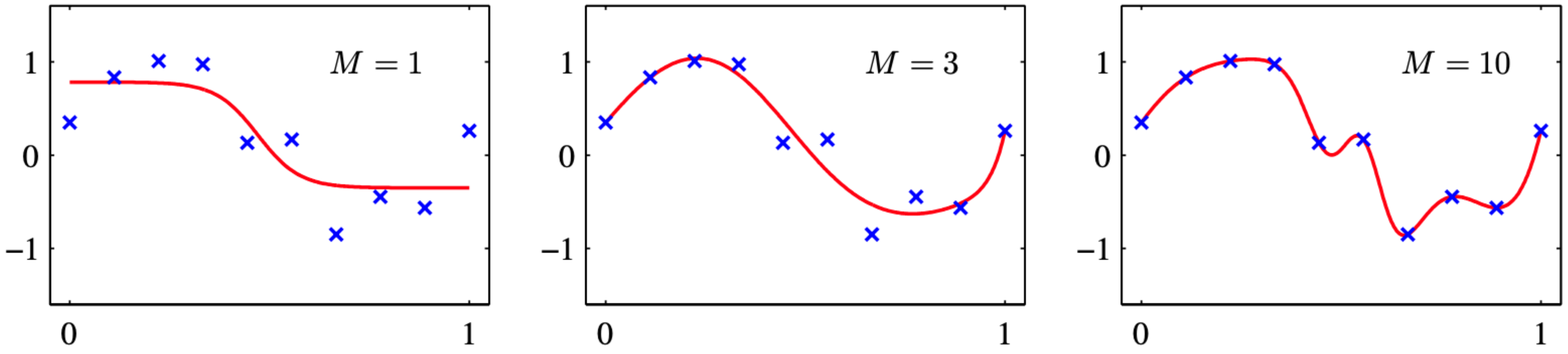
# 目录

- 5.5神经网络的正则化

- 5.6混合密度网络

# 神经网络的正则化

- why?




$$\tilde{E}(\boldsymbol{w}) = E(\boldsymbol{w}) + \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w}$$

# 相容性

- 1.简单正则化的局限性：与网络映射的确定缩放性质不相容。
- 第一层： $z_j = h\left(\sum_i w_{ji}x_i + w_{j0}\right)$ 输出单元 $y_k = \sum_j w_{kj}z_j + w_{k0}$
- 输入变量经过线性变换： $x_i \to \tilde{x}_i = ax_i + b$
- 网络映射不变调整权重和偏置：

$$w_{ji} \to \tilde{w}_{ji} = \frac{1}{a}w_{ji}$$

$$w_{j0} \to \tilde{w}_{j0} = w_{j0} - \frac{b}{a}\sum_i w_{ji}$$

- 输出变量经过线性变换： $y_k \to \tilde{y}_k = cy_k + d$
- 权重和偏置：

$$w_{kj} \to \tilde{w}_{kj} = cw_{kj}$$

$$w_{k0} \to \tilde{w}_{k0} = cw_{k0} + d$$

# 相容性

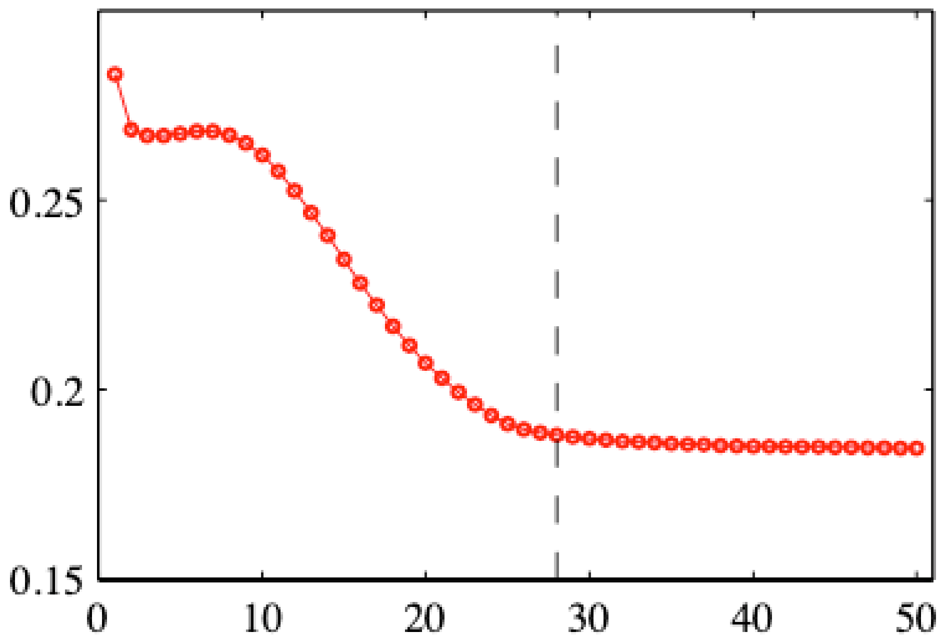
- 简单的正则化： $\tilde{E}(\boldsymbol{w}) = E(\boldsymbol{w}) + \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w}$

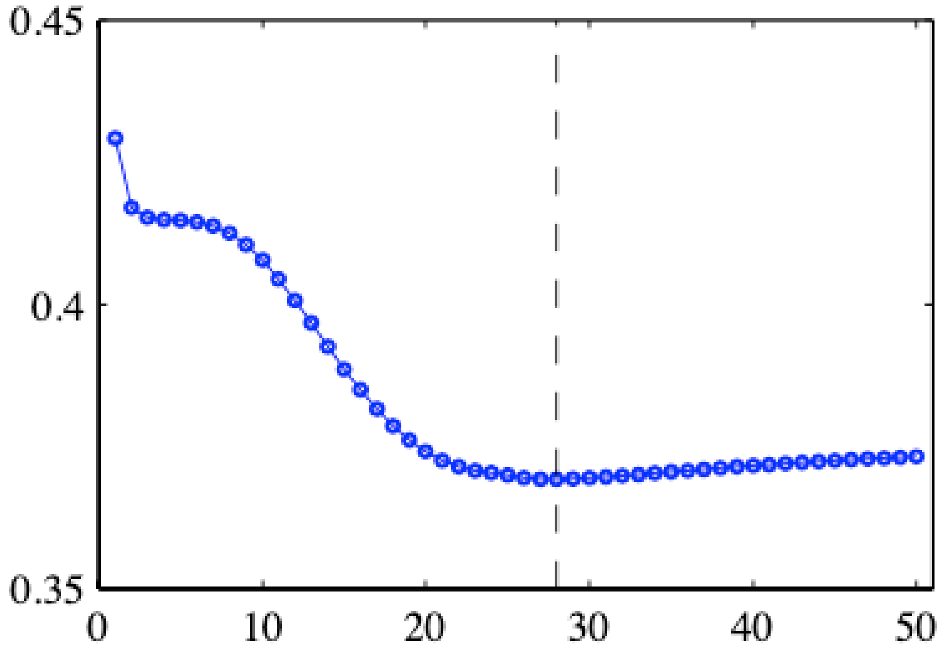
$$\frac{\lambda_1}{2}\sum_{w\in\mathcal{W}_1} w^2 + \frac{\lambda_2}{2}\sum_{w\in\mathcal{W}_2} w^2$$

- 重新缩放： $\lambda_1 \to a^{\frac{1}{2}}\lambda_1$和$\lambda_2 \to c^{-\frac{1}{2}}\lambda_2$.

- 正则化项在权值的变化下不会发生变化。

# early stopping



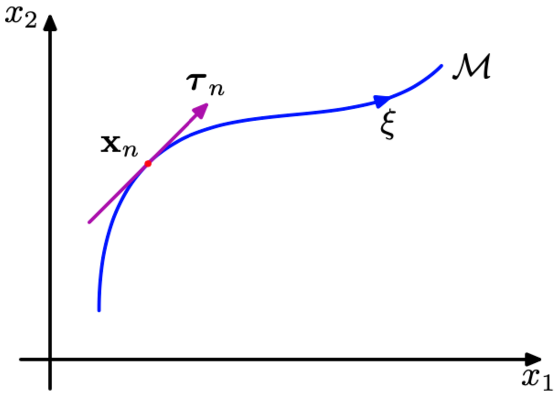




- 在验证集最小的点停止，可以得到一个较好泛化性能的网络。

# 不变性

- 对于一个图像它的类别和它所处的位置是无关的
- 学习到不变性的方法：
- 1.扩展数据集，平移，旋转已有数据，CGAN扩展数据集
- 2.为误差函数增加正则化项，惩罚输入变换时，输出发生的变化，切线传播的方法。

# 切线传播




$$\boldsymbol{\tau}_n = \left.\frac{\partial \boldsymbol{s}(\boldsymbol{x}_n, \xi)}{\partial \xi}\right|_{\xi=0}$$

- 令这个变换作用于xn上产生的，向量为s(xn, ξ)，且s(x, 0) = x
- 变换的效果可以用切向量τ n来近示
- 输出向量 $$\left.\frac{\partial y_k}{\partial \xi}\right|_{\xi=0} = \sum_{i=1}^{D} \left.\frac{\partial y_k}{\partial x_i}\frac{\partial x_i}{\partial \xi}\right|_{\xi=0} = \sum_{i=1}^{D} J_{ki}\tau_i$$

# 切线传播

- 修改误差函数： $\tilde{E} = E + \lambda\Omega$

$$\Omega = \frac{1}{2}\sum_{n}\sum_{k}\left(\left.\frac{\partial y_{nk}}{\partial \xi}\right|_{\xi=0}\right)^2 = \frac{1}{2}\sum_{n}\sum_{k}\left(\sum_{i=1}^{D} J_{nki}\tau_{ni}\right)^2$$

- 通过λ的值确定训练数据和学习不变性之间的平衡

# 用变换后的数据训练

- 单一参数控制的变换，由s(x, ξ)表示。
- 对于未经过变换的输入，误差函数可以写成:

$$E = \frac{1}{2} \iint \{y(\boldsymbol{x}) - t\}^2 p(t \mid \boldsymbol{x}) p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t$$

- 扩展的误差函数：

$$\tilde{E} = \frac{1}{2} \iiint \{y(\boldsymbol{s}(\boldsymbol{x}, \xi)) - t\}^2 p(t \mid \boldsymbol{x}) p(\boldsymbol{x}) p(\xi) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t \, \mathrm{d}\xi$$

- 展开：

$$\begin{aligned} \boldsymbol{s}(\boldsymbol{x}, \xi) &= \boldsymbol{s}(\boldsymbol{x}, 0) + \xi \frac{\partial}{\partial \xi} \boldsymbol{s}(\boldsymbol{x}, \xi) \bigg|_{\xi=0} + \frac{\xi^2}{2} \frac{\partial^2}{\partial \xi^2} \bigg|_{\xi=0} + O(\xi^3) \\ &= \boldsymbol{x} + \xi \boldsymbol{\tau} + \frac{1}{2} \xi^2 \boldsymbol{\tau}' + O(\xi^3) \end{aligned}$$

# 用变换后的数据训练

- 代入：

$$\begin{aligned}\tilde{E} &= \frac{1}{2}\iint \{y(\boldsymbol{x}) - t\}^2 p(t \mid \boldsymbol{x})p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t \\ &+ \mathbb{E}[\xi] \iint \{y(\boldsymbol{x}) - t\}\boldsymbol{\tau}^T \nabla y(\boldsymbol{x}) p(t \mid \boldsymbol{x})p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t \\ &+ \mathbb{E}[\xi^2]\frac{1}{2} \iint \Big[\{y(\boldsymbol{x}) - t\} \left\{(\boldsymbol{\tau}')^T \nabla y(\boldsymbol{x}) + \boldsymbol{\tau}^T \nabla\nabla y(\boldsymbol{x})\boldsymbol{\tau}\right\} \\ &\qquad + \left(\boldsymbol{\tau}^T \nabla y(\boldsymbol{x})\right)^2\Big] p(t \mid \boldsymbol{x})p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t + O(\xi^3)\end{aligned}$$

- 我们把E[ξ2]记作λ：
- 误差函数写成 $\tilde{E} = E + \lambda\Omega$

# 用变换后的数据训练

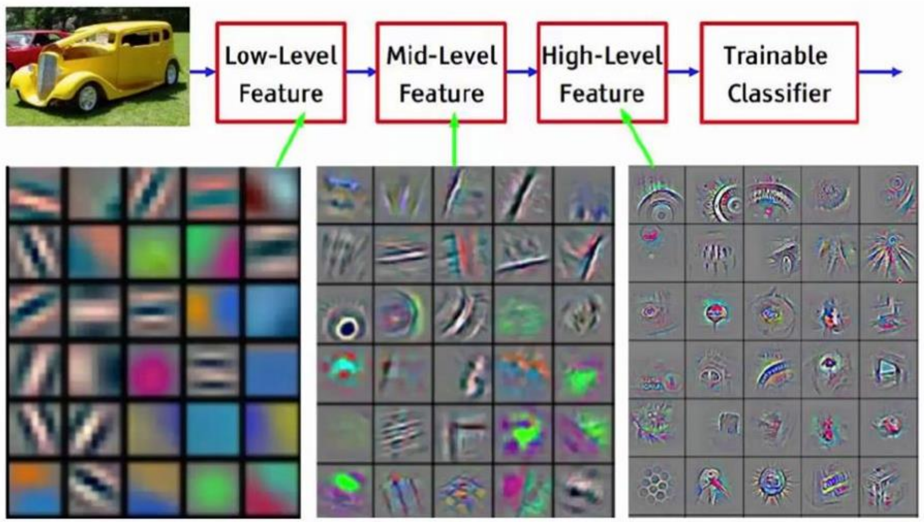
$$\Omega = \frac{1}{2} \int \left( \boldsymbol{\tau}^T \nabla y(\boldsymbol{x}) \right)^2 p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$$

- 和切线传播得到的正则化项等价
- 如果我们考虑一个特殊情况，即输入变量的变换只是简单地添加随机噪声，从而x → x + ξ，那么正则化项的形式为

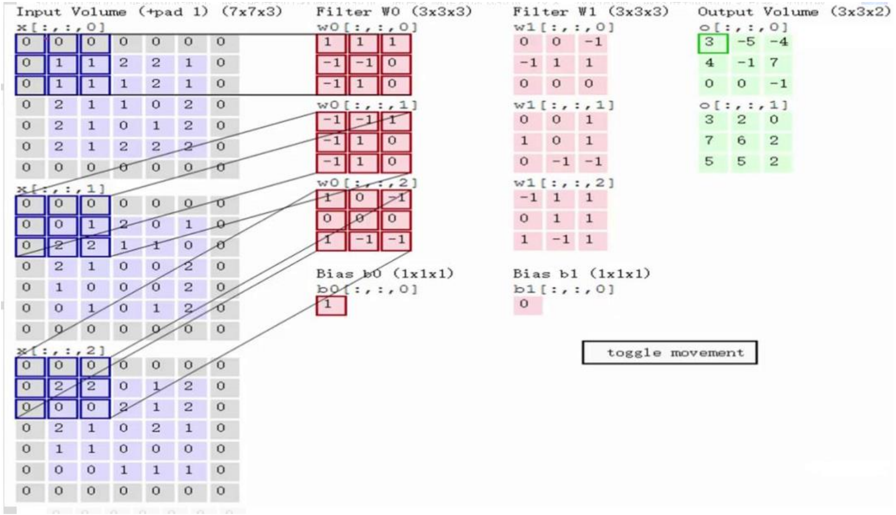
$$\Omega = \frac{1}{2} \int \|\nabla y(\boldsymbol{x})\|^2 p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}$$

# 卷积神经网络

• 1.局部接收场：




• 2.权值共享：



# 软权值共享

- 加入正则化的形式，使权值分组倾向于取更近似的值。采用高斯混合概率分布去分组。
- 概率密度：$$p(\boldsymbol{w}) = \prod_i p(w_i)$$
- 其中：$$p(w_i) = \sum_{j=1}^{M} \pi_j \mathcal{N}(w_i \mid \mu_j, \sigma_j^2)$$
- 取负对数，得到正则化函数：$$\Omega(\boldsymbol{w}) = -\sum_i \ln \left( \sum_{j=1}^{M} \pi_j \mathcal{N}(w_i \mid \mu_j, \sigma_j^2) \right)$$
- 加入正则化的误差函数：$$\tilde{E}(\boldsymbol{w}) = E(\boldsymbol{w}) + \lambda \Omega(\boldsymbol{w})$$

# 软权值共享

- 方便计算{πj }当成先验概率，后验概率：

$$\gamma_j(w) = \frac{\pi_j \mathcal{N}(w \mid \mu_j, \sigma_j^2)}{\sum_k \pi_k \mathcal{N}(w \mid \mu_k, \sigma_k^2)}$$

- 误差关于权值的导数：

$$\frac{\partial \tilde{E}}{\partial w_i} = \frac{\partial E}{\partial w_i} + \lambda \sum_j \gamma_j(w_i) \frac{(w_i - \mu_j)}{\sigma_j^2}$$

- 误差关于均值的导数：

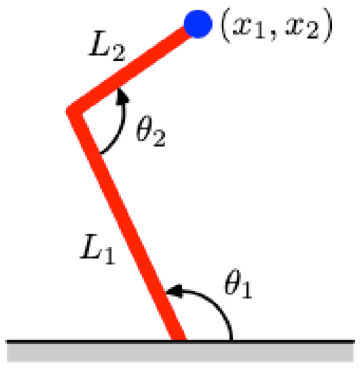
$$\frac{\partial \tilde{E}}{\partial \mu_j} = \lambda \sum_i \gamma_j(w_i) \frac{(\mu_j - w_i)}{\sigma_j^2}$$

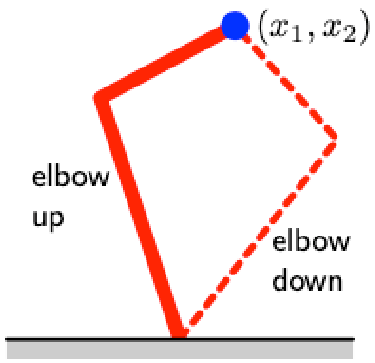
# 软权值共享

- 误差关于方差的导数： $\frac{\partial \tilde{E}}{\partial \sigma_j} = \lambda \sum_i \gamma_j(w_i) \left( \frac{1}{\sigma_j} - \frac{(w_i - \mu_j)^2}{\sigma_j^3} \right)$

# 混合密度网络

- why?
- 正问题和逆问题

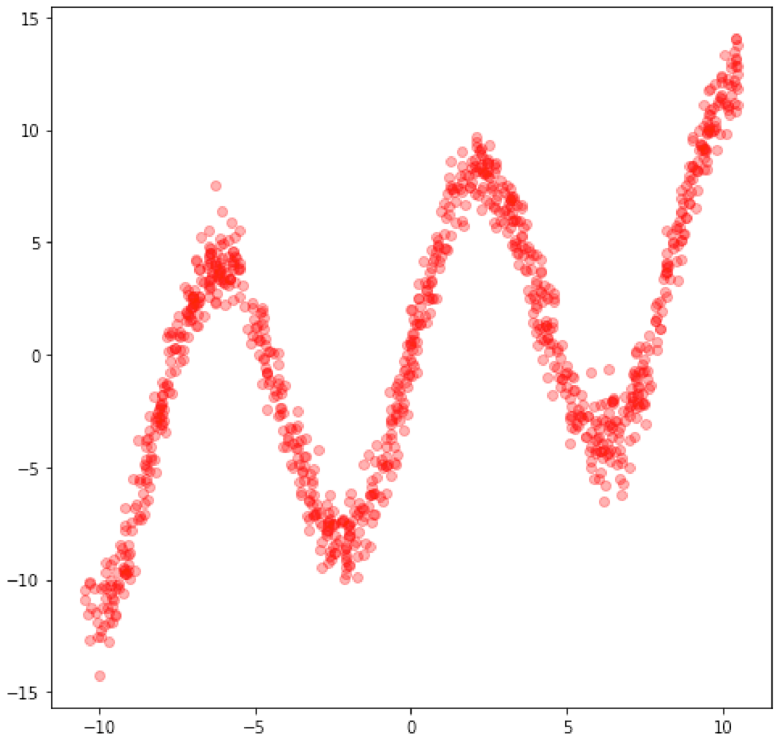


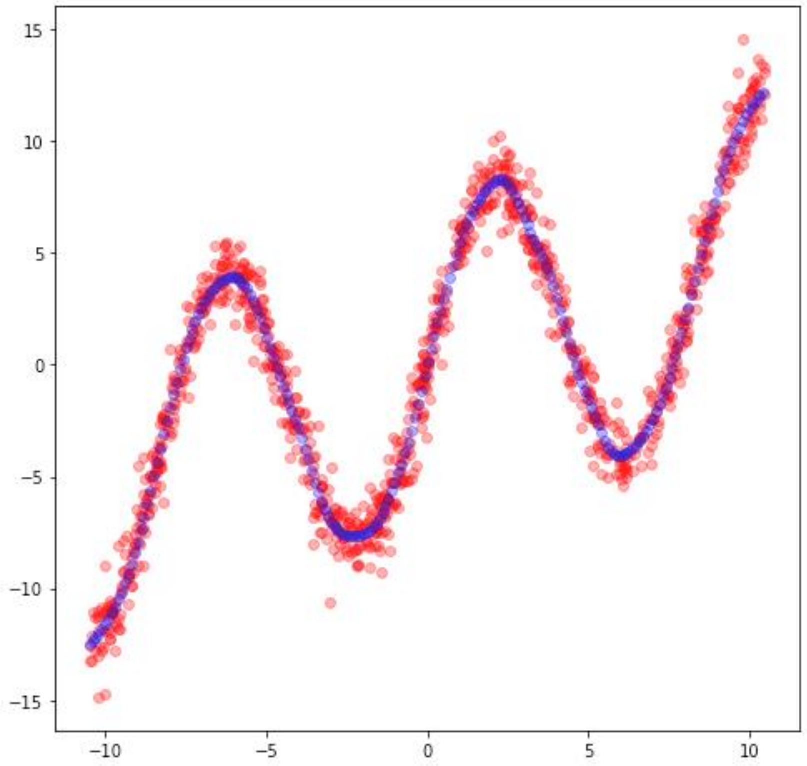




- 1个解，还是多个解

# 混合密度网络

- example:
- 1.单值函数：  $f(x) = 7.0sin(0.75x) + 0.5x$

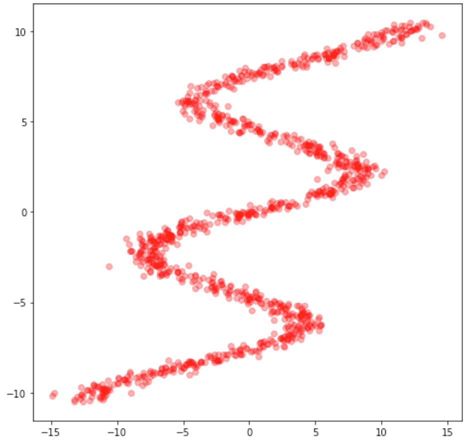


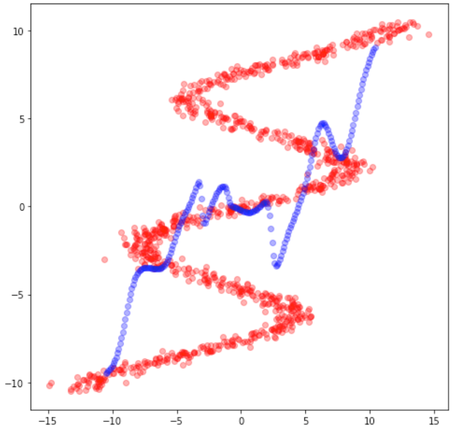



# 混合密度网络

- 交换x,y轴，得到一个多值函数。

$$x = 7.0\sin(0.75y) + 0.5y + \epsilon$$



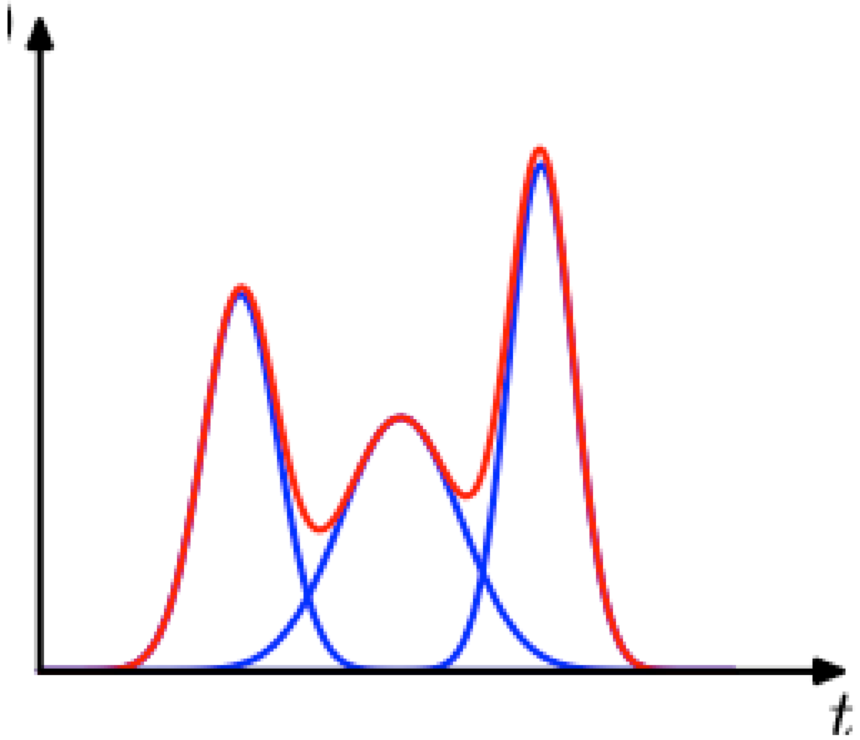




- 引入混合密度网络：
- 高斯分布：

$$p(\mathbf{t} \mid \boldsymbol{x}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{x}) \mathcal{N}(\mathbf{t} \mid \boldsymbol{\mu}_k(\boldsymbol{x}), \sigma_k^2(\boldsymbol{x})\boldsymbol{I})$$

- 混合系数π(x)、均值μ(x)以及方差σ(x)由输入的x确定。

# 混合密度网络




$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

# 混合密度网络

- 混合系数的限制

$$\sum_{k=1}^{K} \pi_k(\boldsymbol{x}) = 1, \quad 0 \leq \pi_k(\boldsymbol{x}) \leq 1$$

$$\pi_k(\boldsymbol{x}) = \frac{\exp(a_k^{\pi})}{\sum_{l=1}^{K} \exp(a_l^{\pi})}$$

- 方差必须满足 $\sigma_k^2(\boldsymbol{x}) \geq 0$

$$\sigma_k(\boldsymbol{x}) = \exp(a_k^{\sigma})$$

- 均值可以表示为： $\mu_{kj}(\boldsymbol{x}) = a_{kj}^{\mu}$

- .误差函数(负对数似然函数) $E(\boldsymbol{w}) = -\sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k(\boldsymbol{x}_n, \boldsymbol{w}) \mathcal{N}(\boldsymbol{t}_n \mid \boldsymbol{\mu}_k(\boldsymbol{x}_n, \boldsymbol{w}), \sigma_k^2(\boldsymbol{x}_n, \boldsymbol{w}) \boldsymbol{I}) \right\}$

# 混合密度网络

- 把混合系数看成与相关的先验概率分布，引入对应的后验概率

$$\gamma_{nk} = \gamma_k(\boldsymbol{t}_n \mid \boldsymbol{x}_n) = \frac{\pi_k \mathcal{N}_{nk}}{\sum_{l=1}^{K} \pi_l \mathcal{N}_{nl}}$$

- 混合系数的导数：

$$\frac{\partial E_n}{\partial a_k^{\pi}} = \pi_k - \gamma_{nk}$$

- 均值的导数：

$$\frac{\partial E_n}{\partial a_{kl}^{\mu}} = \gamma_k \left\{ \frac{\mu_{kl} - t_{nl}}{\sigma_k^2} \right\}$$

- 方差的导数：

$$\frac{\partial E_n}{\partial a_k^{\sigma}} = \gamma_{nk} \left\{ L - \frac{\|\boldsymbol{t}_n - \boldsymbol{\mu}_k\|^2}{\sigma_k^2} \right\}$$

# 混合密度网络

- 应用：
- 手写预测，人体姿态估计等
- cvpr 2019
- Generating Multiple Hypotheses for 3D Human Pose Estimation with Mixture Density Network