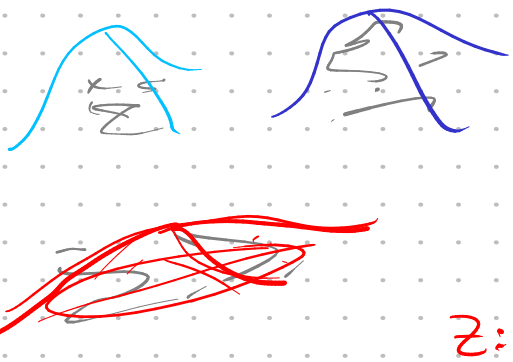


$$p(\bar{x}\mid\bar{\theta}) = \sum_k \pi_k p_k(\bar{x}\mid\bar{\theta}_k)$$

$$p(D\mid\bar{\theta}) = \prod_n \left(\sum_k \pi_k p_k(\bar{x})\right) \xrightarrow[\bar{\theta}]{} \max$$

z: $z_{nk}=1 \Leftrightarrow x_n \in C_k$

$$p(D,z\mid\bar{\theta}) = \prod_n \prod_k \left(\pi_k p_k(\bar{x})\right)^{z_{nk}} \xrightarrow[\theta]{} \max$$

$$\mathbb{E}_{\theta^{(m)}}[z_{nk}] \rightsquigarrow \theta^{(m+1)} = \arg\max p(D, \mathbb{E}[z_{nk}]\mid\theta)$$

EM-algorithm

$$l(\theta) = \log p(X\mid\theta) \xrightarrow[\theta]{} \max$$

latent vars z $\qquad p(X,z\mid\theta)$

$\log p(\theta\mid X) \to \max$?

$$\theta^{(0)} \to \theta^{(1)} \to \theta^{(2)} \to \dots \to \theta^{(m)} \to \theta^{(m+1)} \to \dots$$

$$l(\theta^{(m)}) \le l(\theta^{(m+1)})$$

$$l(\theta) - l(\theta^{(m)}) = \log p(X\mid\theta) - \log p(X\mid\theta^{(m)}) =$$

$$= \log \int p(X,z\mid\theta)\,dz - \log p(X\mid\theta^{(m)}) =$$

$\mathbb{E}[\dots]$

$$= \log \int p(X,z\mid\theta)\cdot \frac{p(z\mid X,\theta^{(m)})}{p(z\mid X,\theta^{(m)})}\,dz - \log p(X\mid\theta^{(m)})$$

$$= \log \mathbb{E}_{p(z\mid X,\theta^{(m)})}\left[\frac{p(X,z\mid\theta)}{p(z\mid X,\theta^{(m)})}\right] - \log p(X\mid\theta^{(m)}) \ge$$

$$f(ax+(1-a)y) \ge a f(x) + (1-a) f(y)$$

$$f(\mathbb{E}_{p(x)}[x]) \ge \mathbb{E}_{p(x)}[f(x)]$$ Jensen's inequality

$$\geq \mathbb{E}_{p(z|x,\theta^{(m)})}\left[\log \frac{p(x,z|\theta)}{p(z|x,\theta^{(m)})}\right] - \log p(x|\theta^{(m)}) =$$

$$= \int p(z|x,\theta^{(m)}) \left(\log \frac{p(x,z|\theta)}{p(z|x,\theta^{(m)})} - \log p(x|\theta^{(m)})\right) dz$$

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

$$\text{“} L(\theta, \theta^{(m)})$$

$+\log p(\theta)$ $\quad -\log p(\theta^{(m)})$

$$l(\theta) - l(\theta^{(m)}) \geq L(\theta, \theta^{(m)})$$

$$l(\theta) \geq l(\theta^{(m)}) + L(\theta, \theta^{(m)})$$

$\theta \longrightarrow \max$

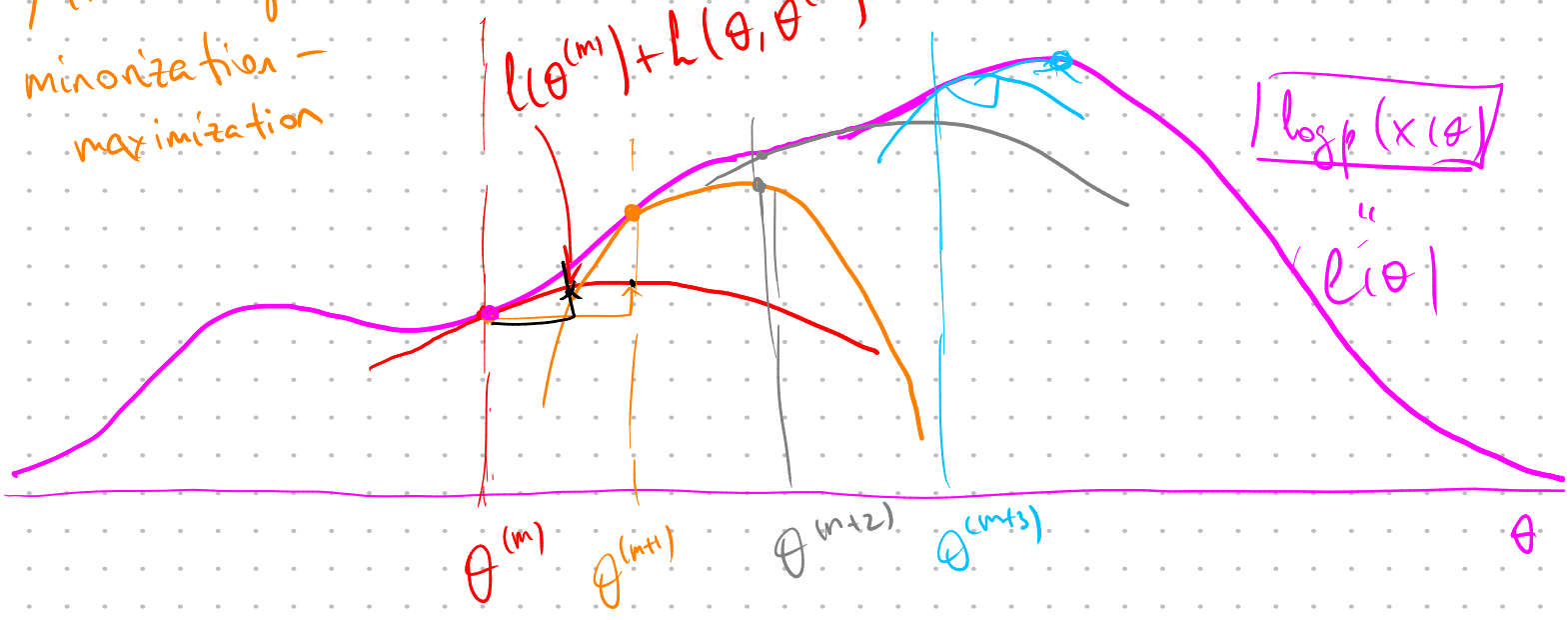
$$\theta^{(m+1)} := \operatorname*{argmax}_{\theta} L(\theta, \theta^{(m)})$$

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

$$L(\theta^{(m)}, \theta^{(m)}) = 0$$

MM-алгоритм

minorization – maximization



$$Q(\theta, \theta^{(m)}) = \int p(z|x,\theta^{(m)}) \log p(x,z|\theta)\, dz =$$

$$= \mathbb{E}_{z|\theta^{(m)}} \left[\log p(x,z|\theta) \right]$$

θ — max

$$X = (\bar{x}_1, \dots \bar{x}_N) \qquad \theta = (\bar{\pi}, \bar{\theta}_1, \dots, \bar{\theta}_K)$$

$$Z = (\bar{z}_1, \dots, \bar{z}_N), \quad z_{nk} = 1 \Leftrightarrow \bar{x}_n \in C_k$$

$\bar{z}_n = (0 \dots 1 \dots 0)$

$$p(X|\theta) = \prod_n \left(\sum_k \pi_k p_k(\bar{x}_n|\bar{\theta}_k) \right) \longrightarrow \max$$

$\log p = l(\theta)$

$$p(x,z|\theta) = \prod_n \prod_k \left(\pi_k p_k(\bar{x}_n|\bar{\theta}_k) \right)^{z_{nk}}$$

$$\log p(x,z|\theta) = \sum_n \sum_k z_{nk} \left(\log \pi_k + \log p_k(\bar{x}_n|\bar{\theta}_k) \right)$$

$$Q(\theta, \theta^{(m)}) = \mathbb{E}_{z|\theta^{(m)}} \left[\log p(x,z|\theta) \right] =$$

$$= \mathbb{E}_{p(z|x,\theta^{(m)})} \left[\sum_n \sum_k z_{nk} \left(\log \pi_k + \log p_k(\bar{x}_n|\bar{\theta}_k) \right) \right] =$$

$$= \sum_n \sum_k \mathbb{E}_{z|x,\theta^{(m)}}[z_{nk}] \cdot \left(\log \pi_k + \log p_k(\bar{x}_n|\bar{\theta}_k) \right] \xrightarrow[\theta]{} \max$$

$$Q(\theta, \theta^{(m)}) = \mathbb{E}_{z|\theta^{(m)}} \left[\log p(x,z|\theta) \right] \xrightarrow[\theta]{} \max$$

① Вместо argmax найдём θ: $Q(\theta, \theta^{(m)}) \geq Q(\theta^{(m)}, \theta^{(m)})$

Generalized EM

② $Q(\theta, \theta^{(m)}) \approx \frac{1}{R} \sum_{r=1}^{R} \log p(x, z_r|\theta)$, где $z_r \sim p(z|x,\theta^{(m)})$

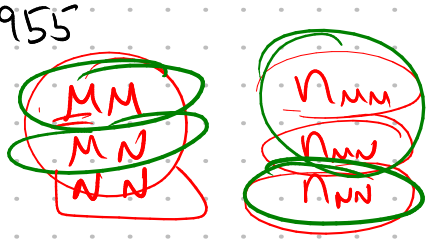
Monte Carlo EM

$(R=1) \rightarrow$ Stochastic EM

Пример Serfling et al., 1955

MN - система

↑ ген.



$$p = \frac{2n_{MM} + n_{MN}}{2(n_{MM} + n_{MN} + n_{NN})}$$

$$q = \frac{2n_{NN} + n_{MN}}{2(n_{MM} + n_{MN} + n_{NN})}$$

Харди-Вайнберга:

MM, NN - с вероятн. $\frac{p^2}{p^2 + 2pq}$

MN — — — $\frac{2pq}{p^2 + 2pq}$

$$\theta_{MAP} = \arg\max_\theta \log p(\theta|X) = \arg\max_\theta \left(\log p(\theta) + \log p(X|\theta) \right)$$

$$\log p(\theta|X) - \log p(\theta^{(m)}|X) = \text{const} + \log p(X|\theta) + \log p(\theta) - \log p(X|\theta^{(m)}) - \log p(\theta^{(m)})$$

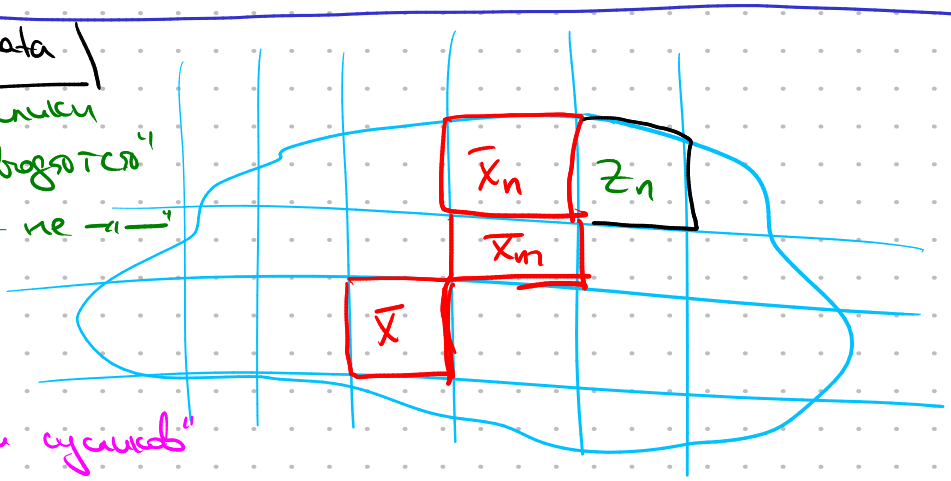
$$\geq L(\theta, \theta^{(m)}) + \log p(\theta) - \log p(\theta^{(m)}) + \text{const}$$

$$\theta^{(m+1)} = \arg\max_\theta \left(Q(\theta, \theta^{(m)}) + \log p(\theta) \right)$$

Presence-only data

$z_n = 1$ ⇔ "кустики водятся"

$z_n = 0$ ⇔ " — " — не — " — "



$y_n = 1$ ⇔ "видели кустики"

$y_n = 1 \Rightarrow z_n = 1$

$y_n = 0 \Rightarrow$?

$p(z|\bar{x})$

$$p(z=1|\bar{x}) = \sigma(\eta(\bar{x})) = \frac{1}{1 + e^{-\eta(\bar{x})}}$$

Prospective vs. retrospective studies

	$d=0$	$d=1$	
$x=0$	π_{00}	π_{01}	π_{0*}
$x=1$	π_{10}	π_{11}	π_{1*}
$p(d=1)$	π_{*0}	π_{*1}	

Prospective

$$p(d=1 \mid x=0) = \frac{\pi_{01}}{\pi_{00}+\pi_{01}} \approx \sigma(\eta(x))$$

$$p(d=1 \mid x=1) = \frac{\pi_{11}}{\pi_{10}+\pi_{11}} \approx 1-\sigma(\eta(x))$$

Retrospective

$$\pi_0 = p(s=1 \mid d=0)$$

$$\pi_1 = p(s=1 \mid d=1)$$

$$\frac{\pi_{01}}{\pi_{00}+\pi_{01}} = p(d=1 \mid x=0, s=1) =$$

$$= \frac{p(d=1 \mid x=0)\,\overbrace{p(s=1 \mid d=1, x=0)}^{\pi_1}}{p(d=1 \mid x=0)\underbrace{p(s=1 \mid d=1, x=0)}_{=\pi_1} + p(d=0 \mid x=0)\underbrace{p(s=1 \mid d=0, x=0)}_{=\pi_0}}$$

$$\frac{\pi_{01}}{\pi_{00}+\pi_{01}} = \frac{\pi_1\, p(d=1 \mid x=0)}{\pi_1\, p(d=1 \mid x=0) + \pi_0\,(1-p(d=1 \mid x=0))}$$

$$\pi_{01}(\pi_1 p + \pi_0 - \pi_0 p) = \pi_1 p(\pi_{00}+\pi_{01})$$

$$\pi_0\pi_{01} - \pi_0\pi_{01}p = \pi_1 p \pi_{00}$$

$$\sigma(\eta(x)) = p = \frac{\pi_0\pi_{01}}{\pi_{00}\pi_1 + \pi_{01}\pi_0}$$

$$\frac{\pi_{01}}{\pi_{00}+\pi_{01}} = \frac{\pi_1\sigma}{\pi_0 + (\pi_1-\pi_0)\sigma} = \frac{\pi_1 / 1+e^{-\eta(x)}}{\pi_0 + \frac{\pi_1-\pi_0}{1+e^{-\eta(x)}}} =$$

$$= \frac{\pi_1}{\pi_0(1+e^{-\eta(x)}) + \pi_1 - \pi_0} = \frac{\pi_1}{\pi_1 + \pi_0 e^{-\eta(x)}}$$

$$\frac{\pi_{01}}{\pi_{00}+\pi_{01}} = \frac{1}{1+\frac{\pi_0}{\pi_1}e^{-\eta(x)}} = \frac{1}{1+e^{-(\eta(x)-\ln\frac{\pi_0}{\pi_1})}}$$

$$\frac{\pi_{01}}{\pi_{00}+\pi_{01}} = \sigma\left(\eta(\bar{x}) + \ln\frac{\pi_0}{\pi_1}\right)$$

$\eta(\bar{x}) \to w_0 + w_1 x$

$$\sigma(\eta^*(\bar{x})) \approx \frac{\pi_{01}}{\pi_{00}+\pi_{01}} \longrightarrow w_0^*, w_1^*$$

$$\Rightarrow w_1 = w_1^*, \quad w_0 = w_0^* - \ln\frac{\pi_0}{\pi_1}$$

$\sigma(\eta(\bar{x})) = p(z=1|\bar{x})$

$s=1$ – выборка

$\pi = p(z=1)$ – вспрещенность

n_p – positive

n_u – unknown

$\sigma(\eta_{naive}(\bar{x})) = p(y=1|\bar{x}, s=1)$

выборка:

$z=1$:	$n_p + \pi \cdot n_u$
$z=0$:	$(1-\pi)n_u$

$$p(y=1|\bar{x}, s=1) =$$

$$= p(y=1|z=1,s=1,\bar{x})\,p(z=1|s=1,\bar{x}) + \underbrace{p(y=1|z=0,s=1,\bar{x})}_{\sim 0}\,p(z=0|s=1,\bar{x}) =$$

$$= p(y=1|z=1,s=1,\bar{x})\,p(z=1|s=1,\bar{x}) = \sigma(\eta(\bar{x}) - \log\frac{\pi_0}{\pi_1})$$

$$= \frac{p(y=1,z=1|s=1)}{p(z=1|s=1)} = \frac{n_p/N}{n_p+\pi n_u/N} = \frac{n_p}{n_p+\pi n_u}$$

предп. перед.

$$\pi_1 = p(s=1|z=1) = \frac{p(z=1|s=1)\,p(s=1)}{p(z=1)} = \frac{\frac{n_p+\pi n_u}{n_p+n_u}\,p(s=1)}{\pi}$$

$\pi_0 = p(s=1|z=0)$

$$\pi_1 = \frac{n_p + \pi n_u}{\pi(n_p + n_u)} \cdot p(s=1)$$

$$\pi_0 = \frac{(1-\pi) n_u}{(1-\pi)(n_p + n_u)} p(s=1)$$

$$\frac{\pi_0}{\pi_1} = \frac{\pi n_u}{n_p + \pi n_u}$$

$$\sigma(\eta_{naive}(\bar{x})) = \frac{n_p}{n_p + \pi n_u} \cdot \sigma\left(\eta(\bar{x}) - \log \frac{\pi n_u}{n_p + \pi n_u}\right)$$

$$p(y | X, \bar{\eta}) = \prod_{i=1}^{n_p+n_u} p(y_i | s_i=1, \bar{x}_i) =$$

$$= \prod_{i=1}^{n_p+n_u} \left(p(y_i=1 | s_i=1, \bar{x}_i)\right)^{y_i} \left(1 - p(y_i=1 | s_i=1, \bar{x}_i)\right)^{1-y_i} =$$

$$= \prod_{i=1}^{n_p+n_u} \left(\frac{n_p}{n_p + \pi n_u} \sigma\left(\eta(\bar{x}) - \log \frac{\pi n_u}{n_p + \pi n_u}\right)\right)^{y_i} \left(1 - \frac{n_p}{n_p + \pi n_u} \sigma(\dots)\right)^{1-y_i} \xrightarrow[\bar{\eta}]{} \max$$

$= \eta_{naive}$

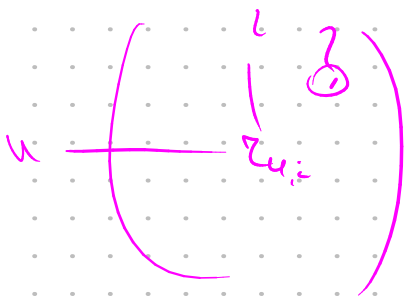
EM-algorithm: $\bar{\eta}^{(0)} \to \bar{\eta}^{(1)} \to \bar{\eta}^{(2)} \to \dots \bar{\eta}^{(m)} \to \bar{\eta}^{(m+1)} \to \dots$

M-шаг: $\sigma\left(\eta(\bar{x}_i) - \log \frac{\pi_0}{\pi_1}\right) \approx z_i^{(m+1)}$, $\eta^{(m+1)} = \eta^* + \log \frac{\pi_0}{\pi_1}$,

E-шаг: $z_i^{(m+1)} = \underset{\eta^{(m)}}{E}[z_i] = \sigma(\eta^{(m)}(\bar{x}_i))$ где $y_i = 0$

$= 1$, $y_i = 1$

pseudolabels



implicit feedback