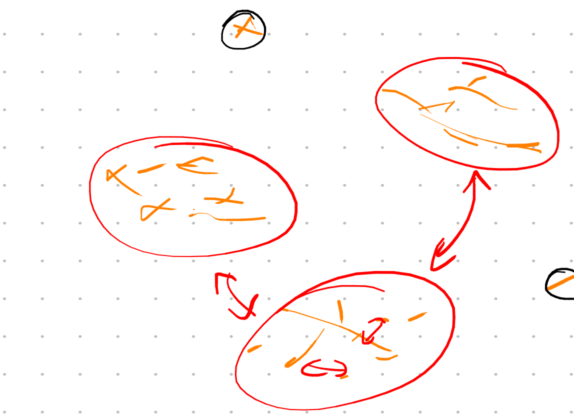
(1) Clustering – unsupervised learning

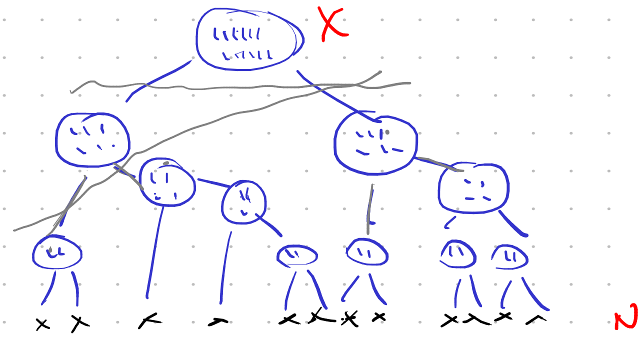
$\bar{x} \in \mathbb{R}^d$

$X = \{\bar{x}_n\}_{n=1}^{N}$



- Hierarchical clustering

Agglomerative Clustering

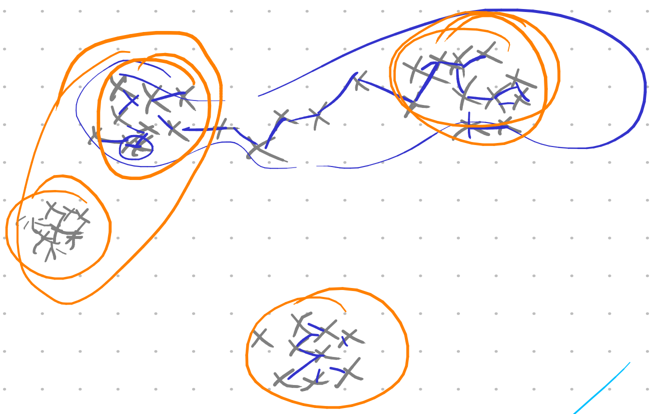


- single-link clustering

$$d(C_1, C_2) = \min_{\substack{x \in C_1 \\ y \in C_2}} d(x, y)$$

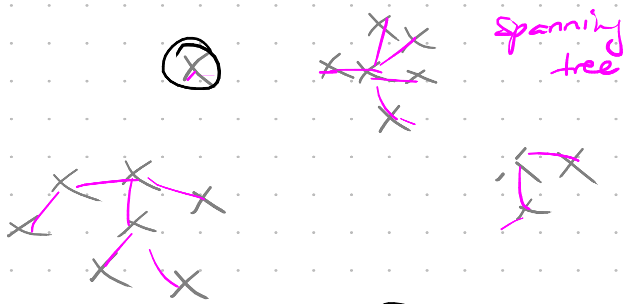
- complete-link clustering

$$d(C_1, C_2) = \max_{x \in C_1,\, y \in C_2} d(x, y)$$

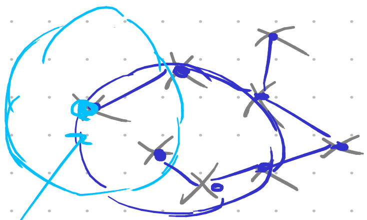


- graph theory

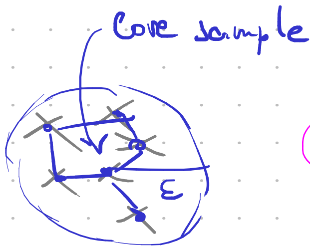
MST
minimal spanning tree

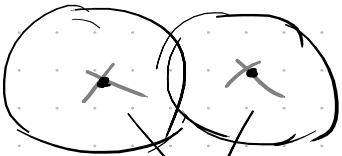


- DBSCAN (ε, min_samples)

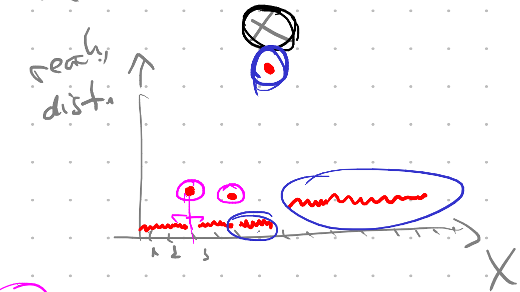


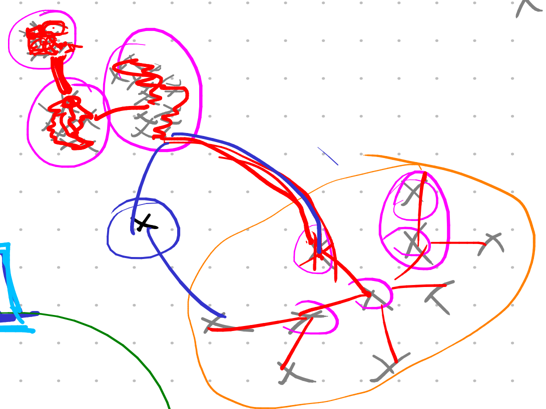
reachable sample

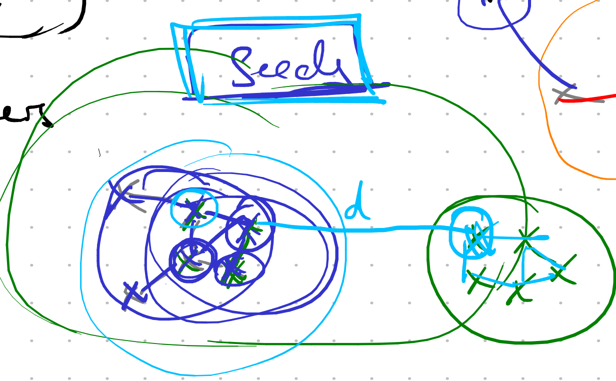




outliers





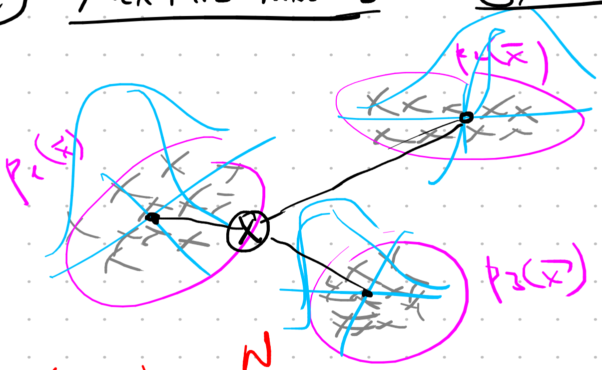


- OPTICS
- BIRCH

② Mixture models   GMM

$$\cancel{\prod_n \left(y_n p_1(\bar x) + (1-y_n) p_2(\bar x_n)\right)}$$

$$\prod_n p_1(\bar x_n)^{y_n}\, p_2(\bar x_n)^{1-y_n}$$



$\bar x \in \mathbb{R}^d$

$p(\bar x)$

$$p(\bar x) = \alpha_1 \underbrace{p_1(\bar x)}_{\bar\theta_1} + \alpha_2 \underbrace{p_2(\bar x)}_{\bar\theta_2} + \alpha_3 \underbrace{p_3(\bar x)}_{\bar\theta_3}$$

$$\alpha_1 + \alpha_2 + \alpha_3 = 1$$

$$X = \{\bar x_n\}_{n=1}^{N}$$

$$p(X \mid \bar\pi, \bar\theta_1, \dots, \bar\theta_K) = \prod_{n=1}^{N} \left(\pi_1 p_1(\bar x_n \mid \bar\theta_1) + \pi_2 p_2(\bar x_n \mid \bar\theta_2) + \dots + \pi_K p_K(\bar x_n \mid \bar\theta_K) \right) \xrightarrow[\bar\pi,\, \bar\theta_1 \dots \bar\theta_K]{} \max$$

$\bar\pi$ $\longrightarrow$ $\underbrace{k - \text{\# кластера}}$ $\longrightarrow$ $\bar x \sim p_k(\bar x \mid \bar\theta_k)$

$$\bar z = (0 \dots \overset{k}{1} \dots 0) \qquad z_k = [\bar x \in C_k]$$

$$p(X, Z \mid \bar\pi, \bar\theta_1 \dots \bar\theta_K) = \prod_{n=1}^{N} p(\bar x_n, \bar z_n \mid \bar\pi, \bar\theta_1 \dots \bar\theta_K) =$$

$$= \prod_{n=1}^{N} p(\bar z_n \mid \bar\pi)\, p(\bar x_n \mid \bar z_n, \bar\theta_1 \dots \bar\theta_K) =$$

$$= \prod_{n=1}^{N} \prod_{k=1}^{K} \underbrace{p(z_n = k \mid \bar\pi)}_{\pi_k}{}^{z_{nk}} \cdot \prod_{k=1}^{K} p_k(\bar x_n \mid \bar\theta_k)^{z_{nk}} =$$

$$= \prod_{n=1}^{N} \prod_{k=1}^{K} \left(\pi_k \cdot p_k(\bar x_n \mid \bar\theta_k) \right)^{z_{nk}}$$

$$\log p(X, Z \mid \bar\pi, \bar\theta_1 \dots \bar\theta_K) = \sum_{n=1}^{N} \sum_{k=1}^{K} \left[\underline{z_{nk} \log \pi_k} + \underline{z_{nk} \log p_k(\bar x_n \mid \bar\theta_k)} \right] =$$

$$\#[z=k] \sim n_k$$

$$= \underbrace{\sum_{k=1}^{K} \left(\sum_{n=1}^{N} z_{nk} \right) \log \pi_k}_{} + \underbrace{\sum_{k=1}^{K} \left(\sum_{n=1}^{N} z_{nk} \log p_k(\bar x_n \mid \bar\theta_k) \right)}_{}$$

$$\xrightarrow[\bar\pi]{} \max \qquad\qquad \xrightarrow[\bar\theta_k]{} \max$$

$$\hat\pi_k = \frac{\sum_n z_{nk}}{N} \qquad\qquad \hat{\bar\theta}_k = \arg\max_{\bar\theta_k} \sum_{n:\, z_{nk}=1} \log p_k(\bar x_n \mid \bar\theta_k)$$

③ Expectation-Maximization Algorithm

X, $\bar\theta$, Z - latent variables: $\quad - \; p(X|\bar\theta) \xrightarrow{\bar\theta} \max$ трудно

$- \; p(X,Z|\bar\theta) \xrightarrow{\bar\theta} \max$ лесно

(E-step) fix $\bar\theta$, find $\overset{\exp}{\mathbb{E}[Z]}$

k-means: $k = \arg\min d(\bar x_n, \bar\mu_k)$

$z_{nk}=1 \iff k = \arg\max_s p(z_{ns}=1|\ldots)$

$$\mathbb{E}[z_{nk}] = p(z_{nk}=1|\underbrace{\bar\pi, \bar\theta_1 \ldots \bar\theta_k}, X) = p(z_{nk}=1|\bar x_n, \bar\pi, \bar\theta_1 \ldots \bar\theta_k) =$$

$$= \frac{p(z_{nk}=1, \bar x_n|\bar\pi, \bar\theta_1 \ldots \bar\theta_k)}{p(\bar x_n|\bar\theta) = \sum_s p(\bar x_n, z_{ns}=1|\bar\theta)} = \frac{\pi_k \cdot p_k(\bar x_n|\bar\theta_k)}{\sum_{s=1}^{K} \pi_s\, p_s(\bar x_n|\bar\theta_s)}$$

(M-step) fix $\mathbb{E}[Z]$, $\quad \mathbb{E}[\log p(X, Z|\bar\theta)] \xrightarrow{\bar\theta} \max$

$$\mathbb{E}_Z[\log p(X,Z|\bar\theta)] = \mathbb{E}_Z\Big[\sum_n \sum_k z_{nk}(\log \pi_k + \log p_k(\bar x_n|\bar\theta_k))\Big] =$$

$$= \sum_n \sum_k \mathbb{E}[z_{nk}](\log\pi_k + \log p_k(\bar x_n|\bar\theta_k))$$

k-means:

$\bar\mu_k := \operatorname{Avg}_{n:\, z_{nk}=1} \bar x_n$

$$\sum_k \Big(\sum_n \mathbb{E}[z_{nk}]\Big)\log\pi_k \xrightarrow{\bar\pi} \max$$

$$\forall k \quad \sum_n \mathbb{E}[z_{nk}] \log p_k(\bar x_n|\bar\theta_k) \xrightarrow{\theta_k} \max$$

$$\sum_n \mathbb{E}[z_{nk}] \log \mathcal{N}(\bar x_n|\bar\mu_k, \Sigma_k) \longrightarrow \max$$

④ EM-алгоритъм в общия вид

X - observables, $\bar\theta$, $p(X|\bar\theta) \xrightarrow{\bar\theta} \max$ - трудно

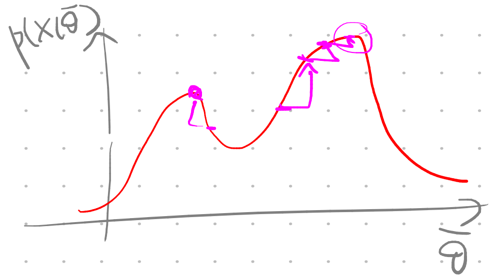
$$p(X|\bar\theta) = \int p(X, Z|\bar\theta)\, dZ \xrightarrow{\bar\theta} \max$$

EM-algorithm: $\quad$ - init $\bar\theta^{(0)}$

- for $m = 0, 1, \ldots$

$$Q(\bar\theta, \bar\theta^{(m)}) = \mathbb{E}_{p(Z|X, \bar\theta^{(m)})}[\log p(X, Z|\bar\theta)]; \quad \bar\theta^{(m+1)} = \arg\max_{\bar\theta} Q(\bar\theta, \bar\theta^{(m)})$$

Yeas: $p(X|\bar\theta^{(m+1)}) \geq p(X|\bar\theta^{(m)})$

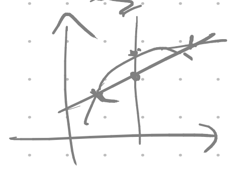


$$\sum_n \sum_{k=1}^{K} \pi_k p_k(\bar x_n|\bar\theta_k)$$

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

$$\log p(x|\bar\theta) - \log p(x|\bar\theta^{(m)}) = \log \int p(x,z|\bar\theta)\,dz - \log p(x|\bar\theta^{(m)}) = f(E_p[x])$$

$$f(\alpha x + (1-\alpha) y)$$



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

$$= \log E_{p(z|x,\bar\theta^{(m)})}\left[\frac{p(x,z|\bar\theta)}{p(z|x,\bar\theta^{(m)})}\right] - \log p(x|\bar\theta^{(m)}) \geq \text{[Jensen's ineq.]}$$

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

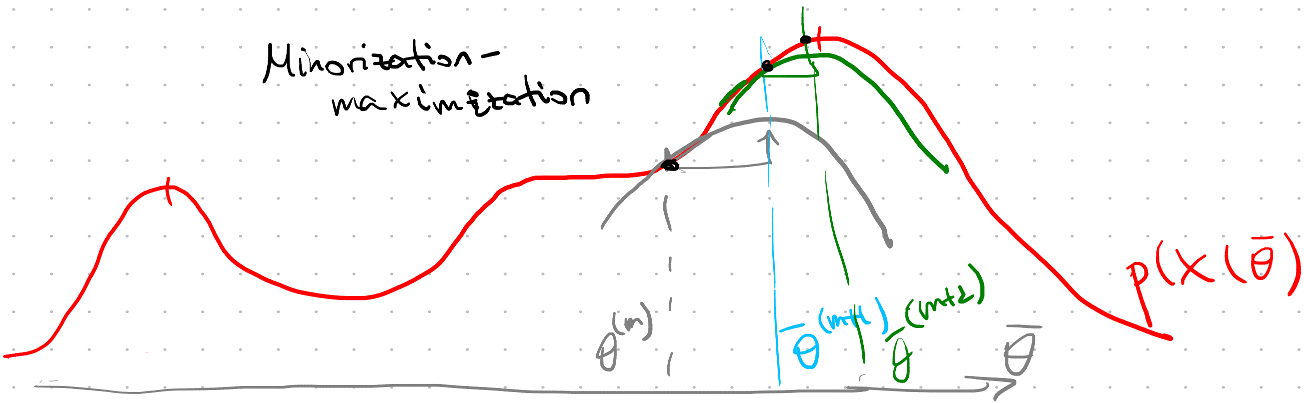
$$= \int p(z|\theta,\bar\theta^{(m)}) \log \frac{p(x,z|\bar\theta)}{p(z|x,\bar\theta^{(m)})\, p(x|\bar\theta^{(m)})}\,dz$$

$$p(z|x,\bar\theta^{(m)})\, p(x|\bar\theta^{(m)}) = p(x,z|\bar\theta^{(m)})$$

$$\log p(x|\bar\theta) - \log p(x|\bar\theta^{(m)}) \geq E_{p(z|\bar\theta,\bar\theta^{(m)})}\left[\log \frac{p(x,z|\bar\theta)}{p(x,z|\bar\theta^{(m)})}\right]$$

$$\log p(x|\bar\theta) \geq L(\bar\theta,\bar\theta^{(m)}) = \log p(x|\bar\theta^{(m)}) + E[\ldots]$$

Minorization-maximization



$$l_{\sim}(\bar\theta, \bar\theta^{(m)}) = \underbrace{\cancel{\log p(x|\bar\theta^{(m)})}}_{\text{const}} + \mathbb{E}_{p(z|x,\bar\theta^{(m)})}\left[\log \frac{p(x,z|\bar\theta)}{\cancel{p(x,z|\bar\theta^{(m)})}}\right] \to \max_{\bar\theta}$$

$$\Longleftrightarrow \quad Q(\bar\theta, \bar\theta^{(m)}) \to \max_{\bar\theta}$$

---

5. EM для mixture models

$$p(\bar x, \bar z|\bar\theta) = \sum \pi_k p_k(\bar x|\bar\theta_k)$$

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

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

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

$$= \sum_n \sum_k \underbrace{\mathbb{E}[z_{nk}]}_{\text{E-step}} \cdot \left(\log \pi_k + \log p_k(\bar x|\bar\theta_k)\right) \xrightarrow[\bar\pi,\, \bar\theta_1 \ldots \bar\theta_K]{\text{M-step}} \max$$

6. Cappellini et al. 1955

домин. $+$
рецесс. $-$

- a: $++$
- b: $+-$, $-+$
- c: $--$

a, b — фенотип A; c — фенотип B

$q = p(+)$, $1-q = p(-)$ $\qquad q = ?$

$$\hat q = \frac{2a+b}{2(a+b+c)} \qquad 1-\hat q = \frac{b+2c}{2(a+b+c)}$$

Дано: $(a+b)$, c

$$p(++) = q^2, \quad p(--) = (1-q)^2, \quad p(+-) = 2q(1-q)$$

$q^{(0)}$,
$$\mathbb{E}_{q^{(0)}}[a] = (a+b)\cdot p(++|\text{фен. } A) = (a+b)\cdot \frac{q^2}{q^2+2q(1-q)} = (a+b)\cdot\frac{q}{2-q}$$

$$\mathbb{E}_{q^{(0)}}[b] = (a+b) - \mathbb{E}[a] = (a+b)\left(1 - \frac{q}{2-q}\right)$$

E-step

M-step:
$$q^{(1)} := \mathbb{E}\left[\frac{2a+b}{2n}\right] = \frac{a+b}{2n}\cdot\left(\frac{2q^{(0)}}{2-q^{(0)}} + \left(1 - \frac{q^{(0)}}{2-q^{(0)}}\right)\right) = \frac{a+b}{\underbrace{2n}_{2(a+b+c)}}\cdot\frac{2}{2-q^{(0)}}$$

7 String clustering

$\{A, C, G, T\}$ — L

L = размер алфавита

$$\begin{array}{l} 1 \qquad\qquad\qquad\qquad\quad M \text{ — длина строки} \\ A\,C\,C\,G\,T\,G\,A\,A\,G\,C \\ A\,C\,G\,C\,T\,G\,A\,T\,G\,C \end{array}$$

$$d = \frac{\#[\cdot \ne \cdot]}{M}$$

$k = 1, \dots, K, \quad \pi_k = p(C_k)$

$C_k: \quad p(x_m = a_l \mid C_k) = \theta_{kml}$

$$p(\bar{x} \mid C_k) = \prod_{m=1}^{M} p(x_m \mid C_k) = \prod_{m=1}^{M} \prod_{l=1}^{L} p(x_m = a_l \mid C_k)^{[x_m = a_l]} = \prod_{m=1}^{M} \prod_{l=1}^{L} \theta_{kml}^{[x_m = a_l]}$$

$\forall k \, \forall m \; \sum_l \theta_{kml} = 1$

latent vars: $\quad z_{nk} = [\bar{x}_n \in C_k]$

$$p(X, Z \mid \bar{\theta}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \left(\pi_k \, p(\bar{x}_n \mid C_k)\right)^{z_{nk}} = \prod_n \prod_k \left(\pi_k \prod_m \prod_l \theta_{kml}^{[x_{nm} = a_l]}\right)^{z_{nk}}$$

E-шаг: $\quad \pi, \theta \to$ мат

$$\mathbb{E}[z_{nk}] = \frac{\pi_k \prod_m \prod_l \theta_{kml}^{[x_{nm} = a_l]}}{\sum_s \pi_s \prod_m \prod_l \theta_{sml}^{[x_{nm} = a_l]}}$$

M-шаг:

$$\mathbb{E}_Z\left[\log p(X, Z \mid \bar{\theta})\right] = \sum_n \sum_k \mathbb{E}[z_{nk}] \cdot \left(\log \pi_k + \sum_m \sum_l [x_{nm} = a_l] \log \theta_{kml}\right)$$

$$= \underbrace{\sum_k \left(\sum_n \mathbb{E}[z_{nk}]\right) \log \pi_k}_{\bar{\pi} \; \max} + \sum_k \sum_m \underbrace{\sum_l \left(\sum_n [x_{nm} = a_l] \cdot \mathbb{E}[z_{nk}]\right) \cdot \log \theta_{kml}}_{\bar{\theta}_{km*} \; \max}$$

$$\theta_{kml}^{(n+1)} = \frac{\sum_n [x_{nm} = a_l] \cdot \mathbb{E}[z_{nk}]}{\sum_{l'} \sum_n [x_{nm} = a_{l'}] \, \mathbb{E}[z_{nk}]} \qquad \sum_{l'} \sum_n [x_{nm} = a_{l'}] \, \mathbb{E}[z_{nk}] = \sum_n \mathbb{E}[z_{nk}]$$

$\log a_1, \log a_2, \dots, \log a_n$

$$\log a_1' = \log \frac{e^{\log a_1}}{e^{\log a_1} + \dots + e^{\log a_n}} = \log a_1 - \underbrace{\log\left(e^{\log a_1} + \dots + e^{\log a_n}\right)}_{\text{logaddexp}}$$