DAGMA无环性表示

□ 提出对数行列式形式的无环性表示：具有3点优势

$$\mathbb{W}^s = \{W \in \mathbb{R}^{d\times d} \mid s > \rho(W \circ W)\}$$

Theorem 1 (Log-determinant characterization). *Let $s > 0$ and let $h^s_{\text{ldet}} : \mathbb{W}^s \to \mathbb{R}$ be defined as* $h^s_{\text{ldet}}(W) \stackrel{\text{def}}{=} -\log\det(sI - W \circ W) + d\log s$. *Then, the following holds:*

(i) $h^s_{\text{ldet}}(W) \geq 0$, *with* $h^s_{\text{ldet}}(W) = 0$ if and only if W *is a DAG.*

(ii) $\nabla h^s_{\text{ldet}}(W) = 2(sI - W \circ W)^{-\top} \circ W$, *with* $\nabla h^s_{\text{ldet}}(W) = 0$ if and only if W *is a DAG.*

Lemma 1. *Let $\mathbb{W}^s$ defined as in (5). Then, for all $s > 0$:*

(i) DAGs $\subset \mathbb{W}^s$. *(ii) $\mathbb{W}^s$ is path-connected.* *(iii) $\mathbb{W}^s \subset \mathbb{W}^t$ for any $t > s$*

Corollary 3. *Let $s > 0$. Then, $h^s_{\text{ldet}}(W)$ is an invex function, i.e., all its stationary points are global minima, and these correspond to DAGs.*

DAGMA首次明确指出$h^s_{ldet}(W)$，$h_{expm}(W)$和$h_{poly}(W)$都是凸函数。

$$h_{\text{expm}}(W) = \operatorname{Tr}(e^{W \circ W}) - d \qquad h_{\text{poly}}(W) = \operatorname{Tr}((I + \tfrac{1}{d} W \circ W)^d) - d$$

[Bello K, Aragam B, Ravikumar P. Dagma: Learning dags via m-matrices and a log-determinant acyclicity characterization[J]. Advances in Neural Information Processing Systems, 2022, 35: 8226-8239.]

对数行列式无环表示优势

□ $h^{s}_{ldet}(W)$不会减弱任意长度环的信息：

➢ 将$h_{expm}(W)$和$h_{poly}(W)$表示成矩阵幂形式：

$$e^{W\circ W}=\sum_{k=0}^{\infty}\frac{1}{k!}(W\circ W)^k$$

$$h_{\text{expm}}(W)=\operatorname{Tr}(e^{W\circ W})-d$$

$$h_{\text{expm}}(W)=\textstyle\sum_{k=0}^{\infty}\boxed{1/k!}\operatorname{Tr}((W\circ W)^k)-d$$

$$\left(I+\frac{1}{d}W\circ W\right)^d=\sum_{k=0}^{d}\frac{\binom{d}{k}}{d^k}(W\circ W)^k$$

$$h_{\text{poly}}(W)=\operatorname{Tr}((I+\tfrac{1}{d}W\circ W)^d)-d$$

$$h_{\text{poly}}(W)=\textstyle\sum_{k=0}^{d}\boxed{\binom{d}{k}/d^k}\operatorname{Tr}((W\circ W)^k)-d$$

$\operatorname{Tr}((W\circ W)^k)$ **表示所有长度为k的环上加权邻接矩阵元素平方乘积之和。**

$h_{expm}(W)$和$h_{poly}(W)$中长为k的环的信息分别被减弱 $\boxed{1/k!}$ **和** $\boxed{\binom{d}{k}/d^k}$**。**

Lemma 4. *For all $W\in\mathbb{W}^{s=1}$, we have $h_{\text{poly}}(W)\le h_{\text{expm}}(W)\le h^{s=1}_{\text{ldet}}(W)$*

理论上可证明$h^{s=1}_{ldet}(W)$是$h_{expm}(W)$和$h_{poly}(W)$的上界，保留更多环的信息。

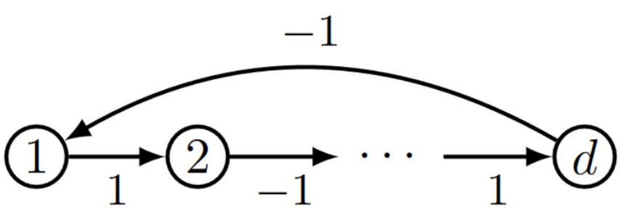
[Bello K, Aragam B, Ravikumar P. Dagma: Learning dags via m-matrices and a log-determinant acyclicity characterization[J]. Advances in Neural Information Processing Systems, 2022, 35: 8226-8239.]

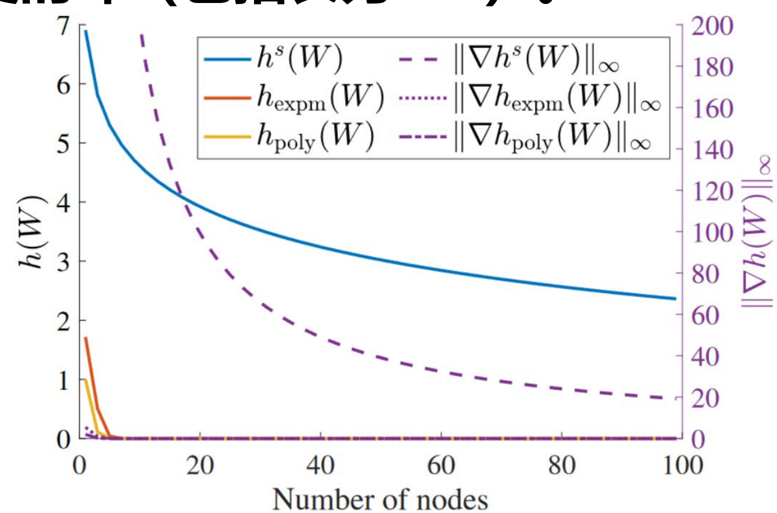
对数行列式无环表示优势

➢ 实际对比$h_{expm}(W)$、$h_{poly}(W)$和$h_{ldet}^{s}(W)$的案例：

- **实验设置：对于d个节点的有环图，边权为1或-1，令$s = 1.001$，作出$h(W)$随d变化的图。**

- **实验结论：当$d = 13$时，$h_{expm}(W) \approx 10^{-9}$，$h_{poly}(W) \approx 10^{-14}$，实践中如果$h(W) \in [10^{-8}, 10^{-5}]$，则被视为0，因此$h_{expm}(W)$和$h_{poly}(W)$无法检测长为13的环；$h_{ldet}^{s}(W)$能检测出更大长度的环（包括长为100）。**





[Bello K, Aragam B, Ravikumar P. Dagma: Learning dags via m-matrices and a log-determinant acyclicity characterization[J]. Advances in Neural Information Processing Systems, 2022, 35: 8226-8239.]

对数行列式无环表示优势

东南大学 SOUTHEAST UNIVERSITY

□ $h_{ldet}^{s}(W)$具有更好的梯度表现：

$$\nabla h_{\text{expm}}(W) = 2(e^{W\circ W})^\top \circ W \qquad e^{W\circ W} = \sum_{k=0}^{\infty} \boxed{\frac{1}{k!}} (W\circ W)^k$$

$$\nabla h_{\text{poly}}(W) = 2((I + \frac{1}{d} W\circ W)^{d-1})^\top \circ W \qquad (I + \frac{1}{d} W\circ W)^{d-1} = \sum_{k=0}^{d-1} \boxed{\frac{\binom{d-1}{k}}{(d-1)^k}} (W\circ W)^k$$

$$\nabla h_{\text{ldet}}^{s}(W) = 2((sI - W\circ W)^{-1})^\top \circ W \qquad (sI - W\circ W)^{-1} = \sum_{k=0}^{\infty} \boxed{\frac{1}{s^{k+1}}} (W\circ W)^k$$

Lemma 5. *For any walk of length k, its contribution to the gradients $\nabla h_{\text{expm}}(W)$ and $\nabla h_{\text{poly}}(W)$ are diminished by $1/k!$ and $\binom{d-1}{k}/(d-1)^k$, respectively. In contrast, $\nabla h_{\text{ldet}}^{s=1}(W)$ does not diminish any walk of any length. This implies that $|\nabla h_{\text{poly}}(W)| \le |\nabla h_{\text{expm}}(W)| \le |\nabla h_{\text{ldet}}^{s=1}(W)|$.*

$\nabla h_{expm}(W)$和$\nabla h_{poly}(W)$容易产生梯度消失问题；当令$s=1$时，$\nabla h_{ldet}^{s}(W)$对任意长度的环的信息不会被减弱，有更好的梯度信息来优化。

理论上可证明$\nabla h_{ldet}^{s=1}(W)$是$\nabla h_{expm}(W)$和$\nabla h_{poly}(W)$的上界，保留更多环的信息。

[Bello K, Aragam B, Ravikumar P. Dagma: Learning dags via m-matrices and a log-determinant acyclicity characterization[J]. Advances in Neural Information Processing Systems, 2022, 35: 8226-8239.]

对数行列式无环表示优势

东南大學 SOUTHEAST UNIVERSITY

实践中$h^s_{ldet}(W)$和$\nabla h^s_{ldet}(W)$计算速度更快：

- 理论上$h_{expm}(W)$、$h_{poly}(W)$和$h^s_{ldet}(W)$的计算复杂度均为$O(d^3)$，但实践中$h^s_{ldet}(W)$和其导数比$h_{expm}(W)$、$h_{poly}(W)$和其导数分别快约10倍。因为$h^s_{ldet}(W)$和$\nabla h^s_{ldet}(W)$的计算分别涉及矩阵的对数行列式和矩阵的逆，这两个问题已有大量工作求解；而$h_{expm}(W)$、$h_{poly}(W)$、$\nabla h_{expm}(W)$和$\nabla h_{poly}(W)$均涉及多次矩阵与矩阵乘法计算（计算更慢），四者性能类似。

$$h_{\text{expm}}(W) = \operatorname{Tr}(e^{W\circ W}) - d \qquad \nabla h_{\text{expm}}(W) = 2(e^{W\circ W})^\top \circ W$$

$$h_{\text{poly}}(W) = \operatorname{Tr}((I + \tfrac{1}{d}W\circ W)^d) - d \qquad \nabla h_{\text{poly}}(W) = 2((I + \frac{1}{d}W\circ W)^{d-1})^\top \circ W$$

$$h^s_{\text{ldet}}(W) \stackrel{\text{def}}{=} -\log\det(sI - W\circ W) + d\log s \qquad \nabla h^s_{\text{ldet}}(W) = 2((sI - W\circ W)^{-1})^\top \circ W$$

实践中，矩阵对数行列式和求逆的计算快于矩阵与矩阵乘法计算

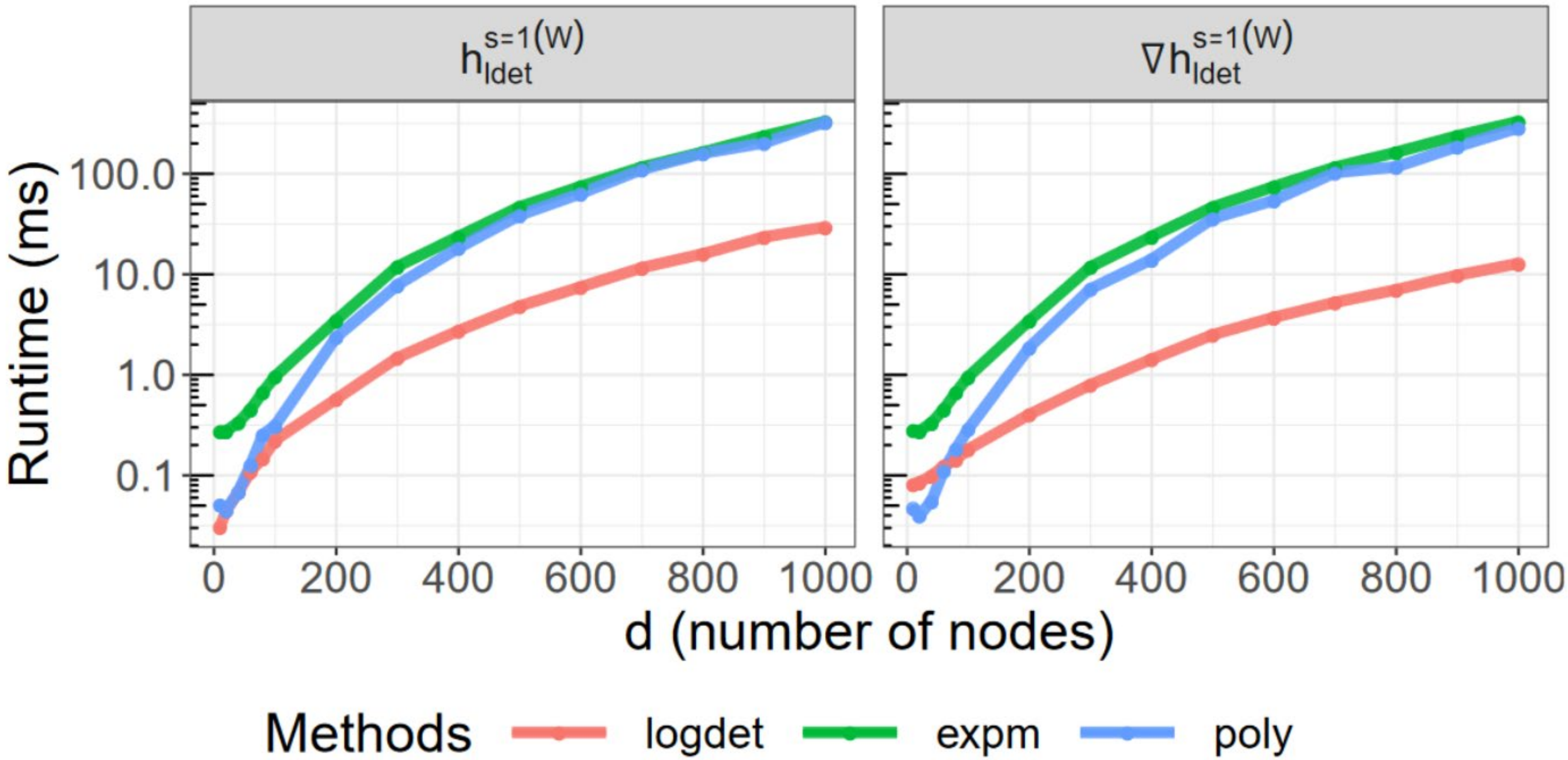
[Bello K, Aragam B, Ravikumar P. Dagma: Learning dags via m-matrices and a log-determinant acyclicity characterization[J]. Advances in Neural Information Processing Systems, 2022, 35: 8226-8239.]

对数行列式无环表示优势

东南大學 SOUTHEAST UNIVERSITY

□ 实践中$h^s_{ldet}(W)$和$\nabla h^s_{ldet}(W)$计算速度更快：

➢ **实验设置：对于每个d，从标准正态分布中随机采样30个矩阵。**



[Bello K, Aragam B, Ravikumar P. Dagma: Learning dags via m-matrices and a log-determinant acyclicity characterization[J]. Advances in Neural Information Processing Systems, 2022, 35: 8226-8239.]

DAGMA优化算法

东南大学 SOUTHEAST UNIVERSITY

Algorithm 1 DAGMA

Require: Data matrix $\boldsymbol{X}$, initial central path coefficient $\mu^{(0)}$ (e.g., 1), decay factor $\alpha \in (0, 1)$ (e.g., 0.1), ℓ_1 parameter $\beta_1 > 0$ (e.g., 0.01), log-det parameter $s > 0$ (e.g., 1), number of iterations T.

1: Initialize $\theta^{(0)}$ so that $W(\theta^{(0)}) \in \mathbb{W}^s$.
2: **for** $t = 0, 1, 2, \ldots T - 1$ **do**
3: Starting at $\theta^{(t)}$, solve $\theta^{(t+1)} = \arg\min_\theta \mu^{(t)}(Q(f_\theta; \boldsymbol{X}) + \beta_1 \|\theta\|_1) + h^s_{\text{ldet}}(W(\theta))$
4: Set $\mu^{(t+1)} = \alpha\mu^{(t)}$

Ensure: $W(\theta^{(T)})$

[Bello K, Aragam B, Ravikumar P. Dagma: Learning dags via m-matrices and a log-determinant acyclicity characterization[J]. Advances in Neural Information Processing Systems, 2022, 35: 8226-8239.]