Fast and Accurate Learning of Probabilistic Circuits by Random Projections

Renato L. Geh
renatolg@ime.usp.br

Denis D. Mauá
ddm@ime.usp.br

Department of Computer Science

Institute of Mathematics and Statistics, University of São Paulo

TPM 2021

Random Projections

Correia et al. [2020]: there is a link between DTs and PCs!

- Well known, established results in DTs;
- What if we apply them to PCs?

Random Projections (RPs):

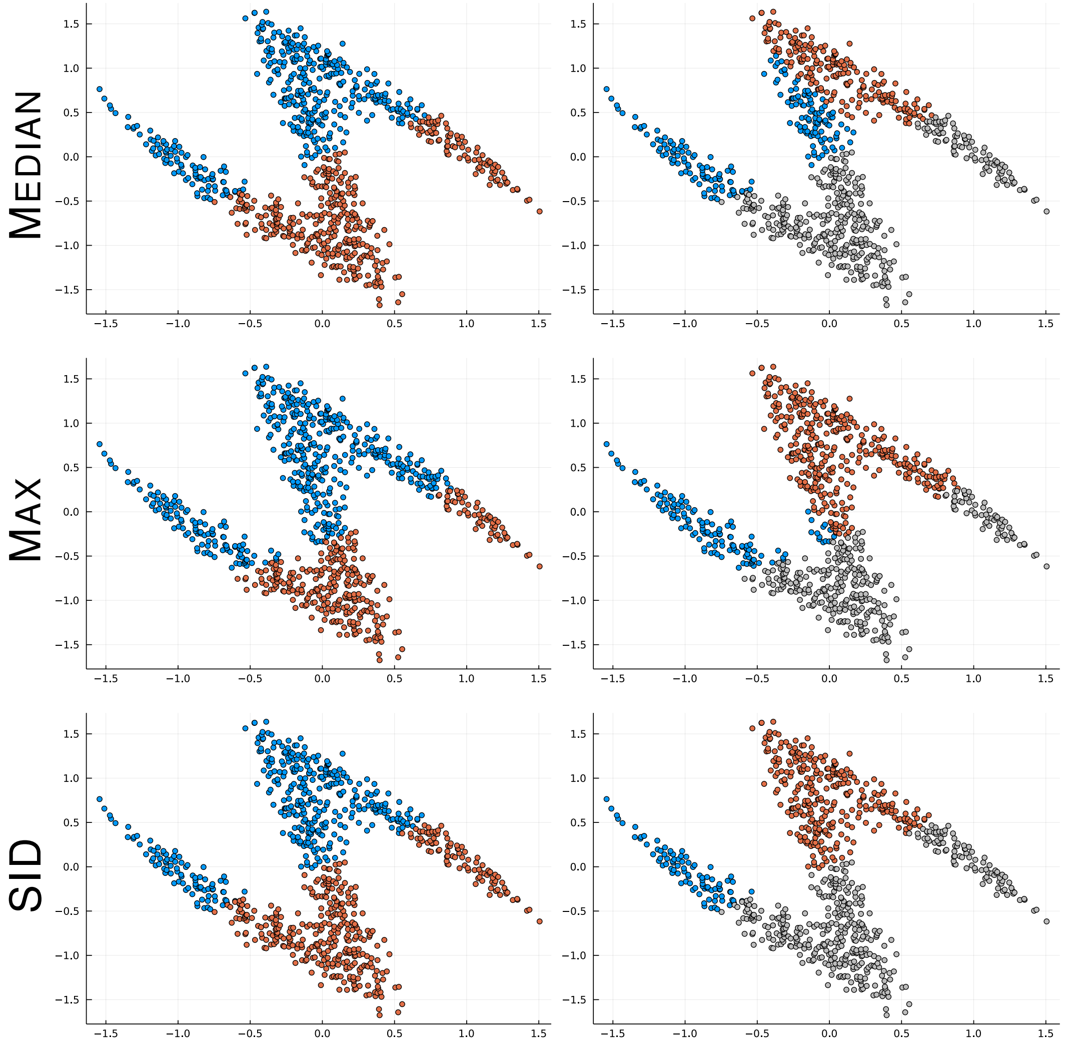
- Axis-aligned projections (kd-trees) are not good enough;
- Splitting by the median is not good enough;
- What about **random projections**?

Dasgupta and Freund [2008] showed that it is!

Idea behind RPs:

- Sample random direction u;
- Construct hyperplane $\mathcal{H}$ with u;
- Apply perturbation δ to $\mathcal{H}$ st. divides data somewhat equally;
- Choice of δ:

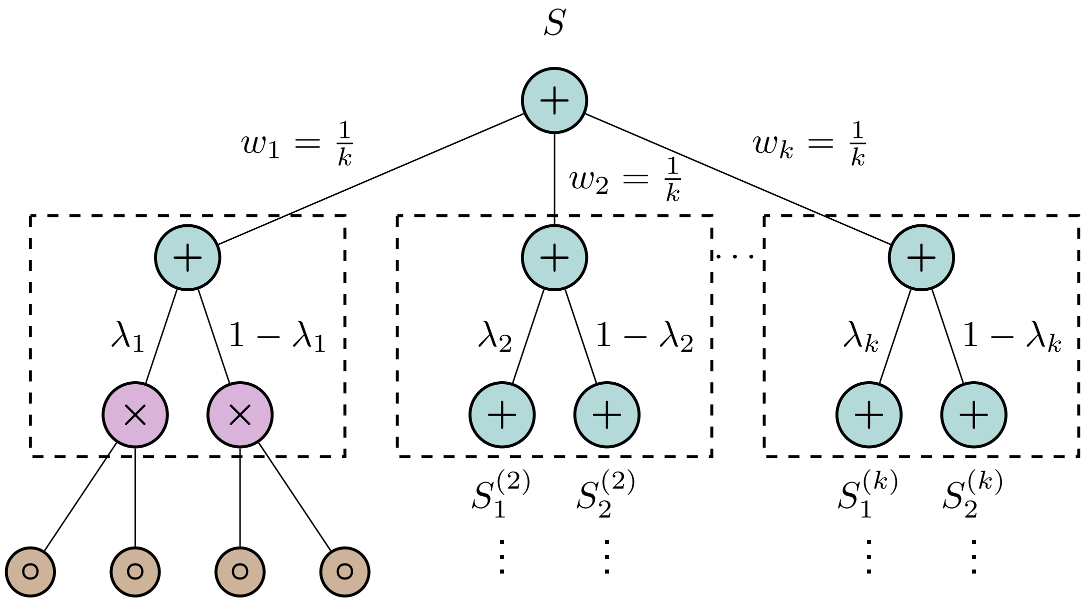
Max: by diameter (maximal distance);
SID: by Square Interpoint Distance.



Learning Probabilistic Circuits by Random Projections

Sum nodes as RPs:

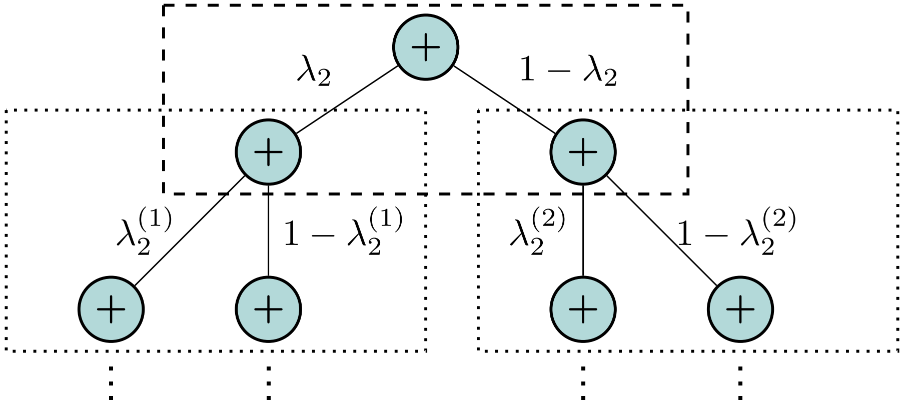
- Directly translate RP Decision Trees to PCs;
- Projections define a clustering of data;
- LearnSPN-like on sums;
- Eventually learn a leaf distribution:
  - (i) Fully factorized circuit;
  - (ii) RAT-SPN (Peharz et al. [2019]).



LearnRP-S

Input Dataset $S \subset \mathbb{R}^n$, no. of trials t, no. of trees k
Output A probabilistic circuit

if it is the first recursion **then**
 return $\sum_{i=1}^{k} \frac{1}{k}$LearnRP-S$(S, t, k)$
else
 Sample t RPs by some criteria
 Select split (S_1, S_2) that minimizes the avg. diam. of S
 if $|S_1|$ is small **then** $P_1 \leftarrow$ LearnDistribution(S_1)
 else $P_1 \leftarrow$ LearnRP-S(S_1, t, k)
 if $|S_2|$ is small **then** $P_2 \leftarrow$ LearnDistribution(S_2)
 else $P_2 \leftarrow$ LearnRP-S(S_2, t, k)
 Set $\lambda \leftarrow |S_1|/|S|$
 return $\lambda \cdot P_1 + (1 - \lambda) \cdot P_2$



Learning Probabilistic Circuits by Random Projections

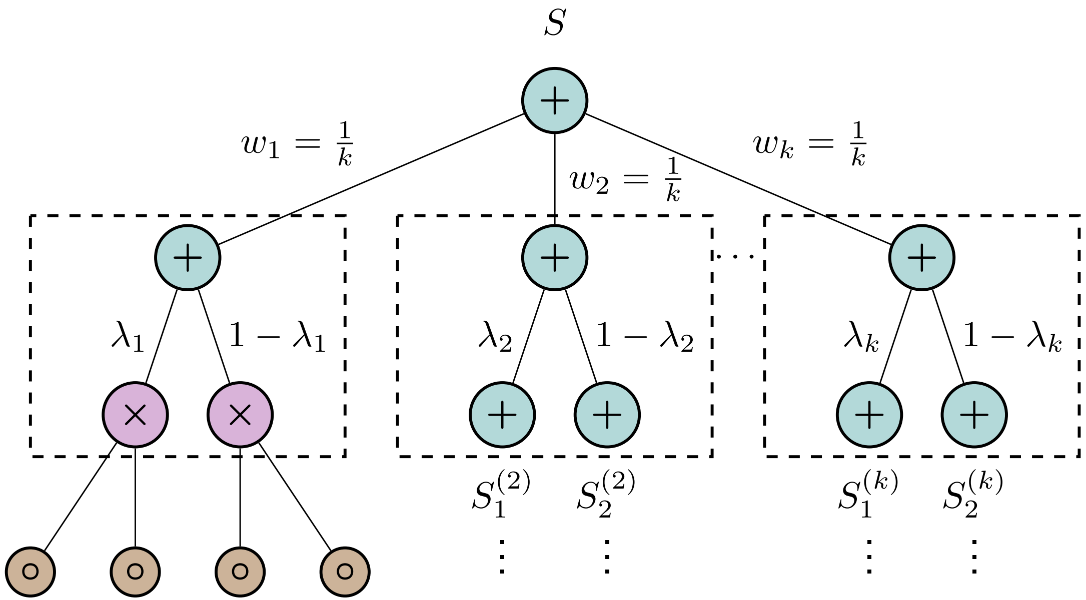
Mixtures of RPs:

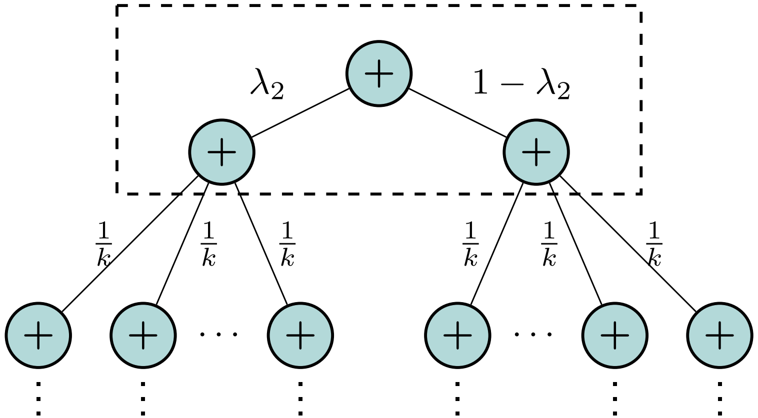
- Same as before, but now with k projections;
- At each data split, creates a “mixture node”;
- Children are projections.

LearnRP

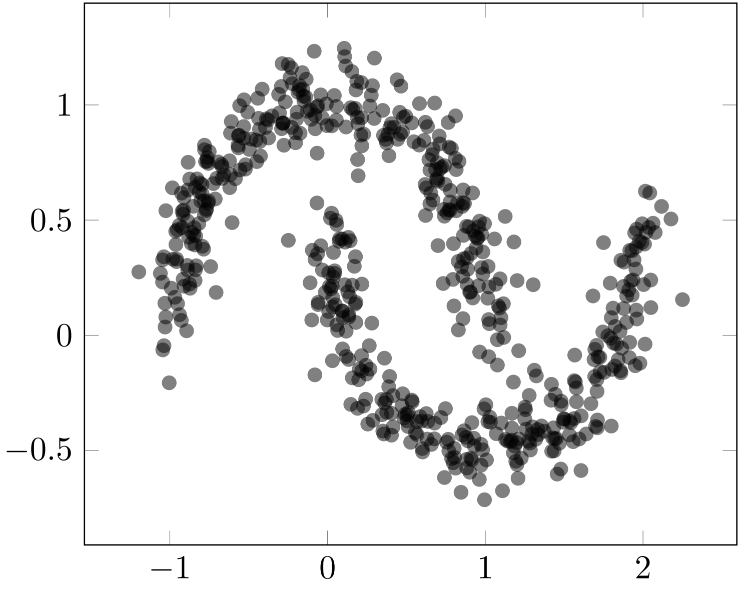
Input Dataset $S \subset \mathbb{R}^n$, no. of trials t, no. of mixtures k
Output A probabilistic circuit

for $i = 1, \ldots, k$ **do**
  Sample t splits by some criteria
  Select split (S_1, S_2) that minimizes the avg. diam. of S
  if $|S_1|$ is small **then**
    $P_1 \leftarrow \textsc{LearnDistribution}(S_1)$
  else $P_1 \leftarrow \textsc{LearnRP}(S_1, t, k)$
  if $|S_2|$ is small **then**
    $P_2 \leftarrow \textsc{LearnDistribution}(S_2)$
  else $P_2 \leftarrow \textsc{LearnRP}(S_2, t, k)$
  Set $\lambda \leftarrow |S_1|/|S|$
  Compute $C_i \leftarrow \lambda \cdot P_1 + (1 - \lambda) \cdot P_2$
return $\sum_{i=1}^{k} \frac{1}{k} C_i$

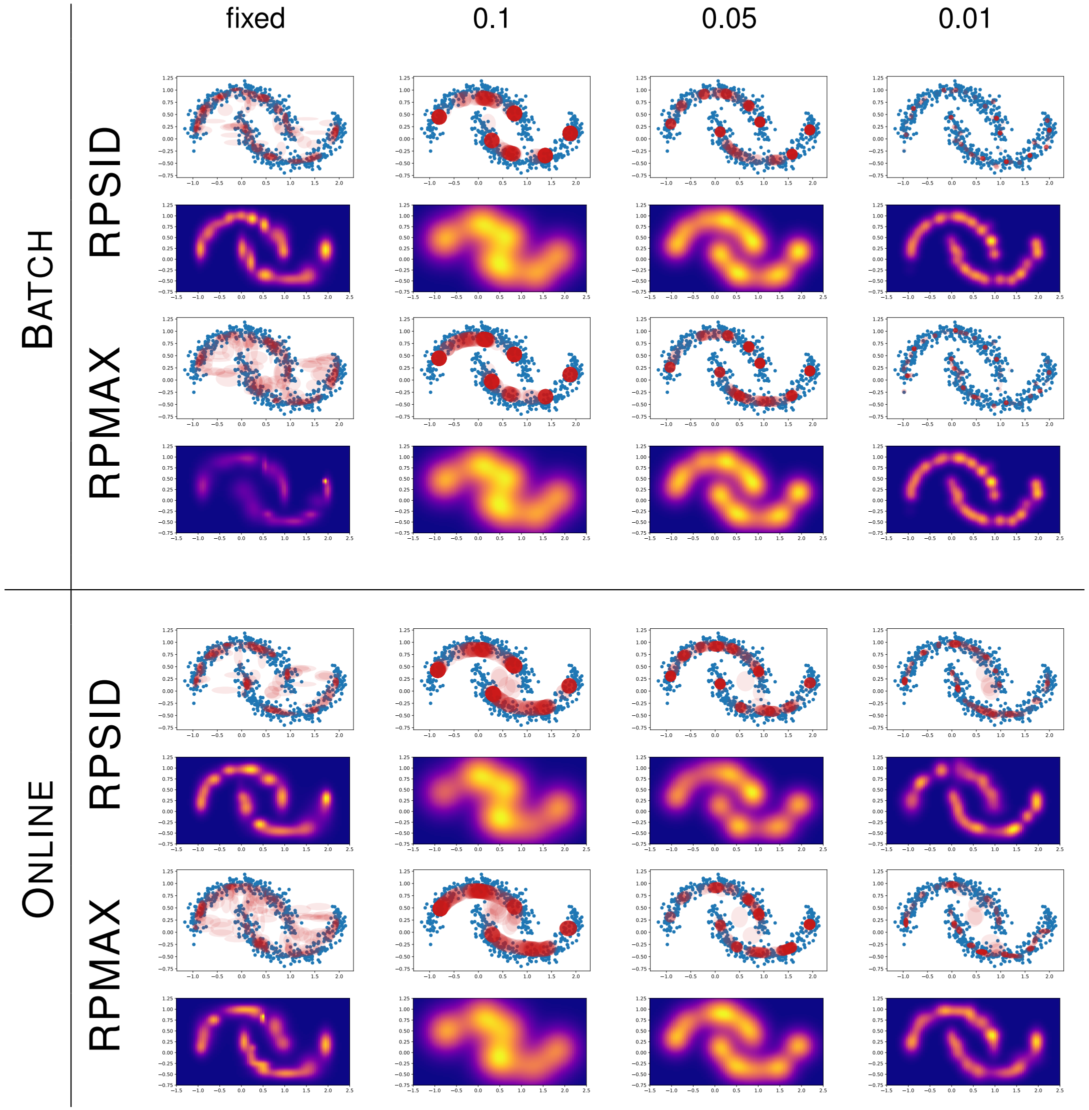




Experiments: `2-moons`



EM	$\sigma^2 \geq$	RPMaxS	RPMax	RPSIDS	RPSID
Batch	fixed	-1.083	-1.139	-1.106	-1.074
	0.10	-1.616	-1.616	-1.616	-1.616
	0.05	-1.387	-1.388	-1.387	-1.385
	0.01	-1.034	-1.041	-1.035	-1.032
Online	fixed	-1.089	-1.119	-1.080	-1.108
	0.10	-1.623	-1.627	-1.624	-1.637
	0.05	-1.416	-1.402	-1.408	-1.412
	0.01	-1.139	-1.080	-1.114	-1.105



Experiments: benchmark datasets

Dataset	#Var.	#Train.	RPMaxSD	RPMaxD	RPSIDSD	RPSIDD	RPMaxS	RPMax	RPSIDS	RPSID	LSPN	Strudel	LPSDD	EXPC
ACCIDENTS	111	12758	-36.77	-37.50	-36.96	-36.87	-37.41	-37.48	-37.16	-37.39	-30.03	**-28.73**	$\lvert$-30.16$\rvert$	-31.02
AD	1556	2461	-36.58	-37.29	-36.14	-38.11	-33.40	-32.83	-34.42	-35.55	$\lvert$-19.73$\rvert$	-16.38	-31.78	**-15.50**
AUDIO	100	15000	-40.25	-40.32	-40.20	-40.25	-40.29	-40.25	-40.28	$\lvert$-40.23$\rvert$	-40.50	-41.50	**-39.94**	-40.91
BBC	1058	1670	-253.15	-254.24	-252.48	$\lvert$-251.19$\rvert$	-254.71	-254.57	-254.74	-254.99	-250.68	-254.41	-253.19	**-248.34**
NETFLIX	100	15000	-57.20	-57.50	-57.22	$\lvert$-57.17$\rvert$	-57.53	-57.40	-57.45	-57.44	-57.02	-58.69	**-55.71**	-57.58
BOOK	500	8700	-34.84	-34.87	-34.87	-34.71	-34.75	-34.85	$\lvert$-34.74$\rvert$	**-34.66**	-35.88	-34.99	-34.97	-34.75
20-NEWSGRP	910	11293	$\lvert$-154.20$\rvert$	-155.36	-153.91	-154.57	-155.39	-155.41	-155.59	-156.08	-155.92	-154.47	-155.97	**-153.75**
REUTERS-52	889	6532	-87.01	-87.64	-86.79	-87.38	-87.58	-87.23	-86.33	-87.28	-85.06	$\lvert$-86.22$\rvert$	-89.61	**-84.70**
WEBKB	839	2803	-157.49	-157.95	$\lvert$-157.06$\rvert$	-157.44	-158.46	-158.72	-158.07	-157.94	-158.20	-155.33	-161.09	**-153.67**
DNA	180	1600	-97.89	-97.21	-97.28	-97.49	-97.45	-97.17	-97.47	-96.68	**-82.52**	-86.22	-88.01	$\lvert$-86.61$\rvert$
JESTER	100	9000	$\lvert$-53.05$\rvert$	-53.09	-53.09	-53.02	-53.21	-53.24	-53.13	-53.06	-75.98	-55.03	**-51.29**	-53.43
KDD	65	180092	-2.17	-2.16	-2.17	-2.15	-2.16	-2.16	-2.16	-2.17	-2.18	-2.13	**-2.11**	$\lvert$-2.15$\rvert$
KOSAREK	190	33375	-11.11	-11.10	-11.14	-11.11	-11.06	-11.07	-11.02	-11.08	-10.98	-10.68	**-10.52**	$\lvert$-10.77$\rvert$
MSNBC	17	291326	-6.24	-6.22	-6.25	-6.32	-6.18	-6.18	-6.20	-6.28	$\lvert$-6.11$\rvert$	**-6.04**	-6.04	-6.18
MSWEB	294	29441	-10.51	-10.38	-10.53	-10.41	-10.25	-10.26	-10.25	-10.29	-10.25	**-9.71**	-9.89	$\lvert$-9.93$\rvert$
NLTCS	16	16181	-6.02	-6.02	-6.03	-6.03	$\lvert$-6.01$\rvert$	-6.01	-6.01	-6.01	-6.11	-6.06	**-5.99**	-6.05
PLANTS	69	17412	-13.94	-14.15	-14.00	-13.91	-14.07	-13.86	-14.02	-13.94	**-12.97**	-12.98	$\lvert$-13.02$\rvert$	-14.19
PUMSB-STAR	163	12262	-33.53	-35.43	-33.55	-33.52	-34.35	-34.24	-34.53	-33.92	-24.78	**-24.12**	-26.12	$\lvert$-26.06$\rvert$
EACHMOVIE	500	4524	-53.03	-53.21	$\lvert$-52.94$\rvert$	-53.22	-53.03	-53.28	-52.88	-53.15	**-52.48**	-53.67	-58.01	-54.82
RETAIL	135	22041	-11.02	-10.95	-11.01	-10.99	$\lvert$-10.93$\rvert$	-10.93	-10.94	-10.93	-11.04	-10.81	**-10.72**	-10.94