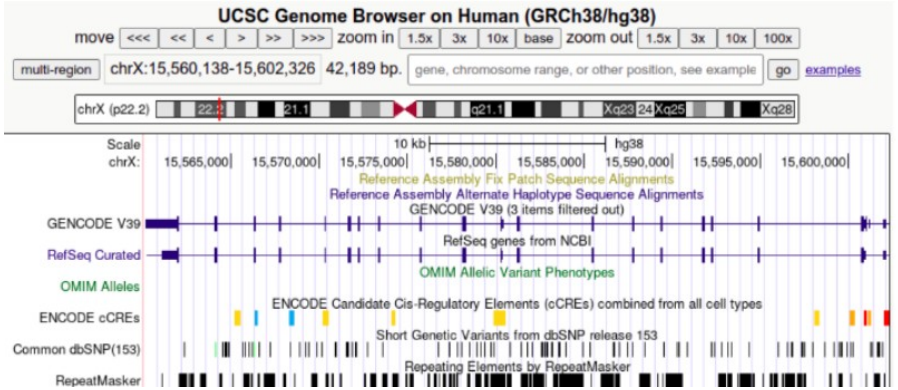
Modelovanie genomických anotácií pomocou diskrétnych distribúcií fázového typu

Hana Derková, školitel: Mgr. Askar Gafurov, PhD.

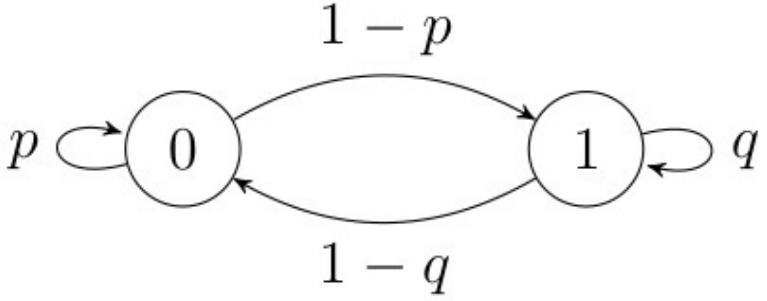
Genomická anotácia

- Kóduje významné časti genómu
- Súbor intervalov
- Významný prekryv -> biologická súvislosť



Miera štatistickej významnosti

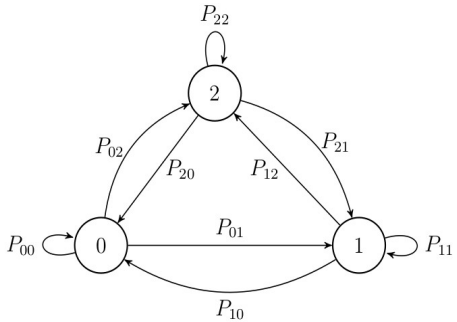
- Dve anotácie *reference* a *query*
- Nulová hypotéza : *query* je generovaná dvojstavovým Markovovským reťazcom
- P-hodnota vypočítaná v O(m^2 + n) - nezávislá od dĺžky genómu
- Problém : distribúcia dĺžok je geometrická



Absorpčný Markovovský reťazec a diskrétna distribúcia fázového typu

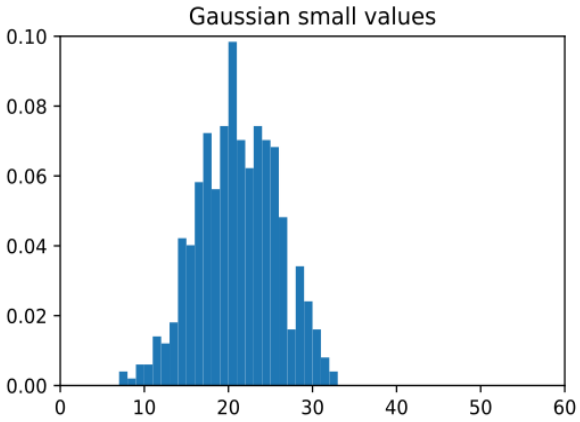
$$P = \begin{bmatrix} P_{00} & P_{01} & P_{02} \\ P_{10} & P_{11} & P_{12} \\ P_{20} & P_{21} & P_{22} \end{bmatrix}$$

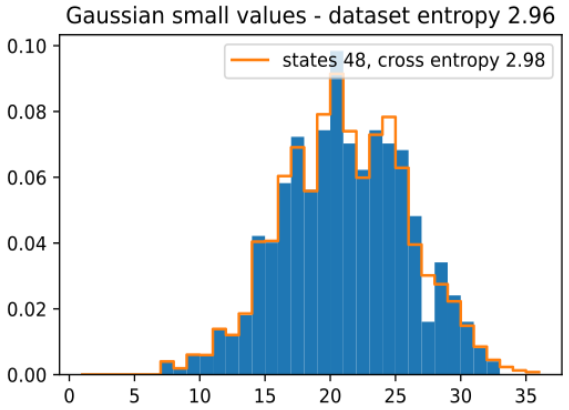
- Markovovský reťazec - stavové automaty
- Absorbčný Markovovský reťazec - každý stav vie dosiahnuť absorpčný stav
- ČAS ABSORPCIE = # krokov pred dosiahnutím absorpčného stavu
  - Tvorí triedu **diskrétnych distribúcií fázového typu**



Úloha

- Pre danú distribúciu dĺžok intervalov / medzier, nájsť hodnoty pravdepodobnosti prechodov tak, aby sa výsledná distribúcia čo najviac podobala distribúcii dĺžok



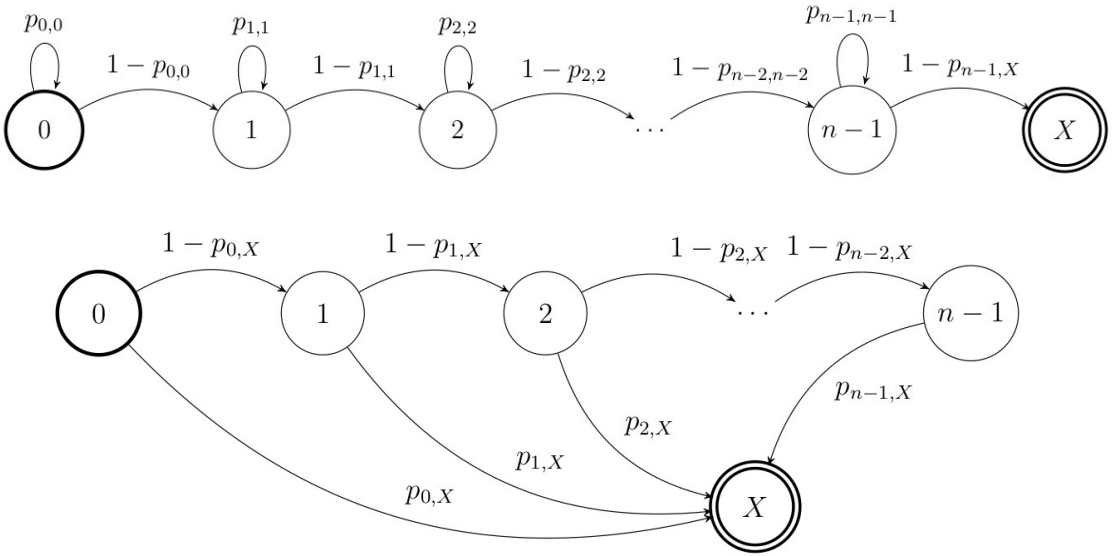


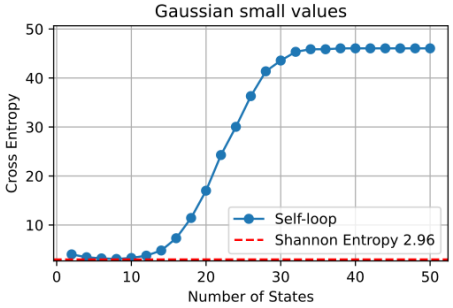
Stratová funkcia a parametre

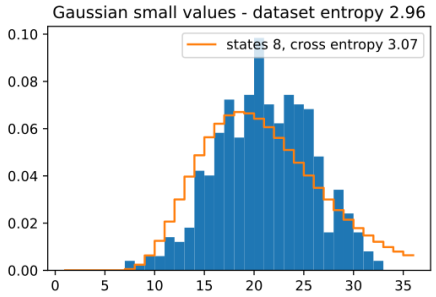
- Minimalizovať negative log likelihood
- Cross entropia - diskrétna fázová distribúcia a distribúcia dát
- Shanonova entropia datasetu
- Scipy - optimizer minimize
- Parametre - pravdepodobnosti prechodov
- Ďalšie parametre - počet stavov Markovovského reťazca, architektúra Markovovského reťazca

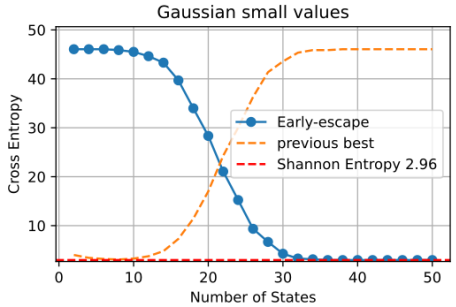
$$-\sum_{d \in data} \ln P(d|\theta)$$

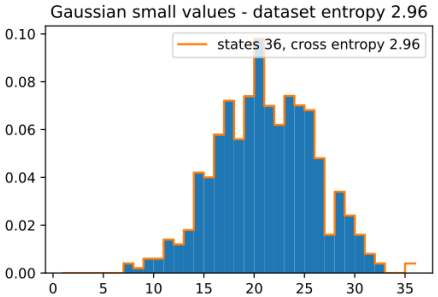
Self-loop a early-escape architektúra

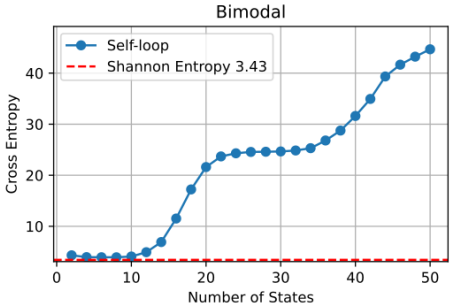


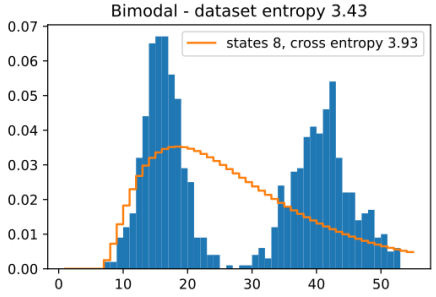

Gaussian small values
Cross Entropy
Number of States
Self-loop
Shannon Entropy 2.96

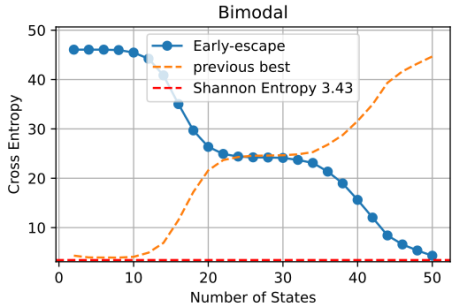

Gaussian small values - dataset entropy 2.96
states 8, cross entropy 3.07

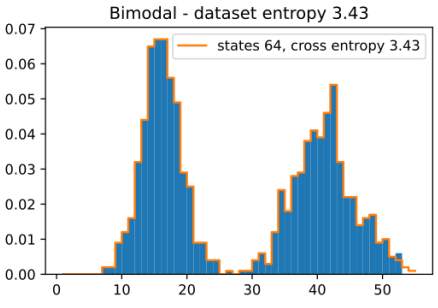

Gaussian small values
Cross Entropy
Number of States
Early-escape
previous best
Shannon Entropy 2.96


Gaussian small values - dataset entropy 2.96
states 36, cross entropy 2.96

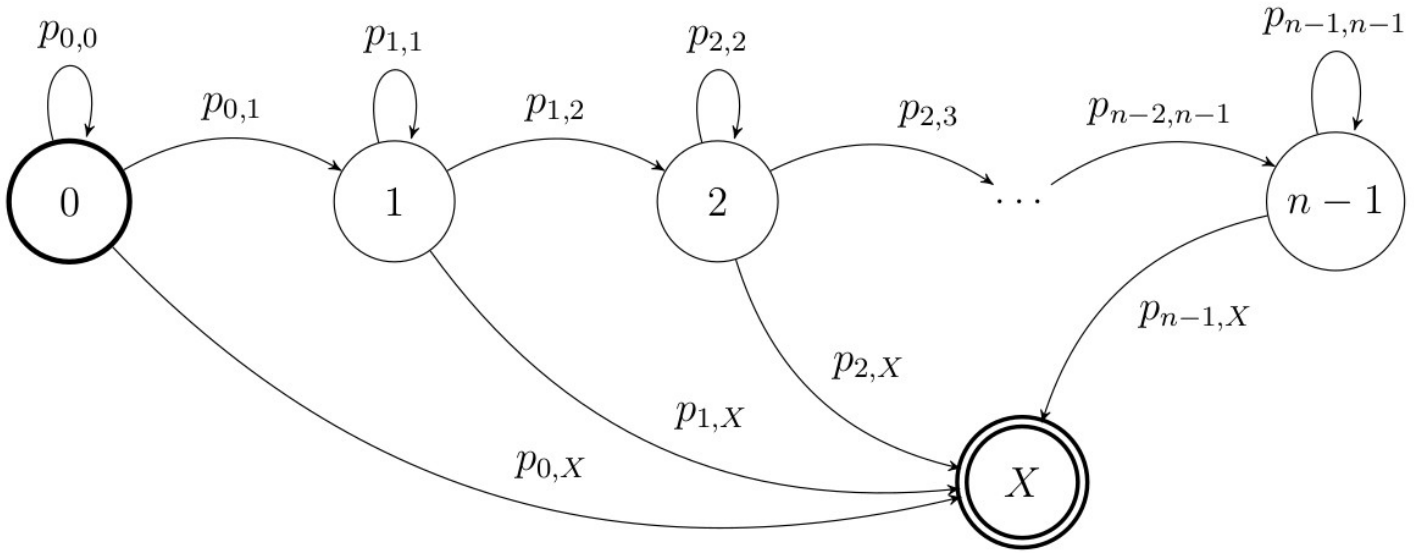

Bimodal
Cross Entropy
Number of States
Self-loop
Shannon Entropy 3.43

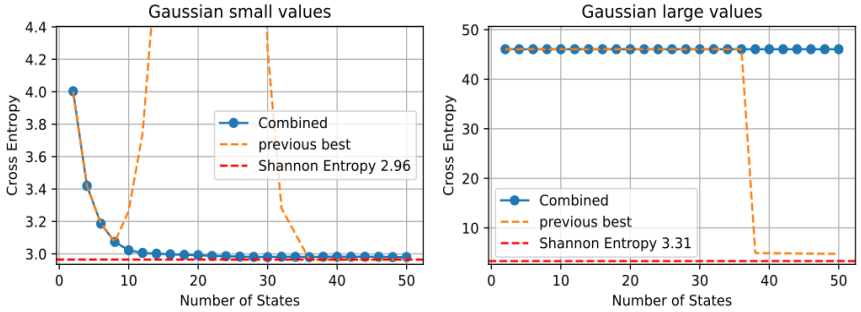

Bimodal - dataset entropy 3.43
states 8, cross entropy 3.93


Bimodal
Cross Entropy
Number of States
Early-escape
previous best
Shannon Entropy 3.43

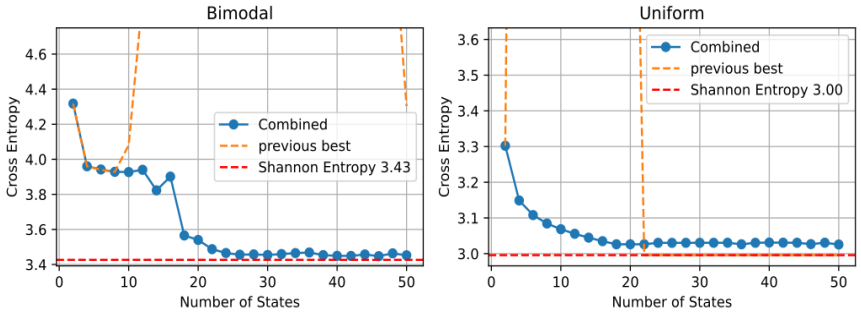

Bimodal - dataset entropy 3.43
states 64, cross entropy 3.43

Combined architektúra

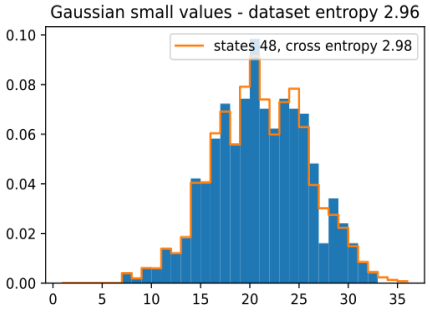


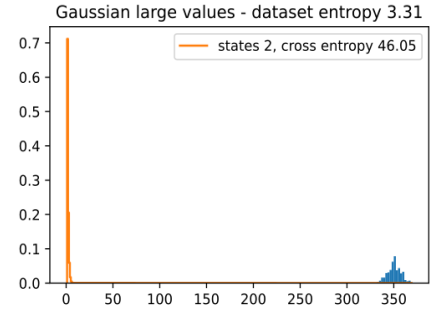

Gaussian small values
Cross Entropy
Number of States
Combined
previous best
Shannon Entropy 2.96

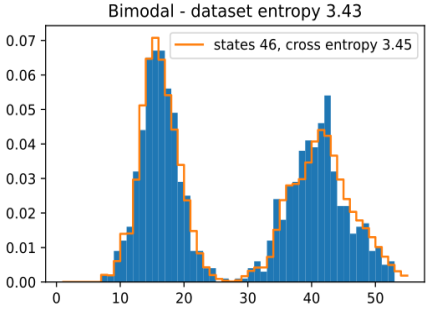
Gaussian large values
Cross Entropy
Number of States
Combined
previous best
Shannon Entropy 3.31

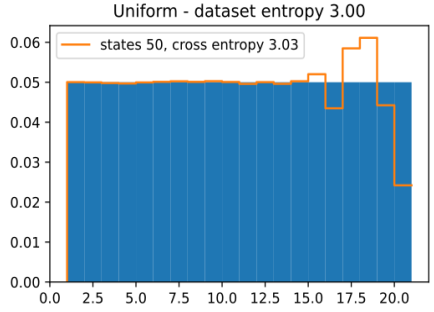

Bimodal
Cross Entropy
Number of States
Combined
previous best
Shannon Entropy 3.43

Uniform
Cross Entropy
Number of States
Combined
previous best
Shannon Entropy 3.00


Gaussian small values - dataset entropy 2.96
states 48, cross entropy 2.98

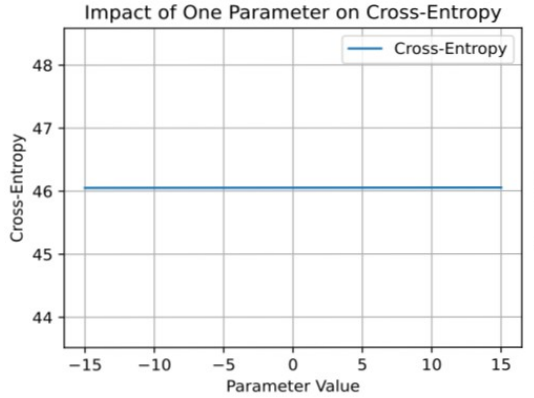

Gaussian large values - dataset entropy 3.31
states 2, cross entropy 46.05

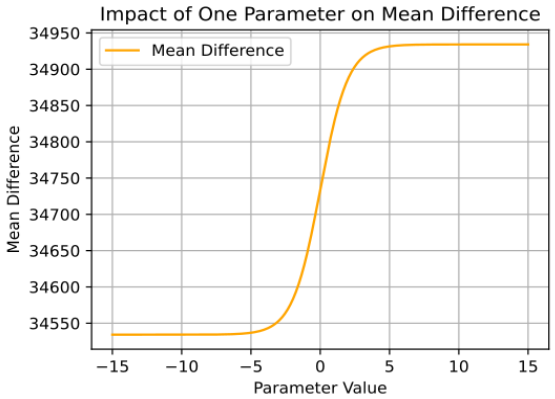

Bimodal - dataset entropy 3.43
states 46, cross entropy 3.45


Uniform - dataset entropy 3.00
states 50, cross entropy 3.03

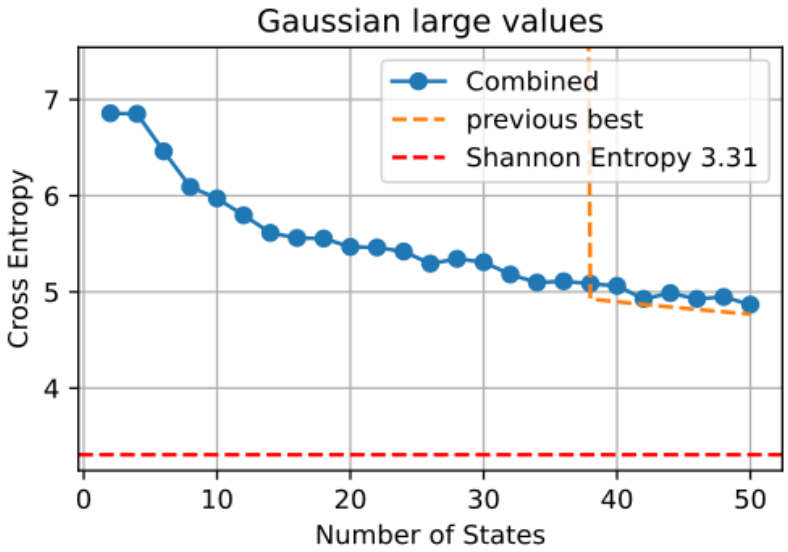
Technika - Mean shifting

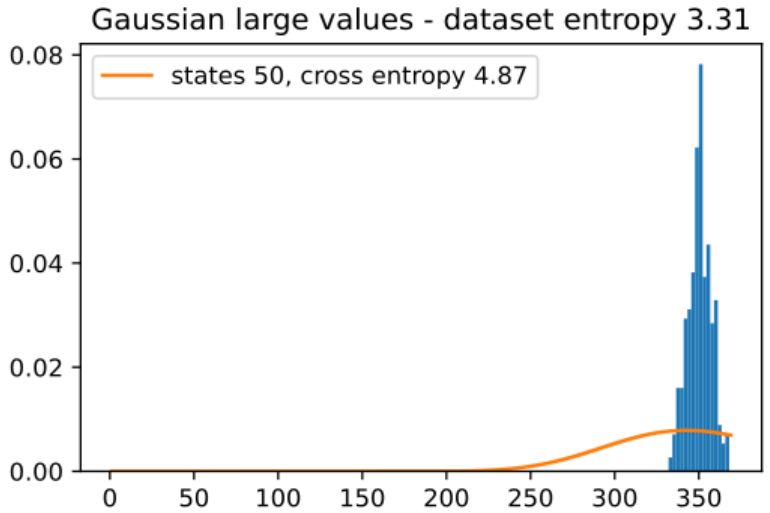
- Nenatrénovaná architektúra pre veľké hodnoty
- Zastavovacia podmienka : projected gradient is too small
- Posunieme masu PHD bližšie ku strednej hodnote distribúcii dát
- Cieľová funkcia je rozdiel stredných hodnôt
- Následne pokračujeme v tréningu s výslednou inicializáciou





Efekt mean shiftingu





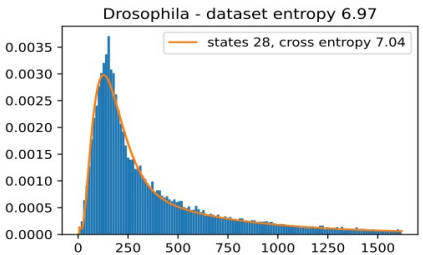
Modelovanie reálnych dát

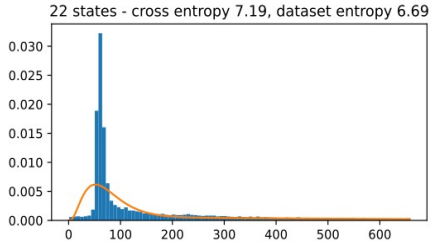
- Combined architektúra + mean shifting
- Subsampling technika - 2000 vzoriek
- Hirt, Gains inc, H3K4me3, drosophila melanogaster exóny

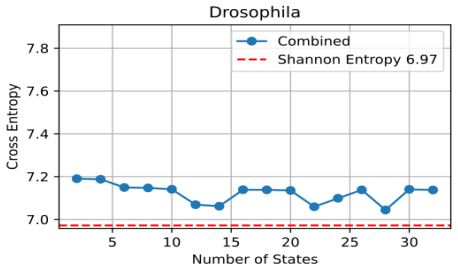
Dataset	Mean Gap Size	Range of Gap Lengths	Total Samples
Drosophila exons	1587	1–710734	63948
Gains inc	952046	4–30821993	3132
H3K4me3	156902	134–30542789	19333
Hirt	16296782	92740–87181600	116

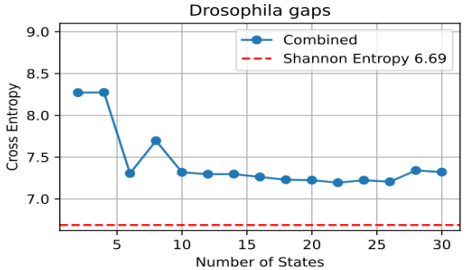
Dataset	Mean	Range of interval lengths	Number of intervals
Drosophila exons	488	1–28074	188467
Gains inc	35581	73–3002725	3132
H3K4me3	1937	124–22590	19358
Hirt	1966196	6–24019400	247

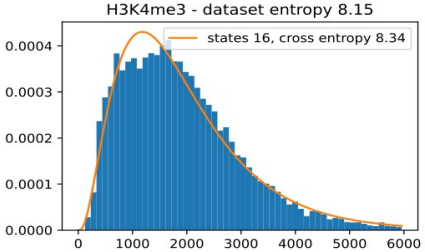
Reálne dáta výsledky

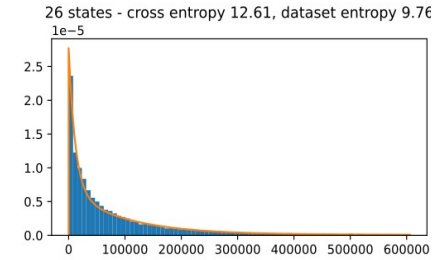


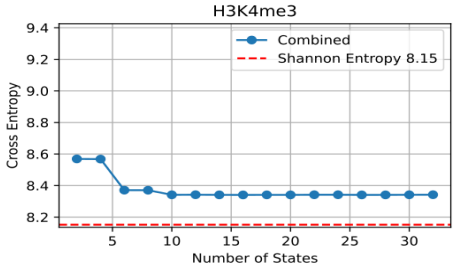


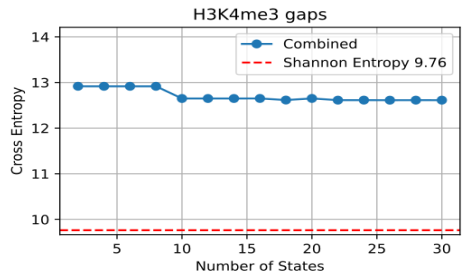




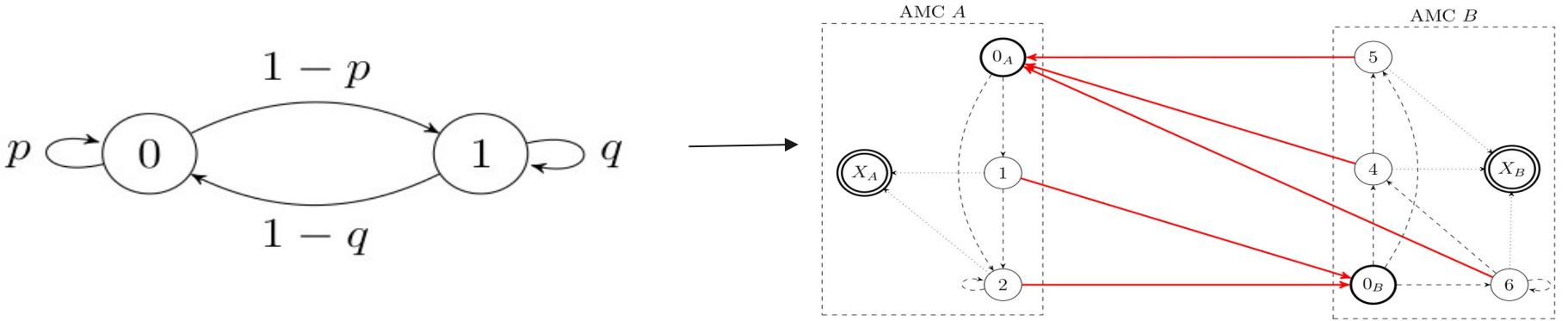




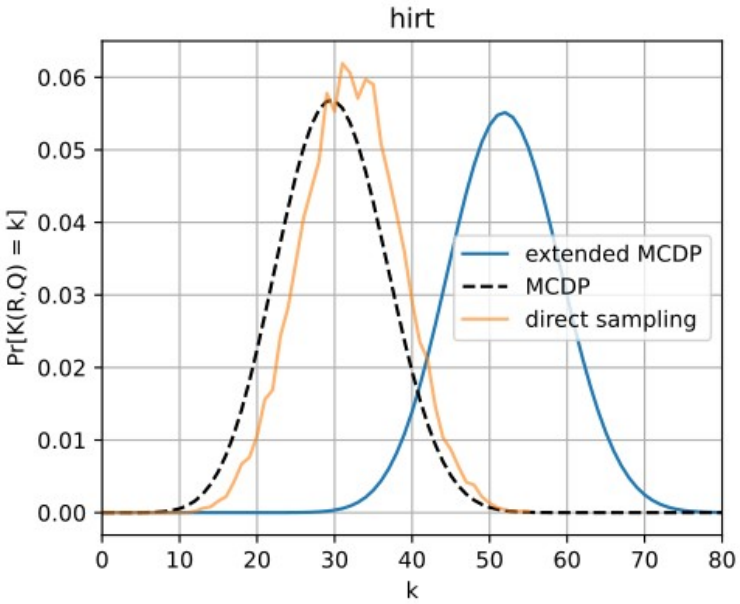


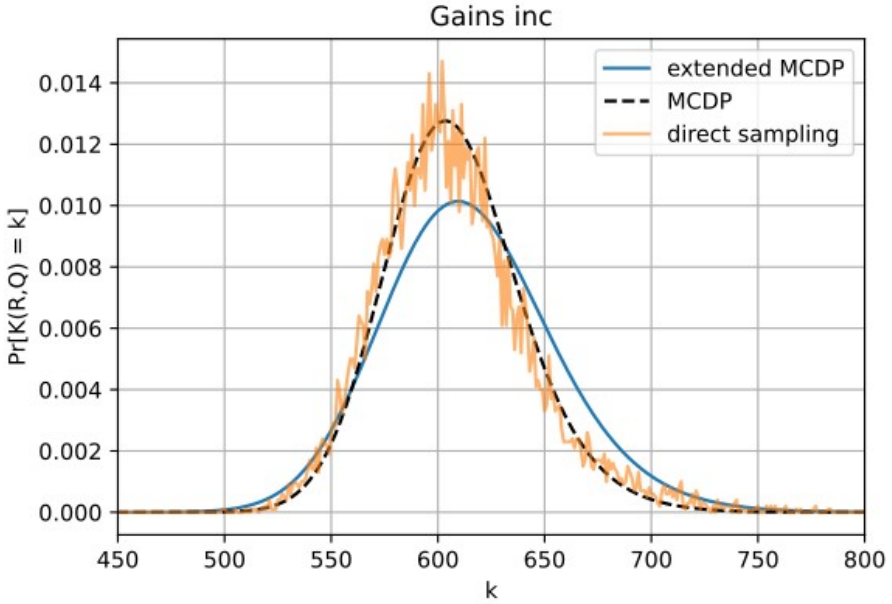


Extended MCDP



Výsledky


hirt
Pr[K(R,Q) = k]
k
extended MCDP
MCDP
direct sampling


Gains inc
Pr[K(R,Q) = k]
k
extended MCDP
MCDP
direct sampling

Ďakujem za pozornosť

Pri práci s reálnymi dátami vzorkujete iba podmnožinu dát. Ako závisí zložitosť výpočtu cross entropie vo vašom kóde od počtu stavov S, počtu dátových bodov N, počtu unikátnych dátových bodov U a maximálnej dĺžky intervalu v dátach L?

$$-\sum_{d \in data} \ln P(d|\theta) \qquad O\left(S^m \cdot U \cdot \log L + N\right)$$

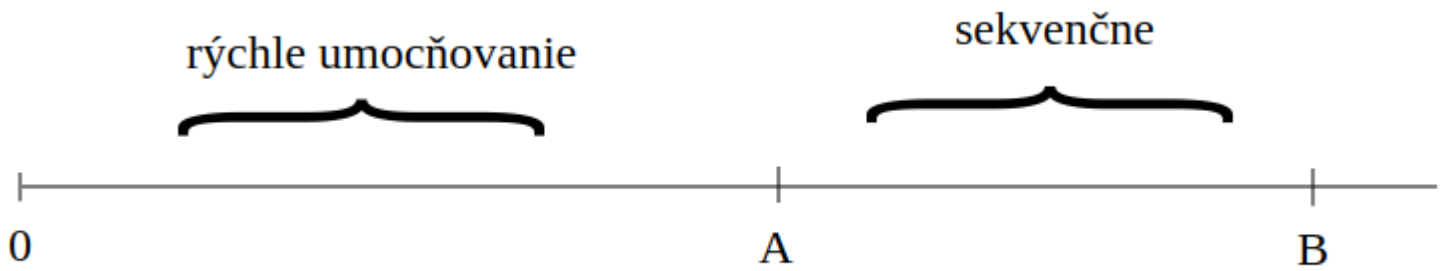
m je zložitosť umocňovania matíc

Ak by ste v predspracovaní zlúčili rovnaké hodnoty v dátach do jednej skupiny, vedeli by ste výpočet cross-entropie upraviť tak, aby nezávisel od N ale len od U?

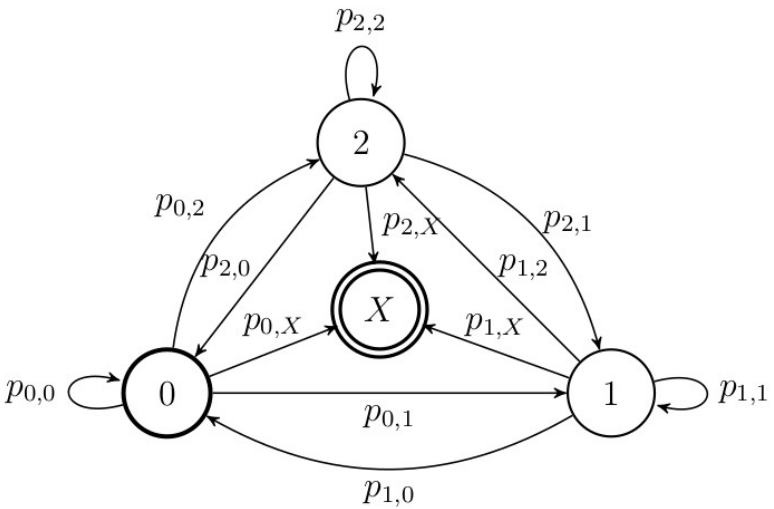
$$O\left(S^m \cdot U \cdot \log L\right)$$

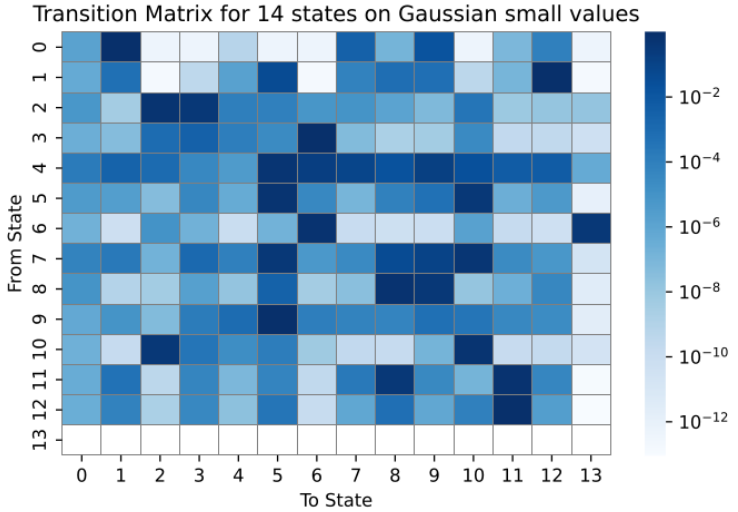
m je zložitosť umocňovania matíc

Ako by ste upravili výpočet, aby vedel rýchlo spracovať prípad, keď vaše dáta pochádzajú z intervalu [A,B], pričom väčšina celých čísel z tohto intervalu sa v dátach nachádza aspoň raz?

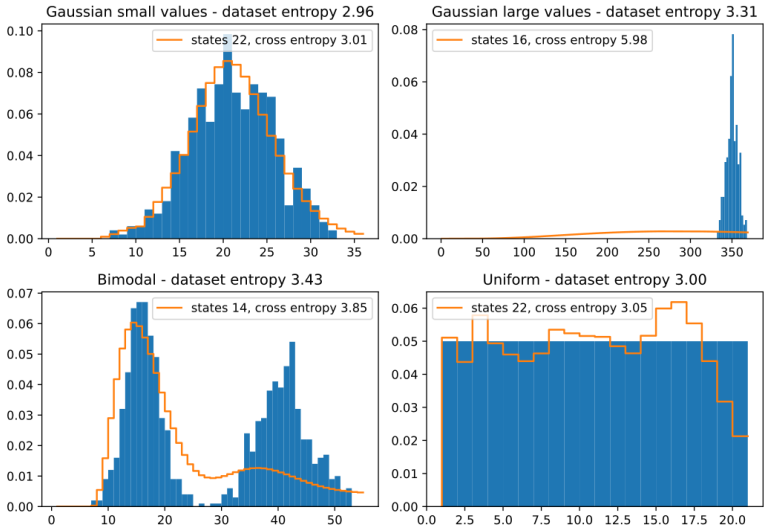


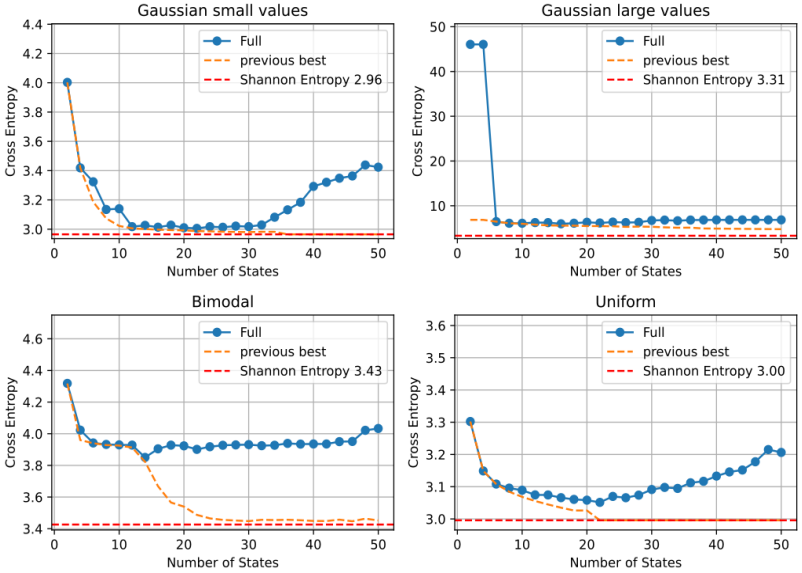
Fully connected architektúra a pruning



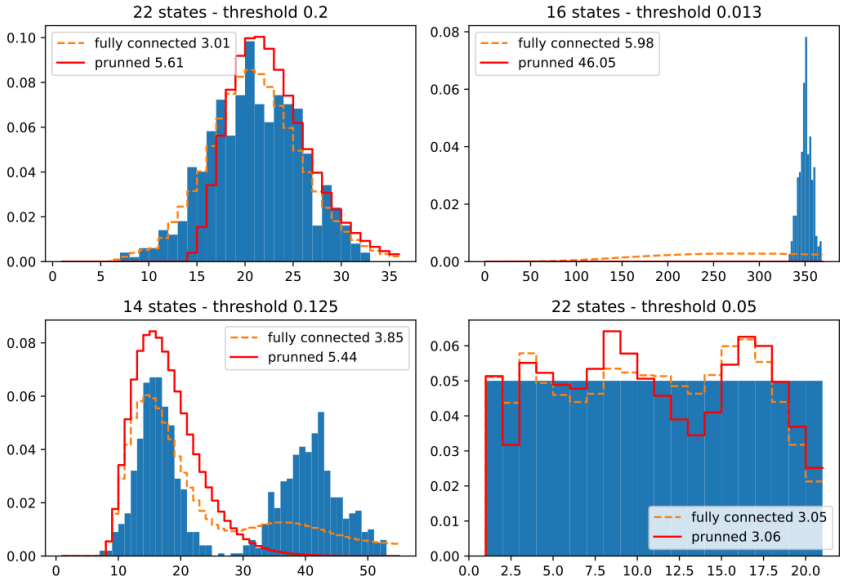


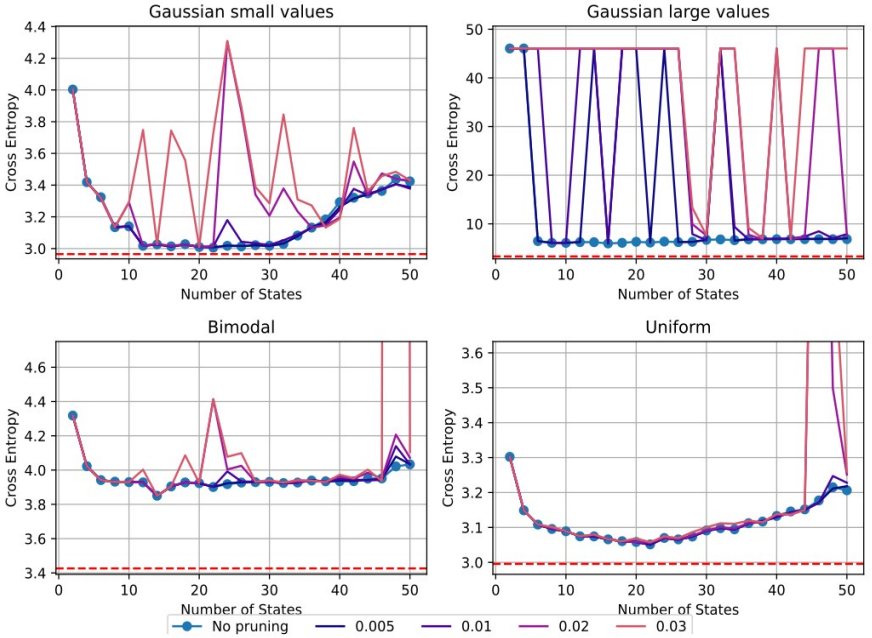
Výsledky pre fully-connected architektúru



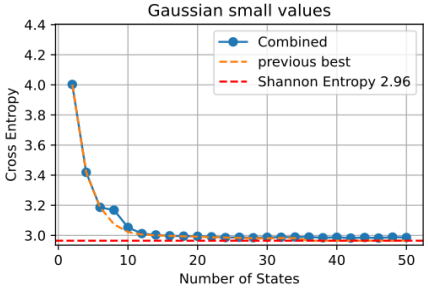


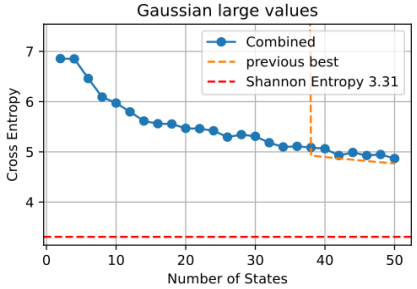
Pruning efekt

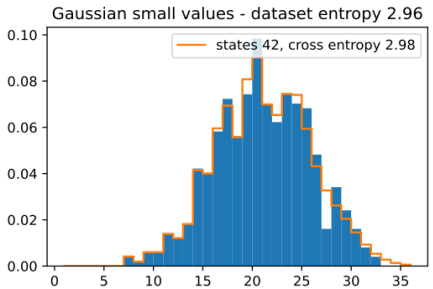


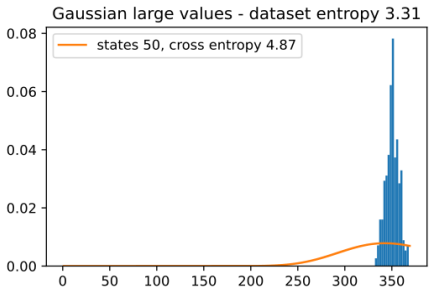


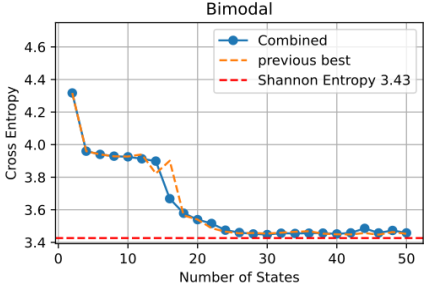
Všetky výsledky pre mean shifting

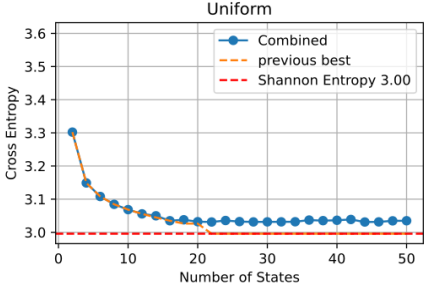


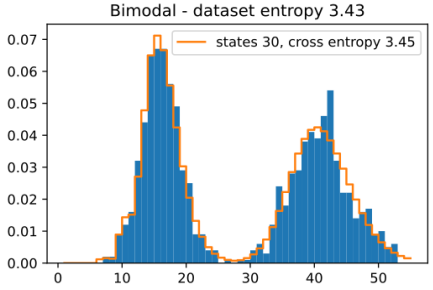


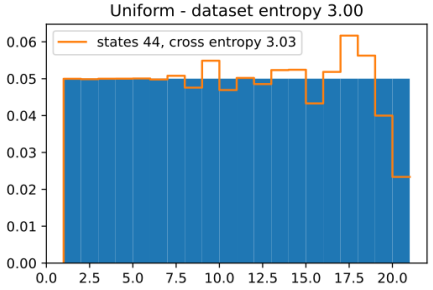




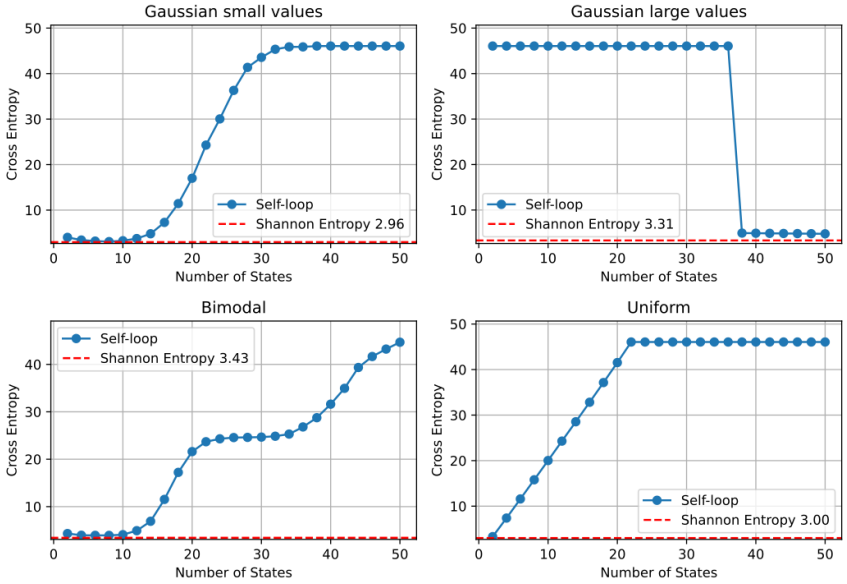


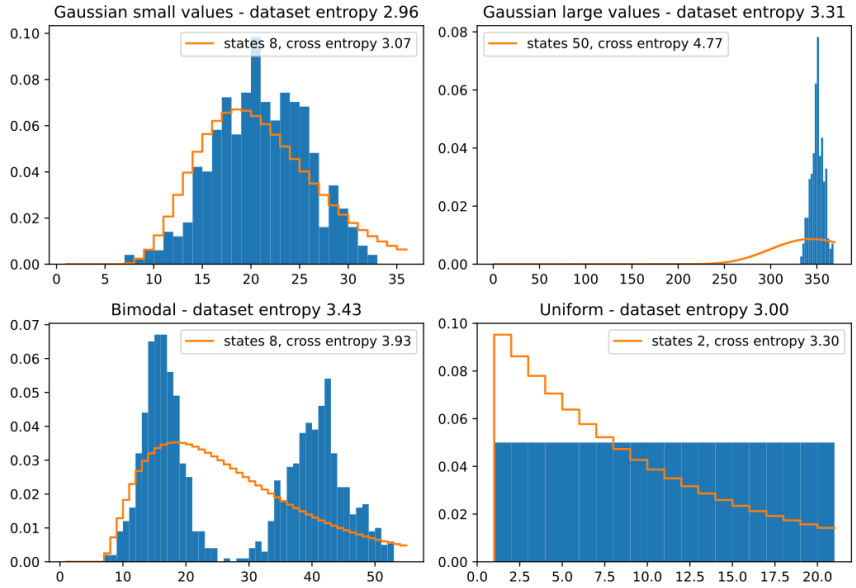






Všetky výsledky pre self-loop architektúru





Všetky výsledky pre early-escape architektúru