Karp–Luby Approximation Algorithm for #DNF

Nimitt

May 18, 2026

Contents

"One sample lies. Many samples converge;
But, what if the sampling space is too sparse?"

DNF

$$F ::= X_i \mid \neg F \mid (F \wedge F) \mid (F \vee F)$$

Literal:

$$X_i \quad \text{or} \quad \bar{X}_i$$

DNF:

$$\phi = C_1 \vee C_2 \vee \cdots \vee C_m$$

where

$$C_i = \bigwedge_{j \in V_i} x_j$$

#DNF

$$\phi : \{0,1\}^n \to \{0,1\}$$

Goal:

$$\#\phi = |\{x : \phi(x) = 1\}|$$

Approximation:

$$\Pr[(1-\varepsilon)\#\phi \le \widetilde{Y} \le (1+\varepsilon)\#\phi] \ge 1-\delta$$

Naive Monte–Carlo

Sample:

$$x_1, \ldots, x_N \sim \{0, 1\}^n$$

Estimator:

$$\widetilde{Y} = \left(\frac{1}{N} \sum_{i=1}^{N} \phi(x_i) \right) 2^n$$

$$\mathbb{E}[\widetilde{Y}] = \#\phi$$

Zero–One Estimator Theorem

Let

$$\mu = \frac{\#\phi}{2^n}$$

If

$$N \geq \frac{4 \ln(2/\delta)}{\mu \varepsilon^2}$$

then

$$\Pr[(1-\varepsilon)\#\phi \leq \widetilde{Y} \leq (1+\varepsilon)\#\phi] \geq 1-\delta$$

Why naive Monte–Carlo fails

$$\mu = \frac{\#\phi}{2^n}$$

can be exponentially small.

Example:

$$\phi = X_1 \wedge X_2 \wedge \cdots \wedge X_n$$

Then:

$$\#\phi = 1$$

and:

$$\mu = \frac{1}{2^n}$$

Need:

$$N = \Omega(2^n)$$

Union structure in DNF

For:

$$\phi = C_1 \vee \cdots \vee C_m$$

define:

$$D_i = \{x : C_i(x) = 1\} \quad \text{and} \quad D = \bigcup_{i=1}^{m} D_i$$

Then:

$$\#\phi = |D|$$

Easy:

sample uniformly from D_i

Lifted space

Define:

$$U = D_1 \oplus D_2 \oplus \cdots \oplus D_m$$

Each element:

$$(x, i) \quad \text{with} \quad x \in D_i$$

Size:

$$|U| = \sum_{i=1}^{m} |D_i|$$

Coverage:

$$\mathrm{cov}(x) = \{(x, i) : x \in D_i\}$$

Duplication

Each satisfying assignment appears:

$$|\mathrm{cov}(x)|$$

times in U.

Hence:

$$|U| = \sum_{x \in D} |\mathrm{cov}(x)|$$

So:

$$|U| = (\text{average duplication}) \cdot |D|$$

Therefore:

$$|D| = \frac{|U|}{\text{average duplication}}$$

Canonical representative

Choose one representative copy:

$$f(x,i) = \begin{cases} 1 & i = \min\{j : x \in D_j\} \\ 0 & \text{otherwise} \end{cases}$$

Then:

$$\sum_{(x,i) \in \mathrm{cov}(x)} f(x,i) = 1$$

Define:

$$G = \{(x,i) \in U : f(x,i) = 1\}$$

Then:

$$|G| = |D|$$

Coverage Monte–Carlo

Sample:

$$(x, i) \in U$$

uniformly.

Estimator:

$$Y = |U| \cdot f(x, i)$$

$$\mathbb{E}[Y] = |U| \cdot \frac{|D|}{|U|} = |D| = \#\phi$$

Sampling procedure

Choose:

$$i$$

with probability:

$$\Pr[i] = \frac{|D_i|}{|U|}$$

Then choose:

$$x \in D_i$$

uniformly.

If clause C_i has k_i literals:

$$|D_i| = 2^{n-k_i}$$

Generating samples

Clause:

$$C_i = X_1 \wedge \bar{X}_3 \wedge X_7$$

Generate:

$$x \in D_i$$

by fixing:

$$X_1 = 1, \quad X_3 = 0, \quad X_7 = 1$$

and sampling remaining variables uniformly.

Density improvement

Naive Monte–Carlo:

$$\mu = \frac{\#\phi}{2^n}$$

Coverage Monte–Carlo:

$$\mu = \frac{|G|}{|U|} = \frac{|D|}{\sum_i |D_i|}$$

Since:

$$|D| \leq \sum_i |D_i| \leq m|D|$$

we obtain:

$$\mu \geq \frac{1}{m}$$

Complexity

By Zero–One Estimator:

$$N = O\left(\frac{m\log(1/\delta)}{\varepsilon^2}\right)$$

samples suffice.

Running time:

$$O\left((nm)\frac{m\log(1/\delta)}{\varepsilon^2}\right)$$

Key idea:

easy sampling + statistical overlap correction

Concentration

Let:

$$Y_1, \ldots, Y_N \in \{0, 1\}$$

with:

$$\mathbb{E}[Y_i] = \mu$$

Estimator:

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i$$

Chernoff:

$$\Pr[\bar{Y} > (1+\varepsilon)\mu] \leq e^{-\mu\varepsilon^2 N/4}$$

$$\Pr[\bar{Y} < (1-\varepsilon)\mu] \leq e^{-\mu\varepsilon^2 N/4}$$

Final bound

$$\Pr[|\bar{Y} - \mu| > \varepsilon\mu] \leq 2e^{-\mu\varepsilon^2 N/4}$$

Choose:

$$N \geq \frac{4\ln(2/\delta)}{\mu\varepsilon^2}$$

Then:

$$\Pr[(1-\varepsilon)\mu \leq \bar{Y} \leq (1+\varepsilon)\mu] \geq 1-\delta$$

Conclusion

$$\#\text{DNF} \in \text{FPRAS}$$

"Many samples do converge;
sample the right space."

Weighted correction

Instead of canonical representatives:

$$f(x,i) = \begin{cases} 1 & \text{canonical copy} \\ 0 & \text{otherwise} \end{cases}$$

use:

$$f(x,i) = \frac{1}{|\mathrm{cov}(x)|}$$

Then:

$$\sum_{(x,i)\in \mathrm{cov}(x)} f(x,i) = 1$$

Estimator:

$$Y = |U| \cdot \frac{1}{|\mathrm{cov}(x)|}$$

Interpretation:

Importance sampling over duplicated assignments

Self-adjusting coverage

Instead of computing:

$$|\mathrm{cov}(x)|$$

estimate it geometrically.

Repeat:

$$\text{sample random clause } C_j$$

until:

$$x \in D_j$$

Waiting time:

$$T_x \sim \mathrm{Geom}\left(\frac{|\mathrm{cov}(x)|}{m}\right)$$

Hence:

$$\mathbb{E}[T_x] = \frac{m}{|\mathrm{cov}(x)|}$$

Self-adjusting coverage

Improves running time:

$$O\left(nm^2\frac{\log(1/\delta)}{\varepsilon^2}\right) \to O\left(nm\frac{\log(1/\delta)}{\varepsilon^2}\right)$$