technische universität
dortmund

Computational Intelligence

Winter Term 2011/12

Prof. Dr. Günter Rudolph

Lehrstuhl für Algorithm Engineering (LS 11)

Fakultät für Informatik

TU Dortmund

Three tasks:

1. Choice of an appropriate problem representation.
2. Choice / design of variation operators acting in problem representation.
3. Choice of strategy parameters (includes initialization).

ad 1) different "schools":

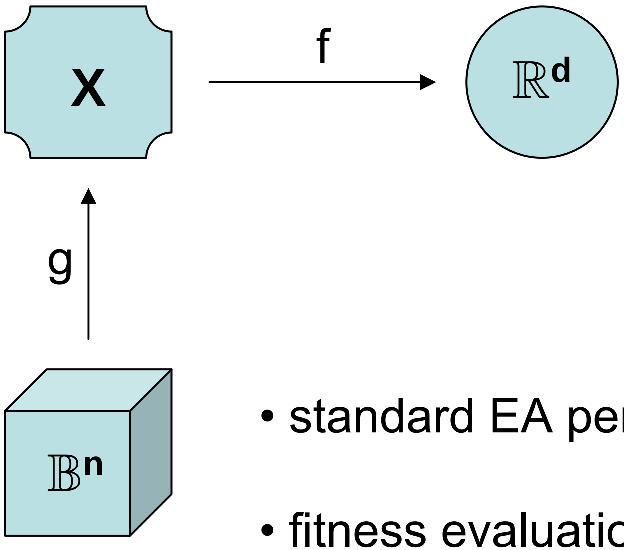
(a) operate on binary representation and define genotype/phenotype mapping
+ can use standard algorithm
– mapping may induce unintentional bias in search

(b) no doctrine: use "most natural" representation
– must design variation operators for specific representation
+ if design done properly then no bias in search

ad 1a) genotype-phenotype mapping

original problem f: $X \to \mathbb{R}^d$

scenario: no standard algorithm for search space X available



- standard EA performs variation on binary strings $b \in \mathbb{B}^n$
- fitness evaluation of individual b via $(f \circ g)(b) = f(g(b))$

  where $g: \mathbb{B}^n \to X$ is genotype-phenotype mapping
- selection operation independent from representation

Genotype-Phenotype-Mapping $\mathbb{B}^n \to [L, R] \subset \mathbb{R}$

- Standard encoding for $b \in \mathbb{B}^n$

$$x = L + \frac{R-L}{2^n - 1} \sum_{i=0}^{n-1} b_{n-i}\, 2^i$$

→ Problem: *hamming cliffs*

000	**001**	**010**	**011**	**100**	**101**	**110**	**111**	← genotype
0	1	2	3	4	5	6	7	← phenotype

1 Bit   2 Bit   1 Bit   3 Bit   1 Bit   2 Bit   1 Bit

↑

Hamming cliff

L = 0, R = 7

n = 3

Genotype-Phenotype-Mapping $\mathbb{B}^n \to [L, R] \subset \mathbb{R}$

- Gray encoding for $b \in \mathbb{B}^n$

Let $a \in \mathbb{B}^n$ standard encoded. Then $b_i = \begin{cases} a_i, & \text{if } i = 1 \\ a_{i-1} \oplus a_i, & \text{if } i > 1 \end{cases}$

$\oplus$ = XOR

000	**001**	**011**	**010**	**110**	**111**	**101**	**100**	← genotype
0	1	2	3	4	5	6	7	← phenotype

OK, no hamming cliffs any longer …

$\Rightarrow$ small changes in phenotype „lead to“ small changes in genotype

since we consider evolution in terms of Darwin (not Lamarck):

$\Rightarrow$ small changes in genotype lead to small changes in phenotype!

but: 1-Bit-change: **000** $\to$ **100** $\Rightarrow$ ☹

Genotype-Phenotype-Mapping $\mathbb{B}^n \rightarrow \mathbb{P}^{\log(n)}$ (example only)

- e.g. standard encoding for $b \in \mathbb{B}^n$

individual:

010	**101**	**111**	**000**	**110**	**001**	**101**	**100**	← genotype
0	1	2	3	4	5	6	7	← index

consider index and associated genotype entry as unit / record / struct;

sort units with respect to genotype value, old indices yield permutation:

000	**001**	**010**	**100**	**101**	**101**	**110**	**111**	← genotype
3	5	0	7	1	6	4	2	← old index

= permutation

ad 1a) genotype-phenotype mapping

typically required: strong causality

→ small changes in individual leads to small changes in fitness

→ small changes in genotype should lead to small changes in phenotype

but: how to find a genotype-phenotype mapping with that property?

necessary conditions:

1) g: $\mathbb{B}^n \to X$ can be computed efficiently (otherwise it is senseless)

2) g: $\mathbb{B}^n \to X$ is surjective (otherwise we might miss the optimal solution)

3) g: $\mathbb{B}^n \to X$ *preserves closeness* (otherwise strong causality endangered)

Let $d(\cdot\,,\cdot)$ be a metric on $\mathbb{B}^n$ and $d_X(\cdot\,,\cdot)$ be a metric on X.

$\forall x, y, z \in \mathbb{B}^n : d(x, y) \le d(x, z) \Rightarrow d_X(g(x), g(y)) \le d_X(g(x), g(z))$

ad 1b) use “most natural“ representation

typically required: strong causality

→ small changes in individual leads to small changes in fitness

→ need variation operators that obey that requirement

but: how to find variation operators with that property?

⇒ need design guidelines ...

ad 2) design guidelines for variation operators

a) reachability

every $x \in X$ should be reachable from arbitrary $x_0 \in X$
after finite number of repeated variations with positive probability bounded from 0

b) unbiasedness

unless having gathered knowledge about problem
variation operator should not favor particular subsets of solutions
$\Rightarrow$ formally: maximum entropy principle

c) control

variation operator should have parameters affecting shape of distributions;
known from theory: weaken variation strength when approaching optimum

ad 2) **design guidelines for variation operators in practice**

binary search space $X = \mathbb{B}^n$

variation by k-point or uniform crossover and subsequent mutation

a) ***reachability***:

regardless of the output of crossover
we can move from $x \in \mathbb{B}^n$ to $y \in \mathbb{B}^n$ in 1 step with probability

$$p(x,y) = p_m^{H(x,y)} (1 - p_m)^{n - H(x,y)} > 0$$

where $H(x,y)$ is Hamming distance between x and y.

Since $\min\{\, p(x,y) : x,y \in \mathbb{B}^n \,\} = \delta > 0$ we are done.

b) ***unbiasedness***

don‘t prefer any direction or subset of points without reason

$\Rightarrow$ use maximum entropy distribution for sampling!

<u>properties:</u>

- distributes probability mass as uniform as possible
- additional knowledge can be included as constraints:
  $\rightarrow$ under given constraints sample as uniform as possible

Formally:

Definition:

Let X be discrete random variable (r.v.) with $p_k = P\{ X = x_k \}$ for some index set K. The quantity

$$H(X) = -\sum_{k \in K} p_k \log p_k$$

is called the ***entropy of the distribution*** of X. If X is a continuous r.v. with p.d.f. $f_X(\cdot)$ then the entropy is given by

$$H(X) = -\int_{-\infty}^{\infty} f_X(x) \log f_X(x) dx$$

The distribution of a random variable X for which H(X) is maximal is termed a ***maximum entropy distribution***. ■

Knowledge available:

Discrete distribution with support $\{ x_1, x_2, \ldots x_n \}$ with $x_1 < x_2 < \ldots x_n < \infty$

$$p_k = \mathrm{P}\{X = x_k\}$$

$\Rightarrow$ leads to nonlinear constrained optimization problem:

$$-\sum_{k=1}^{n} p_k \log p_k \quad \to \text{max!}$$

$$\text{s.t.} \quad \sum_{k=1}^{n} p_k = 1$$

<u>solution:</u> via Lagrange (find stationary point of Lagrangian function)

$$L(p, a) = -\sum_{k=1}^{n} p_k \log p_k + a \left(\sum_{k=1}^{n} p_k - 1 \right)$$

$$L(p,a) = -\sum_{k=1}^{n} p_k \log p_k + a \left(\sum_{k=1}^{n} p_k - 1 \right)$$

partial derivatives:

$$\frac{\partial L(p,a)}{\partial p_k} = -1 - \log p_k + a \overset{!}{=} 0 \qquad \Rightarrow \quad p_k \overset{!}{=} e^{a-1}$$

$$\frac{\partial L(p,a)}{\partial a} = \sum_{k=1}^{n} p_k - 1 \overset{!}{=} 0$$

$$\Rightarrow \sum_{k=1}^{n} p_k = \sum_{k=1}^{n} e^{a-1} = n\, e^{a-1} \overset{!}{=} 1 \quad \Leftrightarrow \quad e^{a-1} = \frac{1}{n}$$

$$p_k = \frac{1}{n}$$

uniform distribution

$\blacksquare$

Knowledge available:

Discrete distribution with support { 1, 2, …, n } with p_k = P { X = k }   **and**   E[X] = ν

$\Rightarrow$ leads to nonlinear constrained optimization problem:

$$-\sum_{k=1}^{n} p_k \log p_k \quad \longrightarrow \text{max!}$$

$$\text{s.t.} \qquad \sum_{k=1}^{n} p_k = 1 \qquad \text{and} \qquad \sum_{k=1}^{n} k\, p_k = \nu$$

<u>solution:</u> via Lagrange (find stationary point of Lagrangian function)

$$L(p,a,b) = -\sum_{k=1}^{n} p_k \log p_k + a \left(\sum_{k=1}^{n} p_k - 1 \right) + b \left(\sum_{k=1}^{n} k \cdot p_k - \nu \right)$$

$$L(p,a,b) = -\sum_{k=1}^{n} p_k \log p_k + a\left(\sum_{k=1}^{n} p_k - 1\right) + b\left(\sum_{k=1}^{n} k\cdot p_k - \nu\right)$$

partial derivatives:

$$\frac{\partial L(p,a,b)}{\partial p_k} = -1 - \log p_k + a + b\,k \stackrel{!}{=} 0 \quad \Rightarrow \quad p_k = e^{a-1+b\,k}$$

$$\frac{\partial L(p,a,b)}{\partial a} = \sum_{k=1}^{n} p_k - 1 \stackrel{!}{=} 0$$

$$\frac{\partial L(p,a,b)}{\partial b} \stackrel{(*)}{=} \sum_{k=1}^{n} k\,p_k - \nu \stackrel{!}{=} 0$$

$$\Rightarrow \sum_{k=1}^{n} p_k = e^{a-1} \sum_{k=1}^{n} (e^b)^k \stackrel{!}{=} 1$$

(continued on next slide)

$$\Rightarrow \quad e^{a-1} = \frac{1}{\sum\limits_{k=1}^{n} (e^b)^k} \qquad \Rightarrow \quad p_k = e^{a-1+b\,k} = \frac{(e^b)^k}{\sum\limits_{i=1}^{n} (e^b)^i}$$

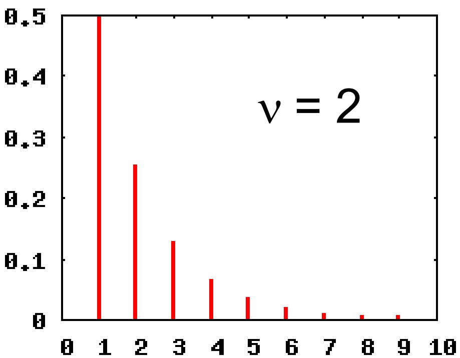
$\Rightarrow$ **discrete Boltzmann distribution**

$$p_k = \frac{q^k}{\sum\limits_{i=1}^{n} q^i} \qquad (q = e^b)$$

$\Rightarrow$ value of q depends on ν via third condition: (*)

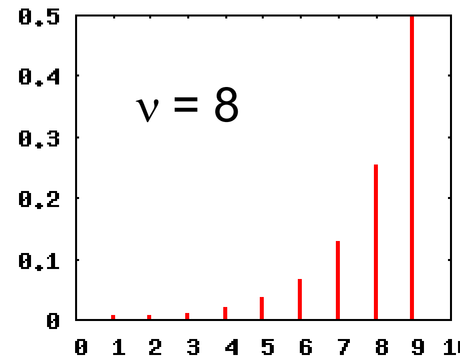
$$\sum_{k=1}^{n} k\,p_k = \frac{\sum\limits_{k=1}^{n} k\,q^k}{\sum\limits_{i=1}^{n} q^i} = \frac{1-(n+1)\,q^n + n\,q^{n+1}}{(1-q)\,(1-q^n)} \overset{!}{=} \nu$$

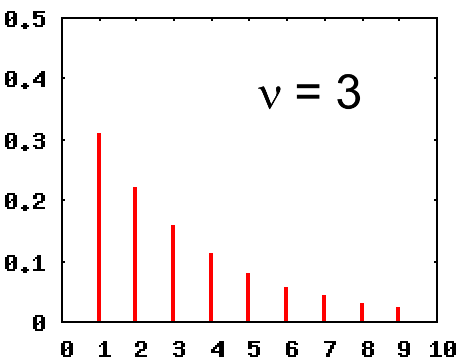
Excursion: Maximum Entropy Distributions



Boltzmann distribution

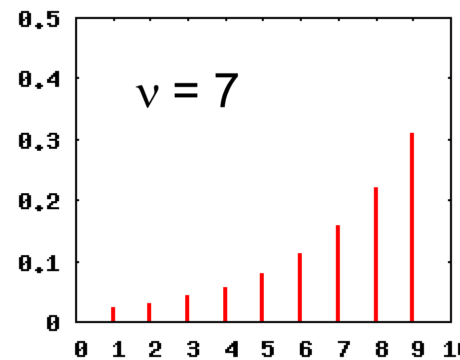
(n = 9)

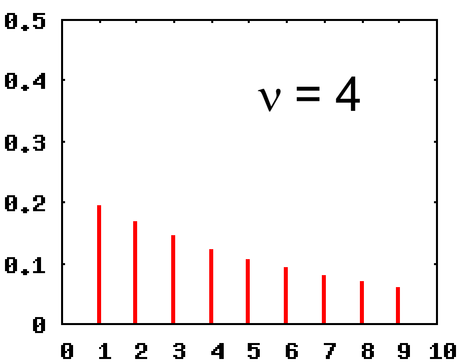


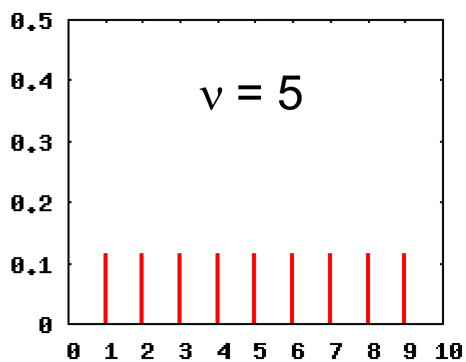


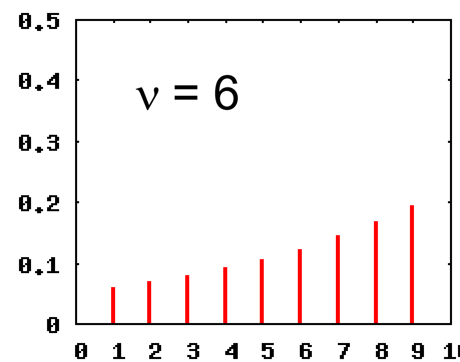
specializes to uniform distribution if $\nu = 5$

(as expected)









Knowledge available:

Discrete distribution with support { 1, 2, …, n } with E[X] = ν **and** V[X] = η^2

⇒ leads to nonlinear constrained optimization problem:

$$-\sum_{k=1}^{n} p_k \log p_k \quad \to \max!$$

$$\text{s.t.} \quad \sum_{k=1}^{n} p_k = 1 \quad \text{and} \quad \sum_{k=1}^{n} k\, p_k = \nu \quad \text{and} \quad \sum_{k=1}^{n} (k-\nu)^2 p_k = \eta^2$$

<u>solution:</u> in principle, via Lagrange (find stationary point of Lagrangian function)

but very complicated analytically, if possible at all

⇒ consider special cases only

note: constraints are linear equations in p_k

Special case: **n = 3** and **E[X] = 2** and **V[X] = η²**

Linear constraints uniquely determine distribution:

$$
\begin{array}{lrcrcrcl}
\text{I.} & p_1 & + & p_2 & + & p_3 & = & 1 \\
\text{II.} & p_1 & + & 2\,p_2 & + & 3\,p_3 & = & 2 \\
\text{III.} & p_1 & + & 0 & + & p_3 & = & \eta^2
\end{array}
$$

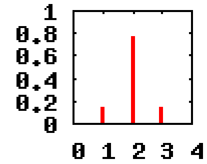
$$
\begin{array}{lrcrcl}
\text{II – I:} & p_2 & + & 2\,p_3 & = & 1 \\
\text{I – III:} & p_2 & & & = & 1-\eta^2
\end{array}
\quad \Bigg\} \quad p_3 = \frac{\eta^2}{2}
$$

insertion in III.

$$p_1 = \frac{\eta^2}{2}$$

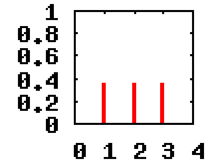
$$\Rightarrow \quad p = \left(\frac{\eta^2}{2}, 1-\eta^2, \frac{\eta^2}{2} \right)$$

$\eta^2 = \frac{1}{4}$



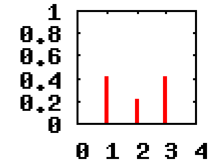
unimodal

$\eta^2 = \frac{2}{3}$



uniform

$\eta^2 = \frac{4}{5}$



bimodal

Knowledge available:

Discrete distribution with unbounded support { 0, 1, 2, … } and E[X] = ν

$\Rightarrow$ leads to infinite-dimensional nonlinear constrained optimization problem:

$$-\sum_{k=0}^{\infty} p_k \log p_k \quad \to \max!$$

$$\text{s.t.} \qquad \sum_{k=0}^{\infty} p_k = 1 \qquad \text{and} \qquad \sum_{k=0}^{\infty} k\, p_k = \nu$$

solution: via Lagrange (find stationary point of Lagrangian function)

$$L(p,a,b) = -\sum_{k=0}^{\infty} p_k \log p_k + a \left(\sum_{k=0}^{\infty} p_k - 1 \right) + b \left(\sum_{k=0}^{\infty} k \cdot p_k - \nu \right)$$

$$L(p,a,b) = -\sum_{k=0}^{\infty} p_k \log p_k + a\left(\sum_{k=0}^{\infty} p_k - 1\right) + b\left(\sum_{k=0}^{\infty} k \cdot p_k - \nu\right)$$

partial derivatives:

$$\frac{\partial L(p,a,b)}{\partial p_k} = -1 - \log p_k + a + b\,k \stackrel{!}{=} 0 \qquad \Rightarrow \quad p_k = e^{a-1+b\,k}$$

$$\frac{\partial L(p,a,b)}{\partial a} = \sum_{k=0}^{\infty} p_k - 1 \stackrel{!}{=} 0$$

$$\frac{\partial L(p,a,b)}{\partial b} \stackrel{(*)}{=} \sum_{k=0}^{\infty} k\,p_k - \nu \stackrel{!}{=} 0$$

$$\Rightarrow \sum_{k=0}^{\infty} p_k = e^{a-1} \sum_{k=0}^{\infty} (e^b)^k \stackrel{!}{=} 1$$

(continued on next slide)

$$\Rightarrow \quad e^{a-1} = \frac{1}{\sum\limits_{k=0}^{\infty} (e^b)^k} \qquad \Rightarrow \quad p_k = e^{a-1+b\,k} = \frac{(e^b)^k}{\sum\limits_{i=0}^{\infty} (e^b)^i}$$

insert

set $q = e^b$ and insists that $q < 1$ $\Rightarrow \sum\limits_{k=0}^{\infty} q^k = \dfrac{1}{1-q}$

$$\Rightarrow \quad p_k = (1-q)\,q^k \quad \text{for} \quad k = 0, 1, 2, \ldots$$

geometrical distribution

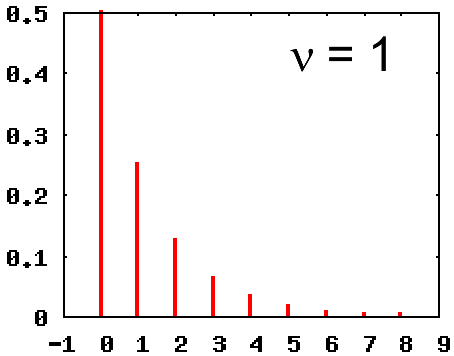
it remains to specify q; to proceed recall that $\sum\limits_{k=0}^{\infty} k\,q^k = \dfrac{q}{(1-q)^2}$

$\Rightarrow$ value of q depends on ν via third condition: (*)

$$\sum_{k=0}^{\infty} k\, p_k = \frac{\sum\limits_{k=0}^{\infty} k\, q^k}{\sum\limits_{i=0}^{\infty} q^i} = \frac{q}{1-q} \stackrel{!}{=} \nu$$

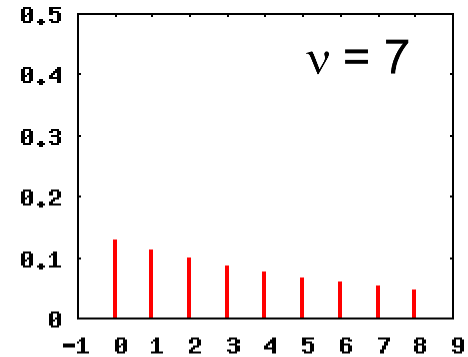
$$\Rightarrow \quad q = \frac{\nu}{\nu+1} = 1 - \frac{1}{\nu+1}$$

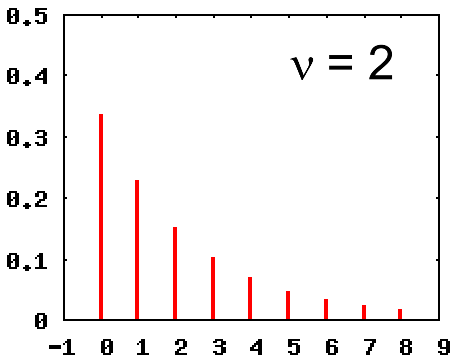
$$\Rightarrow \quad p_k = \frac{1}{\nu+1}\left(1 - \frac{1}{\nu+1}\right)^k$$



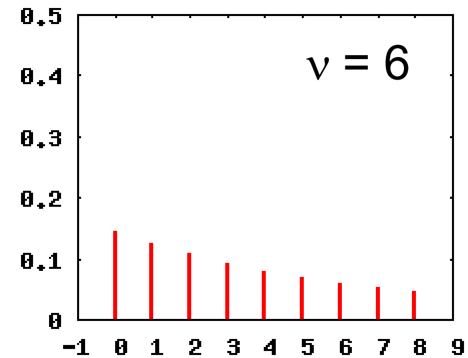
geometrical distribution

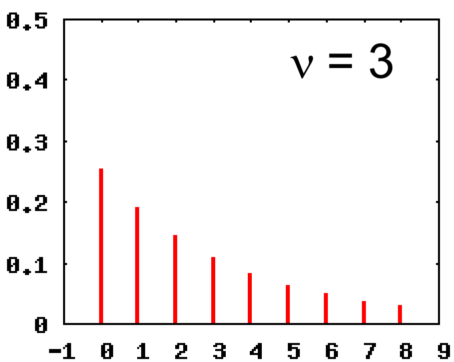
with $E[x] = \nu$

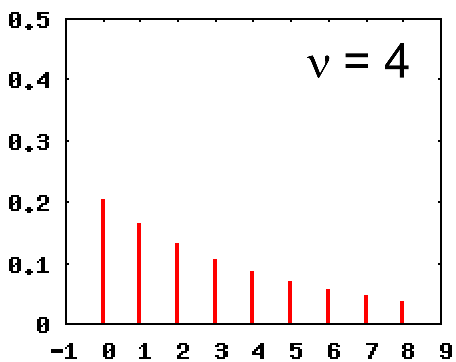


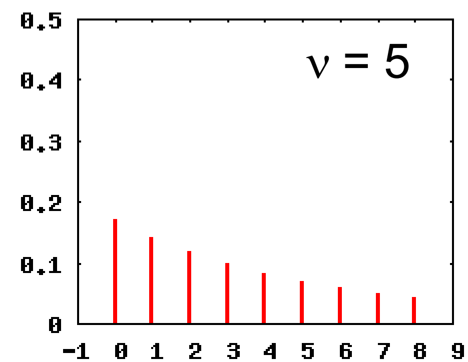


p_k only shown
for $k = 0, 1, \ldots, 8$









Overview:

support { 1, 2, …, n }	$\Rightarrow$ *discrete uniform* distribution
and require $E[X] = \theta$	$\Rightarrow$ *Boltzmann* distribution
and require $V[X] = \eta^2$	$\Rightarrow$ N.N. (**not** Binomial distribution)
support $\mathbb{N}$	$\Rightarrow$ not defined!
and require $E[X] = \theta$	$\Rightarrow$ *geometrical* distribution
and require $V[X] = \eta^2$	$\Rightarrow$?
support $\mathbb{Z}$	$\Rightarrow$ not defined!
and require $E[\|X\|] = \theta$	$\Rightarrow$ *bi-geometrical* distribution (*discrete Laplace* distr.)
and require $E[\|X\|^2] = \eta^2$	$\Rightarrow$ N.N. (*discrete Gaussian* distr.)

Excursion: Maximum Entropy Distributions

support $[a,b] \subset \mathbb{R}$ $\Rightarrow$ uniform distribution

support $\mathbb{R}^+$ with $E[X] = \theta$ $\Rightarrow$ Exponential distribution

support $\mathbb{R}$
with $E[X] = \theta$, $V[X] = \eta^2$ $\Rightarrow$ normal / Gaussian distribution $N(\theta, \eta^2)$

support $\mathbb{R}^n$
with $E[X] = \theta$
and $Cov[X] = C$ $\Rightarrow$ multinormal distribution $N(\theta, C)$

expectation vector $\in \mathbb{R}^n$

covariance matrix $\in \mathbb{R}^{n,n}$

positive definite:
$\forall x \neq 0 : x'Cx > 0$

for permutation distributions ?

→ uniform distribution on all possible permutations

```
set v[j] = j for j = 1, 2, ..., n
for i = n to 1 step -1
  draw k uniformly at random from { 1, 2, ..., i }
  swap v[i] and v[k]
endfor
```

generates permutation uniformly at random in $\Theta(n)$ time

Guideline:

Only if you know something about the problem *a priori* or

if you have learnt something about the problem *during the search*

⇒ include that knowledge in search / mutation distribution (via constraints!)

ad 2) **design guidelines for variation operators in practice**

continuous search space $X = \mathbb{R}^n$

a) reachability

b) unbiasedness

c) control

leads to CMA-ES