Blind No More: Deterministic Partition Crossover and Deterministic Improving Moves

Darrell Whitley
Computer Science, Colorado State University

http://dx.doi.org/10.1145/2908961.2926987

1

Blind No More = GRAY BOX Optimization

Darrell Whitley
Computer Science, Colorado State University

With Thanks to: Francisco Chicano, Gabriela Ochoa, Andrew Sutton and Renato Tinós

2

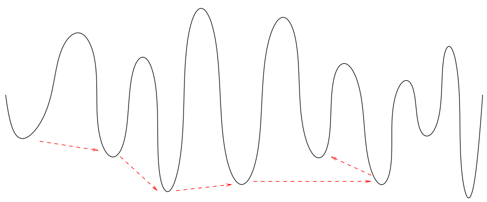
Know your Landscape! And Go Downhill!

3

Know your Landscape! And Go Downhill!

4

What if you could ...

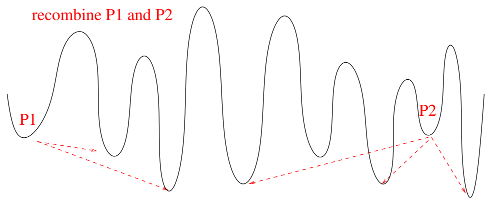


"Tunnel" between local optima on a TSP,
or on an NK Landscape or a MAXSAT problem.

Tunneling = jump from local optimum to local optimum

5

What if you could ...



"Tunnel" between local optima on a TSP,
or on an NK Landscape or a MAXSAT problem
AND go the BEST reachable local optima!

Tunneling = jump from local optimum to local optimum

6

The Partition Crossover Theorem for TSP

Let G be a graph produced by unioning 2 Hamiltonian Circuits.

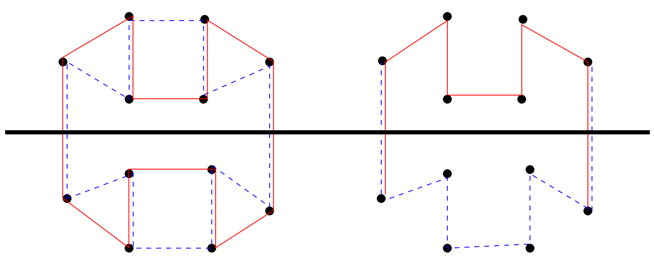
Let G' be a reduced graph so that all common subtours are replaced by a single surrogate common edge.

If there is a partition of G' with cost 2, then the 2 Hamiltonian Circuits that make up G can be cut and recombined at this partition to create two new offspring.

The resulting Partition Crossover is Respectful and Transmits alleles.

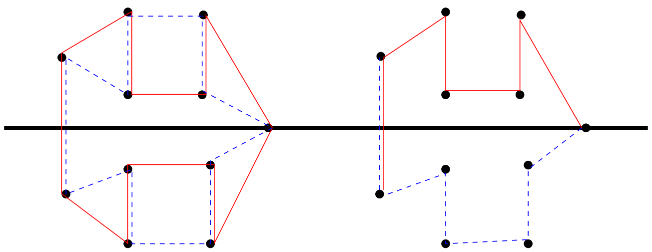
(Using G' makes the proof easier, but is not necessary.)

7



As a side effect: f(P1) + f(P2) = f(C1) + f(C2)

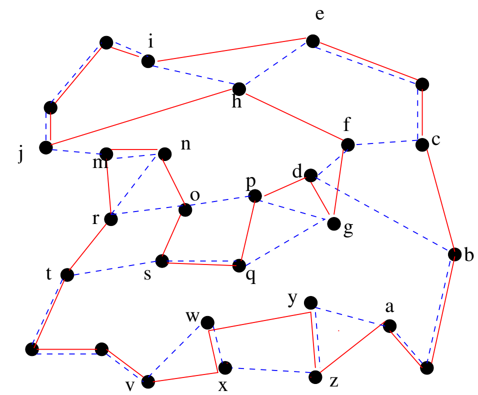
8



With Thanks to Gabriela Ochoa and Renato Tinós

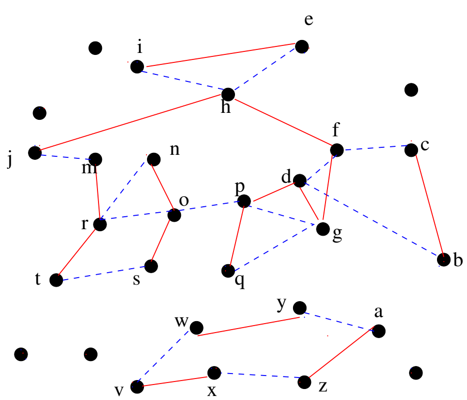
9

Partition Crossover



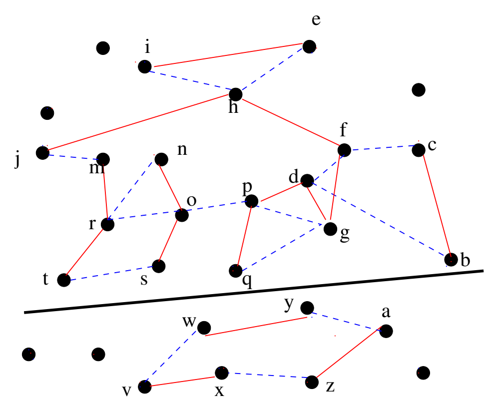
10

Partition Crossover



11

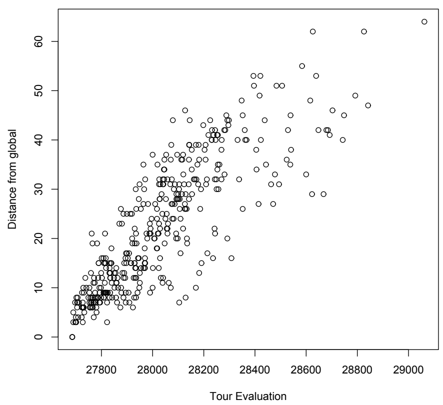
Partition Crossover in O(N) time



12

The Big Valley Hypothesis

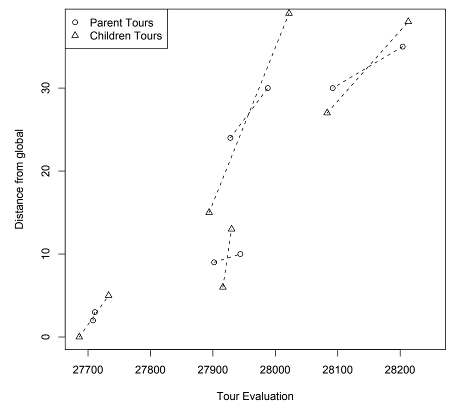
is sometimes used to explain metaheuristic search



13

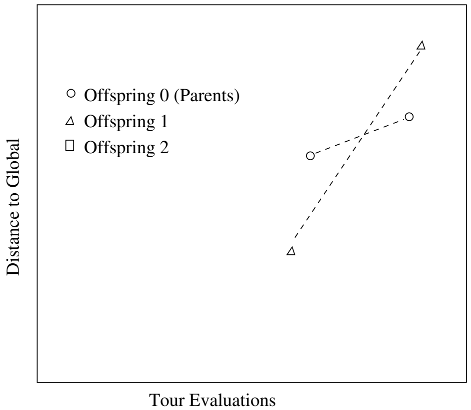
Tunneling Between Local Optima

Local Optima are “Linked” by Partition Crossover



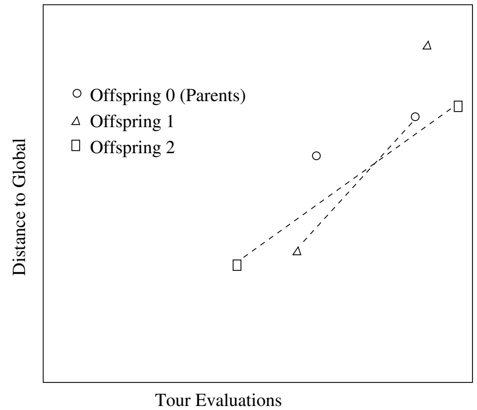
14

Iterated Partition Crossover



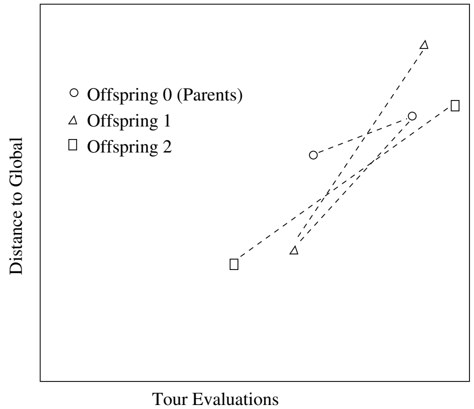
15

Iterated Partition Crossover



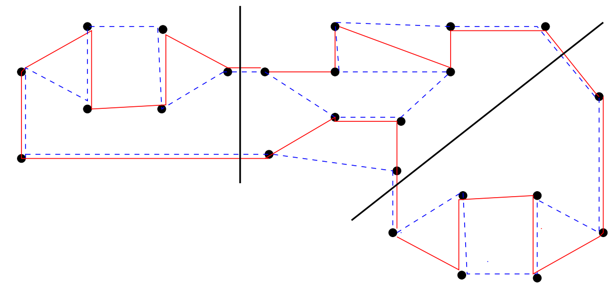
16

Generalized Partition Crossover



17

Generalized Partition Crossover

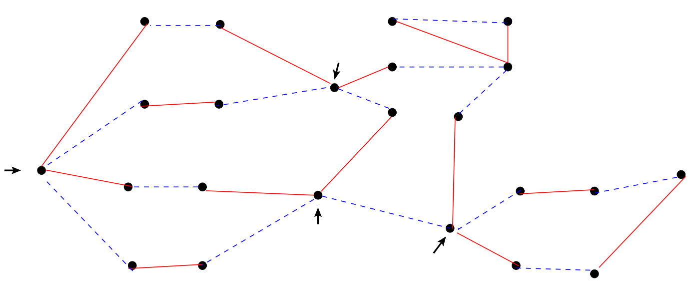


Generalize Partition Crossover is always feasible if the partitions have 2 exits (same color in and out). If a partition has more than 2 exits, the "colors" must match.

This will automatically happen if all of the partitions have cut two.

18

Generalized Partition Crossover with Splitting



19

How Many Partitions are Discovered?

Instance	att532	nrw1379	rand1500	u1817
2-opt	3.3 ± 0.2	3.2 ± 0.2	3.7 ± 0.3	5.0 ± 0.3
3-opt	10.5 ± 0.5	11.3 ± 0.5	24.9 ± 0.2	26.2 ± 0.7
LK-search	5.3 ± 0.2	5.2 ± 0.3	10.6 ± 0.3	13.3 ± 0.4

Table: Average number of *partition components* used by GPX in 50 recombinations of random local optima found by 2-opt, 3-opt and LK-search.

With 25 components, 2^{25} represents millions of local optima.

20

Lin-Kernighan-Helsgaun-LKH

LKH is widely considered the best Local Search algorithm for TSP.

LKH uses deep k-opt moves, clever data structures and a fast implementation.

LKH-2 has found the majority of best known solutions on the TSP benchmarks at the Georgia Tech TSP repository that were not solved by complete solvers: http://www.tsp.gatech.edu/data/index.html.

THE BEST HEURISTIC TSP SOLVERS USE CROSSOVER!

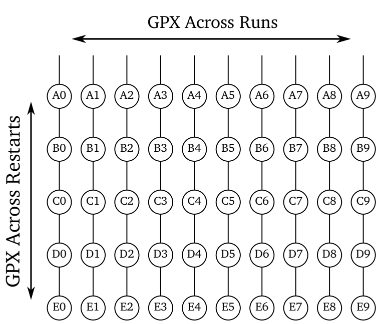
LKH uses"Iterated Partial Transcription"
which is almost the same as GPX but less efficient.

Iterative Partial Transcription and GPX

Instance	C10k.0	C10k.1	C31k.0	C31k.1
LKH-2 no crossover	1.143	1.009	1.489	1.538
LKH-2 w IPT	1.040	0.873	1.280	1.274
LKH-2 w GPX	**1.031**	**0.872**	**1.270**	**1.267**

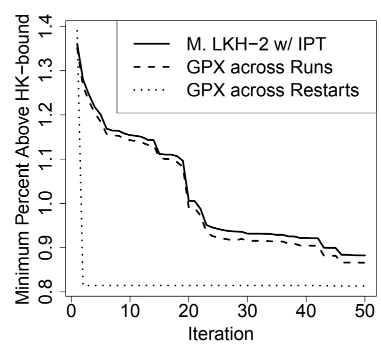
The minimum percentage above the Held-Karp Bound for several clustered instances of the TSP of solutions found by ten random restarts of LKH-2 without crossover, with IPT and with GPX. Best values for each instance are in boldface. Sizes range from 10,000 to 31,000 cities.

GPX Across Runs and Restarts



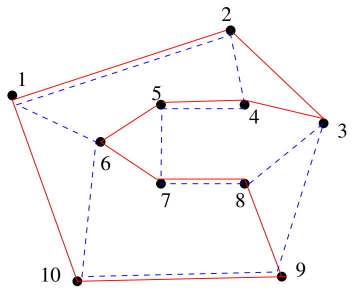
A diagram depicting 10 runs of multi-trial LKH-2 run for 5 iterations per run. The circles represent local optima produced by LKH-2. GPX across runs crosses over solutions with the same letters. GPX across restarts crosses over solutions with the same numbers.

GPX on Clustered Problems

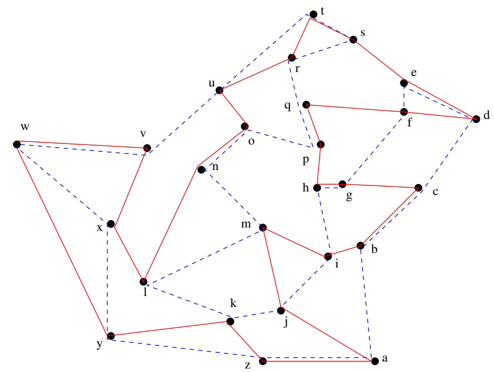


Improvement over time on a 31,000 city Dimacs Clustered Instance.

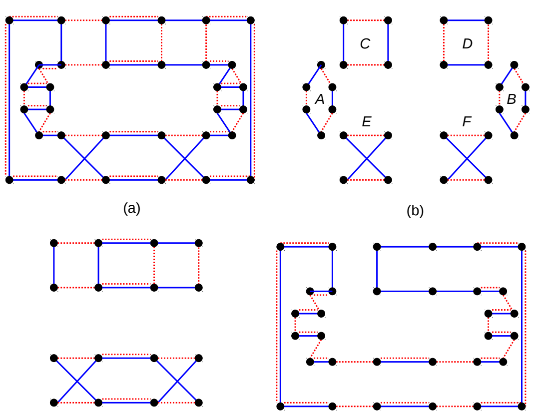
GPX, Cuts Crossing 4 Edges (IPT fails here)



GPX, Complex Cuts



GPX, Complex Cuts



(a) (b)

(c) (d)

NK-Landscapes and MAX-kSAT

For example: A Random NK Landscape: $n = 10$ and $k = 3$.
The subfunctions:

$f_0(x_0, x_1, x_6)$	$f_1(x_1, x_4, x_8)$	$f_2(x_2, x_3, x_5)$	$f_3(x_3, x_2, x_6)$
$f_4(x_4, x_2, x_1)$	$f_5(x_5, x_7, x_4)$	$f_6(x_6, x_8, x_1)$	$f_7(x_7, x_3, x_5)$
	$f_8(x_8, x_7, x_3)$	$f_9(x_9, x_7, x_8)$	

But this could also be a MAXSAT Function,
or an arbitrary Spin Glass problem.

A General Result over Bit Representations

By Constructive Proof: Every problem with a bit representation and a closed form evaluation function can be expressed as a quadratic (k=2) pseudo-Boolean Optimization problem. (See Boros and Hammer)

$$xy = z \;\; iff \;\; xy - 2xz - 2yz + 3z = 0$$
$$xy \neq z \;\; iff \;\; xy - 2xz - 2yz + 3z > 0$$

Or we can reduce to k=3 instead:

$$f(x_1, x_2, x_3, x_4, x_5, x_6)$$

becomes (depending on the nonlinearity):

$$f1(z_1, z_2, z_3) + f2(z_1, x_1, x_2) + f3(z_2, x_3, x_4) + f4(z_3, x_5, x_6)$$

GRAY BOX OPTIMIZATION

We can construct "Gray Box" optimization for pseudo-Boolean optimization problems composed of M subfunctions, where each subfunction accepts at most k variables.

Exploit the *general properties of every* Mk Landscape:

$$f(x) = \sum_{i=1}^{m} f_i(x)$$

Which can be expressed as a Walsh Polynomial

$$W(f(x)) = \sum_{i=1}^{m} W(f_i(x))$$

Or can be expressed as a sum of k Elementary Landscapes

$$f(x) = \sum_{i=1}^{k} \varphi^{(k)}(W(f(x)))$$

GRAY BOX OPTIMIZATION

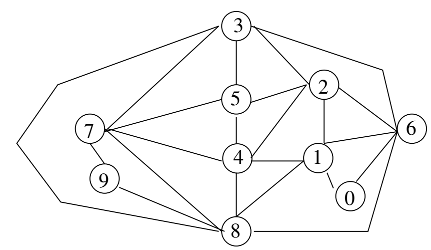
Don't wear a blind fold when crossing a busy street!

But keep the methods as general as possible.

Many methods are not really "Black Box."

For example "Parameterized Complexity"
depends on knowledge of the problem structure.

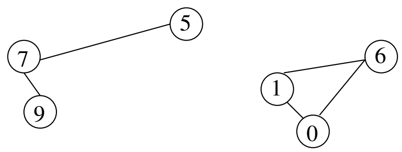
The Variable Interaction Graph



The Variable Interaction Graph

There is a vertex for each variable in the Variable Interaction Graph (VIG). There must be fewer than $2^k\ M = O(N)$ Walsh coefficients. There is a connection in the VIG between vertex v_i and v_j if there is a non-zero Walsh coefficient indexed by i and j, e.g., $w_{i,j}$.

The Recombination Graph: a reduced VIG

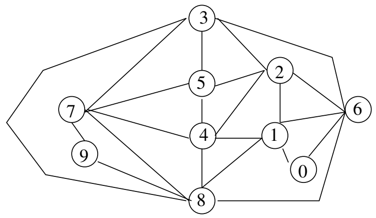


The decomposed Recombination Graph

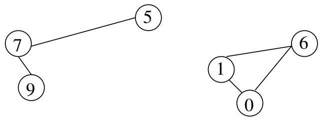
When recombining the solutions 0000000000 and 1100011101, the vertices and edges associated with shared variables 2, 3, 4, 8 are deleted to yield the **recombination graph**.

If the recombination graph of f contains q connected components, then Partition Crossover returns the best of 2^q solutions.

Decomposed Evaluation



The Variable Interaction Graph



The decomposed Recombination Graph

A new evaluation function can be constructed:

$$g(x) = c + g_1(x_5, x_7, x_9) + g_2(x_0, x_1, x_6)$$

where $g(x)$ can be used to evaluate any solution (parents or offspring) that resides in the subspace **000***0*.

Partition Crossover and Local Optima

The Subspace Optimality Theorem: For any k-bounded pseudo-Boolean function f, if Parition Crossover is used to recombine two parent solutions that are locally optimal, then the offspring must be a local optima in the hyperplane subspace defined by the bits shared in common by the two parents.

Example: if the parents 0000000000 and 1100011101
are locally optimal, then the best offspring
is locally optimal in the hyperplane subspace **000***0*.

Partition Crossover and Local Optima

Corolllary: The only possible improving move for offspring generated from parents that are locally optimal must flip a bit that the parents shared in common.

Illustration:

$$g(x) = c + g_1(x_5, x_7, x_9) + g_2(x_0, x_1, x_6)$$

If the parents were locally optimal, every subfunction g_i is locally optimal and offspring cannot be improved by a bit flip.

The only improving moves are on shared bits: **000***0*.

NK-Landscapes

An Adjacent NK Landscape: $n = 6$ and $k = 3$. The subfunctions:

$$\begin{array}{l} f_0(x_0, x_1, x_2) \\ \quad f_1(x_1, x_2, x_3) \\ \qquad f_2(x_2, x_3, x_4) \\ \qquad\quad f_3(x_3, x_4, x_5) \\ \qquad\qquad f_4(x_4, x_5, x_0) \\ \qquad\qquad\quad f_5(x_5, x_0, x_1) \end{array}$$

These problems can be solved to optimality using Dynamic Programming.

37

Percent of Offspring that are Local Optima

Using a Very Simple (Stupid) Hybrid GA:

N	k	Model	2-point Xover	Uniform Xover	PX
100	2	Adj	74.2 ±3.9	0.3 ±0.3	100.0 ±0.0
300	4	Adj	30.7 ±2.8	0.0 ±0.0	94.4 ±4.3
500	2	Adj	78.0 ±2.3	0.0 ±0.0	97.9 ±5.0
500	4	Adj	31.0 ±2.5	0.0 ±0.0	93.8 ±4.0
100	2	Rand	0.8 ±0.9	0.5 ±0.5	100.0 ±0.0
300	4	Rand	0.0 ±0.0	0.0 ±0.0	86.4 ±17.1
500	2	Rand	0.0 ±0.0	0.0 ±0.0	98.3 ±4.9
500	4	Rand	0.0 ±0.0	0.0 ±0.0	83.6 ±16.8

38

Number of partition components discovered

N	k	Model	Paired PX	
			Mean	Max
100	2	Adjacent	3.34 ±0.16	16
300	4	Adjacent	5.24 ±0.10	26
500	2	Adjacent	7.66 ±0.47	55
500	4	Adjacent	7.52 ±0.16	41
100	2	Random	3.22 ±0.16	15
300	4	Random	2.41 ±0.04	13
500	2	Random	6.98 ±0.47	47
500	4	Random	2.46 ±0.05	13

Paired PX uses Tournament Selection. The first parent is selected by fitness. The second parent is selected by Hamming Distance.

39

Optimal Solutions for Adjacent NK

		2-point	Uniform	Paired PX
N	k	Found	Found	Found
300	2	18	0	100
300	3	0	0	100
300	4	0	0	98
500	2	0	0	100
500	3	0	0	98
500	4	0	0	70

Percentage over 50 runs where the global optimum was Found in the experiments of the hybrid GA with the Adjacent NK Landscape.

40

But a Hybrid Genetic Algorithm is NOT how we should solve these NK Landscape Problems.

We can exactly know the location of improving moves in constant time. No *enumeration of neighbors* is needed.

Under well behaved conditions,
we can exactly know the location of improving moves
for r steps ahead in constant time.

Walsh Analysis

Every n-bit MAXSAT or NK-landscape or P-spin problem is a sum of m subfunctions, f_i:

$$f(x) = \sum_{i=1}^{m} f_i(x)$$

The Walsh transform of f is a sum of the Walsh transforms of the individual subfunctions.

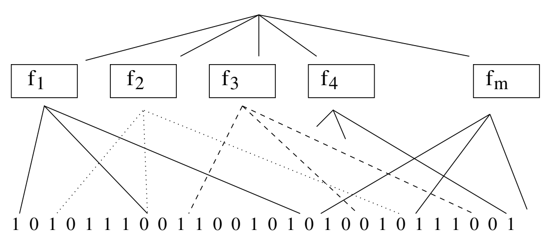
$$W(f(x)) = \sum_{i=1}^{m} W(f_i(x))$$

If m is $O(n)$ then the number of Walsh coefficients is $m\ 2^k = O(n)$.

"Mk Landscapes" (GECCO paper on Monday)

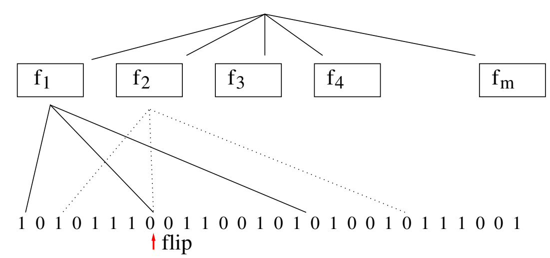
A General Model for all bounded Pseudo-Boolean Problems

$$f(x) = \sum_{i=1}^{m} f_i(x, \text{mask})$$



When 1 bit flips what happens?

$$f(x) = \sum_{i=1}^{m} f_i(x, \text{mask}_i)$$



Constant Time Steepest Descent

Assume we flip bit p to move from x to $y_p \in N(x)$.
Construct a vector $Score$ such that

$$Score(x, y_p) = -2 \left\{ \sum_{\forall b,\ p \subset b} -1^{b^T x} w_b(x) \right\}$$

In this way, all of the Walsh coefficients whose signs will be changed by flipping bit p are collected into a single number $Score(x, y_p)$.

The GSAT algorithm has done this for 23 years (Thanks to H. Hoos).

NOTE: Hoos and Stützle have claimed a constant time result, but without proof. An average case complexity proof is required to obtain general constant time complexity results (Whitley 2013, AAAI). Also it does not matter if the problems are uniform random or not.

45

Best Improving and Next Improving moves

"Best Improving" and "Next Improving" moves cost the same.

GSAT uses a Buffer of best improving moves

$$Buffer(best.improvement) =< M_{10}, M_{1919}, M_{9999} >$$

But the Buffer does not empty monotonically: this leads to thrashing.

Instead uses multiple Buckets to hold improving moves

$$Bucket(best.improvement) =< M_{10}, M_{1919}, M_{9999} >$$

$$Bucket(best.improvement - 1) =< M_{8371}, M_{4321}, M_{847} >$$

$$Bucket(all.other.improving.moves) =< M_{40}, M_{519}, M_{6799} >$$

This improves the runtime of GSAT by a factor of 20X to 30X.
The solution for NK Landscapes is only slightly more complicated.

46

The locations of the updates are obvious

$$\begin{aligned}
Score(y_p, y_1) &= Score(x, y_1) \\
Score(y_p, y_2) &= Score(x, y_2) \\
Score(y_p, y_3) &= Score(x, y_3) - 2(\sum_{\forall b,\ (p \wedge 3) \subset b} w'_b(x)) \\
Score(y_p, y_4) &= Score(x, y_4) \\
Score(y_p, y_5) &= Score(x, y_5) \\
Score(y_p, y_6) &= Score(x, y_6) \\
Score(y_p, y_7) &= Score(x, y_7) \\
Score(y_p, y_8) &= Score(x, y_8) - 2(\sum_{\forall b,\ (p \wedge 8) \subset b} w'_b(x)) \\
Score(y_p, y_9) &= Score(x, y_9)
\end{aligned}$$

47

What if we could look R Moves Lookahead?

Consider R=3

Let $Score(3, x, y_{i,j,k})$ indicate we move from x to $y_{i,j,k}$ by flipping the 3 bits i, j, k. In general, we compute $Score(r, x, y_p)$ when flipping r bits.

$$f(y_i) = f(x) + Score(1, x, y_i)$$

$$\begin{aligned}
f(y_{i,j}) &= f(y_i) + Score(1, y_i, y_j) \\
f(y_{i,j}) &= f(x) + Score(2, x, y_{i,j})
\end{aligned}$$

$$\begin{aligned}
f(y_{i,j,k}) &= f(y_{i,j}) + Score(1, y_{i,j}, y_k) \\
f(y_{i,j,k}) &= f(x) + Score(3, x, y_{i,j,k})
\end{aligned}$$

With thanks to Francisco Chicano!

48

Why Doesn't this exponentially EXPLODE???

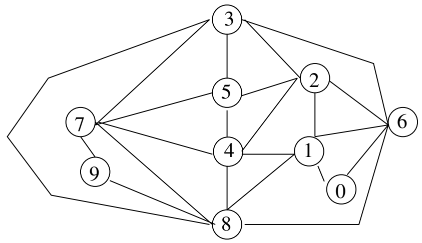
$$
\begin{aligned}
f(y_{i,j,k}) &= ((f(x) + Score(1, x, y_i)) + Score(1, y_i, y_j)) + Score(1, y_{i,j}, y_k) \\
Score(3, x, y_{i,j,k}) &= Score(2, x, y_{i,j}) + Score(1, y_{i,j}, y_{i,j,k})
\end{aligned}
$$

If there is no Walsh Coefficient $w_{i,j}$ then $Score(1, y_i, y_{i,j}) = 0$.

Assume we have already moves of length shorter than 3.
If there are no Walsh Coefficients "linking" i, j, k then $Score(3, x, y_{i,j,k}) = 0$.

49

The Variable Interaction Graph



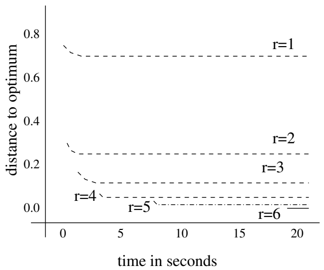
The Variable Interaction Graph

Assume all distance 1 moves are taken.

There cannot be a move flipping bits 4, 6, 9 that yields an improving move because there are no interactions and no Walsh coefficients.

50

Multiple Step Lookahead Local Search



In this figure, N = 12,000, k=3, and q=4. The radius is 1, 2, 3, 4, 5, 6.
At r=6 the global optimum is found.

51

What's (Obviously) Next?

- Local Search with **r** Move Lookahead PLUS Partition Crossover.
- Apply **r** Move Lookahead and Partition Crossover to MAX-kSAT.
- Use Deterministic Improving Moves.
- Use Deterministic Recombination.

52

What's (Obviously) Next?

- Put an End to the domination of Black Box Optimization.
- Wait for Tonight and Try to Take over the World.

53

Introduction to Elementary Landscapes

What is a landscape?

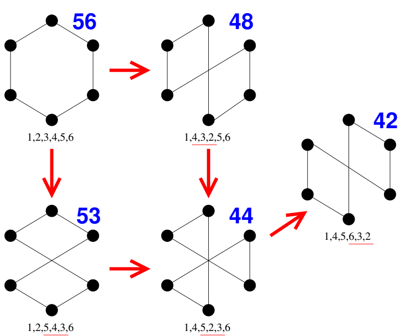
- Many different intuitive definitions
- A mathematical formalism of the *search space* of a combinatorial optimization problem

Definition: a landscape is a tuple (X, N, f)

A set of *states*	X
A *neighborhood* operator	$N : X \mapsto \mathcal{P}(X)$
A *fitness* function	$f : X \mapsto \mathbb{R}$

54

Introduction: a landscape



X	set of *states*
$N : X \mapsto \mathcal{P}(X)$	neighborhood operator
$f : X \mapsto \mathbb{R}$	objective function

55

Preliminaries

$$G(X, E)$$

is the underlying graph induced by N.
We assume G is regular with vertices of degree d.

$$\mathbf{A} \in \mathbb{R}^{|X| \times |X|}$$

is the *adjacency matrix* of G.
If x_1 and x_2 are neighbors, $A(x_1, x_2) = 1$.

$$\mathbf{\Delta} = \mathbf{A} - d\mathbf{I}$$

is the Laplacian of G.

56

The Wave Equation: definition 1

On an arbitrary landscape

- f and N are *unrelated*

On an elementary landscape

The wave equation

$$\mathbf{\Delta} f = \lambda f$$

- where λ is a scalar
- In other words, f is an eigenvector of the Laplacian

57

The Wave Equation: definition 1

Average change

$$\mathbf{\Delta} f = (\mathbf{A} - d\mathbf{I})f = k(\bar{f} - f)$$

$$\mathbf{\Delta} f(x) = \sum_{y \in N(x)} (f(y) - f(x)) = k(\bar{f} - f(x))$$

Average value

$$\begin{aligned} \underset{y \in N(x)}{\text{avg}} \{f(y)\} &= \frac{1}{d} \sum_{y \in N(x)} f(y) \\ &= f(x) + \frac{1}{d} \left(\sum_{y \in N(x)} f(y) - f(x) \right) \\ &= f(x) + \frac{1}{d} \Delta f(x) \\ &= f(x) + \frac{k}{d} (\bar{f} - f(x)) \end{aligned}$$

58

The Wave Equation: definition 2

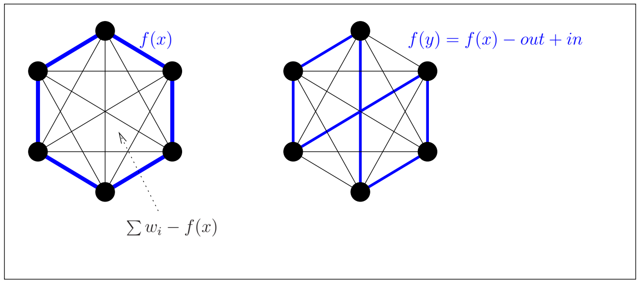
$$f(x) = \sum \text{ a subset of "components"}$$

Starting from average...

$$\underset{y \in N(x)}{\text{avg}} \{f(y)\} = f(x) + \underset{y \in N(x)}{\text{avg}} \{\text{components in} - \text{components out}\}$$

59

Example: TSP under 2-opt



- Components: set of edge weights $w_{i,j}$
- $f(x) =$ sum of edge weights induced by tour x
- There are $n(n-1)/2 - n$ weights not in tour x
- Average value of components out: $\frac{2}{n} f(x)$
- Average value of components in: $\frac{2}{n(n-3)/2} (\sum w - f(x))$

60

The Components and $\bar{f}$

Let C denote the set of components

$0 < p_3 < 1$ is the proportion of the components in C that contribute to the cost function for any randomly chosen solution

$$\bar{f} = p_3 \sum_{c \in C} c$$

For the TSP:

$$\bar{f} = \frac{n}{n(n-1)/2} \sum_{w_{i,j} \in C} w_{i,j}$$

$$\bar{f} = \frac{2}{n-1} \sum_{w_{i,j} \in C} w_{i,j}$$

61

The Wave Equation: definition 2

$$\begin{aligned}
\operatorname*{avg}_{y \in N(x)} \{f(y)\} &= f(x) + \frac{2}{n(n-3)/2}\left(\sum w - f(x)\right) - \frac{2}{n} f(x) \\
&= f(x) + \frac{2}{n(n-3)/2}\left((n-1)/2\bar{f} - f(x)\right) - \frac{2}{n} f(x) \\
&= f(x) + \frac{(n-1)}{n(n-3)/2}(\bar{f} - f(x)) \\
&= f(x) + \frac{k}{d}(\bar{f} - f(x))
\end{aligned}$$

62

For a 5 city TSP

	ab	bc	cd	de	ae	ac	ad	bd	be	ce
ABCDE	1	1	1	1	1	0	0	0	0	0
ABEDC	1	0	1	1	0	1	0	0	1	0
ABCED	1	1	0	1	0	0	1	0	0	1
ABDCE	1	0	1	0	1	0	0	1	0	1
ACBDE	0	1	0	1	1	1	0	1	0	0
ADCBE	0	1	1	0	1	0	1	0	1	0

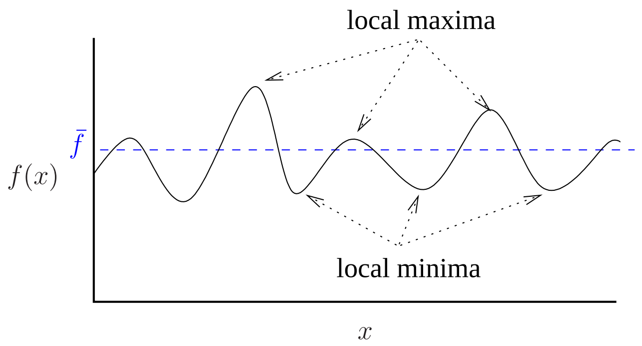
63

Looking at the neighbors in aggregate.

ab	bc	cd	de	ae	ac	ad	bd	be	ce
1	1	1	1	1	0	0	0	0	0
1	1	1	1	1	0	0	0	0	0
1	1	1	1	1	0	0	0	0	0
0	0	0	0	0	1	1	1	1	1
0	0	0	0	0	1	1	1	1	1

64

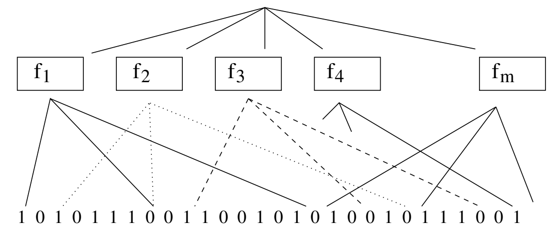
Properties



A General Model for all bounded Pseudo-Boolean Problems

This obviously applies to MAXSAT and NK Landscapes

$$f(x) = \sum_{i=1}^{m} f_i(x, \text{mask})$$



Superpositions of Elementary Landscapes

$$f(x) = f1(x) + f2(x) + f3(x) + f4(x)$$

$$f1(x) = f1_a(x) + f1_b(x) + f1_c(x)$$
$$f2(x) = f2_a(x) + f2_b(x) + f2_c(x)$$
$$f3(x) = f3_a(x) + f3_b(x) + f3_c(x)$$
$$f4(x) = f4_a(x) + f4_b(x) + f4_c(x)$$

$$\varphi^{(1)}(x) = f1_a(x) + f2_a(x) + f3_a(x) + f4_a(x)$$
$$\varphi^{(2)}(x) = f1_b(x) + f2_b(x) + f3_b(x) + f4_a(x)$$
$$\varphi^{(3)}(x) = f1_c(x) + f2_c(x) + f3_c(x) + f4_a(x)$$

$$f(x) = \varphi^{(1)}(x) + \varphi^{(2)}(x) + \varphi^{(3)}(x)$$

MAX-3SAT decomposition

MAX-3SAT is a superposition of 3 elementary landscapes

Walsh span of order p

$$\varphi^{(p)} = \sum_{\{i\ :\ bc(i)=p\}} w_i \psi_i$$

The p^{th} Walsh span is an elementary landscape

$$\Delta\varphi^{(p)} = -2p\varphi^{(p)}$$

With Thanks to Andrew Sutton!

MAX-3SAT decomposition

Recall that we can express f as:

$$f(x) = \sum_{i=1}^{m} \sum_{j=1}^{2^k} w_{m(i,j)} \psi_{m(i,j)}(x)$$

Grouping the Walsh decomposition results in

$$f(x) = \sum_{p=0}^{3} \varphi^{(p)}(x)$$

Thus MAX-3SAT is a superposition of 3-elementary landscapes

69

This allows us to compute Statistical Summaries

We can compute statistics efficiently over Hyperplanes

We can compute statistics efficiently over Hamming Balls

70

THANK YOU

Take Home Message:

PROBLEM STRUCTURE MATTERS.

Black Box Optimizers can never match the performance of an algorithm that efficiently exploits problem structure.

But we need only a small amount of information:
Gray Box Optimization.

For Mk Landscapes , we can use
Deterministic Moves and Deterministic Crossover.

71