STOCHASTIC SEARCH FOR NONLINEAR CONSTRAINED GLOBAL OPTIMIZATION

Benjamin W. Wah

Department of Electrical and Computer Engineering
and the Coordinated Science Laboratory
University of Illinois at Urbana-Champaign
Urbana, Illinois 61801, USA
b-wah@uiuc.edu
http://manip.crhc.uiuc.edu

Contributors: Prof. Y. N. Chang, Prof. Y. Shang, Y. X. Chen, T. Wang, and Z. Wu

Outline

- Introduction
  - Problem definition
  - Previous work
- Theory of discrete Lagrange multipliers
  - Necessary and sufficient conditions for constrained local minimization
- Stochastic search algorithms (SSAs) for constrained global minimization
  - Provable asymptotic convergence
- Theory of SSAs
  - Optimal schedules
  - Limits of parallel processing
- Some sample results
- Conclusions

Motivations

- Abundant applications in nonlinear constrained optimization
  - Mathematics
  - Machine vision
  - Robotics
  - Database systems
  - Text processing
  - Computer graphics
  - Security
  - Artificial intelligence
  - Integrated circuit design automation
  - Autonomous control
  - Autonomous planning
  - Computer networks
  - Mechanical designs
- NP-hard
  - Complete methods and parallel processing cannot handle large problem instances
  - Heuristic methods have difficulties with nonlinear constraints

Nonlinear Constrained Optimization

- Nonlinear constrained minimization problem

$$\begin{array}{lll} \text{minimize} & f(x) & \\ \text{subject to} & g(x) \leq 0 & x = (x_1, x_2, \ldots, x_n) \\ & h(x) = 0 & \text{is a vector of variables} \end{array}$$

- Assumptions
  - Feasible local minimum exists
    - Variable space needs not be bounded
    - Variables may be continuous, mixed-integer, or discrete
    - $f(x)$ should be lower bounded ($\Rightarrow$ existence of minimum)
    - Constraint functions need not be bounded
  - No differentiability and continuity requirements on objectives and constraints

Aerospace Applications

- Helicopter blade design
  - Minimize blade vibration by choosing optimal mass, stiffness, etc. within certain ranges subject to an upper bound on total mass
- Wing platform design
  - Minimize operating costs while meeting constraints on minimum range, maximum weight, takeoff length, landing approach speed, etc.
- Engine nozzle design
  - Determine 10 nozzle geometry parameters such that energy losses in an aero engine are minimized

Industrial Engineering Applications

- Calibration of pipes
  - Automatic calibration of water distribution network
- Traffic equilibrium
  - Constrained variational problems that arise in traffic equilibrium
- Energy-conversion systems
  - Calculation of mass flows, thermodynamic properties (temperature, pressure, enthalpy, entropy) and composition of gases in large energy-conversion systems (like power stations)

Previous Work on Derivative-Free Methods

- *Direct-solution methods*
  - *Reject, discard, repair*
  - *Enumerative branch-and-bound* needs linearized constraints
  - *Branch-and-reduce* has difficulties with highly nonlinear functions
- *Transformations into nonlinear constrained 0-1 programming* are restricted to small problems, due to increased number of variables
- *Lagrangian relaxation* works for linear integer minimization
- *Transformations into unconstrained penalty functions*
  - Many existing unconstrained optimization algorithms
  - Unless suitable penalties are chosen, a local minimum of an unconstrained penalty function is only a necessary but not a sufficient condition for the point to be a constrained local minimum of the original problem

Previous Work Requiring Differentiability

- *Lagrange-multiplier methods*
  - Require differentiability and continuity of functions
- *Interval methods* work for differentiable continuous functions
- *Generalized Benders decompositions* and *outer approximations*
  - Work for MINLPs when the number of decomposed convex subproblems is manageable
- *Penalty methods* have the same limitations as before
- Not for discrete and mixed-integer global optimization or for problems with non-differentiable continuous functions

Our Approach

- Discretized continuous variables

  $\Rightarrow$ Unified representation in discrete constrained NLPs

- Penalty formulations

  $\Rightarrow$ New necessary and sufficient conditions (easy to implement)

- Stochastic search algorithms

  $\Rightarrow$ Provable asymptotic convergence to constrained global minima

  $\Rightarrow$ Optimal schedules and limits on parallel processing

THEORY OF DISCRETE LAGRANGE MULTIPLIERS

Nonlinear Optimization Problems with Equality Constraints

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & h(x) = 0 \end{aligned}$$

$$x = (x_1, x_2, \ldots, x_n)$$
$$h(x) = (h_1(x), \ldots, h_m(x))$$

Continuous Space

x is a vector of continuous variables

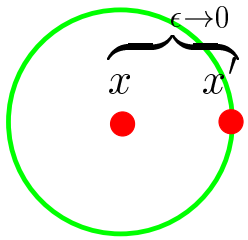
Discrete Space

x is a vector of discrete variables

Neighborhood $\mathcal{N}(x)$ of Point x

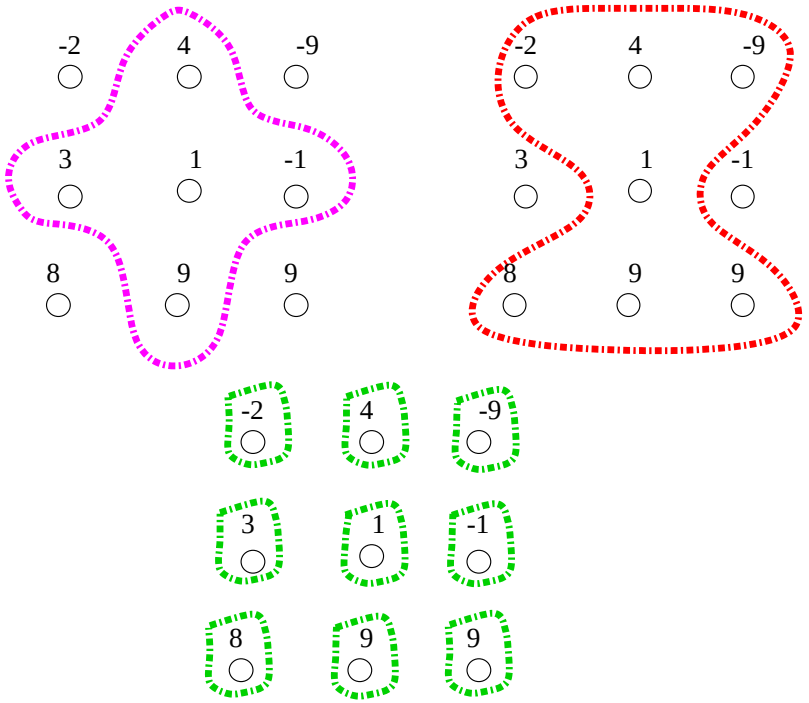
Continuous Space: $\mathcal{N}_{cn}(x)$

Defined by open sphere



Discrete Space: $\mathcal{N}_{dn}(x)$

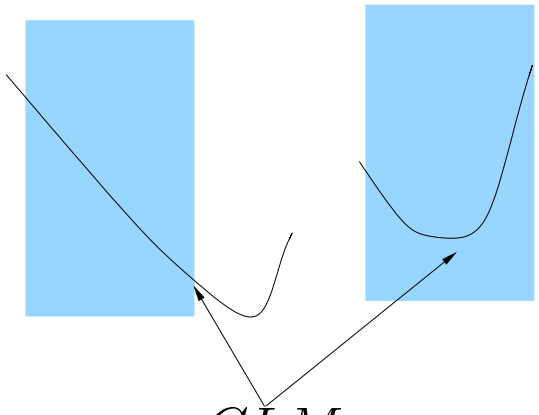
User defined



Constrained Local Minimum (CLM)

Continuous Space: CLM_{cn}

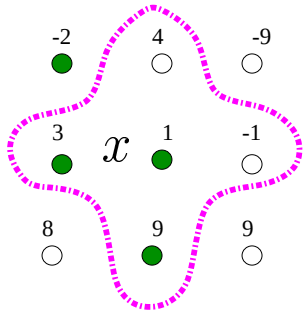
- Feasible local minimum when compared to feasible points inside an open sphere
- Whether point x is a CLM_{cn} is well defined



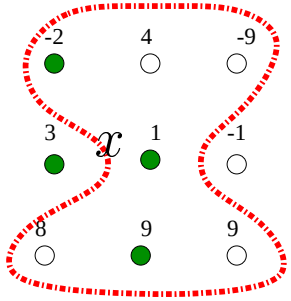
CLM_{cn}

Discrete Space: CLM_{dn}

- Feasible local minimum with respect to neighboring feasible points
- Whether point x is a CLM_{dn} depends on $\mathcal{N}_{dn}(x)$



CLM_{dn} at x



No CLM_{dn} at x

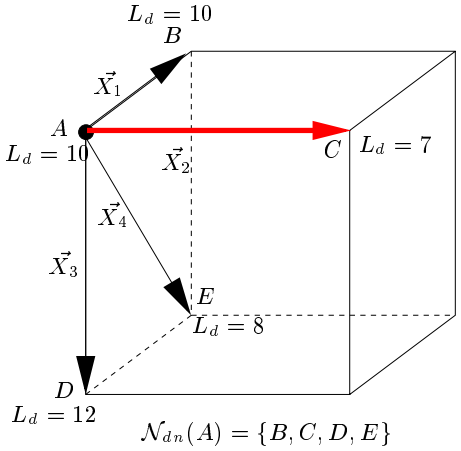
Descent Directions

Continuous Space

- Gradient (∇) provides direction of descents
- Composition, addition, and multiplication of gradients
- Chain rule

Discrete Space

- Direction of Maximum Potential Drop (DMPD)
  - $\Delta(x)$ is a vector that points in the direction of maximum function-value drop in the neighborhood of x
- No similar operations on DMPDs
- DMPD of x depends on $\mathcal{N}_{dn}(x)$



Lagrangian Function

Continuous Space: $L_c(x, \lambda)$

- $L_c(x, \lambda) = f(x) + \lambda^T h(x)$

Discrete Space: $L_d(x, \lambda)$

- $L_d(x, \lambda) = f(x) + \lambda^T H(h(x))$
  - $H(h(x))$ transforms $h(x)$ into a non-negative function
  - Examples: $H(y) = y^2$, $H(y) = |y|$, $H(y) = \max(y, 0)$

Saddle Points

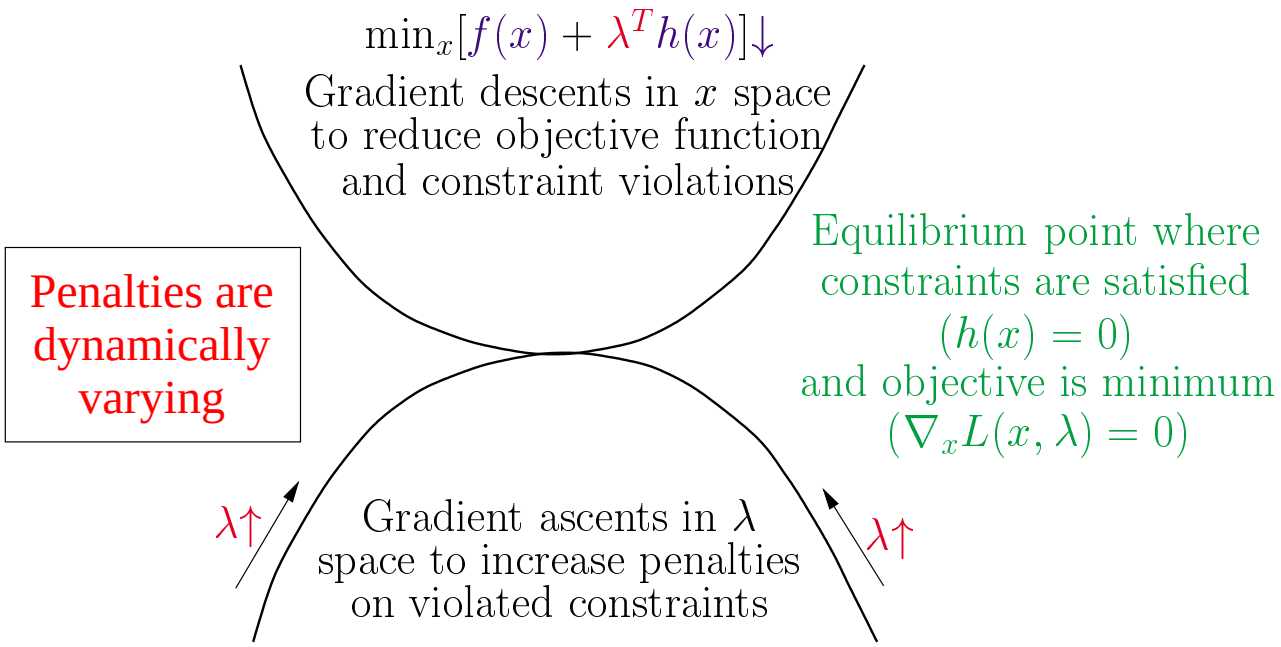
Continuous Space: $SP_{cn}(x^*, \lambda^*)$

- $L_c(x^*, \lambda) \leq L_c(x^*, \lambda^*) \leq L_c(x, \lambda^*)$
  $\forall \lambda$, x sufficiently close to (x^*, λ^*)
- Not useful concept in continuous space
  - A point can be verified to be a SP_{cn} after it is found
  - No systematic way to find SP_{cn}

Discrete Space: $SP_{dn}(x^*, \lambda^*)$

- $L_d(x^*, \lambda) \leq L_d(x^*, \lambda^*) \leq L_d(x, \lambda^*)$
  $\forall \lambda \in R^m$ and $\forall x \in \mathcal{N}_{dn}(x^*)$
- Feasible point + local minimum of $L_d(x, \lambda^*)$ in x space
- Core of discrete Lagrange-multiplier theory

Intuitive Meaning Behind Saddle Points



Theory of Lagrangian Multipliers

Continuous Space

- First-order necessary conditions
  - If x^* is a CLM_{cn} & *regular point*, then $\exists\ \lambda$ such that

    $$\nabla_x L_c(x, \lambda) = 0, \ \ \nabla_\lambda L_c(x, \lambda) = 0$$

  - Proved using implicit function theorem and chain rule
- Second-order sufficient conditions

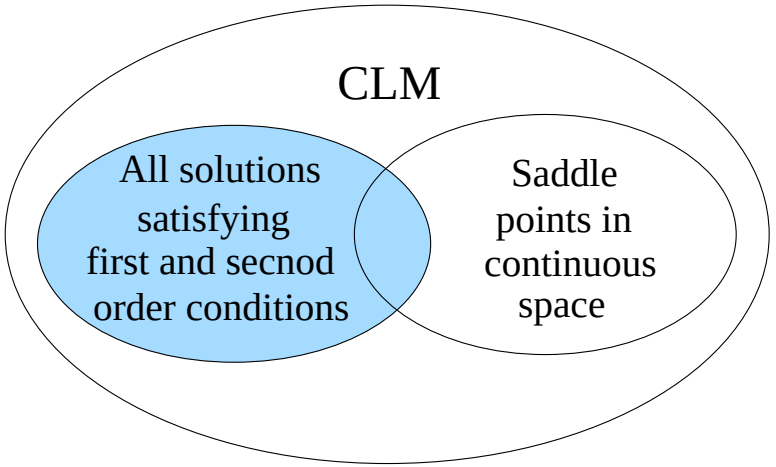
Discrete Space

- First-order necessary and sufficient conditions
  - x^* is a CLM_{dn}, iff x^* is a SP_{dn}
  - Proved through the concept of SP_{dn}

CLM, Saddle Points, and First-Order Conditions

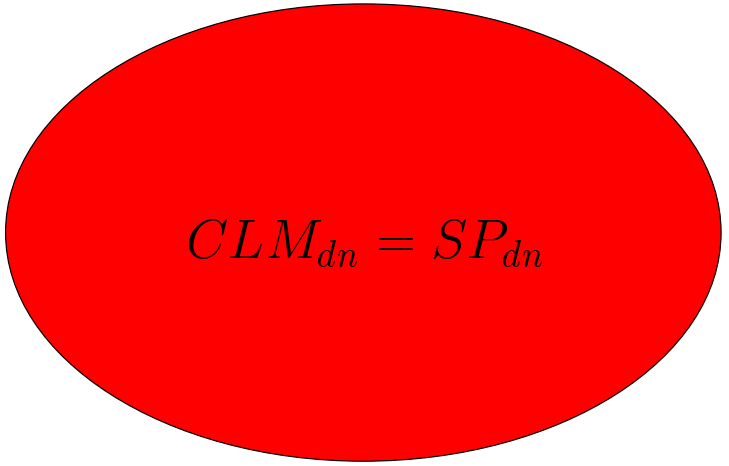
Continuous Space

- $SP_{cn} \Rightarrow CLM_{cn}$
- Set of $SP_{cn} \neq$ set of points satisfying first- and second-order conditions
- Global optimization not meaningful



Discrete Space

- $CLM_{dn} \Leftrightarrow SP_{dn}$
- Global optimization is meaningful



Non-negative Constraints

Handling Inequality Constraints

Continuous Space

- Adding slack variable to convert

  $g_i(x) \leq 0$ into $g_i(x) + z_i^2 = 0$

- Differentiable at $g_i(x) = 0$

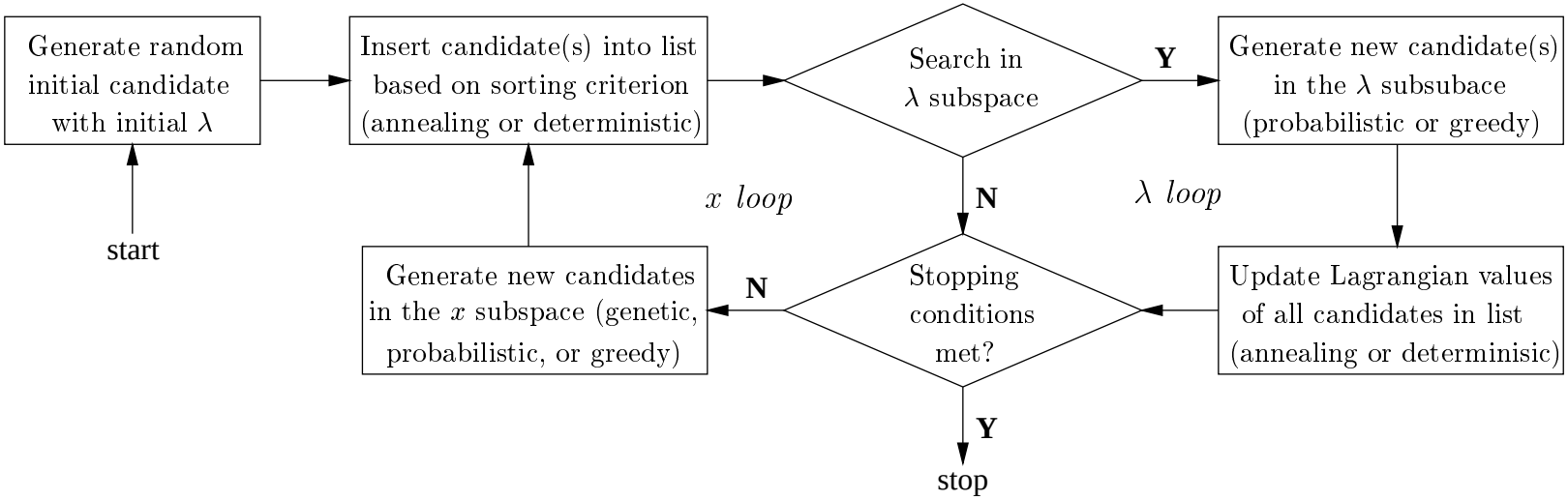
Discrete Space

- $g_i(x) \leq 0 \Rightarrow \max(g_i(x), 0) = 0$
- Not differentiable at

  $\max(g_i(x), 0) = 0$

STOCHASTIC SEARCH ALGORITHMS (SSAS) FOR CONSTRAINED GLOBAL MINIMIZATION

General Framework to Look for SP_{dn}



- Stop at either feasible points or at CLM_{dn} if all $x \in \mathcal{N}_{dn}(x)$ can be enumerated
- Algorithms studied: DLM, CSA, CGA, CSAGA

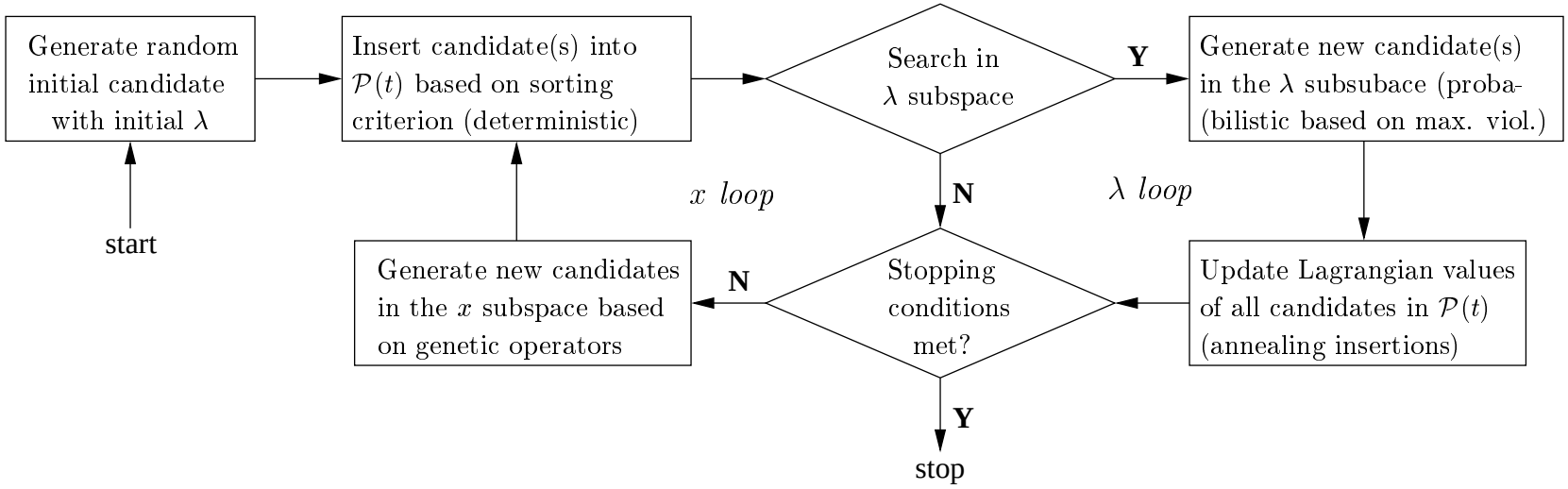
Discrete Lagrangian Method (DLM)

- Given a user-defined neighborhood $\mathcal{N}_{dn}(x)$, $\lambda_0 \leftarrow 0$, and $\mathbf{x}_0 = (x_0, \lambda_0)$
  - Generate trial point $\mathbf{x}'$ using $G(\mathbf{x}, \mathbf{x}')$,
  - Accept $\mathbf{x}'$ if $L_d(\mathbf{x}')$ is reduced wrt x or increased wrt λ when compared to $L_d(\mathbf{x})$
  - Stop search when no further improvements can be found
- If $\mathcal{N}_{dn}(x)$ is small enough to be enumerated in each iteration, then search can find CLM_{dn}

Constrained Simulated Annealing (CSA)

- Given a user-defined neighborhood $\mathcal{N}_{dn}(x)$, $\lambda_0 \leftarrow 0$, $\mathbf{x}_0 = (x_0, \lambda_0)$,
  T_0: initial temperature,
  N_T: number of trials per temperature,
  α: cooling rate $(0 < \alpha < 1)$
  - Generate trial point $\mathbf{x}'$ using $G(\mathbf{x}, \mathbf{x}')$,
  - Accept $\mathbf{x}'$ using Metropolis probability $A_T(\mathbf{x}, x')$ if $L_d(\mathbf{x}')$ is reduced wrt x or increased wrt λ when compared to $L_d(\mathbf{x})$
  - Reduce T by α and repeat steps until $T \leq T_\infty$
- Asymptotic Convergence Theorem
  - Markov chain modeling CSA converges asymptotically to a constrained global minimum (CGM_{dn}) with probability one
  - Result is of theoretical interest only

Constrained Genetic Algorithm (CGA)



- Descents in original x subspace using genetic algorithm
- Using L_d as fitness function
- Necessary condition for CGA to converge
  - All candidates are feasible solutions to the original problem

Combined Constrained SA and GA (CSAGA)

1. **procedure** CSAGA(P, N_g)
2. set $t \leftarrow 0$, T_0, $0 < \alpha < 1$, and $\mathcal{P}(t)$;
3. **repeat** /* over multiple generations */
4. **for** $i \leftarrow 1$ **to** P **do** /* SA in Lines 5-10 */
5. **for** $j \leftarrow 1$ **to** f **do**
6. generate $\mathbf{x}'_j$ using $G(\mathbf{x_j}, \mathbf{x'_j})$;
7. accept $\mathbf{x'_j}$ with probability $A_T(\mathbf{x_j}, \mathbf{x'_j})$
8. **end_for**
9. **end_for**
10. set $T \longleftarrow \alpha \times T$; /* set T for the SA part */
11. **repeat** /* by GA over probes in x subspace */
12. $y \leftarrow GA(select(\mathcal{P}(t)))$;
13. evaluate $L_d(y, \lambda)$ and insert y into $\mathcal{P}(t)$;
14. **until** sufficient number of probes in x subspace;
15. $t \leftarrow t + f$; /* update generation number */
16. **until** $(t \geq N_g)$
17. **end_procedure**

Design Issues of SSAs for Constrained Global Minimization

- Duration of each run
- Population size
- Many tunable parameters
- Parallel processing
- Anytime search
- Optimality of evaluations

THEORY OF SSA

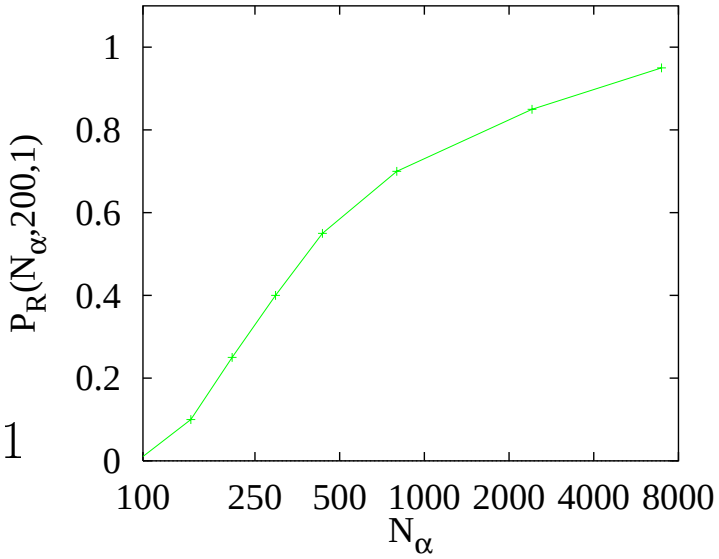
Reachability Probabilities

- SSA with N_α probes
- $p_j(Q)$ = probability that SSA finds solution of quality Q in the j^{th} probe
- Reachability probability is the probability that any of the N_α probes succeeds in finding a solution

$$P_R(N_\alpha, Q, 1) = 1 - \prod_{j=0}^{N_\alpha}(1 - p_j(Q)) \leq 1$$

  - All probes are independent (simplistic assumption)

SSA optimizing a 10-D constrained Rastrigin function

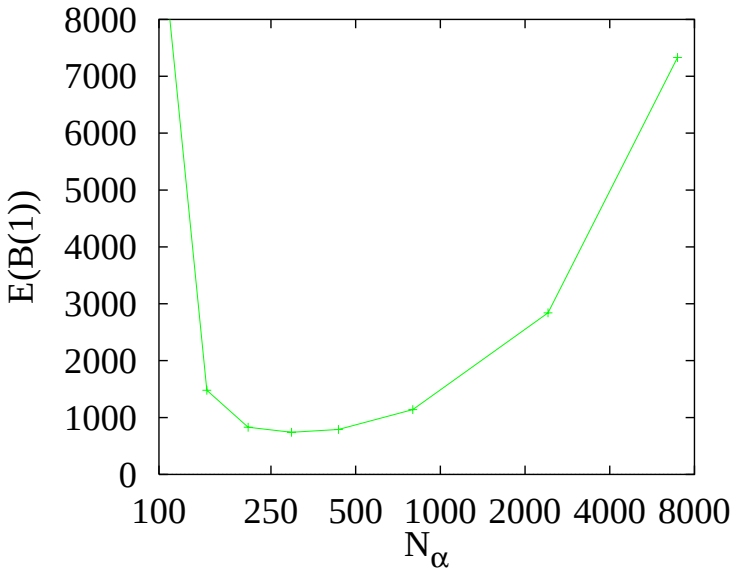


Expected Sequential Time With Multiple Runs

- SSA with fixed N_α can be run multiple times from different starting points to improve its success probability

- $B(1)$ = total sequential probes in multiple independent runs from random starting points to find a solution of quality Q

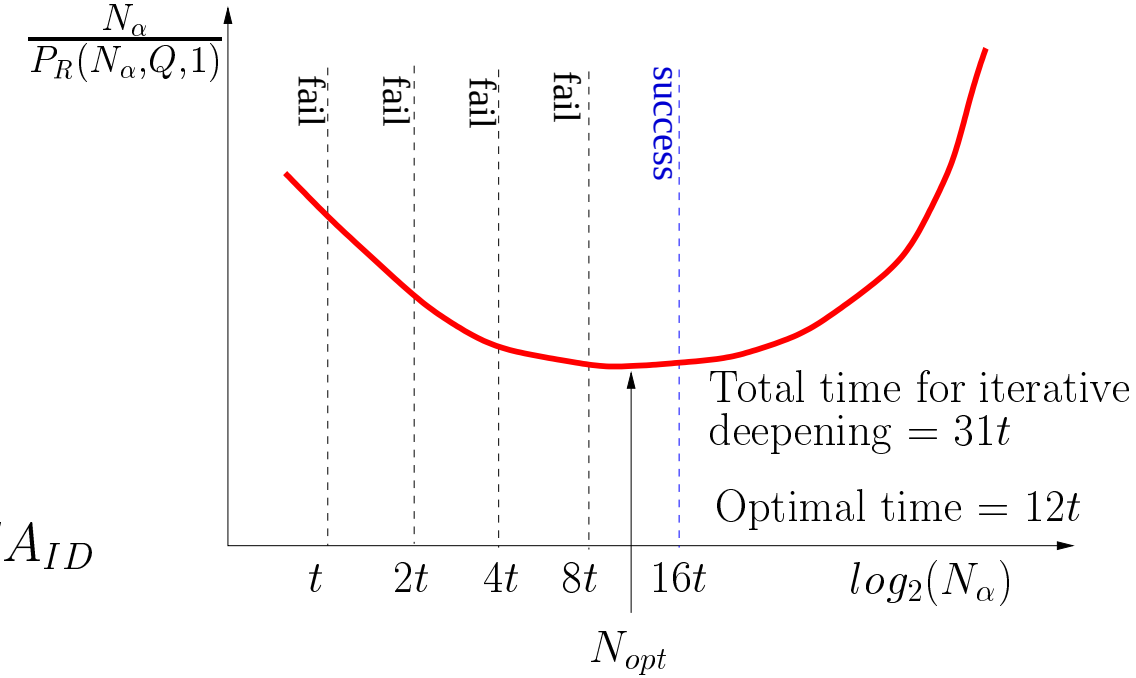
SSA optimizing a 10-D constrained Rastrigin function



$$E(B(1)) = \sum_{j=1}^{\infty} j \times N_\alpha \left[(1 - P_R(N_\alpha, Q, 1))^{j-1} P_R(N_\alpha, Q, 1)\right] = \frac{N_\alpha}{P_R(N_\alpha, Q, 1)}$$

SSA_{ID}: Sequential SSA with Iterative Deepening

$\frac{N_{opt}}{P_R(N_{opt},Q,1)}$ is a tight lower bound for any sequential SSA



Procedure Sequential SSA_{ID}

1. Set small N_α;
2. Execute SSA sequentially K times with N_α probes; If success, exit;
3. $N_\alpha \longleftarrow N_\alpha \times \rho$ (typically $\rho = 2$) ; Goto Step 2;

$\mathcal{B}_{ID}(1)$ = total time taken by SSA_{ID} to find solution of quality Q

Sufficient Conditions for $E(\mathcal{B}_{ID}(1)) = O\left(\frac{N_{opt}}{P_R(N_{opt},Q,1)}\right)$

1. $P_R(N_\alpha, Q, 1)$ is monotonically non-decreasing with respect to N_α in $(0, \infty)$
2. $\frac{N_\alpha}{P_R(N_\alpha,Q,1)}$ versus N_α curve satisfies sufficient conditions for existence of absolute minimum in $(0, \infty)$

   2.1. $P_R(0, Q, 1) = 0$ and $\lim_{x\to\infty} P_R(x, Q, 1) = 1$

   2.2. $\left.\frac{\partial^2 P_R(x,Q,1)}{\partial x^2}\right|_{x=0} > 0$

   (Rate of change of reachability curve at $x = 0$ is > 0;
   conditions not satisfied by random probing)

3. $\rho \times (1 - P_R(N_{opt}, Q, 1))^K < 1$ where ρ is typically 2

Optimality Condition of Sequential SSA_{ID} when $K = 1$

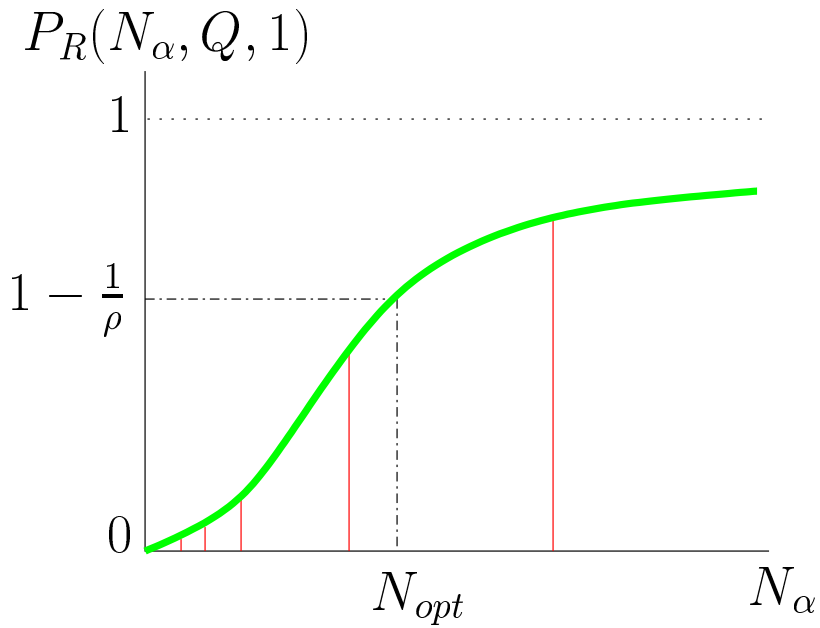
- Basic scheme with $K = 1$
  - One run of SSA at each N_α
  - $\rho = 2$
- Average iterative deepening time

$$E(\mathcal{B}_{ID}(1)) = O\left(\frac{N_{opt}}{P_R(N_{opt}, Q, 1)}\right)$$

when N_{opt} satisfies

$$\rho \times (1 - P_R(N_{opt}, Q, 1))^1 < 1$$

$$\Longrightarrow P_R(N_{opt}, Q, 1) > 1 - \frac{1}{\rho}\bigg|_{\rho=2} = 0.5$$



Optimality Condition of SSA_{ID} when $P_R(N_{opt}, Q, 1) \not> 0.5$

- $P_R(N_{opt}, Q, 1) \not> 50\%$

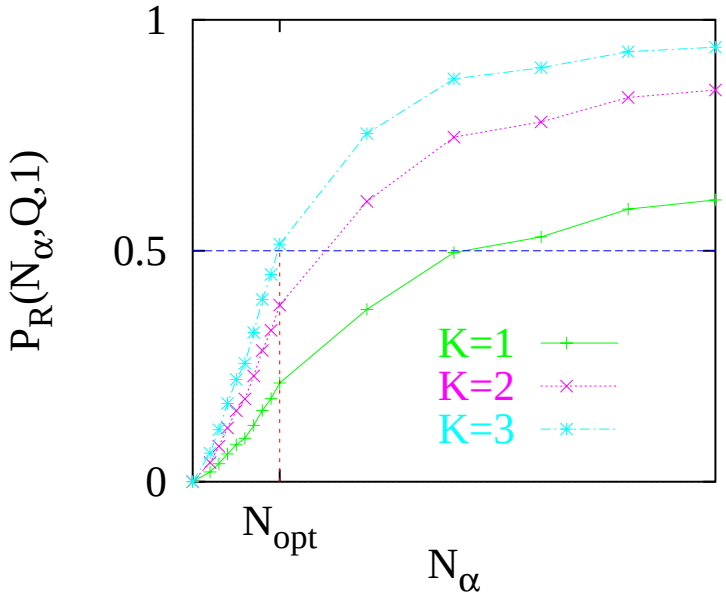
  $\Rightarrow$ Potential overshot

- Solution
  - Raise P_R curve to $P'_R = 1-(1-P_R(N_\alpha, Q, 1))^K$ by making K runs at each N_α so that

$$P'_R(N_{opt}, Q, 1) > 0.5$$

$$E(\mathcal{B}_{ID}(1)) = O\left(\frac{N_{opt}K}{P'_R(N_{opt}, Q, 1)}\right) = O\left(\frac{N_{opt}}{P'_R(N_{opt}, Q, 1)}\right)$$

SSA optimizing a 10-D constrained Rastrigin function



Parallel SSA: Goal

- Minimize search time by parallelizing SSA_{ID} on n processors
  - Assumption
    - Each SSA is scheduled as a unit without partitioning
  - Reachability probability: $P'_R(N_\alpha, Q, n) = 1 - (1 - P_R(N_\alpha, Q, 1))^n$
  - Average completion time of single identical runs on n processors

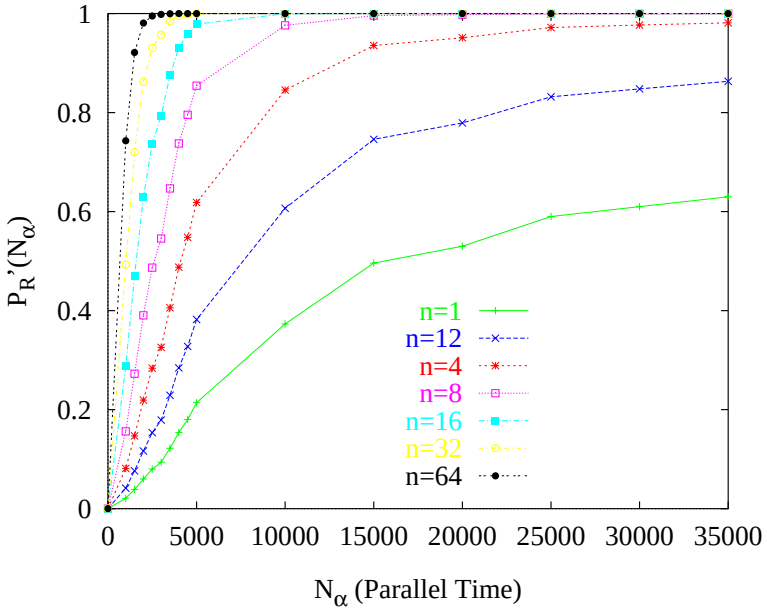
$$E(B(n)) = \frac{N_\alpha}{P'_R(N_\alpha, Q, n)}$$

  - Goal of parallel SSA_{ID}

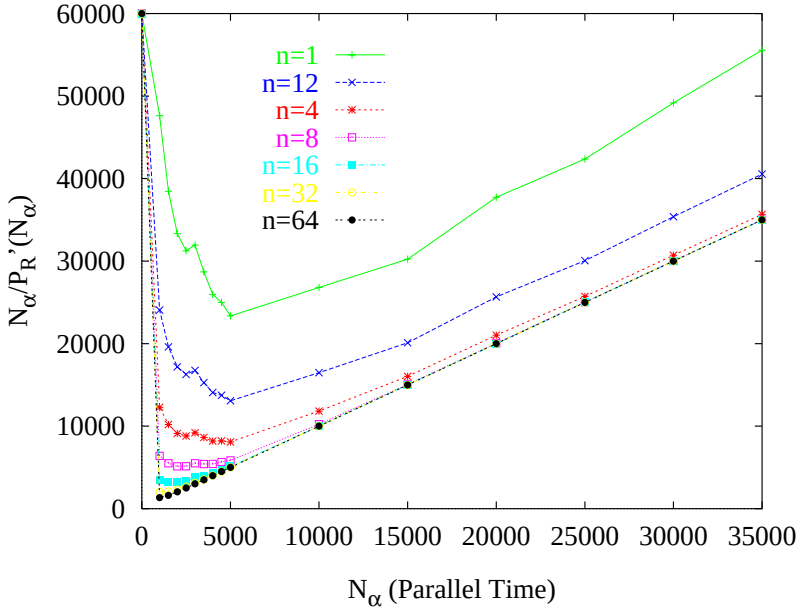
$$E(\mathcal{B}_{ID}(n)) = O\left(\frac{N_{opt}}{P'_R(N_{opt}, Q, n)}\right)$$

Example: 10-D Constrained Rastrigin Problem

$P'_R(N_\alpha, Q, n)$ curves



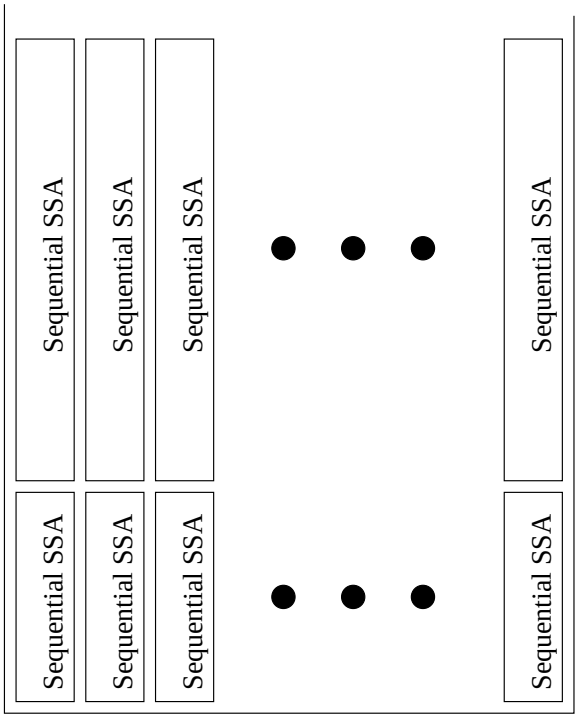
$\frac{N_\alpha}{P'_R(N_\alpha,Q,n)}$ curves



Naive Parallel SSA_{ID}

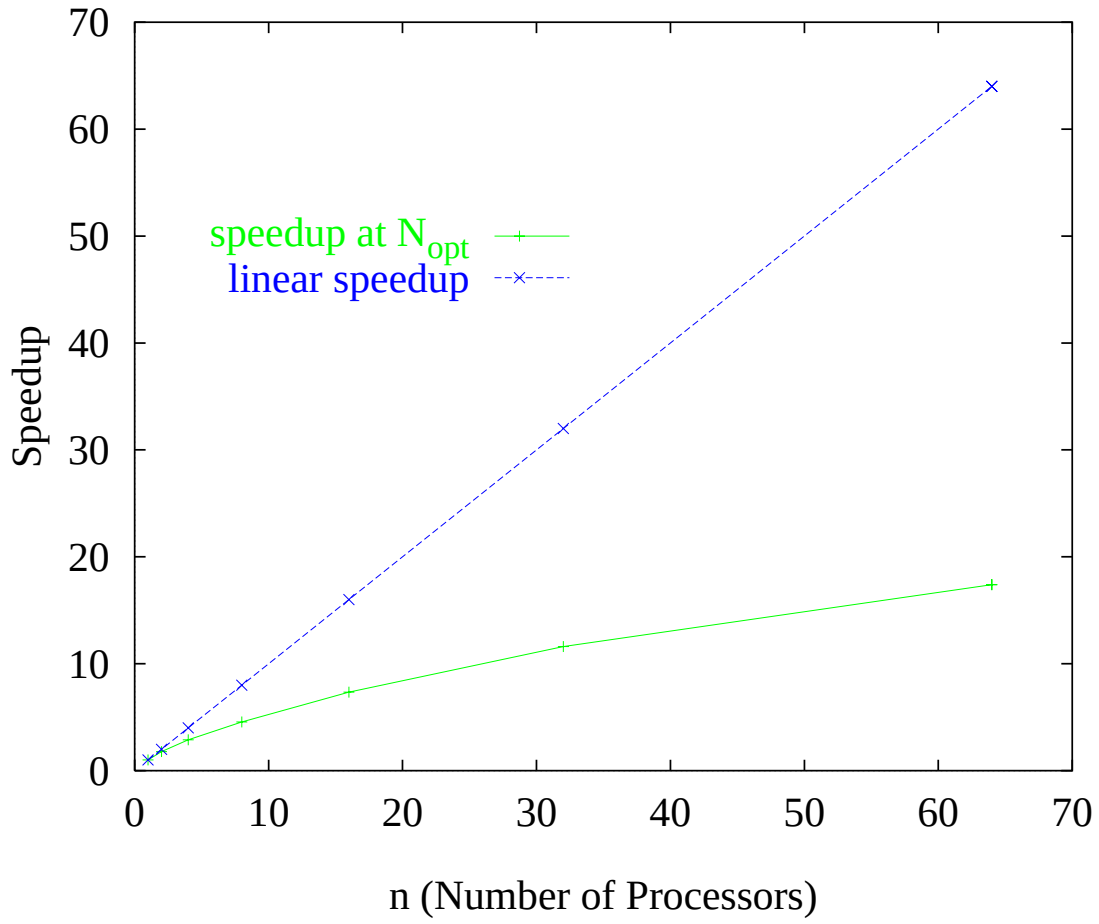
- Procedure *Parallel* SSA_{ID}

  1. Set small N_α;
  2. Execute sequential SSA with N_α on each processor
     If success, exit;
  3. $N_\alpha \longleftarrow N_\alpha \times \rho$ (typically $\rho = 2$) ; Goto Step 2;



- Naive parallel SSA_{ID} has $E(\mathcal{B}_{ID}(n)) = O\left(\frac{N_{opt}}{P'_R(N_{opt},Q,n)}\right)$
- Better schedules will allow $E(\mathcal{B}_{ID}(n)) \longrightarrow \frac{N_{opt}}{P'_R(N_{opt},Q,n)}$

Speedups on 10-D Constrained Rastrigin Problem



SOME SAMPLE RESULTS

Experimental Results on G1-G10

Problem ID	Global Solution f^*	EAs		CSA_{ID}	CGA_{ID}		$CSAGA_{ID}$			
		Best Sol.	Method	$\bar{\mathcal{T}}(f^*)$	P_{opt}	$\bar{\mathcal{T}}(f^*)$	P	$\bar{\mathcal{T}}(f^*)$	P_{opt}	$\bar{\mathcal{T}}(f^*)$
G1 (min)	-15	-15	Genocop	1.65	40	5.49	3	1.64	2	1.31
G2 (max)	-0.80362	0.803553	S.T.	7.28	30	311.98	3	5.18	3	5.18
G3 (max)	1.0	0.999866	S.T.	1.07	30	14.17	3	0.89	3	0.89
G4 (min)	-30665.5	-30664.5	H.M.	0.76	5	3.95	3	0.95	3	0.95
G5 (min)	4221.9	5126.498	D.P.	2.88	30	68.9	3	2.76	2	2.08
G6 (min)	-6961.81	-6961.81	Genocop	0.99	4	7.62	3	0.91	2	0.73
G7 (min)	24.3062	24.62	H.M.	6.51	30	31.60	3	4.60	4	4.07
G8 (max)	0.095825	0.095825	H.M.	0.11	30	0.31	3	0.13	4	0.10
G9 (min)	680.63	680.64	Genocop	0.74	30	5.67	3	0.57	3	0.57
G10 (min)	7049.33	7147.9	H.M.	3.29	30	82.32	3	3.36	3	3.36

- Algorithms derived from the framework can find optimal solutions without problem-dependent strategies and tuning used in EA
- CGA_{ID} is not competitive as compared to CSA_{ID} and $CSAGA_{ID}$
- Fixed $P = 3$ in $CSAGA_{ID}$ leads to minor performance loss

Experimental Results on Floudas and Pardalos' Problems

- Selected large ($n_v > 10$) problems, $CSAGA_{ID}$ with fixed $P = 3$

Problem		$f(x)$	CSA_{ID}	$CSAGA_{ID}$
ID	Best Sol.	n_v	$\bar{\mathcal{T}}(f^*)$	$\bar{\mathcal{T}}(f^*)$
2.7.1(min)	-394.75	20	35.11	14.86
2.7.2(min)	-884.75	20	53.92	15.54
2.7.3(min)	-8695.0	20	34.22	22.52
2.7.4(min)	-754.75	20	36.70	16.20
2.7.5(min)	-4150.4	20	89.15	23.46
5.2(min)	1.567	46	3168.29	408.69
5.4(min)	1.86	32	2629.52	100.66
7.2(min)	1.0	16	824.45	368.72
7.3(min)	1.0	27	2323.44	1785.14
7.4(min)	1.0	38	951.33	487.13

- Substantial improvements:
  - Average time $CSAGA_{ID}$ takes to find an optimal solution is 1.3 to 26.3 times less than that of CSA_{ID}

Conclusions

- Necessary and sufficient conditions for constrained satisfaction and local minimization
  - Easy to implement conditions
- Stochastic search algorithms for constrained global minimization
  - Provable asymptotic convergence, although not practical
- Optimal schedules in (parallel) stochastic search procedures
  - Overcoming one of the major limitations in existing algorithms
  - Optimal parallel schedules have sub-linear speedups